Towards Annotating Narrative Segments

Nils Reiter
Institute of Natural Language Processing,
Stuttgart University
nils.reiter@ims.uni-stuttgart.de

Abstract

We report on first annotation experiments on narrative segments. Narrative segments are a pragmatic intermediate layer that allows studying more complex narratological phenomena. Our experiments show that segmenting on limited context information alone is difficult. High inter-annotator agreement on this task can be achieved by coupling the segmentation with summarization and aligning parts of the summaries to segments of the text.

1 Introduction

In this paper, we present ongoing work and first insights into the manual annotation of narrative segments. We introduce the notion of narrative segments as a pragmatic intermediate layer, that is a first step towards annotation of more complex narratological phenomena and has the prospects of being identifiable automatically. Furthermore, narrative segments can serve as an abstraction layer for applications such as social network extraction. If narratives describe connected events (Mani, 2012), we define a narrative segment as a coherent and separable sub-sequence of the events in a full narrative. A narrative segment ends, e.g., when place or time of the events change.

(1) [...] With a whirl of skirts and with the brilliant sparkle still in her eyes, she cluttered out of the door and down the stairs to the street.

Where she stopped the sign read: "[...]"

In (1), (O. Henry: *The Gift of the Magi*), an undefined amount of time passes between the character running down the stairs and stopping at the sign. Since the time and place of the events change, this would be the beginning of a new segment. Coincidentally, there is also a paragraph boundary at this position.

Working quantitatively with a specific theory requires annotations of text(s). Unfortunately, instantiating a theory such that it is annotatable is challenging (Hovy and Lavid, 2010), especially within Digital Humanities. The annotation process, however, can also be a productive way of validating and objectifying a theory. In this paper, we showcase how to systematically explore different ways of formalizing and annotating narrative segments, a category that is implicitly present in narratological theory, but not spelled out in detail.

2 Related Work

Related work to this paper falls in three areas: Segmentation of narrative texts (and the corresponding annotation efforts), narratology-driven annotation and discourse annotation. In the project Heurecléa[1], a corpus of German and English literary texts is being annotated (Gius and Jacke, 2014), following closely the narratological theories (Genette, 1980; Lahn and Meister, 2008). To our knowledge, annotations are still work in progress and not yet released.

There are publications about topical segmentation of narratives, for which annotated data has been created. Kazantseva and Szpakowicz (2014) have used a novel that has been annotated with topical segments by 3-6 people (differing by chapters). The authors report a mean pairwise segmentation similarity of 0.79. The evaluation data set used by Kauchak and Chen (2005) consists of two novels and is based on the chapter segmentation done by the authors of the novels. To our knowledge, there are no previous works on segmenting narrative texts into plot parts, which does not presume a topical shift.

There are annotated news corpora in the area of discourse, (Carlson et al., 2002; Prasad et al., 2008) that feature fine-grained discourse relations

[1] http://heureclea.de

Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 34–38,
Beijing, China, July 30, 2015.

between relatively small text spans. Although larger structures have been discussed in the literature (J. Grosz and L. Sidner, 1986) but not (yet) annotated. Move analysis (Biber et al., 2007) provides a framework for corpus-based study of discourse structures, but assumes discourse moves to be defined functionally and not by plot content.

3 Annotating Narratological Theory

3.1 Narratological Theory

Three time-related phenomena can be discerned: Order, duration and frequency (Genette, 1980). Narratives often deviate from the chronological *order* and include anachronies, e.g., flash-forwards (prolepsis). The emphasized sentence in (2), from *Chris Farrington: Able Seaman* (Jack London) shows a flash-back.

(2) The boats could not be back before midnight. *Since noon the barometer had been falling* [. . .], [and signs were ripe for a storm.]

Many narratives contain slow and fast parts since the *duration* of different parts of the narrative varies. This is formalized as the relation between story time (ST, the time that passes within the story) and narrating time (NT, the time "consuming" the story takes). The phenomena *pause* ($ST = 0$), *slow down* ($ST < NT$), *scene* ($ST = NT$), *summary* ($ST > NT$) and *ellipsis* ($NT = 0$) are straightforwardly distinguished. The emphasized sentence in (2) is also a summary, because the falling of the barometer (ST) has been taking a lot longer than to read the sentence (NT).

The term *Frequency* is used to describe the relation between the number of times an event happens (n) within the story and the number of times it is narrated (m). Schematically, one can distinguish five cases: (i) $n = 1 = m$, (ii) $n = 1, m > 1$, (iii) $n > 1, m = 1$, (iv) $n = m > 1$ and (v) $n > 1, m > 1, m \neq n$.

These categories – anachrony, pause, . . . – are not categories of the entire text, but of specific "narrative segments" (Genette, p. 35). These implicitly assumed narrative segments are not defined in any way by Genette. However, the detection of such segments is a prerequisite in order to investigate these phenomena.

3.2 Annotation Setup

To formalize the notion of narrative segments, there are a number of aspects to consider. Our aim is developing a formalization that is both theoretically motivated and can be annotated reliably.

