Modeling Plots of Narrative Texts as Temporal Graphs

Leonard Konle, Fotis Jannidis

Julius-Maximilians-Universität Würzburg, Würzburg, Germany

Abstract

The paper outlines a formal model of plot (and syuzhet) for narrative texts. The basic unit are scenes and the motif repertoire instantiated in the scene. The motif repertoire consists of three sets of (closely related) elements: character stereotypes, types of verbal actions and action types. It is assumed that the motif repertoire is highly dependent on the corpus which is analyzed, in our case a corpus of romance and horror novels published as pulp fiction. The resulting information is represented in a temporal graph which in turn is used to compute relevant information on the scenes and characters. Scenes are also characterized by their valence and their arousal value. A second representation which offers with a topic model of the direct speech and the narrative text a simple proxy for the types of verbal actions and the action types is also created. To assess the ability of these information structures to indicate changes in the temporal structures three evaluation methods are used based on artificial data. We can confirm that a very abstract representation of the plot is able to do so, but contrary to our expectations the more information-rich model which makes use of the topic model is not better in doing so. The main contribution of this paper is its attempt to integrate different research proposals into one integral model. We offer a descriptive framework and a proposal for the formal model of plot, which makes it possible to identify research problems and align existing approaches.

Keywords

plot, temporal graphs, scenes, characters, modeling

1. Introduction

For our understanding of narrative literature character and plot are basic and central categories. Though computational literary studies already have a rich landscape of character models, it is not yet as advanced when it comes to analyzing plot. The main reason is the complexity not only of a generic model of plot, but of the subproblems involved. Most of the contributions to the discussion of plot and event in recent years have tried to map the myriad of elements which can be found in plot descriptions to one or a very small set of textual phenomena. [14] uses sentiment values as indicators for plot fluctuation, [3] map from different groups of function words to three concepts: staging, plot progression and cognitive tension, [29] classify verbs to four types of event (changes of state, process events, stative events and non-events).[1] Alternatively [27] basically do without any abstraction and map almost each verb to itself. Our main goal in this paper is to discuss the outline of a model which could offer a more complex

CHR 2022: Computational Humanities Research Conference, December 12 – 14, 2022, Antwerp, Belgium

leonard.konle@uni-wuerzburg.de (L. Konle); fotis.jannidis@uni-wuerzburg.de (F. Jannidis)
0000-0001-5833-0414 (L. Konle); 0000-0001-6944-6113 (F. Jannidis)

[1] A detailed presentation of earlier computational research on plot can be found in Elsner [7].

representation than those mentioned, and to delineate what kind of problems the CLS community has to solve to reach this point. All in all, this paper is more of a modeling study with some attempts at implementation and evaluation than a typical CLS paper concentrating on the details of a specific implementation. But in our view the discussion on how to model plot has reached a point in recent years where that, which is described as plot in CLS, has only a very vague resemblance with what people in literary studies and beyond mean, when they use the term. But it stands to reason that only a solid model of plot, nearer to this established use of the term, can be the basis for understanding genre systems, historical developments of literature and many other aspects of literary communication.

It would be misleading to assert that there is one meaning of 'plot' in literary studies. The term has many layers of meaning, not the least because many different analytical traditions use this term in their English translations (details see [18]). We use the term here to refer to the structure constituted by the sequence of events. The term 'event' refers here as usual to the (inter-)actions of characters. But the term structure does not imply a reduction to some shape or outline, but rather a feature-rich representation which nevertheless can be abstract enough to recognize patterns and based on that similarity between texts. In the discussion of the term 'plot' very different levels of abstraction are used. But in our understanding only rarely is the term used in such an abstract way, that only the amount or intensity of action is measured, as some have interpreted Freytag's famous five-stage model of plot [9]. Most uses of 'plot' include more concrete aspects of the actions depicted in the text. This is closer to what is represented in a summary which concentrates on the main plot points.

The need for abstraction is confirmed by the narratological work on 'event'. [20] (see also [12]) has shown that the concept is so encompassing, that basically anything can be an event for someone under specific circumstances. In other words, a computational model of events in this understanding would need to encompass a complete model of the world. Add to this another observation, made first as far as we know by [15], that summaries of literary texts use more generalizations and abstractions compared to those of non-literary texts, which is also confirmed by our own work. In other words, the depiction of characters and events in literary texts is usually concrete, and a summary will compact this information using different terms including more generalizations and abstractions. This constellation is, so we believe, the main reason why real progress in this field has been stalled.

So abstraction is necessary, but how much? In the first step, we use a model which is richer than most models for representing plot. Its components are chosen based on earlier research. We also start, as the model by Elsner [8], with the characters but we add three components: First, we model plot as a sequence of scenes in which characters and events are nested. So first, we segment the text into scenes. This is based on work on scene segmentation [30] and is similar to the proposal in [23] where the relation between plot elements is described as 'event-scene-level-plotline-plot' (p.302). For each scene, we construct a character network. Second, recent work by linguists [5] and also work on speech rendition in narrative texts [4] has shown that a novel consists of speech rendition and of narration and these can be regarded as two text types or registers which have to be treated separately. Third, for each of these three components - character, speech rendition, narrative text - we define a very small set of generalizations and map the text to these.

Based on the insight mentioned above, that almost everything can be an event in some texts, we acknowledge that it is probably not possible to generate these generalizations independent from the corpus you wish to analyze. In other words, similar to Propp in his analysis of fairy tales [24], we do not define generalizations which are valid everywhere, but only for those texts we wish to analyze.

On the other hand, even if we acknowledge this dependency on the text corpus, it is unclear on what level these generalizations should be established and how. Probably this can only be answered by taking the corpus into consideration. We are interested in analyzing pulp fiction ('Heftromane'), literature written for entertainment, which is published in thin volumes on cheap paper. From this pulp fiction or dime novels we look at two genres, two, because this literature is from its very beginnings rather strictly binary gendered in relation to reader expectations by its publishers; in our selection, one of the genres - romance - is addressing women and the other - horror - is addressing men.

Work on plot may include information about characters [7] and work on character stereotypes often includes information about plot aspects. [1] for example include actions of which characters are agents or patients. [13] use Propp's plot functions as features to cluster the characters and identify character roles. It seems rather obvious that these aspects, character stereotypes, and actions/events are closely related. We therefore propose to use the term 'motif repertoire' for those character/plot/event elements which are typically present in a given corpus (usually a genre, a series etc.). As described above, we think it is useful for the analysis of narrative texts to distinguish between plot / event elements in narrative text and in direct speech. Thus we have three classes in the motif repertoire of a corpus which closely interact: 1) character stereotypes / roles, 2) verbal actions (somewhat more concrete than the usual linguistic speech acts, for example 'the [stereotype X] tells the [stereotype: heroine] that her [stereotype: beloved] wants to marry [stereotype: rival]', 3) action types and events ('the [stereotype: antagonist] attacks the [stereotype: hero]'). Establishing this motif repertoire fully (or even fully automated) is a very hard problem and beyond the scope of this paper. In this paper we will discuss two models for these generalizations. Our first approach was to define these generalizations based on our reading experience. In a second approach, we abstracted less and kept more information of the specific text corpus.

