CA Journal of Cultural Analytics

A Shared Task for the Digital Humanities Chapter 1: Introduction to Annotation, Narrative Levels and Shared Tasks

Nils Reiter, Marcus Willand, Evelyn Gius

08.20.19

Article DOI: 10.22148/16.048

Journal ISSN: 2371-4549

Cite: Nils Reiter, Marcus Willand, Evelyn Gius, "A Shared Task for the Digital Humanities Chapter 1: Introduction to Annotation, Narrative Levels and Shared Tasks," Journal of Cultural Analytics. November 4, 2019. doi: 10.22148/16.048

Annotation guidelines for literary phenomena are a clear desideratum within the field of text-oriented digital humanities. Creating guidelines that are widely applicable, however, is almost only possible in large annotation projects, which are naturally expensive. Moreover, scholars interested in large-scale analyses of literary texts are required to perform a lot of tasks that are outside of their core expertise, while researchers from computer science interested in method development for literary texts are required to create annotated data by themselves. Shared tasks, a workshop and research format that is popular in natural language processing, are a way to address both issues at the same time. This volume documents the setup and the results of the first shared task conducted within the digital humanities. The shared task started in May 2018 and is the first one that has the development of annotation guidelines as its main goal.

This special issue comes in two volumes. The first one is structured as follows: In this introduction (Chapter 1), we will cover the goals and underlying motivations of the project, describe basic assumptions on (this kind of) annotation, give background on narratological theory and on the role of narrative levels in text analysis and introduce our shared task procedure. Chapter 2 ("Evaluating Annotation

Guidelines") explains how the submitted annotation guidelines have been evaluated. This also includes a description of the metric used for inter-annotator agreement. Chapter 3 ("Annotation Guidelines Overview and Evaluation Results") provides a structured overview and comparison of the guidelines and presents the evaluation results. The remaining chapters document the annotation guidelines and contain an introductory rationale for each guideline and a review. The guidelines are published as they were submitted (besides layout and minor language editing). Thus, the evaluation results are based on the guidelines you find in this volume.

In sum, this first volume documents the preparatory work and the results of the workshop we held to complete the first shared task. Since the discussions and insights of this workshop gave rise to a large number of improvements to the guidelines, we decided to publish the revised guidelines as well. The improved guidelines, which document the final outcomes of the first shared task, will be published in the second volume of the special issue.

Please note that this shared task (called SANTA, for "Systematic Analysis of Narrative levels Through Annotation") will be followed by a second one, with the goal of automatic detection of narrative levels.

Motivation

This project addresses two issues prevalent in digital humanities and computational literary studies: The distribution of labor, competences, and tasks in the interdisciplinary research field of digital humanities and the inter-subjective manual and reliable automatic recognition of narrative levels in narrative texts.

Distribution of Labor, Tasks and Competences

Given the current state of computational analysis of narrative texts[1] digital humanities projects that aim at analyzing content-related aspects of such texts *on a large scale* need to make technical-methodological progress in order to automatically detect the phenomena of interest. Therefore, many such projects are collaborative projects between researchers from computer science/natural language processing and literary or cultural studies. Although there is a growing number

[1] Performance on narrative texts is not systematically evaluated, but can be expected to be less than what is considered the state of the art: https://nlpprogress.com

of tutorials, how-tos and textbooks for various digital humanities topics,[2] the daily organization of such digital humanities projects remains challenging for a number of reasons:

Developing a shared language and common understanding of the subject at hand is one of the first tasks that new digital humanities projects often have to tackle. At times, computer scientists are only interested in the methodological part (without interpreting the results in reference to the texts under examination), while humanities scholars typically focus on conceptual issues or interpretation of the results. Thus, the individual goals of partners might be different even within the same project.

We believe that formats such as this adapted shared task offer unique opportunities to members of the digital humanities community with both backgrounds. In such a shared task, participants can focus on what they do best. Literary scholars can focus on the literary phenomenon that they are interested in and experienced with. Given their disciplinary routines and text experience, they are best qualified for exploring, defining, and exemplifying the narratological concepts without worrying about the implementability of their concepts or about making their findings automatable. Restricting oneself to simpler concepts just because one thinks they might be easier to detect automatically is a dead end for methodological innovation, as the limitations of computers are often only hearsay and constantly evolving. Moreover, if the conceptual complexity has been included in an annotation guideline that can be applied inter-subjectively, and a corpus with annotated concepts has been created, any computer scientist and/or machine learning expert can work on the automatic detection of the concepts, even if they are not experts in narratology (because a "ground truth" is available in the annotations). Similarly, as the shared task provides an empirical evaluation that can be trusted, machine learning models do not have to be transparent or explainable. When applying machine learning in a digital humanities scenario, there is often a trade-off between performance and transparency: Machine learning models that achieve better performance (e.g. neural networks) may be less transparent, while transparent models (e.g. decision trees) often lack in performance. In this case, because of the empirical evaluation, computer scientists can opt for the best performance.

[2] Among many others, cf. Susan Schreibman, Ray Siemens and John Unsworth, eds., *Companion to Digital Humanities* (Blackwell, 2004), http://www.digitalhumanities.org/companion/; Ray Siemens and Susan Schreibman, eds., *A Companion to Digital Literary Studies* (Oxford: Wiley-Blackwell, 2008); Matthew Lee Jockers, *Text Analysis with R for Students of Literature* (Cham: Springer, 2014); Fotis Jannidis, Hubertus Kohle, and Malte Rehbein, eds. *Digital Humanities. Eine Einführung* (Stuttgart: Metzler, 2017); Anandi Silva Knuppel and Maria José Afanador Llach, eds., "Programming Historian," accessed January 17, 2019, https://programminghistorian.org.