Exp.	Context	Autom. Task	Annotators
1	10 sent.	classification	non-experts
2	full text	segmentation	students
3	full text	summ. align.	students

Table 1: Experiment Overview

Context Knowledge In contrast to most linguistic concepts, which are done with a limited amount of context, full text knowledge is an underlying assumption in literary studies. Requiring annotators to have full text knowledge makes the annotation process slower. In crowd sourcing, it is hard to control whether annotators will have read the entire (possibly long) text.

Annotation Unit and Task The annotation unit is the text portion that is annotated, i.e., assigned to a given category. In NLP, these are usually defined in linguistic terms, e.g., sentences, phrases, or tokens. The theoretical literature in narratology does not presume a fixation on a linguistic unit, but instead allows freedom on the selection of the actual unit. The examples shown by Genette (1980) range from noun phrases ("the prospect of a war", prolepsis) to multiple sentences. The decision on the annotation unit also influences the task this problem can be cast as for automatization: Annotating full sentences would allow casting as a classification task in the future, while allowing free spans to be annotated would lead to a segmentation task.

Annotator Selection Crowd sourcing experiments allow asking non-experts for their intuitions. This requires to break down the annotation task such that knowledge of theory or terminology are no longer required. Also, the amount of time that workers spend can not be fully controlled. Expert annotations are harder to organize, but ideally allow annotating higher level concepts and use of domain terms.

4 Annotation Experiments

We conducted three experiments to explore different ways of setting up the task regarding the aspects discussed above. In all experiments we ask annotators to detect narrative segments and calculate inter-annotator agreement as a measure of

Does the yellow sentence start a new narr. unit?
A narrative unit starts, whenever

- the speed of narration changes (e.g., more time passing than before as in "Ten days later, . . . "),
- time and place change (e.g., flashbacks as in "Ten years ago, I was a successful businessman in . . . "), or
- the narrator changes (e.g., longer segments of direct or indirect speech, attributed to a character in the narration; internal monologue).

Figure 1: Worker instructions in Exp. 1

the "annotatability". Table 1 shows a schematic overview of the experiments.

4.1 Experiment 1: Crowd Sourcing

The first experiment was conducted as a crowd sourcing classification task using CrowdFlower[2]. The workers were presented a sentence (in yellow) within a context of ten sentences before and after. They were given a yes/no question, but with an additional "I can't tell" option. The workers annotated all sentences from two narrative texts, *Chris Farrington: Able Seaman* (J. London) and *The Winepress* (J. Essberger), in random order. The exact definitions are shown in Fig. 1. Due to difficulties in automatic parsing, we opted for annotating full sentences in this experiment.

Results and Discussion In total, we collected 1,763 ratings from 315 different workers, for $ 64. Of these ratings, 1,406 (79.8%) are of the non-new class, 339 (19.2%) of the new class. Our data set included eleven test questions and the following results are based on the five ratings for each item from the most trust-worthy workers (measured against the test questions).

We evaluate the workers' performance using inter-annotator agreement Fleiss' κ (Fleiss, 1971) and show the fraction of different kinds of majority cases. The results can be seen in Table 2. The workers achieve a κ-agreement of 0.27 and 0.21. In part, the low score can be explained by skewedness of the task – most sentences are of the same category (not starting a new segment), which makes the chance-agreement very high (0.67 and 0.73). There is a large portion (57.8% and 44.7%) of sentences where all workers are in agreement.

[2] https://www.crowdflower.com

Text	*Seaman*	*Winepress*
Fleiss' κ	0.27	0.21
5|5-agreement	57.8%	44.7%
4|5-agreement	28.3%	34.9%
3|5-agreement	13.9%	20.3%

Table 2: Quantitative analysis results of Exp. 1

Manual inspection revealed that most disagreement cases are sentences involving direct speech and thoughts representation or giving background information (3). These cases were not covered by the guidelines.

(3) The *Sophie Sutherland* was a seal-hunter, registered out of San Francisco, [. . .].

4.2 Experiment 2: Student Annotators

In this experiment, we collected two annotations for each of 19 short stories from (paid) students of German literature. As a general design change, we asked the annotators to first read the entire text and only make boundary annotations in a second step. We also made several definitions for cases that were difficult in previous experiments: a) Dramatic scenes (dialogues) typically belong to a narrative segment, b) encyclopedic parts (e.g., landscape descriptions) and c) events that are not "really" happening in the narrative (e.g., thoughts, possibilities) can constitute segments on their own.

Additionally, we allowed the annotators to mark segment boundaries on different levels. This allows finer distinction between segment boundaries of different granularities. We asked the annotators to first mark the most clear, top-level segmentations and in a second (and third) step subdivide the segments into smaller pieces. A boundary of level n is also a boundary of level $n + 1$.

Corpus The stories have been selected randomly out of the TextGrid[3] corpus, the only restrictions being on the genre (narratives) and length ($2k$ – $12k$ tokens). In total, the corpus contains 4.692 sentences (avg. length: 21.9 tokens).

