# Detecting Scenes in Fiction: A new Segmentation Task

**Albin Zehe**[*]
University of Würzburg
zehe@informatik.uni-wuerzburg.de

**Leonard Konle**[*]
University of Würzburg
leonard.konle@uni-wuerzburg.de

**Lea Dümpelmann**[†], **Evelyn Gius**[‡], **Andreas Hotho**[§], **Fotis Jannidis**[§], **Lucas Kaufmann**[§]
**Markus Krug**[§], **Frank Puppe**[§], **Nils Reiter**[¶], **Annekea Schreiber**[‡], **Nathalie Wiedmer**[‖]

## Abstract

This paper introduces the novel task of scene segmentation on narrative texts and provides an annotated corpus, a discussion of the linguistic and narrative properties of the task and baseline experiments towards automatic solutions. A scene here is a segment of the text where time and discourse time are more or less equal, the narration focuses on one action and location and character constellations stay the same. The corpus we describe consists of German-language dime novels (550 k tokens) that have been annotated in parallel, achieving an inter-annotator agreement of $\gamma = 0.7$. Baseline experiments using BERT achieve an F1 score of 24 %, showing that the task is very challenging. An automatic scene segmentation paves the way towards processing longer narrative texts like tales or novels by breaking them down into smaller, coherent and meaningful parts, which is an important stepping stone towards the reconstruction of plot in Computational Literary Studies but also can serve to improve tasks like coreference resolution.

## 1 Introduction

Text segmentation is a long standing issue in the area of natural language processing (NLP) encompassing different tasks like segmenting a text into sentences or finding the boundaries between different topics. In this paper, we introduce the task of *scene segmentation*. A scene can be understood as a segment of a text where the story time and the discourse time are more or less equal, the narration focuses on one action and space and character constellations stay the same. Scenes can be found predominately in narrative texts like novels or biographies, which can be understood as a sequence of segments, where some of the segments are scenes and others are not. Scene segmentation is of great interest for the high-level analysis of longer texts, for example the reconstruction of plot, but also for many areas of NLP that deal with longer narrative texts, since even modern methods struggle with processing text longer than a couple of sentences or paragraphs. As an example, the memory requirements of state of the art coreference resolution models scale with $\mathcal{O}(n^4)$ (input length $n$) (Lee et al., 2017; Joshi et al., 2020) and their performance deteriorates on longer texts. Therefore, it is very helpful to break down the texts into smaller pieces where the character constellation remains the same, enabling us to perform coreference resolution within a scene and then match the characters identified across multiple scenes. Additionally, scene segmentation can also be used to facilitate the summarization of long texts: Since Reiter (2015) has shown that parts of a human written summary correspond well to their similar notion of coherent segments in the original text, it is reasonable to assume that segmenting a text into scenes and then summarizing these scenes is a promising way towards the summarization of long texts. Finally, the number and length of scenes in a text defines a kind of "narrative rhythm", which we briefly analyze in Section 4.3. This rhythm might serve as a characteristic of an author, or be used as a metric for recommending books to readers.

While the objective of scene segmentation is structurally similar to topic segmentation, there are some important differences: Scenes are defined as narrative units, where each unit has a coherent and stable structure in respect to time, place, character constellation and plot. Narrated time and narrative time in a scene (Scheffel et al., 2019) are more or less equal, which can be seen, for example, in

[*] Equal Contribution
[†] University of Heidelberg
[‡] TU Darmstadt
[§] University of Würzburg
[¶] University of Cologne
[‖] University of Stuttgart

the rendition of verbal communication as direct speech. Scenes in fiction are thus not based only on the topic covered in the narrative. For example, a new scene may cover the same topic as the previous one, but take place in a different location or with a different set of characters. Thus, commonly used segmentation algorithms fail at our task. Even a fine-tuned BERT-based model does not perform well, as we show in Section 5.1. We see this as a sign that scene segmentation requires a large amount of natural language understanding and can, along with other tasks related to an in-depth analysis of narrative structures, serve as a challenge and benchmark for future NLP models.

Our contributions in this paper are as follows: We present and publish a new data set of German fictional texts annotated with scenes and provide an extensive discussion of the guidelines used in the annotation process (which are also included in the release). Additionally, we show that established baselines for text segmentation fail to capture the notion of a narrative scene, necessitating the development of new methods for this task. Our main goal is to introduce the task of scene segmentation and provide resources as well as guidelines to enable research towards this task, which will in turn improve the possibility of processing long, narrative texts in the future.

## 2 Related Work

Other segmentation tasks have been discussed in NLP for a while, mostly with the goal of identifying regions of news or other non-fictional texts discussing certain topics. The task of topic segmentation is then to identify points in the text in which the topic under discussion changes. Early work to this end uses similarity of adjacent text segments (such as sentences or paragraphs) with a manually designed similarity metric in order to produce the resulting segments. One of the most well known systems of this manner is TextTiling (Hearst, 1997), which was applied to science magazines. Similarity based on common words (Choi, 2000; Beeferman et al., 1999) was superseded with the introduction of Latent Dirichlet Allocation (Blei et al., 2003), which allowed to segment the text into coherent text snippets with similar topic distributions (Riedl and Biemann, 2012; Misra et al., 2011). This procedure was extended by the integration of entity coherence (John et al., 2016) and Wanzare et al. (2019) have used it on (very short) narrative texts in an attempt to extract scripts. Recently, many approaches making use of neural architectures deal with the detection and classification of local coherence (e. g. Li and Jurafsky, 2016; Pichotta and Mooney, 2016; Li and Hovy, 2014), which is an important step for a text summarization of high quality (Xu et al., 2019). Text segmentation using neural architectures was conducted on Chinese texts and it was shown that recurrent neural networks are able to predict the coherence of subsequent paragraphs with an accuracy of more than $80\,\%$ (Pang et al., 2019). Lukasik et al. (2020) compare three BERT based architectures for segmentation tasks: Cross-Segment BERT following the NSP Pretraining-Task and fine-tuned on segmentation, a Bi-LSTM on top of BERT to keep track of larger context and an adaption of a Hierarchical BERT network (Zhang et al., 2019).