The two shared tasks we are organizing focus on two different sides of annotation. The first task, which consists of guideline development, forms the basis for an independent and reliable empirical evaluation of the automatic detection systems later on. Thus, a machine learning model that performed well in the second task may be safely used for new texts of the same kind as the test data (which is again transparent to scholars).

In addition to allowing everyone to focus on their field of expertise, this setup also renders a decoupling of the conceptual from the technical work possible. Scholars can focus on the development of annotation guidelines. This includes conceptual work, as well as a first step to operationalize scholarly concepts (to the extent of being applicable in an intersubjective manner). In this shared tasks model, the scholars do not have to be in the same project, at the same university or even on the same continent as the researchers developing the automatic detection tools (which includes technical work). This lowers entry barriers, as one does not have to work in a well-funded, interdisciplinary project in order to contribute to the overarching goals. Instead, scholars and researchers can contribute to the shared task at their own pace and integrate this single contribution more easily into their own research agendas. Moreover, this is possible without an augmentation of the workload in interdisciplinary collaborations.

Annotation Guidelines for Narrative Levels

The detection of narrative levels and through that the identification of coherent text parts is required for the analysis of narrative texts to facilitate subsequent, content-related literary research based on the data obtained (about plot, characters, narrated world, etc.). While there is no exact statistics on this, narrative levels are such a common phenomenon that they are very often not even explicated in literary studies. Thus, automatically detecting narrative levels is a crucial contribution and groundwork research in the field of computational literary studies. Moreover, narrative levels can be a mediator connecting hermeneutic and automatic text analysis. Even though the complexity of narrative levels is considered comparably low from a literary studies point of view and comparably high from a natural language processing perspective, it is potentially relevant for text analysis of all sorts. Additionally, in comparison to other phenomena narrative levels are a rather little disputed phenomenon within literary studies. Finally, the definitions of narrative levels are usually based on textual and narrative features. For example, verbs of utterance and subsequent direct speech can be textual signals for narrative levels as well as the presence of a different story world that can be identified through the analysis of space or other narrative phenomena. Narrative

levels are therefore useful for the analysis of texts displaying a divergence between their textual structure and the structure of the narrated.

In sum, we consider narrative levels a good choice for a shared task. Its most important quality for our purposes is that it can bridge the gap between the theoretical discussion of a phenomenon and the application in text mining.

Most of today's text processing software is based on machine learning of various types. Given their interdependence with other phenomena as well as surface and content level characteristics, machine learning is the prime technique to automatically detect levels in texts. Machine learning, however, can only be successfully applied if training and testing data is available in high quantities. Such data needs to be annotated, i.e., it is a necessity to have texts in which narrative levels are already marked. These annotated texts can then be used to train models to detect levels in new, not previously annotated texts.

Annotation guidelines are needed not only to ensure the coherence of the annotations, but also to deal with unusual cases and to allow the annotation to be done by non-experts. Since annotation processes are expensive and time-consuming, it is unrealistic that different theoretical approaches to one concept (e.g. narrative levels) will be used as a basis for the annotation of this concept. Thus, a certain conceptual agreement within the community needs to be reached beforehand. Ideally, it should be one that leads to annotations which are useful for as many scholars as possible, even the ones with a different theoretical understanding of the concept. In the case of narrative levels, this can be shown by the question of whether simple or complex level concepts should be used. On the one hand, a less complex concept might reach higher inter-annotator agreement, but on the other hand, it might also lead to annotations that are less interesting for literary scholars to work with, as more differentiated concepts are thought to cover more complex literary phenomena.

In sum, annotation guidelines are a core ingredient towards automatic recognition of a concept. To ensure the scholarly usefulness of the resulting automatic recognition tools, experts in narratology need to be involved in the process of guideline creation.

Annotation

The term "annotation" is used with different meanings within the digital humanities community. In our project, the term is used for the process of marking segments of a text to belong to a defined category. We also assume that such

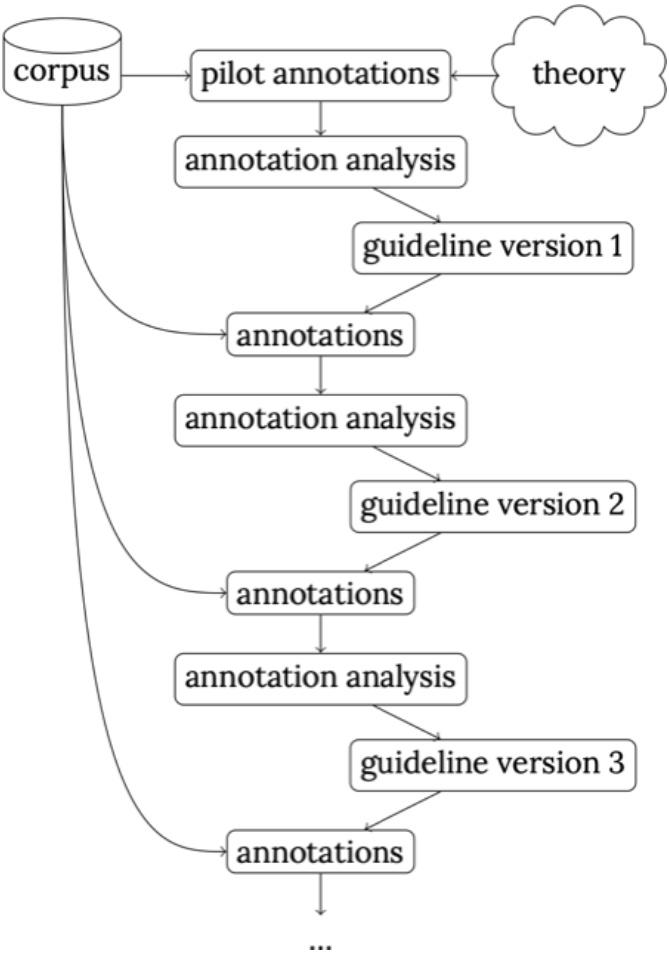
categories are determined beforehand (ruling out exploratory or explanatory annotation) and that their detection is based on the contents of the text and not on structure or formatting (ruling out the annotation of e.g. text structure in TEI XML).[3] This also entails that detecting these categories is not trivial and requires text understanding and a certain level of text interpretation.