Agreement We calculated κ agreement using boundary similarity (Fournier, 2013) as a measure for observed agreement [4]. Boundary similarity is

[3] http://www.textgridrep.de

[4] Since chance agreement is very low (< 0.1) the numbers in the table are almost identical to boundary similarity.

Text	κ-Agreement per level 1	2	3
1009	0.409	0.517	0.478
14	0.393	0.375	0.28
weighted avg.	0.263	0.384	0.347

Table 3: Annotator agreement in Experiment 2

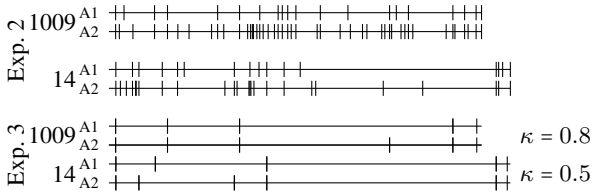

Figure 2: Segmentation Annotations

based on an edit distance measure and penalizes near misses less than full misses. We used a near miss window of two avgerage sentences (44).

Results and Discussion The κ agreement scores for two individual and all texts can be seen in Table 3, separated by level. In general, we take these results as an indicator that narrative segments are something that annotators can agree upon, but that there is some room for improvement of our guidelines and definitions. Regarding the different levels of segmentation, we have to note that the annotators did have different understandings of these levels and used them very differently. This can be seen in the fact that the agreement on level 3 is higher than on level 1.

Interestingly, the total number of boundaries annotated by the annotators do not differ that much: A1 added 3.825 boundaries, while A2 added 4.293. Although it was not required or suggested, the majority of boundaries fall on sentence boundaries (A1: 67.4%, A2: 80.2%). Most of the remaining boundaries are annotated on clause boundaries.

4.3 Experiment 3: Summary Annotations

In the third experiment, we asked the annotators from the second experiment to summarize the text and then align parts of the summary with specific text segments. The idea behind this experiment was to couple the segmentation task with a "real" task that makes sense outside of the annotation task and guides decisions on granularity. We evaluated only the (now implicit) segmentation of the texts, using the same measures as before. An advantage of this setup is that the summaries allow insight into the annotators' intentions.

Results and Discussion Figure 2 shows the resulting segmentations of the two annotators and the corresponding agreement scores on the right side. In terms of the scores, the agreement is much higher than in Exp. 2. All annotated segment boundaries fall on sentence boundaries. Since the annotators have participated in Exp. 2, they are more trained than before. As the two stories were not discussed in group meetings, they should not be biased towards specific segmentations.

In the figure, we can also compare the segmentations of the same texts in Exp. 2 for each annotator. As can be seen, the annotators produced much larger segments in the third experiment, while annotator A2 still created a finer segmentation. The only remaining difference among the segmentations on text 1009 can be explained with the help of the summaries: An event that is deemed important by one annotator is not even mentioned by the other and therefore, not summarized separately. In this way, the two segmentations reflect also on different literary interpretations of the texts.

5 Conclusions and Future Work

We presented first annotation experiments on narrative segmentation. We see it as i) a prerequisite step towards quantitative analyses of complex phenomena from narratological theory and ii) useful for applications (e.g., social network extraction). Furthermore, systematically exploring different possibilities in formalizing concepts from humanities theories in this way can help bridge the gap between theoretical concepts and annotatable categories. Although events play a major role in narratives, we are aiming for pragmatic annotations that tap into intuitive understanding of narratives without presuming event annotations.

Our annotation experiments indicate that annotating segment boundaries in isolation is difficult. However, when coupled with a more involved task (like summarizing a narrative), higher agreement can be achieved and also allows insight into the intention of annotators. In the future, we will extend these experiments on annotation and use these annotations as test and training data for automatic segmentation of narrative texts.

Acknowledgements

We thank our annotators Hannah Franz and Dominik Wabersich and our collaboration partner Marcus Willand.

References

Douglas Biber, Ulla Connor, and Thomas A. Upton. 2007. *Discourse on the Move*. Number 28 in Studies in Corpus Linguistics. John Benjamins Publishing Company, Amsterdam.

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

J.L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):420–428.

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. 1980. *Narrative Discourse*. Cornell University Press, Ithaca, New York. Translated by Jane E. Lewin.

Evelyn Gius and Janina Jacke. 2014. Zur annotation narratologischer kategorien der zeit. Annotation Manual 1.0, Hamburg University, January.

Eduard Hovy and Julia Lavid. 2010. Towards a 'science' of corpus annotation: A new methodological challenge for corpus linguistics. *International Journal of Translation Studies*, 22(1), January.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, 12(3):175–204, July.

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, June. 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, August. Dublin City University and Association for Computational Linguistics.

Silke Lahn and Jan Christoph Meister. 2008. *Einführung in die Erzähltextanalyse*. Metzler, Stuttgart, Germany.

Inderjeet Mani. 2012. *Computational Modeling of Narrative*, volume 5 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers, December.

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.