The main contributions of our paper are a more detailed analysis of a content-rich plot model and the difficulties involved. In some important aspects it is indebted to [7], but adds more modern ways to model temporality using temporal graphs and is based on scenes as basic units. We offer a descriptive framework and a proposal for the formal model of plot, which makes it possible to identify research problems and we describe some ways to evaluate these models. So, it is not our goal to find a very sophisticated and highly performant implementation for a specific task, but rather to investigate how a complex and feature-rich model of plot can be constructed and evaluated. In the practical parts of this paper, we rely, where possible, on existing tools and only add our own implementations where we need to fill specific gaps to reach our goal. These implementations are usually only simple place-holders for more sophisticated solutions to be found in the future.

1.1. Plot models for entertainment literature

The basic outline of our modeling approach has three levels. On the basis we have single texts which belong to a corpus. The plot models we discuss are meant to represent simple literature written and read for the purpose of entertainment. In doing so we follow our belief, that the domain of literature is too heterogeneous and especially 'high' literature too complex to construct models in this early stages of research in Computational Literary Studies which can cover literature in general. So we start our research with highly formulaic literature published as pulp fiction on the German speaking markets, specifically we work with two genres romance and horror. So even if we look at a specific single text, we look at it through the lens of an information system based on the structure of the corpus the text comes from.

The texts are segmented into scenes. Each scene can be represented abstractly as character stereotypes communicating and interacting. The character stereotypes, that is the kind of stereotype and also the elements of these stereotypes, are specific for a corpus. The same is true for the types of communication and the types of action rendered in a scene. So while each specific scene is represented abstractly, the elements of this abstraction are obtained through an analysis of the whole corpus – usually based on the genre –, the text belongs to. Figure 1 shows this basic outline. The specific components of character stereotypes, types of communication and types of action and events, we chose here are very simplified, in our empirical studies, described below, we used slightly more complex representations.[2] In our model the smallest

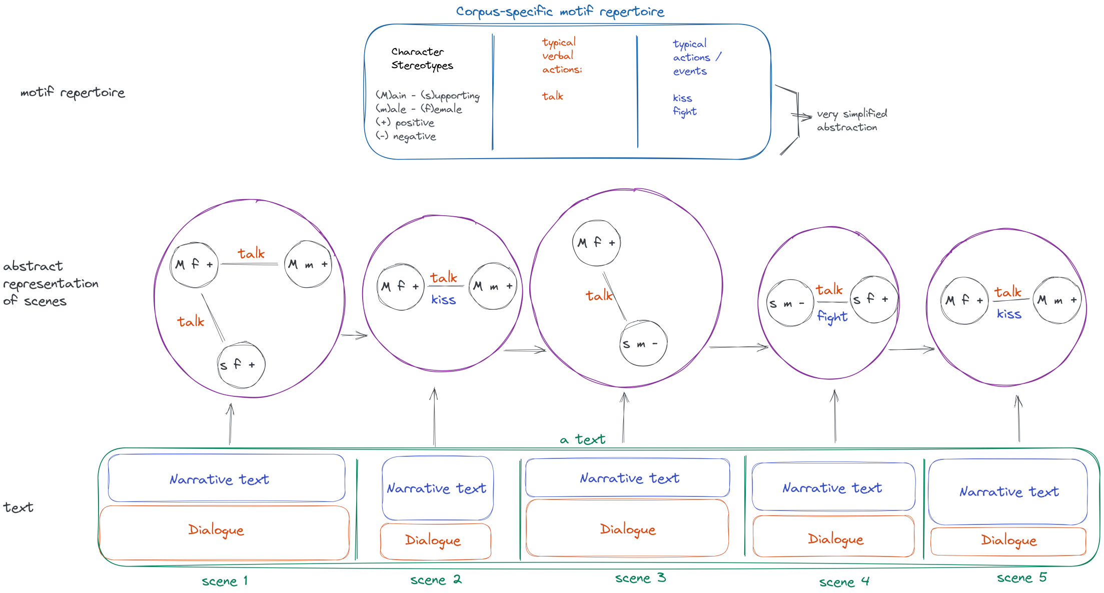


Figure 1: The temporal graph, an abstraction of the scene-segmented text, consists of elements of the motif repertoire.

[2] We do net deal with another important aspect which can be understood as additional part of the corpus-specific motif repertoire and which is instantiated in each scene: setting and space in general.

segment is a scene. 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 stays the same. All of these conditions are not absolute but rather relative, that is, small changes in either of them do not necessarily lead to a scene change but can rather be seen as indicators." [30]. In a scene we can find characters and events. An event is usually the action of one or more characters, often the interaction between them.

Based on our understanding of plot as a chronologically and ideally causally ordered sequence of events, we would now reorder the scenes accordingly. With the current state of the analytical tools in CLS, it is not feasible to do this automatically. Therefore we use the sequence as given by the text. At a later stage such a reordering could be added to the processing steps described below without any larger impact on the later steps. It is only necessary to remind oneself, that our model shows similarity between texts not on the histoire level alone, but on both levels: what happened and in which sequence was it narrated. In other words, we are not really talking about plot here, but rather about *syuzhet*, the plot as it is narrated. We also ignore the problem of narrative level, because none of the novels we read from these genres uses different narrative levels.

Not all scenes are equally important. There are always scenes which would never be mentioned in a summary while others are crucial. Even if the criteria for this weighting are hard to represent exactly, rough indicators like the level of valence and arousal could suffice for the time being.

In modeling the dime novels for our empirical research we tried to be as simple as possible:

1. Characters are described along three dimensions: main character vs. supporting character, positive vs. negative, male vs. female[3]. We considered using the actantial model proposed by Greimas [11] which in turn is an abstraction of the corpus-based classification developed by Propp [24]. Greimas distinguishes between subject, object, helper, opponent, sender and receiver. But it seems to us that our dimensions allow us to capture the intuitions which are also the basis for Greimas. The first positive main character is usually the subject, while negative main characters are usually the opponents. Our approach avoids the classification problems which usually arise especially from the last two concepts.
2. To determine the interaction types relevant for the description of events in entertainment literature is probably the most challenging aspect. We start with the simple fact that a high proportion of these texts consists of direct speech. Add to this reported and free indirect speech, and depiction of communication comprises around 40-50% of narrative texts, depending on the genre. So the first type of interaction is (usually verbal) communication. In our first model, we don't distinguish between different verbal actions like love declaration and death threat. In the second, we use a representation which covers some aspects without making it necessary to explicitly construct the motif repertoire ourselves.

[3] The social construction of gender is a complex phenomenon, but entertainment literature usually simplifies this into a binary system; cf. the extensive discussion in Koolen [16].