Some work has been done on segmenting narrative texts, but aiming at identifying topical segments – which, as we have pointed out above, is different from scene segmentation. With a set of handcrafted features, Kauchak and Chen (2005) achieve a WindowDiff score (Pevzner and Hearst, 2002) of about $0.5$, evaluated on two novels. Kazantseva and Szpakowicz (2014) have annotated the novel *Moonstone* with topical segments, and presented a model to create a hierarchy of topic segments. They report about $0.3$ WindowDiff score. Most closely related to our task are the papers by Reiter (2015), which documents a number of annotation experiments, and Kozima and Furugori (1994), which presents lexical cohesiveness based on the semantic network Paradigme (Kozima and Furugori, 1993) as an indicator for scene boundaries and evaluates their approach qualitatively on a single novel. However, neither of them provide annotation guidelines, annotated data or a formal definition of the task.

A related area of research is discourse segmentation, where the goal is also to find segments that are not necessarily defined by topic, and are also assigned labels in addition to the segmentation. There are annotated news corpora in this area featuring fine-grained discourse relations between relatively small text spans (Carlson et al., 2002; Prasad et al., 2008). Although larger structures have been discussed in literature (Grosz and Sidner, 1986), no annotated corpora have been released.

## 3 Task: Scene Segmentation

In this section we will first present a description of the task placing it in narratological tradition,

and then we will describe it more formally as an atypical instance of a segmentation task.

### 3.1 Scenes in Narrative Text

In narratology, the analysis of narrative texts usually distinguishes between *discours* and *histoire* (Genette, 1983). *Discours* stands for the text or the representation of the narrative, while *histoire* concerns the narrated world, including characters and plot. In an ideal-typical view, it is assumed that plot is composed of several transformation processes from the smallest, spatiotemporal units – the so-called events.

Operationalizing plot is a challenging problem, because it involves natural language understanding, inferencing and interpretation on a high level (Meister, 2003). Nevertheless, different approaches are conceivable and have been discussed in the literature. Jockers (2015) approximates plot as an emotional arc of a narrative by assigning sentiment scores to each sentence and applying a Fourier transformation to derive an overall arc. Another way of modeling plot is to detect individual events in a text and then combining those to larger units and finally to a representation of the plot. There have been advances on the detection of events (Sprugnoli and Tonelli, 2019; Sims et al., 2019; Aldawsari and Finlayson, 2019) in texts. However, the definition of an event is unclear, with large possible differences in the level of granularity, making it an unstable starting point for analyzing plot.

Our approach to action in narratives is grounded in narratology, but by focussing on scenes it tackles the phenomenon on a less granular level than events. In narratology, the notion of scene has been introduced by Genette (1983) as a concept concerning the so-called pace of narration, i. e., the relation between the amount of time that passes in the narrative (story time, or *histoire*) and the amount of time covered by its narration (narrated time, or *discours*). Genette defines a scene as follows: "scene, most often in dialogue, which, as we have already observed, realizes conventionally the equality of time between narrative and story" (Genette, 1983, p. 94). It is important to note that the equality is put as "only a kind of conventional equality between narrative time and story time" (Genette, 1983, p. 87). Defining scenes based only on one feature, time, is useful in the context of Genette's theory, but it lacks descriptive power when the concept is supposed to be used to analyze plot, because the concept of plot is always implying aspects like character (and character constellation) and event sequences. In addition, Genette's definition of scene leads to two notions – story and narrative time – that are not easier to operationalize. Therefore, we adopt a more general understanding of scenes that includes characters, space and action. This is closer to our everyday understanding of scenes and similar to the understanding of scenes in plays as "a division [. . . ] during which the action takes place in a single place without a break in time" or "a part of a play, movie, story, etc., in which a particular action or activity occurs" (Learner's Dictionary).

In order to capture this fuller notion we follow Gius et al. (2019) in defining scenes: A scene is a segment of the *discours* (presentation) of a narrative which presents a part of the *histoire* (connected events in the narrated world) such that (1) time is equal in *discours* and *histoire*, (2) place stays the same, (3) it centers around a particular action, and (4) the character constellation is equal. All of these conditions are not absolute, there can be small changes in either component, as detailed below.

In media like film or plays usually one scene follows another. Non-scenes, in which the progress of time is narrated in a compressed way can be found, but are relatively rare. In narrative texts passages which are not scenes can be found more often between scenes. The boundary between scenes can be clear cut, often indicated by phrases like 'at the next morning', 'in the meantime' etc, but can also be vague, for example when reflections of the narrator or a character are bridging two scenes or when the narrated time is accelerated at the end of one scene and then slowed again for the next.