This notion of annotation is most similar to the linguistic notion of annotations of, for instance, coreference chains or semantic roles.[4] There are, however, a number of properties of narrative annotations that need to be taken into account for the annotation workflow: A narrative level may need to be annotated in a large portion of the entire text, while semantic role fillers are typically constrained to single noun phrases. In addition, the relevant context is typically much larger. For the linguistic annotation tasks that have been subject of shared tasks in the past, a context window of a single sentence is sufficient. Annotating coreference chains is the exception here, as it is typically considered a document level task and requires full document knowledge. Narrative annotations regularly consider the entire document as relevant co(n)text, thus requiring full text knowledge of the annotators. It is entirely conceivable (but not easy to implement in reality) to also consider text-external sources as relevant context (e.g., socio-historic conditions). This larger context has the potential to make narrative annotations more interpretative than linguistic ones.

[3] http://www.tei-c.org

[4] Annotating coreference chains is the task of identifying which mentions of some entity refer to the same one (e.g., in "A house was bought by Mary. Peter loves her", the pronoun "her" refers to Mary). Identifying semantic roles would tell us that Mary is the agent of the first sentence, and "a house" is the patient or theme (i.e., the thing that has been bought).

Annotation Process



The annotation process that we have in mind is iterative and tightly connected to the development of a guideline. This iterative process is depicted in Figure 1 and is clearly related to the MATTER cycle.[5] In each step, we not only increase the amount of annotated texts, but the annotation guideline is improved as well. Of course, changes in the annotation guideline need to be reflected: They might change how previous portions of the texts should have been annotated, which

[5] J. Pustejovsky and A. Stubbs, *Natural Language Annotation for Machine Learning* (Sebastopol, CA: O'Reilly Media, Inc., 2013), 23ff.

should then be updated as well. The core idea in this annotation process—with the goal of producing coherent and inter-subjective annotations—is to have multiple annotators annotate the same texts in parallel, at least for some of the data. This allows the inspection and comparison of annotations and thus the identification of issues in the guideline. While asking the annotators for their impressions on the annotation process is valuable, not all issues are easily noticeable by annotators. Comparing annotations of the same texts with the same annotation guideline quickly reveals these possible issues. This annotation workflow has been employed by Evelyn Gius and Janina Jacke[6] on narrative time phenomena and it has obvious parallels to the "hermeneutic circle" that describes a general epistemological pattern in the humanities. If used in this way, the annotation workflow (and the iterative refinement of the annotation guidelines) has repercussions on the theoretical level and can be used productively for the development and refinements of theoretical concepts.[7]

Annotation Guidelines

The goal of this annotation process is to produce coherent and systematic annotations. To this end, the annotations are done with the help of annotation guidelines. Annotation guidelines mediate between a specific theoretical understanding of concepts (like that of a narrative level) and the practical annotation of the concept in texts. They have multiple purposes, all of them directed towards the explication of theoretical concepts and/or the process of annotation:

1. Fill the gaps: Theories are often not specific enough to be used directly. In order to be as abstract as possible, they typically neglect many details and leave them underspecified (e.g. how to handle dashes marking insertions). These cannot be decided individually by annotators during the annotation process and thus need to be defined beforehand.
2. Provide examples: Ideally, an annotation guideline makes it possible for non-experts in narratology to also perform annotation. To this end, examples are provided, and/or replacement/insertion tests are formulated.
3. Make text-specific adaptations: Even for relatively simple phenomena in linguistics (e.g. parts of speech), existing annotation guidelines cannot be

[6] "The Hermeneutic Profit of Annotation: On Preventing and Fostering Disagreement in Literary Analysis," *International Journal of Humanities and Arts Computing* 11, no. 2 (October 2017): 233-54.

[7] Cf. Janis Pagel, Nils Reiter, Ina Rösiger, and Sarah Schulz, "A Unified Annotation Workflow for Diverse Goals," in *Proceedings of the Workshop on Annotation in Digital Humanities, co-located with ESSLLI 2018*, ed. Sandra Kübler and Heike Zinsmeister (Sofia, Bulgaria, 2018) for a general-purpose workflow description.

expected to be all-encompassing, because the variability and creativity of human language production is enormous, and new text types are appearing constantly (consider part of speech tagging on twitter data). Annotation guidelines are a means to address phenomena that are text or genre specific.