3. This leaves the narrative text which is not conveying information about communication, but about other types of events. From this, for the first approach we only use two categories: The non-verbal expressions of positive affection (especially erotic interaction) and of antagonistic action (fight) are typical interaction types for dime novels. Again, for the second approach we used a simple more content-rich representation without making it necessary to explicitly construct the motif repertoire ourselves.

In short, scenes are identified and values for valence and arousal are computed for each scene. Then for each scene a character graph is constructed which represents the character dimensions and the interaction types. These scene graphs are then integrated into a temporal graph according to the sequence of scenes. The temporal graph allows to compute sequence-sensitive measures for characters which are added summarily to the scene (more complex representations of these informations are thinkable, but it is not easy to integrate them into the representation of a whole novel, see discussion).

This information is complemented with information on the scene, valence and arousal and the averaged centrality measure for the characters involved in the scene ('personal weight'). In a second approach we added to this general scene information the specific distribution of topics for direct speech and for the narrative text to add more concrete information about the genre specific interaction and event types. Using a topic model is a valid, but probably relatively crude way to construct a motif repertoire for the interaction and event types based on a corpus. This is one of the many points in this paper where we can only point to future research.

2. Corpus

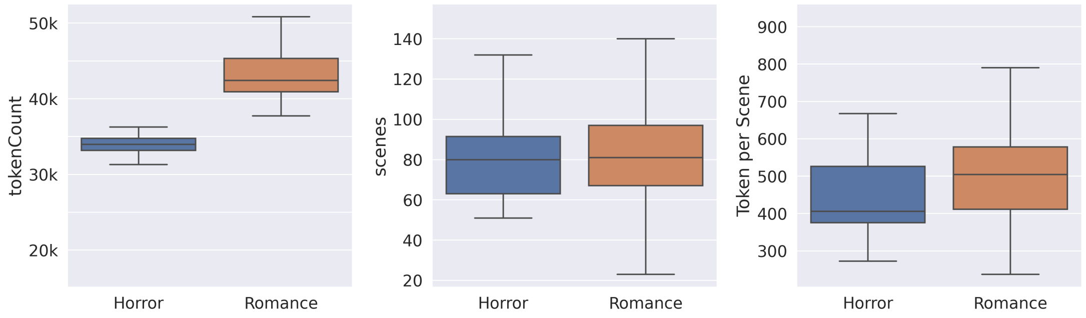


Figure 2: Corpus statistics.

The corpus consists of 192 dime novels from the genres horror (39) and romance (153). The novels are relatively short with an average length of 39k tokens. Despite being longer than horror novels, romances show the same amount of scenes. This is due to shorter scenes in horror novels (Fig. 2).

3. Methods

The next sections describe how to obtain the information to create the graph on a technical level[4]. For more details on pre- and post-processing, please see Appendix A.

3.1. Preprocessing

The foundation for the enrichment of our corpus is a pipeline containing a set of state-of-the-art NLP tools for the German language [6]. More precisely: Tokenization, Lemmatization, Sentence Splitting, Part-of-Speech Tagging, Morphological Analysis, Dependency Parsing, Named Entity Recognition, detection of direct, indirect, reported and free-indirect speech and Coreference Resolution. Scene segmentation is done outside of this pipeline with [19], the best contribution in the shared task 'scene annotation' 2021 [30].

3.1.1. Character extraction

The easiest way to determine if a character is present in a scene is to check if its name is mentioned. But characters are often mentioned even though they are not present. For the most common possibilities, we have created a filter so that only mentions are considered that a) are outside of verbal actions and b) outside sentences with past perfect tense. In addition, a character must perform at least one action (be the subject of a sentence), to be considered present. For the special case of first-person narration, we had to use an extra routine, since the narrator's name is mentioned only rarely. Therefore, if it is a first person narrative, all pronouns of the 1st person singular, which fulfil the above conditions, are added to the character "narrator". We treat the information whether it is a first person narrative and the name of the narrator as given metadata.

3.1.2. Action extraction

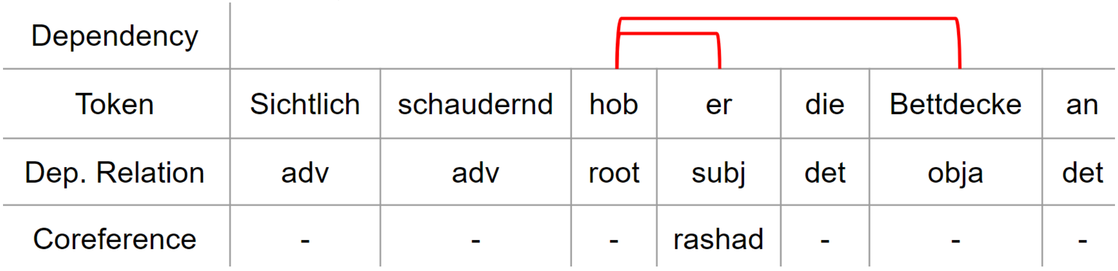
To capture actions of a character in a scene, all of its mentions are filtered by those the dependency parser has labelled as subject of a sub-sentence. The dependency tree is searched for the corresponding predicate and, if available, object of the clause (see Tab. 1). The query can resolve active and passive constructions. Auxiliary verbs are skipped in the dependency tree. If a sub-sentence is followed by a sub-sentence of the same order, which does not contain a new subject, the subject of the first sub-sentence is retained. The result is a set of subject-verb-object triples associated with a character and a scene. Sentences in past perfect tense or direct speech are ignored. If the object of a triple is also a character mention, it is replaced in the triple by its name.

3.1.3. Valence and Arousal

Valence and arousal are assessed using an affective norms word list [17]. The values for characters are calculated from the average of the values of all tokens in triples in the novel with

[4] Code and Data: https://github.com/LeKonArD/character_temp_graphs

Table 1: Example of Action Triple Extraction. Result: (rashad, heben, Bettdecke).transl: “visibly shuddering, he lifted the bed cover” -> (rashad, lift, bed cover)



Dependency							
Token	Sichtlich	schaudernd	hob	er	die	Bettdecke	an
Dep. Relation	adv	adv	root	subj	det	obja	det
Coreference	-	-	-	rashad	-	-	-

which they are associated. The values for scenes are calculated from the values of all triples within it.

3.1.4. Interaction Types

We identify three types of actions: Fighting, Erotic Actions and Talk. Combat and eroticism are determined by matching word lists on the subject-verb-object triples of a scene. How much is spoken in a scene can be directly determined by the output of speech recognition. Since scenes do not necessarily have only one interaction type, a score (e.g. relative share of words) for each type is calculated.