### 3.2 Formal Task Definition

After defining scenes from a narrative perspective, we can formalize the task of *scene segmentation*: We are given a narrative text (e. g., a novel) as input and derive a segmentation that additionally labels each segment of the text either as a scene or as a non-scene. This is a notable difference to other segmentation tasks with no further distinction between types of segments.

There are multiple possible operationalizations of this task. In this paper, we frame scene segmentation as a sentence-level classification task where we find borders between segments and additionally classify the borders, representing whether there is a scene before and af-

ter them. More specifically, a border can fall into one of the three classes SCENE–SCENE, SCENE–NONSCENE, NONSCENE–SCENE.[1]

Other operationalizations include, for example, only providing a simple segmentation in a first step and then additionally classifying each segment as either scene or non-scene using a text classification model or directly using a sequence labeling model to assign each token or sentence an IOB tag. While these are also valid approaches to scene segmentation, we focus on the first method here, since it provides the easiest end-to-end operationalization: We can train one model that simultaneously detects borders and classifies them into one of the border types. One possible drawback of this method is that a model might predict incompatible scene borders (e.g., SCENE-NONSCENE followed by SCENE-SCENE). This problem can be alleviated by the use of a CRF-based classifier, where such a sequence would be recognized as very unlikely/impossible.

## 4 Corpus

### 4.1 Annotation

We annotated 15 dime novels from diverse genres (love, horror, adventure, etc.) in German language with a total of 36 k sentences and 550 k tokens.[2] We decided to use dime novels because preliminary studies have shown that the task is quite challenging even in this literary medium, which is more accessible to human readers than highbrow literature. Moreover, the length of an individual dime novel (Ø 36 k tokens) allows to annotate a reasonable number of full novels with reasonable effort.

The annotation of the corpus was performed by two annotators, an additional curator established the gold standard.[3] Annotators were also asked to document the reason for each scene change. The guidelines are the result of two iterations, incorporating feedback from and discussion with the annotators.

**Overview** In the following, we provide an overview of the central aspects of our annotation guidelines.[4] The guidelines are based on four main components: *time*, *space*, *action* and *characters*. In short, a change in any of these components (e.g., a large jump in time) is a signal for a scene change. The following paragraphs describe our guidelines in more detail, specifically which of these signals are most important for determining whether there is a scene change at a given position. We also detail how we deal with *contradictory signals and corner cases*. We conclude this section with a discussion of certain *typographical markers* found in our texts and the *difficulties* encountered in the annotation process.

**Time** With regard to time, the default for a scene is a chronological narration with a uniform pace whereas scene boundaries are indicated by changes in chronology (i. e. anachronies like flashbacks or flash-forwards), temporal omission (i. e., ellipsis) or major changes of the narrative pace. For all candidates for scene changes, the impact of the temporal phenomenon in question needs to be weighted with regard to its context within the narration. For example, we generally assume that the greater a time leap is in relation to the general granularity of time in the narration, the more likely the scene is changing. Therefore, if the general narration speed is rather low and action is for example narrated more or less on an hourly base, a leap of one day probably indicates a boundary between two scenes. On the contrary, if action is narrated on a day-to-day-base, a leap of one day is probably not a scene boundary but rather part of an ongoing scene.

**Space** With regard to space, the default for a scene is to take place within the same space whereas a change in space indicates a scene boundary. Space, similar to time, is analyzed with regard to the granularity of space within the narrative. The general principle adopted for the detection of relevant space changes is a container principle, i. e. space can be composed from smaller units (spaces, rooms). For example, the following passage is considered to take place within the same space, since the rooms in question (the corridor and the breakfast room) are parts of a hotel, i. e. the same space container:

> Auf dem Weg zum Frühstückszimmer meinte mein Partner: “Ich habe mir die halbe Nacht den Kopf darüber zerbrochen, wie wir unserem geheimnisvollen Gegner die Maske herunterreißen könnten.” “Und?”, fragte ich. “Ich war nicht gerade mit

[1] We do not segment non-scenes further.

[2] We use SpaCy (Honnibal et al., 2020) for tokenization and sentence-splitting.

[3] The annotators are students with backgrounds in computer science, digital humanities and/or German studies with prior experience in annotating the same or similar tasks.

[4] The full guidelines are available in Gius et al. (2021).

Geistesblitzen gesegnet", sagte Suko. Wir betraten das Frühstückszimmer.

(*Der Turm der 1000 Schrecken*)[5]

**Action** A container principle has also been adopted for the analysis of action. Within a scene, action is assumed to be coherent and continuous. Generally, according to the container principle, we have to decide whether actions can be counted as belonging to the previous one. We also introduced a test based on the more intuitive understanding of scenes discussed above, where annotators imagined the passage in question as a movie and asked themselves whether it could be transposed to one movie scene. The boundaries of the scene are the points where a fade out (or fade in) could be inserted.

**Characters** The fourth aspect of scene, characters, again is supposed to stay stable within a scene whereas a change in character constellation can indicate a change of scene. Here it is important to examine both the role of the character that joins or leaves and the course of action. The more important the character is and/or the more the narrative focuses on a different action after the change in constellation, the more likely we have a scene boundary.