4. Provide a log: Finally, most annotation processes accumulate a lot of procedural knowledge, as decisions on edge cases have to be made on a daily basis. An annotation guideline also serves the purpose of a log to document these decisions and make them traceable by other researchers.

Annotation Analysis

Agreement between annotators is a major goal of this kind of annotation: Two annotators, who annotate the same text with the same annotation guideline are *generally* expected to produce the same annotations.[8] Inspecting annotations with respect to their achieved agreement is consequently a major component of the annotation analysis step in Figure 1.

The regular **discussion of annotation decisions** with the actual annotators is an effective way of learning about issues in the guideline. Asking annotators to explain their decisions (in particular if they have been diverging or difficult) not only keeps their attention up, it also reveals misunderstandings and/or highlights areas in which the guideline can be improved.

In addition, the amount of agreement between annotators can be quantified. This is known as inter-annotator agreement, and numerous metrics have been proposed for different kinds of annotation tasks.[9] All metrics aim at striking a balance between observed agreement and expected agreement. While the former

[8] There are exceptions, in particular regarding literary texts. In these cases, polyvalent text readings might lead to different annotations which constitute a case of justified disagreement. cf. Evelyn Gius and Janina Jacke, "The Hermeneutic Profit of Annotation: On Preventing and Fostering Disagreement in Literary Analysis" *International Journal of Humanities and Arts Computing* 11, no. 2, (2017), 233-54.

[9] Joseph L. Fleiss, "Measuring Nominal Scale Agreement Among Many Raters," *Psychological Bulletin* 76, no. 5 (1971): 420-28; Jacob Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement* 20, no. 1 (1960): 37-46; Chris Fournier, "Evaluating Text Segmentation Using Boundary Edit Distance," *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Sofia, Bulgaria: Association for Computational Linguistics, 2013), 1702-12, http://aclweb.org/anthology/P13-1167; Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier, "The Unified and Holistic Method Gamma (⊠) for Inter-Annotator Agreement Measure and Alignment," *Computational Linguistics* 41, no. 3 (2015): 437-79; see Ron Artstein and Massimo Poesio, "Inter-Coder Agreement for Computational Linguistics," *Computational Linguistics* 34, no. 4 (2008) for an overview.

expresses how well real annotators agree, the latter expresses how much annotations would overlap if they were done at random. Thus, the actual, observed agreement is set in relation to the difficulty of the annotation task. The reasoning behind this is that, for instance, it is much easier to achieve agreement if there are only two categories than if there are 25 categories. Thus, the expected agreement (a.k.a. chance agreement) for two categories is higher than for 25 categories, which lowers the inter-annotator agreement if the observed agreement remains stable. Most inter-annotator agreement metrics are in the interval [-∞:1], in which values above zero express that the annotators agree more than chance agreement.

Measuring inter-annotator agreement for higher level tasks properly is not as easy as it sounds. This is due to the fact that many such tasks are actually composed of multiple subtasks and require the annotators to make multiple decisions in sequence. Annotating named entities, for instance, requires annotators to first find a segment that is a named entity, and secondly, to categorize this segment into a specific named entity category, such as person or location. The inter-annotator agreement metric needs to either take both decisions into account, which makes the exact calculation complex, or employ simplifying assumptions (e.g., to ignore overlapping spans).

In natural language processing, inter-annotator agreement is also often considered an upper boundary for machine performance. If humans only agree to a certain extent, we cannot expect machines to do better.

The Subject of Analysis: Narrative Levels

The target concept in this shared task is the concept of narrative levels. Narrative levels are a ubiquitous phenomenon in narratology that is well-known to readers (and watchers) with and without an academic interest in narrative structure. They are a central element of every narrative. In some cases they are even a fundamental feature of a narration, as in the book *Arabian Nights* where Scheherazade tells a story every night in order not to be executed; or in Boccaccio's *Decameron* where a group of ten people having escaped from the great plague in Florence to the countryside help pass the time by telling stories. Even the TV show *How I Met Your Mother* consists of episodes narrated by the protagonist who tells stories from the past that eventually lead to his marriage. Very generally speaking, a narrative level is a separable part of a story within a story-narrative. The 'within' in the 'story within a story-narrative' is typically, but not necessarily, thought of as a subordination.

While the question of the status of these stories within stories as narrative levels depends on the actual definition of narrative levels and is thus disputable, all examples mentioned before show that narrative levels are a fundamental part of narratives, or, more precisely, that narratives can be seen as being constructed entirely of narrative levels. This also holds for narratives where the plot is not dependent on the integration of narrative levels as in these examples.

The introduction of additional narrators and their narrations is a typical aspect of the natural practise of storytelling and is therefore a frequent phenomenon in narratives, regardless of their function and mediality. Narrative levels are present in all narratives, in fictional texts as much as in self-narrations, journalistic writing, jokes, and many other text types. Additionally, narrative levels are not restricted to written text, but can also be observed in oral storytelling as well as in moving images, again both in fictional and non-fictional forms. Thus, narrative levels are highly relevant for all types of narrative analyses.

In our shared tasks, we focus on narrative levels in fictional texts, since the concept of narrative levels was originally developed for these, and they still constitute the major area of research. In addition, computational analysis in the context of narratives so far works best for written texts.

Please note: The remainder of this section provides an orientation for those who are interested in the role of narrative levels in manual and computational analysis and (literary) theory as well as our approach to it. The specific handling of narrative levels in the submitted guidelines and its evaluation will be discussed in Chapter 3 ("Annotation Guidelines Overview and Evaluation Results").