3.1.5. Character features

How to detect character appearances and thus also who appears alongside is already discussed above under ‘Character Extraction’. This representation is complemented by the valence and arousal values at the character level (see: Valence and Arousal). To differentiate between major and minor characters we use Temporal Closeness Centrality[5] [22]. As a sanity test, we identified the protagonists and their love interests in 20 novels and checked their values. The result shows: In all cases the protagonist has highest centrality and the love interest is second.

3.1.6. Topic Model

In order to add semantic information as a proxy for the motif repertoire to the predominantly structural model we resort to topic modeling [2]. Since our research corpus is not large enough to create our own topic model, we use a background corpus consisting of 10k other dime novels divided into segments of 500 tokens.[6] Following the reasoning that there is a fundamental difference between text and dialogue in scenes, we divide each scene into two documents based on this criterion. To underline this assumption we try to classify dialogue and text based on topic distribution. A logistic regression achieves a stable performance of an accuracy of .86 (std: 0.008).

[5] We used the python library Teneto [28] for the representation of the temporal graph and the computation of the centrality measure; for an explanation of temporal graphs and the measure see below.

[6] 1.7m documents, 4000 iterations, 150 topics

We used a temporal graph to represent the scene and character information and computed the Temporal Closeness Centrality (details see Appendix).

3.2. Evaluation

The evaluation of plot models proves to be especially challenging, because it is so time consuming. Ideally we would have for each text 3 or more structured summaries which cover all scenes. They would list the important characters and the important events (separately for direct speech and narrative text) for each scene, but would also indicate which scene could be left out as not or less relevant. Usually we base our evaluation on data sets with a few hundred instances, but in this case the compilation would take - even with pulp fiction novels which are only 64 pages long - almost prohibitively long. (In this context the data set described in [27] which has event annotations for 100 novels is especially noteworthy). Therefore we think that for some time at least research on plot has to use proxies. In this paper we use three approaches.

1. Because plot schemas for very different genres are usually easy to distinguish, the task to distinguish genre based on a structural plot representation can be used as a proxy. Basically we measured the average distance between texts of a genre and between all texts and we expect texts which belong to a genre to show a marked lower distance.
2. Similar to [7] and [25] we construct a second set of text representations where we randomly change the sequence of scenes. Here the task is to distinguish real novels from the artificial ones, in other words real novels should be more similar to each other than the artificial counterparts. We also inform about the distances between real and artificial texts split after genres to capture genre specific differences.
3. Formulaic genres often have recurring scenes which can be found in almost or all text instances. In romances, for example, there is always a scene in which the lovers meet for the first time. In pulp fiction horror, there is almost always a scene where the protagonist fights the evil antagonist. We take half of the romances in our corpus and identify those scenes, which describe the first meeting. Then we replace these scenes in 60% of the texts with another scene (B1), in 20% of the texts we don't change anything (B2), and in 20% of the texts we move the scene to the last third of the text (B3). Then we compare our text representations with the other half of the texts, which haven't been changed (A1). If the representation is capturing temporal information, we should see a higher similarity of B2 with A while B1 and B3 are less similar.

4. Experiments

The first experiment uses the evaluation task for genre differentiation. Four approaches (see Fig. 3) to plot representation are tested:

tf-idf. Word frequencies over the entire novel, weighted by tf-idf are the de facto standard for representing long texts and serve as a baseline. More specifically, we use the 5000 most frequent content words (nouns, verbs and adjectives). Similarity is calculated with euclidean distance of tf-idf vectors.

Global Characteristics. The second approach is based on properties of the entire novel, which are generated by queries on the temporal graph. Following properties are included: Number of characters, the average of fight score, erotic score, share of speech, arousal, valence, character centrality and proportion of characters genders over all scenes. Similarity is calculated with euclidean distance of all features.

Time Series. This representation models the plot of a novel as a multidimensional time series, where scenes are used as timesteps. Each timestep consists of the information on: number of characters, fight score, erotic score, share of speech, arousal, valence, character centrality and the proportion of characters genders. We measure similarity by applying multidimensional dynamic time warping with euclidean distance [26].

Temporal Graph. To measure similarity of temporal graphs directly, without condensing the available information to other formats (e.g. time-series), we make use of dynamic temporal graph warping (dtgw) introduced by [10]. Unfortunately, this measure does not use the node and edge weights and attributes in its calculation of similarity, only distances between unweighted edges are covered. Therefore, only the information about who appears in which scene is included in this calculation.

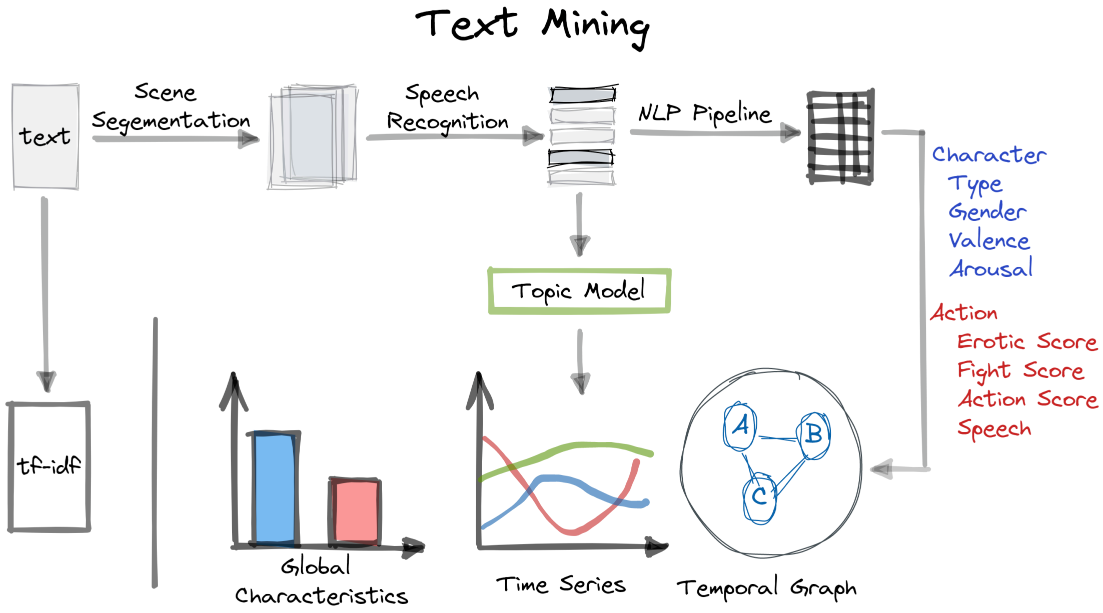


Figure 3: Overview on plot representation approaches. For details on the NLP Pipeline see Appendix A.

Figure 4 shows the results of the first evaluation task[7]. As expected, both genres are easily

[7] To avoid bias due to different group sizes, each experiment is repeated 500 times with ten randomly drawn novels from each group.