**Contradictory Signals and Corner Cases** Since these aspects are often not consistent with each other, we weight them according to their observed relevance for a scene change. Most relevant is a change in the event sequence, followed by character constellation, time and finally space.

In addition to the four relevant aspects in narrations for scenes, we included procedures for unclear cases in our guidelines. The most frequent ones are short passages of reflection, as for example inner monologues, that can be found directly before and after passages clearly qualifying as a scene. If they are shorter than the scene itself, they are also considered part of the scene.

**Typographical Markers** With regard to typical typographical markers of scenes in some texts (e.g., ***), we changed our handling during the annotation.[6] We started with the stars included, but then decided to erase them in order to rely on content-related aspects only due to their lack of generalizability: While *** is used somewhat consistently in German dime novels, this is not the case for other types of narrative texts, e.g. novels in book form. Since we want to keep our task, dataset and, in the future, solutions as general as possible, we did not rely on these markers. Changing our handling of these markers did not have an influence on our inter-annotator agreement. Despite this, it is an open question whether to include such typographical markers in future work. These markers, as well as chapters and paragraphs,[7] are standardized ways of signaling the segmentation of narratives. Therefore, the start of a scene directly after a marker may differ systematically from a start in running text.

**Annotation Difficulties** Most persisting annotation difficulties are caused by the fact that most criteria in the guideline have a relative nature. It is not trivial to decide whether a change in time, space or character configuration is determining the beginning (or end) of a scene, since its importance depends on the granularity of time, space or character configuration in the specific narrative. These granularities cannot be specified on a general level. The second issue is the lack of an operationalization of action applicable for the analysis of literary narratives.

Nevertheless, with the described approach to scene annotation we seem to tackle most of the issues the annotators and the curators came up with and improved the agreement between the annotators considerably. Overall, the combination of operationalized narratological categories with the more intuitive test of the transposability to a movie scene seems to be a fruitful approach.

To illustrate the annotation task for an international audience, we annotated a copyright-free novel and machine translated it using DeepL.[8]

### 4.2 Measuring Inter-Annotator Agreement

Several measures to evaluate segmentation systems have been proposed in the past. According to a survey by Kazantseva and Szpakowicz (2012) these are also widely used to quantify inter-annotator agreement. When choosing a measure for the evaluation of scene segmentation, the following characteristics must be considered: (1) Near misses

[5] Our translation: On the way to the breakfast room, my partner said, "I spent half the night worrying about how we could rip the mask off our mysterious opponent." "And?", I asked. "I wasn't exactly blessed with flashes of inspiration", Suko said. We entered the breakfast room.

[6] The markers are removed from text in the release, but there is information on their presence in metadata.

[7] Only 43% of all scene boundaries align with paragraphs.

[8] https://professor-x.de/gsd/viewer.html

are not critical, since scene boundaries tend to be fuzzy. (2) High variance of segment length in documents and between documents. (3) The existence of non-scenes leads to two classes of segments.

A major objective in the research on the evaluation of segmentation is to overcome influences from the field of classification, which model segmentation as token or character classification, and to replace these with measures that calculate a penalty dependent on the distance between true and predicted boundaries. After an evaluation of existing metrics (Pk: Beeferman et al. 1997, WindowDiff: Pevzner and Hearst 2002, Segmentation Similarity: Fournier 2013), we opted for using $\gamma$ (gamma) (Mathet et al., 2015) to measure inter-annotator agreement (and prediction performance, see below).

Values of $\gamma$ range from $-\infty$ to 1 theoretically, empirically often from 0 to 1; with 1 meaning there are no disagreements. The basic idea of gamma is to combine aligning and comparing the annotations into a single metric. Once an alignment between the annotations is established, near misses and category disagreements can be measured in a straightforward way, configurable by the user. Because there is no way to calculate this alignment a priori, gamma selects the alignment that leads to the least overall disagreement, which is then considered the observed disagreement $\gamma_o$. Expected disagreement $\gamma_e$ is calculated by sampling from the existing annotations such that random annotations can be compared. The final gamma score is calculated as $\gamma = 1 - \frac{\gamma_o}{\gamma_e}$, as for other metrics based on disagreements. Because the measured disagreement is dependent on both boundary positions and segment categories, segmentation tasks producing gaps or tasks that include unitizing and categorization (like named entity detection or topic segmentation) are supported by $\gamma$.

On all novels in our corpus, annotators reach an agreement of $\gamma = 0.7$ (with a standard deviation of $\sigma = 0.07$). The developers of $\gamma$ do not provide an explicit interpretation scale. However, since its value lies in the interval of $[-\infty; 1]$, with 0 representing agreement purely by chance and $\gamma$ is a disagreement metric similar to Krippendorff's $\alpha$, a similar scale can be applied to it. Thus, the reported agreement of $0.7$ is acceptable, given the fact that the task is new and very complex. Figure 1 shows aligned scene annotations of two annotators together with $\gamma$ scores.

| Category | Portion |
|---|---|
| Scenes starting with direct speech | 14 % |
| Scenes ending with direct speech | 13 % |
| Sentences containing direct speech | 55 % |

Table 1: Information on direct speech in the corpus

### 4.3 Corpus Analysis

This section gives an overview of various quantitative analyses that we have conducted on the annotated corpus.