A Brief Narratological Background

Generally speaking, the major aspects of narrative levels are the narrator(s) and the horizontal and vertical embedding. The notion of narrative levels, as many other concepts in narratology, was introduced by Gérard Genette, one of the most famous narratologists. The phenomenon had already been described by others,[10] but it was Genette who coined the term *narrative level.*[11] In *Narrative Discourse*, he discusses a passage in Proust's *A la recherche du temps perdu* (1913-1927), in which a character tells stories of their past loves to another character in an inn.

[10] For example by Bertil Romberg, *Studies in the Narrative Technique of the First-Person Novel* (Stockholm: Almqvist & Wiksell, 1962).

[11] Gérard Genette, "Discours Du Récit," *Figures III*, 67-282. (Paris, 1972) (English translation: Gérard Genette. *Narrative Discourse: An Essay in Method* (Ithaca, N.Y: Cornell University Press, 1980).

Genette points out that it is not a distance in time or space that separates the narrated episodes from the inn, but rather "a sort of threshold represented by the narrating itself, a difference of *level*" and provides the following definition: "We will define this difference in level by saying that *any event a narrative recounts* is *at a diegetic level immediately higher than the level at which the narrating act producing this narrative is placed*".[12] According to this, narrative levels are produced by a narrating act, i.e., a new narrator is introduced in the narrative, recounting something as a new narrative. Or, as Pier puts it, "Narrative levels are most accurately thought of as diegetic levels, the levels at which the narrating act and the narratee are situated in relation to the narrated story."[13]

Many theorists have thought of narrative levels in terms of narrative framing or narrative embedding, some of them changing the terminology for the levels on which embedding occurs while doing so. Within narratology, there are several unresolved issues with the concept, ranging from terminological to categorical problems. For example, Wolf Schmid not only introduced a simpler nomenclature for the possible levels of embedding,[14] but also claimed that embedding can occur on all levels (Schmid, *Narratology*). In contrast, in Genette's view the introduction of a narrator is crucial for an embedded level and thus embedding can occur only within the so-called intradiegetic level. Another issue that is pointed out by Pier is that "intercalation" would be the more appropriate term for describing the relation between narrative levels. Framing and embedding are operations that involve inclusion, whereas levels are distributed vertically (Pier, Narrative Levels, 4). Some theorists include these seemingly contradictory concepts in their narrative level concept. Starting with Mieke Bal's approach,[15] a series of models[16] was developed that added horizontal embedding (where no change

[12] Genette, "Discours Du Récit," 228, emphasis in original.

[13] John Pier, "Narrative Levels," (revised version; uploaded 23 April 2014), Paragraph 1. *The Living Handbook of Narratology*. Hamburg: Hamburg University. http://www.lhn.uni-hamburg.de/article/narrative-levels-revised-version-uploaded-23-april-2014 [last accessed 12 Feb 2019]. See also for a detailed discussion of narrative level concepts from a historical and a systematic perspective. The following overview in the text outlines the most important aspects Pier points out. Additional information can be found in Manfred Jahn, „N2.4. Narrative Levels," Manfred Jahn. *Narratology: A Guide to the Theory of Narrative*, (2017), and William Nelles „Embedding," David Herman, Manfred Jahn, and Marie-Laure Ryan (eds.). *Routledge Encyclopedia of Narrative Theory*. (London; New York: Routledge, 2005), 134-135.

[14] Schmid replaced Genette's terms extra-, intra- and metadiegetic, by primary, secondary and tertiary level of embedding (cf. Wolf Schmid, *Narratology. An Introduction*. (Berlin: de Gruyter, 2010), 67-70.

[15] Mieke Bal, *Narratology: Introduction to the Theory of Narrative*. (Toronto: U of Toronto Press, 1997), 43-66.

[16] Among other: William Nelles. *Frameworks: Narrative Levels and Embedded Narrative*. (New York: P. Lang, 1997), 121-158, and Marie-Laure Ryan. *Possible Worlds, Artificial Intelligence, and Narrative Theory*. (Bloomington: Indiana University Press, 1991), 175-200.

of level takes place) to the Genettian vertical embedding or shift between levels. Therefore, horizontal embedding means that narratives are narrated by different narrators on the same level.[17] Based on its claimed background in artificial intelligence, Marie Laure Ryan's account would seem to be the most relevant for the context of this project (Ryan, *Possible Worlds*). Ryan approaches the question of narrative levels in terms of boundaries, frames and stacks. Still, her usage of terms is often not in their computational sense proper and thus does not provide a straightforward approach to operationalization and automation. It is rather her introduction of the now well-established concepts of ontological (semantic) boundaries and illocutionary (speech act) boundaries that seems promising for computational approaches and more general operationalization.

There are certainly other approaches to narrative levels in narratology that could be added to this overview. But it should have become clear that narrators and concepts related to embedding are the most important aspects to consider. Even though narrative levels have been debated for over 50 years, there are still open issues connected to the concept such as its relation to frame theory or its deployment in cognitive narratology (Pier, *Narrative Levels*, 31-32).

Relevance of Narrative Levels for Text Analysis

We consider narrative levels as highly relevant for many text analysis projects because they are constitutive for narrative texts. This constitutivity makes them virtually ubiquitous in texts with narrative portions.