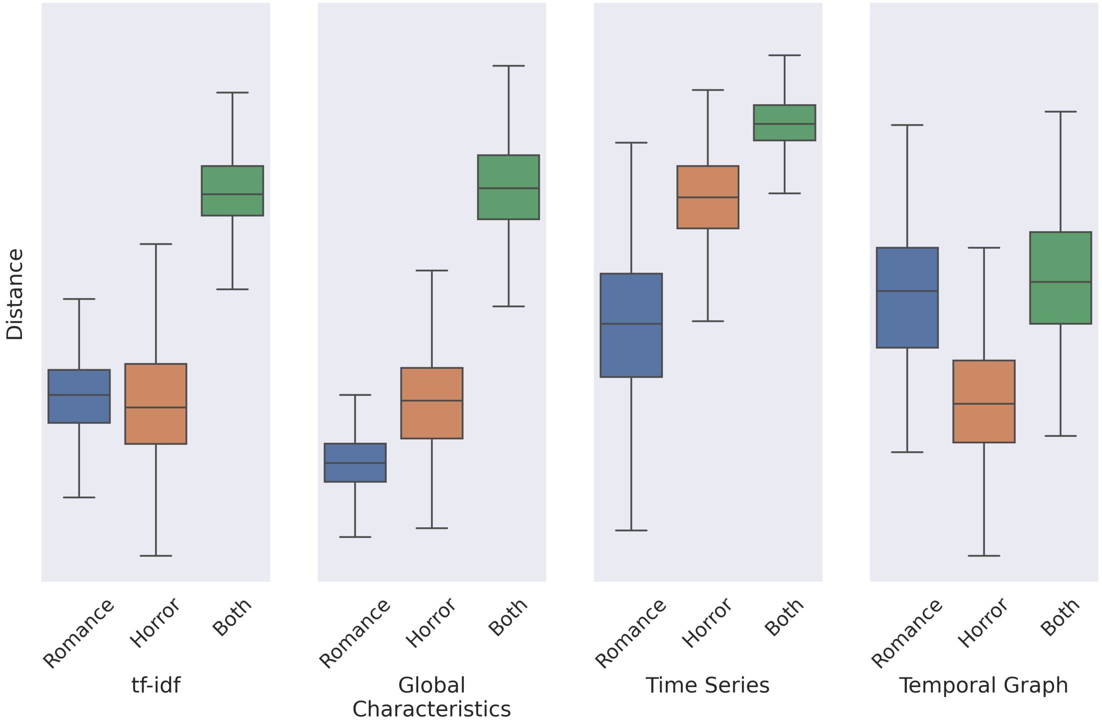


Figure 4: Results of evaluation task 1. Romance: Distance between Love novels; Horror: Distance between Horror novels; Both: Distance between Horror and Love novels. To pass the test, the distance between Love and Romance should be smaller than between both. The y-axis is not labelled because only the relations of the distances are relevant for the experiment and not their absolute values.

distinguished using tf-idf and Global Characteristics. The Time Series data is more blurry, but still passes the test, while the Temporal Graph representation fails.

The second and third evaluation tasks involve altering the sequence of scenes. Therefore it is not reasonable to test representations lacking sequential information. This limits us to the use of Time Series and Temporal Graph representation. Since the temporal graph has already failed at the first task, only Time Series is tested. In addition to the variant already used in test 1, we test whether the performance can be increased by supplementing the structural information with semantic information, our proxy for the motif repertoire. For this purpose, the distribution of topics in scenes (separated into narration and speech) is reduced to 4 dimensions and used as an additional feature of the time series. We also try to reduce the number of scenes by using only the 10 scenes[8] with the highest arousal value within a novel.

Figure 5 shows the performance of this setup in evaluation test 2. The reduction to essential scenes is clearly a harmful preprocessing step. The enrichment with information from the topic model has only a very small influence on the result. The same conclusions are valid for Evaluation Task 3 (see Fig. 6).

[8] We also tested 5 and 20 scenes, without noticing any big differences.

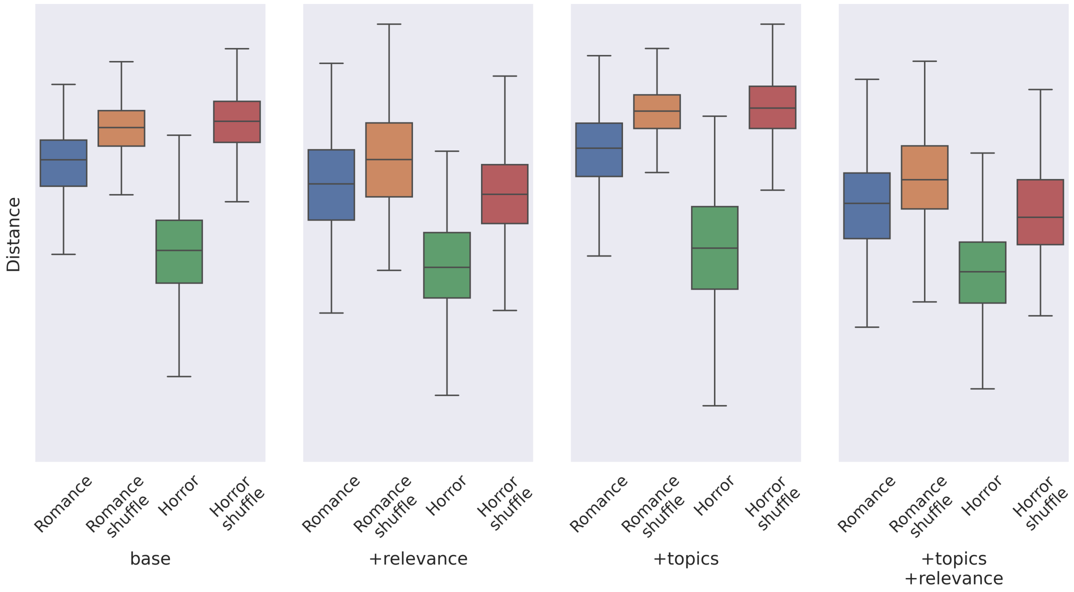


Figure 5: Results of evaluation task 2. Romance: Distance between Love novels; Horror: Distance between Horror novels; Romance shuffle: Distance between Love novels and shuffled love novels; Horror shuffle: Distance between Horror novels and shuffled Horror novels. To pass this test, the distance of Romance shuffle and Horror shuffle needs to be higher than their non-shuffled counterparts. (+topics: Topic Model Features included; +relevance: reduction to essential scenes)

5. Discussion

Most importantly, the result of the first experiment shows that the temporal structure even of the very reduced information we used to model plot is part of an overall plot shape which can be used to measure similarity of texts. The even more reduced version, in which we computed the similarity directly on the temporal graph, did not contain enough information. This validates the approach to represent plot based on the temporal information of the text, but it also indicates that temporal graphs are a useful way to represent the information but at the moment are not a good way to compute the similarity between texts.