First, we find that scenes are much more common in the texts than non-scenes, as the corpus contains 971 segments marked as scenes and 34 marked as non-scenes. This results in 937 SCENE-SCENE, 30 SCENE-NONSCENE and 23 NONSCENE-SCENE boundaries with respect to our task definition (see section 3.2). Figure 2 shows the lengths of scenes and non-scenes in comparison. As can be seen clearly, scenes are typically much longer than non-scenes, although with a quite high spread.

As direct speech plays an important role in narrative texts in general, and has already been established as a core ingredient of scenes, we use the supervised STRW Recognizer (Brunner et al., 2020) to detect the distribution of direct speech in the texts, and their relation to the scene boundaries. As Table 1 shows, direct speech does not serve well as a marker to detect scene boundaries.

Another interesting aspect of the scenes is their distribution along the text flow. Figure 3 shows for each sentence of a novel how many scenes haven been found up to this point. While it can be directly observed that average scene length seems to be a discriminatory property of the stories (*Fürstenkinder*, *Jason Dark* and *Sophienlust* contain nearly the same amount of sentences but differ highly in scene count), there is no pattern along the x-axis discernible. Denser clusters may appear at any point within the text. Nevertheless, we can highlight writing style differences between the stories. Figure 3 shows that some stories (*2012*, *Tausend Pferde*) start with a rapid succession of scenes, while others begin with longer sequences (*Fürstenkinder*, *Verschmäht*).

As we have summarized above, time, place and characters are important constituents of scenes. We therefore analyzed in which form they are expressed shortly after a new scene has begun. Figure

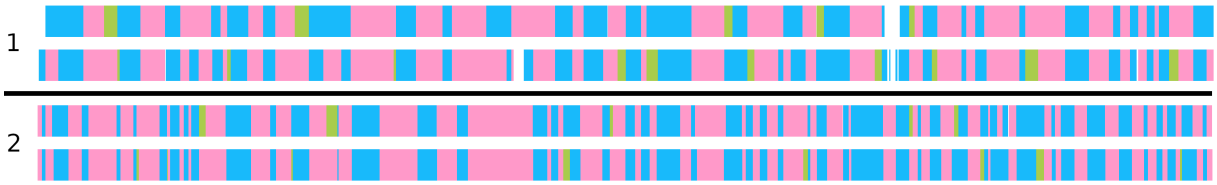


Figure 1: Scene annotations of *Dr. Bergen*, (1, $\gamma = 0.57$, worst agreement in the dataset) and *Tausend Pferde* (2, $\gamma = 0.83$, best agreement in the dataset). Three colors (red, blue and green) are used for scenes to make it easier to identify differences and matches. White gaps indicate non-scenes.

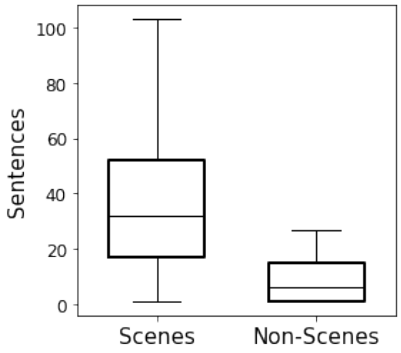


Figure 2: Length of scenes and non-scenes

4 shows the distribution of the occurrence of locational and temporal adverbials for each text in first sentences in scenes and all other sentences, as well as their coincidence with annotated reasons for a scene change. We used a list of adverbials according to (Eisenberg, 2006) as well as a self-made list. We found that there is high variation with respect to the scene-starting sentences. Temporal adverbials occur more frequently in the first sentence of a scene, but the difference for locational adverbials is much smaller. It becomes clear that the information about space and time changes with adverbials was often not perceived by the annotators as decisive for scene changes.

Additionally, we also analyzed the distribution of explicit character references (i.e., referencing them by their name rather than, e.g., pronouns). To this end, we used the method proposed by Jannidis et al. (2017) to extract all such references and plotted their position in a scene (Figure 5). We find that, as we would expect, the first 5 % of sentences in a scene contain more explicit character references on average than all other segments.

### 4.4 Corpus Release

The corpus is available on our website.[9] Since the texts are copyrighted, we cannot publish them directly. Instead, we provide the EAN of the epubs and a script that merges text and standoff annotations.

## 5 Baseline Experiments

### 5.1 Setup

Similar to measuring inter-annotator agreement, it is not trivial to define a metric for evaluating a task like scene segmentation. For the evaluation here, we provide two **metrics**: (1) precision, recall, F1-score are measured as a sentence-wise classification task, with different granularities as described below (two classes vs. four classes). (2) In addition, we employ the observed part of gamma $\gamma_o$ as a prediction performance metric. This metric is calculated from the alignment between gold and system output with the least disagreement. The reason for also reporting this metric is that it is in line with the annotation experiments and captures the task more directly. After all, the annotators are asked to create units in the context of the entire discourse, and not to classify individual sentences.

In order to assess the difficulty of scene segmentation, we evaluate multiple baselines on our proposed dataset: Two unsupervised standard segmentation techniques (TextTiling and TopicTiling) and two additional supervised baselines based on BERT. Note that the unsupervised techniques cannot perform the full scene task defined in Section 3.2, but only the first part, that its, detecting scene borders.