Text analysis refers to research steps that conduct a manual or automatic analysis of textual properties in relation to a specific research goal. Manual text analysis is usually a prerequisite for a hermeneutic interpretation of a literary text (and is typically not perceived as a distinct work step). Automatic text analysis employs methods from natural language processing (such as the detection of grammatical structure). On top of that, more "high-level" processing steps are typically added, one of them might be the detection of narrative structures such as levels. It is the ultimate research question that governs the kind and number of processing steps that need to be conducted for automatic text analysis. As we argue here, any text analysis with a focus on plot or character (i.e., analysis of the narrated content) requires the detection of narrative levels, as do some linguistically oriented preprocessing steps.

[17] Even though, in our view, this is already inherent in Genette's conception of voice that, among other, includes narrative levels and person. Therefore, even a shift of addressee may be interpreted as possible level change (cf. Evelyn Gius, *Erzählen über Konflikte. Ein Beitrag zur digitalen Narratologie* (Berlin: De Gruyter, 2015), 165-66).

The structure of narratives in terms of narrative levels can already be an interesting topic on its own. For example, one could be interested in the functions of frame stories in a certain period or the degree of nesting of narrative levels in fairytales in comparison to social novels. The way narrative levels are organized could in general may be a constituting element for a literary style in a broad sense,[18] and would be very hard to detect by existing stylometric approaches.

Still, the detection of narrative levels as a preparatory step for the analysis of a narrative is even more common and thus important. Gaining an understanding of the narrative levels present in a narration is a necessary prerequisite for its analysis. This applies to analyses concerned with the phenomena of the fictional world and to ones looking at the textual representation of the narrative, i.e., at phenomena related to what happens in the narrative (also known as the *what* of narration or *histoire*) or the very text (also known as the *how* of narration or *discours*).[19]

For the analysis of narratives one often needs to correctly conjoin narrative parts. Thus, it is necessary to have a prior understanding of the narrative levels present in a text. For example, when looking at character constellations, one should consider only characters in a more or less coherent space-time continuum, since interactions between characters are usually confined to coherent parts of the fictional world or story world and thus do not cross temporal or spatial borders. The identification of the narrative levels in a narrative and the analysis of their spatio-temporal features are therefore prerequisites to a proper character analysis.

Temporal or spatial coherence may also be relevant for the analysis of narrative representation. Thus, an analysis of the temporal relation between fictional world (*histoire*) and its representation (*discours*) is only sensible after having identified which narrative levels belong to which space-time continuum. The fictional world within a narrative is not necessarily coherent and can exhibit parts that are not connected to the main setting temporally or spatially as, for example, the world of a dream.[20]. Therefore, a reconstruction of the order of events in the fictional world needs to first analyze which narrative levels belong to which parts of the fictional world and then analyze the temporal order only for the connected ones.

There are currently no published systems that detect narrative levels automatically. While there certainly is a need for text segmentation as a preparatory

[18] Berenike J. Herrmann, Karina van Dalen-Oskam, and Christof Schöch, "Revisiting style, a key concept in literary studies," *Journal of Literary Theory* 9, no. 1 (2015): 25-52.

[19] The terms *historie* and *discours* have been coined by Genette (*Narrative Discourse*).

[20] Cf. Gius and Jacke "The Hermeneutic Profit of Annotation", 247-248.

step for subsequent processing of other phenomena,[21] segmentation is currently mainly accomplished by using textual surface phenomena. Features that can be derived directly from a text or its markup (e.g. in XML) are used as basis for segmentation (e.g., paragraphs, or, where available, chapters or other structural information encoded in the text). However, for more complex tasks the segmentation is more helpful the more meaningful it is. A division into chapters cannot be assumed to respect the structure of the events in the fictional world, as chapters are introduced for various reasons and some of them may have nothing to do with the plot of the narrative. Even worse, a division into, say, ten equal parts obviously is not related to the fictional world at all. As some segmentation is required for certain text analysis tasks, texts are often just segmented into parts of equal length, which is clearly not a usual procedure in literary studies. Hence, segments should be anchored in the narrated events rather than in chapter boundaries or, even worse, completely arbitrary segments of equal length in order to maximize their value for the analysis.

There are, however, some approaches to a content-related segmentation of texts in natural language processing.[22] Approaches in discourse analysis/processing[23] or topical segmentation[24] clearly feature related aspects to the ones needed to detect narrative levels, but a full-fledged and genuine automatic detection of narrative levels remains a desideratum. Moreover, existing approaches are typically tested and developed on texts such as news or wikipedia articles. As these texts differ in key areas from literary texts (fictionality, narrativity), the approaches cannot

[21] Cf. Nils Reiter, "Towards Annotating Narrative Segments," Kalliopi Zervanou, Marieke van Erp, and Beatrice Alex eds.*Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, (Beijing, China: Association for Computational Linguistics, 2015), 34-38.

[22] Among others: Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant, "Text Segmentation as a Supervised Learning Task," *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, (Association for Computational Linguistics, 2018), 469-73, https://doi.org/10.18653/v1/N18-2075. Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto, "Unsupervised Text Segmentation Using Semantic Relatedness Graphs," *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, (Berlin, Germany: Association for Computational Linguistics, 2016), 125-30, http://anthology.aclweb.org/S16-2016.)

[23] For an introduction cf. Manfred Stede, *Discourse Processing* (San Rafael, California: Morgan & Claypool, 2012).