Contrary to our expectations the addition of more concrete information about the motif repertoire of direct speech and narrative text in the form of topic models did not increase the similarity. It is unclear to us whether this is caused by an unsatisfying representation, in other words maybe the topic models did not capture the motif repertoire, for example because it lacks generalization. Anecdotal evidence suggests that this is the case for some motifs. In the romance novels we were able to identify a retarding plot element, namely the heroine's doubt as to whether the beloved is seriously interested in her at all. But the reasons for these doubts and the concrete ways these doubts are articulated are very different and have little in common on the surface of the text. Another reason for the low performance increase could be that the integration of the information about the motif repertoire into our scene representation was

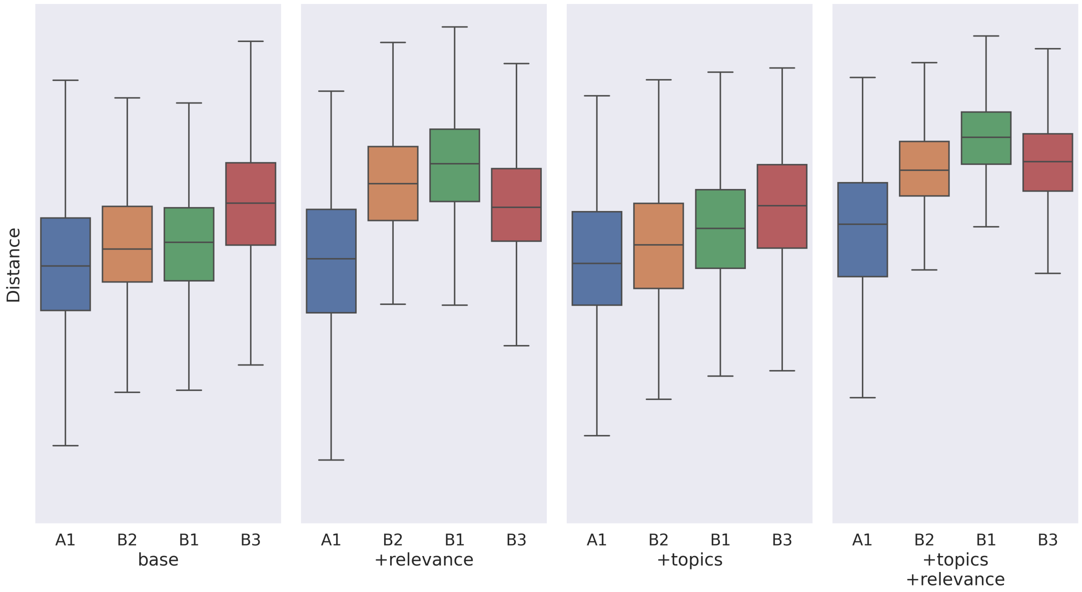


Figure 6: Results of evaluation task 3. A1: Distance between A1 and A1 (intra group distance). B2: Distance between A1 and B2; B1: Distance between A1 and B1; B3: Distance between A1 and B3. To pass this test B2 needs to be higher than A1 and lower than B1 and B3.

subpar, for example because the valid information is drowned in the noise of all the scenes and topics which a reader would filter out.

Also our attempt to detect the relevant scenes has not worked as intended. As the concept of relevance is also part of the more general problem of detecting the main elements of the plot, this problem is probably closely related to the problem of generalizing and abstracting the event information. There is a challenging relationship between the text specific use of the motif repertoire and the generalization necessary to allow the comparison of texts and the evaluation of similarity. The concreteness of the instantiation of the motif repertoire basically leads to an information overload.

We evaluated our scene representation by using a distance metric based on a similarity measure using dynamic time warping. It is unclear to us whether this measure is the best way to proceed. It looks at the whole time series allowing for differences in the temporal extension of the patterns. But most of the information may be actually noise under the perspective of reconstructing the human perception of similarity of narratives.

To proceed further in this direction the following research problems have to be solved in a more satisfying way:

- What is the best graph representation to include all relevant information and derive simpler views for computational purposes. A temporal graph alone is unsatisfactory, because then the information about the scenes has to be handled externally. So a bipartite graph may be a useful model, where one set of nodes and edges represent the temporal graph as in our approach and another set of nodes represent the scenes.

- Identification of those scenes which are crucial for the plot. A relevance score for each scene could be used to filter the relevant ones based on the level of abstraction intended.
- Abstraction and generalization of events. This is probably the hardest problem of all and can only be approached by annotating the motif repertoire for one genre more extensively. On this level also patterns of scene n-grams could be extracted, like 'captured-freed'.

- Abstraction and generalization of events. This is probably the hardest problem of all and can only be approached by annotating the motif repertoire for one genre more extensively. On this level also patterns of scene n-grams could be extracted, like 'captured-freed'.

- We need a clearer understanding of what uses the term 'plot' in literary studies (beyond the meta discussion in narratology) really has, for example in the construction of genres. Similarity of complex phenomena usually happens by comparing them under a specific perspective which ignores a lot of given information. To achieve this level of abstraction and generalization we should analyze how it is done in literary studies.
- In the long run, a real evaluation will have to be based on human judgment, that is annotations: Structured summaries of a genre corpus which will also create the motif repertoire for this specific corpus. These annotations could also be the ground truth for derived text formats as we used them in this paper (we basically just assumed that they work as intended). As each genre will have to create its own motif repertoires, working with these automatically derived formats will be unavoidable and needs to be put on a solid basis.

Additionally, the problems we did not touch upon in this paper have to be solved too, for example the temporal reordering of the scenes and the detection of narrative levels. As already mentioned in the introduction, the main contribution of this paper is not a solution to a problem, but a more extensive description of the aspects involved in the rather complex problem of plot. Its main purpose is to be used as the basis for the communication in CLS and to drive research in the many subproblems we outlined.

References

[1] D. Bamman, T. Underwood, and N. A. Smith. "A Bayesian Mixed Effects Model of Literary Character". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: Association for Computational Linguistics, 2014, pp. 370–379. doi: 10.3115/v1/P14-1035. url: http://aclweb.org/anthology/P14-1035.

[2] D. Blei, A. Y. Ng, and M. I. Jordan. "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3 (2003), pp. 993–1022.

[3] R. L. Boyd, K. G. Blackburn, and J. W. Pennebaker. "The narrative arc: Revealing core narrative structures through text analysis". In: *Science Advances* 6.32 (2020), eaba2196. doi: 10.1126/sciadv.aba2196. url: https://www.science.org/doi/10.1126/sciadv.aba2196.

[4] A. Brunner, S. Engelberg, F. Jannidis, N. D. T. Tu, and L. Weimer. “Corpus REDEWIEDERGABE”. In: *Proceedings of The 12th Language Resources and Evaluation Conference, Marseille*. Marseille, 2020, pp. 796–805. URL: http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.100.pdf.