**Unsupervised Baselines** We use the TextTiling (Hearst, 1997) implementation from nltk (Bird et al., 2009) and evaluate these hyper-parameters: $w, k \in \{5, 10, ..., 45\}$, smoothing_width, smoothing_rounds $\in \{1, ..., 4\}$. Reported results are for the best configuration.[10] For TopicTiling (Riedl and Biemann, 2012), we train an LDA (Blei et al.,

[9] https://professor-x.de/german-scene-dataset

[10] We used the best possible configuration on the entire data set to provide an upper bound for their performance.

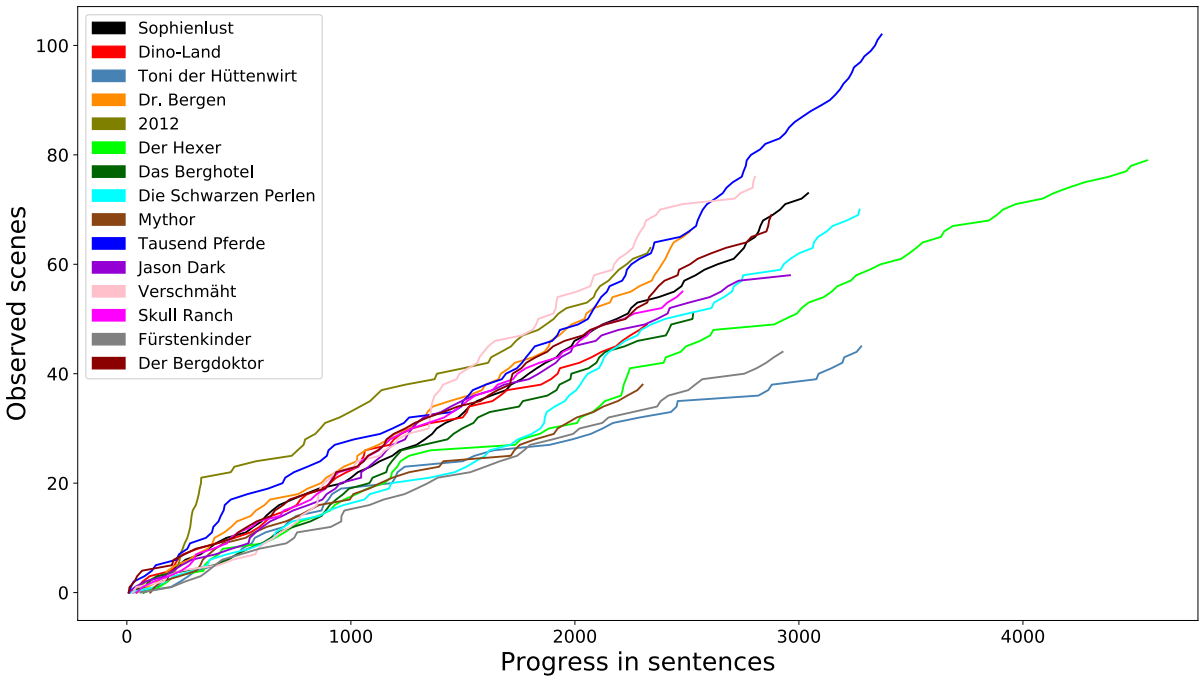


Figure 3: Relationship between the amount of sentences passed (x axis) and the number of scenes (y axis) for every novel. Higher slope indicates shorter scenes. An interactive version of this figure, suited for color blind people can be found online.[a]

[a] https://professor-x.de/gsd/sentences-per-scene.html

2003) model on a corpus of 870 dime novels[11] (appr. 2 million words)[12] and evaluate on our scene segmentation dataset using the recommended parameters.[13] As seen in Table 2, both unsupervised baselines perform very poorly at the task, reaching an F1-score of 4 % and 5 %, respectively.

**BERT Baseline** Since the standard unsupervised methods do not perform well for our task, we build a simple supervised baseline. To this end, we fine-tune a pre-trained BERT model[14] to binary scene segmentation in the following way: We construct a training sample as a triple of *(sentence, context, label)*, where *sentence* is a target sentence from a text, *context* concatenates the two previous and following sentences with the target sentence and *label* is BORDER, if there is a scene border before the target sentence and NOBORDER otherwise. In order to capture the distinction between scenes and non-scenes, we also evaluate a more fine-grained 4-label classification task with the BORDER label split into its three possible subclasses (SCENE-SCENE, SCENE-NONSCENE, NONSCENE-SCENE), as defined in Section 3. We fine-tune BERT using FARM[15] for a default duration of ten epochs in a leave-one-out style, training on all texts except one and evaluating on the remaining. Table 2 shows that, while performing much better than the unsupervised baselines from the previous paragraph, BERT is not capable of providing a satisfactory segmentation.

| Model (class) | Prec. | Rec. | F1 | $\gamma_o$ |
|---|---|---|---|---|
| TextTiling | 0.02 | 0.97 | 0.04 | 0.01 |
| TopicTiling | 0.03 | 0.22 | 0.05 | 0.02 |
| BERT (binary) | 0.49 | 0.15 | 0.24 | 0.15 |
| BERT (S-S) | 0.43 | 0.13 | 0.2 | 0.15 |
| BERT (S-NS) | 0.85 | 0.24 | 0.38 | |
| BERT (NS-S) | 0.0 | 0.0 | 0.0 | |

Table 2: Results from all baselines. BERT (binary) denotes the performance scores for the BORDER class. S-S denotes borders between two scenes, NS-S and S-NS denote borders between scenes and non-scenes.