[24] Anna Kazantseva and Stan Szpakowicz, "Hierarchical Topical Segmentation with Affinity Propagation," *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, (Dublin, Ireland: Dublin City University and Association for Computational Linguistics, 2014), 37-47. http://www.aclweb.org/anthology/C14-1005. Anna Kazantseva and Stan Szpakowicz. "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*, (Association for Computational Linguistics, 2012), 211-20. http://aclweb.org/anthology/N12-1022.

be directly applied to literary texts.

As we have already argued above, a segmentation based on the textual content is especially important when aiming for an analysis of more complex phenomena as addressed in literary studies, such as character constellation, point of view or the temporal structure of the plot. These are often only indirectly connected to the textual surface of narratives and it is a clear desideratum to base automatic segmentation more heavily on the content of narrations. For an application in literary studies or story-related analysis, narrative levels are the more adequate criterion for segmentation.

Generally speaking, the research goals related to text analysis often differ widely with literary scholars and computer scientists, but for research in both areas, narrative levels are an important asset. Most importantly, the analysis of narrative levels allows for a subsequent analysis of text qualities that may be structural, surface-related, or in the realm of narrative phenomena as well. Due to the integration of the latter, i.e., more content-related aspects of texts, a narrative level-based approach is much more adequate for text segmentation and subsequent analyses than a merely structural approach.

Narrative Levels in the Shared Task

For this shared task, we did not specify any theoretical background for the narrative level concepts, thus not providing one of the specific narratological approaches discussed above, nor did we disclose our personal preference for a specific approach. Instead, the participants were encouraged to choose the approach they considered adequate. We provided a basic explanation as well as reading suggestions (categorized as introductory, basic, or advanced) on the homepage of the shared task,[25] but we did not intentionally prioritize any approach. Therefore, participants could use any (or even no) narratological theory as a basis for the operationalization in their guideline.

There are several reasons for this decision:

a) Even though there are only few well-established approaches to narrative levels in narratology and most of them overlap, there is no consensus about *the* narrative level concept. Narratologists tend to have strong and diverse opinions about the nature of narrative levels, and there are good arguments that can be made for

[25] Some of them have been discussed in section 3 above, for the comprehensive list see: https://sharedtasksinthedh.github.io/levels/

most available theories. Therefore, there was no way to select the most suitable concept for level annotation.

b) As in many humanities' disciplines, there is no established procedure of identifying the 'right' theory among coexisting approaches. The idea of something being right, true, objective etc. is hardly compatible with the humanities' disciplinary paradigm or matrix. Within the humanities paradigm, theories and interpretations typically exist alongside each other and may even contradict each other. This plurality is owed to the humanities and their often heavily interpretative analysis of ambiguous and multifaceted human artifacts. Since the overall process of understanding is rather complex and its parts are not completely intelligible, limiting the analysis of an artifact to the usage of specific theories can lead to a premature exclusion of approaches yielding relevant insights. Therefore, limiting the narrative level analysis in the shared task to one approach would have meant ignoring the process through which theoretical or methodological approaches were and are developed in literary theory.

c) Annotation guidelines barely play a role in contemporary narratology, and annotatability is—at this moment—not a criterion regularly considered. From a narratological point of view, the pure guideline creation is likely not that interesting, compared to a discussion/comparison of narratological theories. However, it was clear from the beginning that the participation of narratology experts would be of utmost importance to this shared task. Therefore, allowing different theoretical "flavors" to compete would also spark the interest of narratologists who may be new to the development of annotations.

Against this background, allowing for all possible theories was advantageous to the research process on multiple levels. Most importantly, it allowed us to stick to the humanities' paradigm and at the same time provide a framework for the exploration and testing of theories in this first shared task on guideline creation. This ensures a higher relevance of the automation's outcomes to their users.

The Shared Task

Shared Tasks in Natural Language Processing

Shared tasks are an established research format within the community of natural language processing (NLP) with the core idea being that multiple participants try to solve the same task given by the organizers (e.g. automatic prediction of

part of speech tags). The solutions are then evaluated on the same data set with the same metric and thus directly comparable. Generally, a shared task works as follows: The organizers publish a call for participation in the task, describing the task as well as the associated data set in some detail. Shortly thereafter, the organizers publish a development and/or training data set. The dataset contains gold information, i.e., the categories to be identified are already annotated. This data set is then used by the participants to develop/train systems to automatically solve the defined task. After several months of development time, the organizers publish a second data set without the annotations: the test data. The participants apply their systems to the test data set (typically within a week) and send/upload the predictions made by their systems to the organizers. The organizers then evaluate all systems' predictions with the same evaluation script and against the same reference data. After this, a ranking of the systems can be generated, and a workshop is conducted to present the different systems and discuss the outcome.

History

Within natural language processing, shared tasks have their roots in the Message Understanding Conference (MUC) community.[26] In this context, the goal has been to extract snippets of information from news reports (covering incidents of terrorist attacks in South America) or naval messages. The major contributions of the shared tasks in the context of MUC are categorized into three different categories by Beth M. Sundheim and Nancy A. Chinchor:[27] The first category, progress evaluation, refers to the progress in terms of raw system performance with a clearly defined evaluation metric, which can be used to express the current state of the art, for comparison to a previous performance or to measure progress towards human performance (given the same metrics can be applied to human and machine performance). The second category, adequacy evaluation, expresses the adequacy of the evaluation metrics: "it is not possible to translate the evaluation results directly into terms that reflect the specific requirements of any particular real-life applications."[28] By applying evaluation metrics and scientific discourse about them, fostered by the MUC challenges, the community

[26] Cf. Beth M. Sundheim, "The Message Understanding Conferences," *Proceedings of the Tipster Text Program: Phase I* (Fredericksburg, Virginia, USA: Association for Computational Linguistics, 1993), 5, https://doi.org/10.3115/1119149.1119153.