[5] J. Egbert and M. Mahlberg. “Fiction – one register or two?” In: *Register Studies* 2.1 (2020), p. 72. URL: https://www.academia.edu/42908069/Fiction%5C%5Fone%5C%5Fregister%5C%5For%5C%5Ftwo%5C%5FSpeech%5C%5Fand%5C%5Fnarration%5C%5Fin%5C%5Fnovels.

[6] A. Ehrmanntraut, L. Konle, and F. Jannidis. *LLpro, A Literary Language Processing Pipeline for German Narrative Texts*. 2022. URL: https://github.com/aehrm/LLpro.

[7] M. Elsner. “Abstract Representations of Plot Structure”. In: *Linguistic Issues in Language Technology, Volume 12, 2015 - Literature Lifts up Computational Linguistics*. CSLI Publications, 2015. URL: https://www.aclweb.org/anthology/2015.lilt-12.5.

[8] M. Elsner. “Character-based kernels for novelistic plot structure”. In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Eacl ’12. Usa: Association for Computational Linguistics, 2012, pp. 634–644.

[9] G. Freytag. *Freytag’s Technique of the Drama: An Exposition of Dramatic Composition and Art. 4th ed. Chicago: Scott, Foresman and Co.* 4th. Chicago: Scott, Foresman and Co., 1908.

[10] V. Froese, B. Jain, R. Niedermeier, and M. Renken. “Comparing Temporal Graphs Using Dynamic Time Warping”. In: *Complex Networks and Their Applications VIII*. Ed. by H. Cherifi, S. Gaito, J. F. Mendes, E. Moro, and L. M. Rocha. Studies in Computational Intelligence. Cham: Springer International Publishing, 2020, pp. 469–480. DOI: 10.1007/978-3-030-36683-4_38.

[11] A. J. Greimas. *Structural Semantics: An Attempt at a Method*. Lincoln: University of Nebraska Press, 1983.

[12] P. Hühn. *Event and Eventfulness*. Ed. by P. Hühn, J. Pier, W. Schmid, and J. Schönert. Hamburg, 2013. URL: https://www-archiv.fdm.uni-hamburg.de/lhn/node/39.html.

[13] L. Jahan, R. Mittal, and M. Finlayson. “Inducing Stereotypical Character Roles from Plot Structure”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 492–497. URL: https://aclanthology.org/2021.emnlp-main.39.

[14] M. Jockers. *» A Novel Method for Detecting Plot Matthew L. Jockers*. 2014. URL: https://www.matthewjockers.net/2014/06/05/a-novel-method-for-detecting-plot/.

[15] A. Kazantseva and S. Szpakowicz. “Summarizing Short Stories”. In: *Computational Linguistics* 36.1 (2010), pp. 71–109. DOI: 10.1162/coli.2010.36.1.36102. URL: https://aclanthology.org/J10-1003.

[16] C. Koolen. *Reading beyond the female : The relationship between perception of author gender and literary quality*. Amsterdam: Institute for Logic, Language and Computation, 2018.

[17] M. Köper and S. Schulte im Walde. "Automatically Generated Affective Norms of Abstractness, Arousal, Imageability and Valence for 350 000 {G}erman Lemmas". In: Portorož, Slovenia: Erla, 2016, pp. 2595–2598.

[18] K. Kukkonen. *Plot.* Ed. by P. Hühn, J. Pier, W. Schmid, and J. Schönert. Hamburg, 2014. URL: https://www-archiv.fdm.uni-hamburg.de/lhn/node/115.html.

[19] M. Kurfalı and M. Wiren. "Breaking the Narrative: Scene Segmentation through Sequential Sentence Classification". In: *Proceedings of the Shared Task on Scene Segmentation.* Düsseldorf, 2021. URL: http://ceur-ws.org/Vol-3001/paper6.pdf.

[20] J. C. Meister. *Computing Action: A Narratological Approach.* 1 edition. Berlin ; New York: De Gruyter, 2003.

[21] O. Michail. *An Introduction to Temporal Graphs: An Algorithmic Perspective.* 2015. DOI: 10.48550/arXiv.1503.00278. URL: http://arxiv.org/abs/1503.00278.

[22] R. K. Pan and J. Saramäki. "Path lengths, correlations, and centrality in temporal networks". In: *Physical Review E* 84.1 (2011), p. 016105. DOI: 10.1103/PhysRevE.84.016105. URL: https://link.aps.org/doi/10.1103/PhysRevE.84.016105.

[23] A. Piper, R. J. So, and D. Bamman. "Narrative Theory for Computational Narrative Understanding". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.* Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 298–311. URL: https://aclanthology.org/2021.emnlp-main.26.

[24] V. Propp. *Morphology of the Folktale. Austin: University of Texas P.* Austin: University of Texas Press, 1968.

[25] N. Reiter, J. Sieker, S. Guhr, E. Gius, and S. Zarrieß. "Exploring Text Recombination for Automatic Narrative Level Detection". In: *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022).* 2022, pp. 3346–3353.

[26] M. Shokoohi-Yekta, B. Hu, H. Jin, J. Wang, and E. Keogh. "Generalizing DTW to the multi-dimensional case requires an adaptive approach". In: *Data Mining and Knowledge Discovery* 31.1 (2017), pp. 1–31. DOI: 10.1007/s10618-016-0455-0. URL: https://doi.org/10.1007/s10618-016-0455-0.

[27] M. Sims, J. H. Park, and D. Bamman. "Literary Event Detection". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* Florence, Italy: Association for Computational Linguistics, 2019, pp. 3623–3634. DOI: 10.18653/v1/P19-1353. URL: https://aclanthology.org/P19-1353.

[28] W. H. Thompson, P. Brantefors, and P. Fransson. "From static to temporal network theory: Applications to functional brain connectivity". In: *Network Neuroscience* 1.2 (2017), pp. 69–99. DOI: 10.1162/NETN_a_00011. URL: https://doi.org/10.1162/NETN%5C%5Fa%5C%5F00011.

[29] M. Vauth, H. O. Hatzel, E. Gius, and C. Biemann. "Automated Event Annotation in Literary Texts". In: *CHR 2021: Computational Humanities Research Conference, November 17–19, 2021, Amsterdam, The Netherlands*. Amsterdam, 2021, p. 13. URL: http://ceur-ws.org/Vol-2989/short%5C%5Fpaper18.pdf.

[30] A. Zehe, L. Konle, S. Guhr, A. Hotho, F. Jannidis, L. Kaufmann, M. Krug, F. Puppe, N. Reiter, and A. Schreiber. "Shared Task on Scene Segmentation". In: *Stss Konvens*. 2021, p. 21.