[11] Detailed list of novels: https://professor-x.de/german-scene-dataset/list

[12] Using the implementation from http://gibbslda.sourceforge.net/.

[13] https://github.com/riedlma/topictiling

[14] https://deepset.ai/german-bert.

[15] https://github.com/deepset-ai/FARM.

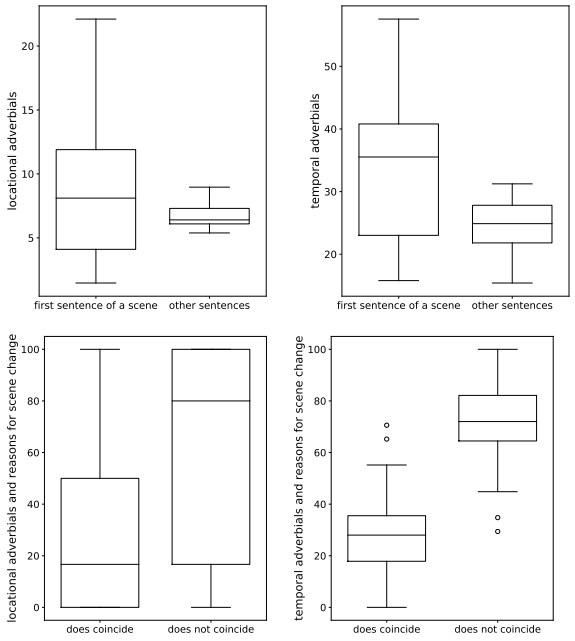


Figure 4: Distribution of locational (left) and temporal (right) adverbials at the beginning of a scene, as well as their coincidence with annotated reasons for a scene change.

### 5.2 Discussion

Our results show that standard unsupervised methods for text segmentation are not applicable to the task of scene segmentation. This is unsurprising, as our definition of scenes is not based on topical coherence, but on other aspects of the text, which are not considered by TextTiling and TopicTiling.

Additionally, our supervised baseline suggests that a BERT model is capable of picking up some signals for scene changes, but applying a standard model is not sufficient without additional modifications, such as specialized model architectures. On average, the system divides the novels into four times more scenes than the annotation specifies. A qualitative evaluation of the predicted borders shows that it is almost always possible to find one of the reasons defined in the guidelines for a scene change. Thus, the reason for the hypersensitivity of the baseline does not seem to lie in the inability to recognize markers, but rather in their contextualization. Typical errors are caused by (a) mentions of characters and places without actually appearing/becoming the place of action, (b) metaphors (e.g. “Maybe they only needed a narrow bridge to get back together?”), (c) indirect or reported speech, (d) different forms of referencing characters and (e) moving through places in a scene (e.g., “The visitor nodded and followed her into the kitchen.”)

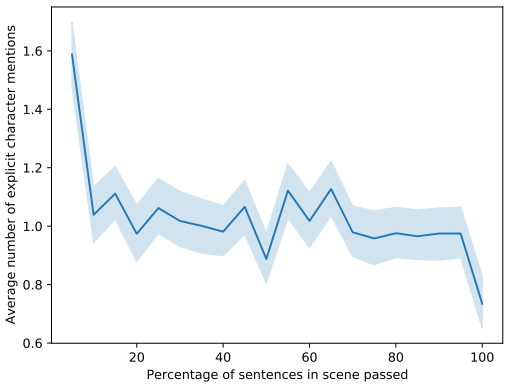


Figure 5: Distribution of explicit character references over positions in a scene.

The BERT model is also not capable of finding any borders from non-scenes to scenes, leading to a score of 0 in this setting. This, in combination with the task’s relevance for the analysis of long texts, motivates further research. A possible direction for future work would be using the information from our analysis, like the presence of temporal markers and character references.

## 6 Conclusion

This paper describes the task of detecting scenes in narrative texts and introduces a corpus annotated for this task. The corpus consists of a number of German dime novels annotated according to guidelines describing the specifics of the task in detail. The inter-annotator agreement indicates that the task is challenging, but feasible for humans. As the analysis of the corpus shows, the information about character constellation, time, space and action is informative, but only an integral understanding of the text makes it possible to fully solve the task. Thus, apart from the many applications of a scene segmentation itself, it also provides an interesting challenge for natural language processing: Due to the high level of natural language understanding required, it will likely necessitate the development of novel approaches to be solved satisfactorily.

## References


Mohammed Aldawsari and Mark Finlayson. 2019. Detecting subevents using discourse and narrative features. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4780–4790, Florence, Italy. Association for Computational Linguistics.

Doug Beeferman, Adam Berger, and John Lafferty. 1997. Text segmentation using exponential models. *arXiv preprint cmp-lg/9706016*.

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. *Machine learning*, 34(1-3):177–210.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Annelen Brunner, Ngoc Duyen Tanja Tu, Lukas Weimer, and Fotis Jannidis. 2020. To BERT or not to BERT - comparing contextual embeddings in a deep learning architecture for the automatic recognition of four types of speech, thought and writing representation. In *Proceedings of the 5th Swiss Text Analytics Conference and the 16th Conference on Natural Language Processing, SwissText/KONVENS 2020, Zurich, Switzerland, June 23-25, 2020 [online only]*, volume 2624 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2002. Rst discourse treebank, ldc2002t07. Technical report, Philadelphia: Linguistic Data Consortium.