A. Pre and Postprocessing

Preprocessing. The output of the different preprocessing tools (Tokenization, Lemmatization, Sentence Splitting, Part-of-Speech Tagging, Morphological Analysis, Dependency Parsing, Named Entity Recognition, detection of direct, indirect, reported and free-indirect speech and Coreference Resolution) is carefully aligned and saved in conll-format. Scene segmentation is not (yet) part of this pipeline, therefore we tested both passing novels through the pipeline and segment afterwards or segment first and processing the segments individually. After reviewing the results, we conclude that a priori segmentation is preferable. From a theoretical perspective, the segmentation can only affect the pipeline steps NER, Speech detection and Coreference Resolution, since the other tools work on sentence and word level. The impact on NER and Speech Detection is negligible, considering the size of the context windows these tools use, since scenes are much longer. Coreference resolution, on the other hand, operates on the entire text. The idea that more text and thus more information about characters (alternative names, appellatives, gender) increases performance is obvious. However, according to our findings, it is beyond the corefenece model's capabilities to exploit this information over a long text. This agrees with the original authors' assessment that the memory capacity of the model is not sufficient for long texts. For example, we see that despite matching names, new corefence clusters are created or even worse all mentions of a paragraph are assigned to one cluster regardless of differing gender and names. This behavior is suppressed by a-priori segmentation. This is not surprising, considering that the definition of scenes in the dataset which was used to train the segmentation tool is strongly tied to stable character constellations.

Postprocessing. Both tools for scene segmentation (y: 0.17) and coreference resolution (F1: 64.72) are far from perfect. Nevertheless, we think they are good enough to work with. To improve the results a bit more we apply a number of post-processing steps. The biggest source of error in scene segmentation is over-segmentation, which leads to arbitrarily short scenes. To mitigate this, we set a lower limit for scene length of 200 words. If this is underrun, we merge a scene with its following one.

Coreference postprocessing is a bit more complex. First, the most frequent proper name of a cluster is set as its identity. Then all other proper names in this cluster are checked, if they have already been present in a previous scene, the mentions are assigned to this cluster. Then the grammatical gender is used. For example, if there are male mentions (pronouns) in a female cluster, they are assigned to the nearest cluster in the text with the appropriate gender. In this way, coreference resolution benefits from both: Information from preceding text and meaningful segmentation. In the case of first-person narratives, all first person pronouns (ich, mein, meiner, meine, etc.) are assigned to the predefined entity of the narrator. This is required since the model is not trained for this type of text and the narrator's name is rarely mentioned and if mostly inside of direct speech. Mentions of groups and clusters without proper names are ignored completely.

B. Modeling Temporality with Temporal Graphs

Temporal Graphs are an interesting extension to graph theory which has developed methods to represent and analyze static graphs - and in recent years an increasing amount of research is looking into the much more complicated situations of graphs which develop over time[21]. Temporal graphs add the dimension of time. Figure 7 shows a temporal graph as a sequence of static graphs. Each time step represents nodes and their links, in our use case the character constellation in one scene.

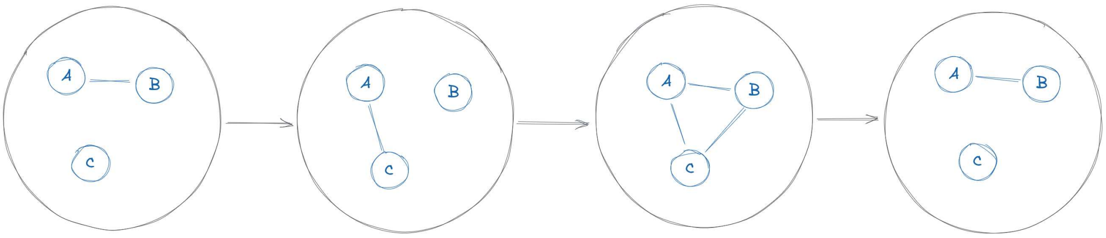


Figure 7: A simple temporal graph as a sequence of static graphs.

Figure 8 shows a variant of this visualization, where we substituted the explicit depiction of the edges with an implicit representation: Character nodes are only shown for those scenes, in which they are present in a scene and the interaction of the co-present characters in a scene is implied. Bill, Sheila, Suko and Jane are friends of the protagonist John. Harris is the antagonist, Clou, his helper, and Martha, Peter and Wayne (victims). The story is told mainly from the perspectives of Sheila and John.

Based on this representation as a temporal graph, we calculated the temporal closeness centrality for each character. Temporal closeness centrality [22] is a generalization of static closeness centrality. A high value of Ct indicates that other nodes can be easily reached from the node i.

Obviously it would be the best representation to add this centrality information to each character node. But similarity measures for temporal graphs are not able yet to handle node attributes but only work on the basic network structure. So we averaged the centrality measures for all characters and used it as a scene attribute.

Temporal graphs, which have been intensely researched in recent years, provide a rich medium to model all aspects of plot we are interested in. In our case the information described above can be transformed into a complex temporal graph. In order to realize its full potential, the graph needs several types of nodes (character types, scenes) and edges (interaction types),

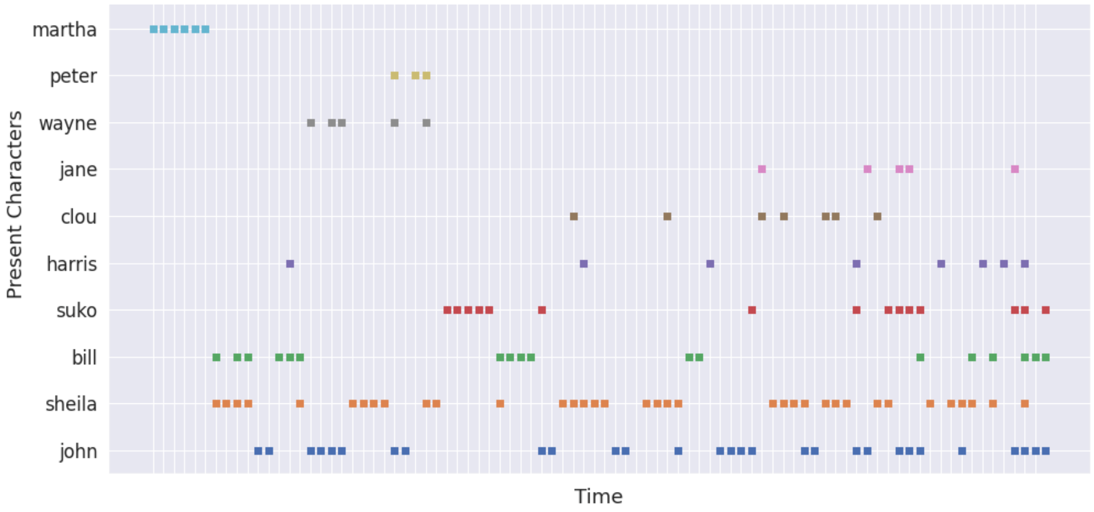


Figure 8: Temporal Graph of the horror novel "John Sinclair Nr.6: Anruf aus dem Jenseits" (Call from the beyond.)

as well as weighting of these edges. Unfortunately, the goal of representing the entire complexity leads to a model to which no methods are applicable. Therefore computation of measures will then be done on simplified views of these integral graphs.