Freddy YY Choi. 2000. Advances in domain independent linear text segmentation. *arXiv preprint cs/0003083*.

Peter Eisenberg. 2006. *Der Satz*, volume 2 of *Grundriss der deutschen Grammatik*. J. B. Metzler, Stuttgart, Weimar.

Chris Fournier. 2013. Evaluating text segmentation using boundary edit distance. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1702–1712. Association for Computational Linguistics.

Gérard Genette. 1983. *Narrative discourse: An essay in method*. Cornell University Press.

Evelyn Gius, Fotis Jannidis, Markus Krug, Albin Zehe, Andreas Hotho, Frank Puppe, Jonathan Krebs, Nils Reiter, Nathalie Wiedmer, and Leonard Konle. 2019. Detection of scenes in fiction. In *Proceedings of Digital Humanities 2019*.

Evelyn Gius, Carla Sökefeld, Lea Dümpelmann, Lucas Kaufmann, Annekea Schreiber, Svenja Guhr, Nathalie Wiedmer, and Fotis Jannidis. 2021. Guidelines for detection of scenes. https://doi.org/10.5281/zenodo.4457177.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Marti A Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Fotis Jannidis, Isabella Reger, Lukas Weimer, Markus Krug, Martin Toepfer, and Frank Puppe. 2017. Automatische Erkennung von Figuren in deutschsprachigen Romanen. *Proceedings of DhD*.

Matthew L. Jockers. 2015. Revealing Sentiment and Plot Arcs with the Syuzhet Package.

Adebayo Kolawole John, Luigi Di Caro, and Guido Boella. 2016. Text segmentation with topic modeling and entity coherence. In *International Conference on Hybrid Intelligent Systems*, pages 175–185. Springer.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

David Kauchak and Francine Chen. 2005. Feature-based segmentation of narrative documents. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, pages 32–39, Ann Arbor, Michigan. Association for Computational Linguistics.

Anna Kazantseva and Stan Szpakowicz. 2012. Topical segmentation: a study of human performance and a new measure of quality. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 211–220. Association for Computational Linguistics.

Anna Kazantseva and Stan Szpakowicz. 2014. Hierarchical topical segmentation with affinity propagation. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 37–47, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Hideki Kozima and Teiji Furugori. 1993. Similarity between words computed by spreading activation on an english dictionary. *In Proceedings of the European Association for Computational Linguistics*.

Hideki Kozima and Teiji Furugori. 1994. Segmenting narrative text into coherent scenes. *Literary and Linguistic Computing*, 9(1):13–19.

Merriam-Webster Learner's Dictionary. https://learnersdictionary.com/definition/scene. Accessed: 2021-01-20.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Jiwei Li and Eduard Hovy. 2014. A model of coherence based on distributed sentence representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2039–2048.

Jiwei Li and Dan Jurafsky. 2016. Neural net models for open-domain discourse coherence. *arXiv preprint arXiv:1606.01545*.

Michal Lukasik, Boris Dadachev, Gonçalo Simões, and Kishore Papineni. 2020. Text segmentation by cross segment attention.

Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The unified and holistic method gamma ($\gamma$) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479.

Jan Christoph Meister. 2003. *Computing action a narratological approach*. Walter De Gruyter, Berlin; New York.

Hemant Misra, François Yvon, Olivier Cappé, and Joemon Jose. 2011. Text segmentation: A topic modeling perspective. *Information Processing & Management*, 47(4):528–544.

Yihe Pang, Jie Liu, Jianshe Zhou, and Kai Zhang. 2019. Paragraph coherence detection model based on recurrent neural networks. In *International Conference on Swarm Intelligence*, pages 122–131. Springer.

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

Karl Pichotta and Raymond J Mooney. 2016. Learning statistical scripts with lstm recurrent neural networks. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Rashmi Prasad, Alan Lee, Nikhil Dinesh, Eleni Miltsakaki, Geraud Campion, Aravind Joshi, and Bonnie Webber. 2008. Penn Discourse Treebank Version 2.0 LDC2008T05. Web download, Linguistic Data Consortium, Philadelphia.

Nils Reiter. 2015. Towards Annotating Narrative Segments. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 34–38, Beijing, China. Association for Computational Linguistics.

Martin Riedl and Chris Biemann. 2012. Topictiling: a text segmentation algorithm based on lda. In *Proceedings of ACL 2012 Student Research Workshop*, pages 37–42. Association for Computational Linguistics.

Michael Scheffel, Antonius Weixler, and Lukas Werner. 2019. Time. In Peter et al. Hühn, editor, *The living handbook of narratology*. Hamburg.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary Event Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634, Florence, Italy. Association for Computational Linguistics.

Rachele Sprugnoli and Sara Tonelli. 2019. Novel Event Detection and Classification for Historical Texts. *Computational Linguistics*, 45(2):229–265.

Lilian Diana Awuor Wanzare, Michael Roth, and Manfred Pinkal. 2019. Detecting everyday scenarios in narrative texts. In *Proceedings of the Second Workshop on Storytelling*, pages 90–106, Florence, Italy. Association for Computational Linguistics.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Discourse-aware neural extractive model for text summarization. *arXiv preprint arXiv:1910.14142*.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5059–5069, Florence, Italy. Association for Computational Linguistics.