[27] "Survey of the Message Understanding Conferences," in *HUMAN Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey* (Plainsboro, New Jersey, 1993), http://aclweb.org/anthology/H93-1011.

[28] Beth M. Sundheim and Nancy A. Chinchor, "Survey of the Message Understanding Conferences," in *HUMAN Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey* (Plainsboro, New Jersey, 1993), 58.

gains insight into potential weaknesses of the evaluation metrics. Finally, the third category, diagnostic evaluation, refers to the fact that the MUC challenges also generate insights into reasons for over- and underperformance of certain systems. By participating in the challenges and inspecting the prediction errors, the system developers gain insight into possible bottlenecks and can find ways for improvements of the system. All three categories have been present in shared tasks in the years following Sundheim and Chinchor's publication.

Starting with the year 2000, the Conference on Natural Language Learning (CoNLL) has been the home for a series of shared tasks on various topics: Chunking,[29] clause identification,[30] language-independent named entity recognition,[31] various forms of syntactic parsing either multilingually or for specific languages[32] and semantic representation/role labeling.[33] Other conferences and venues have taken up the shared task concept as well, for instance, the PASCAL Recognizing Textual Entailment challenge,[34] which ran for eight years until

[29] Erik F. Tjong Kim Sang and Sabine Buchholz, "Introduction to the CoNLL-2000 Shared Task: Chunking," *Proceedings of Fourth Conference on Computational Natural Language Learning and of the Second Learning Language in Logic Workshop* (Lisbon, Portugal: acl, 2000).

[30] Erik F. Tjong Kim Sang and Hervé Déjean, "Introduction to the CoNLL-2001 Shared Task: Clause Identification," *Proceedings of the ACL 2001 Workshop on Computational Natural Language Learning*, (2001), http://www.aclweb.org/anthology/W01-0708.

[31] Erik F. Tjong Kim Sang and Fien de Meulder, "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition," Walter Daelemans and Miles Osborne eds.*Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*„ (2003), 142-47, http://www.aclweb.org/anthology/W03-0419.

[32] Sabine Buchholz and Erwin Marsi, "CoNLL-X Shared Task on Multilingual Dependency Parsing," *Proceedings of the Tenth Conference on Computational Natural Language Learning (Conll-X)* (New York City: Association for Computational Linguistics, 2006), 149-64, http://www.aclweb.org/anthology/W/W06/W06-2920; Joakim Nivre et al., "The CoNLL 2007 Shared Task on Dependency Parsing," *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007* (Prague: Association for Computational Linguistics, 2007), 915-32, http://www.aclweb.org/anthology/D/D07/D07-1096; Sandra Kübler, "The PaGe 2008 Shared Task on Parsing German," *Proceedings of the Workshop on Parsing German* (Columbus, Ohio: Association for Computational Linguistics, 2008), 55-63, http://www.aclweb.org/anthology/W/W08/W08-1008; Jan Hajič et al., "The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages," *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task* (Boulder, Colorado: Association for Computational Linguistics, 2009), 1-18, http://www.aclweb.org/anthology/W09-1201.

[33] Xavier Carreras and Lluís Màrquez, "Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling," Hwee Tou Ng and Ellen Riloff eds. *HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning*, (Boston, Massachusetts, USA: Association for Computational Linguistics, 2004), 89-97; Xavier Carreras and Lluís Màrquez, "Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling," *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)* (Ann Arbor, Michigan: Association for Computational Linguistics, 2005), 152-64, http://www.aclweb.org/anthology/W/W05/W05-0620; Johan Bos and Rodolfo Delmonte, eds., *Semantics in Text Processing: STEP 2008 Conference Proceedings* 1, *Research in Computational Semantics* (London, UK: College Publications, 2008).

[34] Ido Dagan, Oren Glickman, and Bernardo Magnini, "The PASCAL Recognising Textual Entail-

2013. Having started under the label SensEval in 2000,[35] the SemEval initiative now hosts many shared tasks every year concerning the evaluation of semantic analysis tools. For the first time, an open call was issued to propose shared tasks for the SemEval roof to be organized in 2018. No less than twelve shared tasks were offered by SemEval in 2018.[36]

Carla Parra Escartín et al.[37] discuss several reasons for the popularity and success of shared tasks in natural language processing: Apart from fostering development in a certain field, they also allow for direct comparison between systems. A number of de facto standards have evolved in shared tasks (e.g., the widely used CoNLL format for storing annotated data). Finally, curated data sets have been created along with the shared tasks and subsequently made available:[38] "shared tasks have proven themselves to be very effective in incentivising research in specialised areas."[39]

Finally, ethical considerations about shared tasks have been pointed out by Escartín et al., mainly due to their competitive nature. Competition may lead to secretive behavior, hurt the relations of researchers with colleagues and lead to a general disregard for ethics. Escartín et al. identify a number of concrete scenarios which could directly impact the success story of shared tasks and might be a consequence of the competition. They propose that organizers follow a certain framework to minimize the negative impact of the competitive nature of shared tasks such as declaring early and explicitly if organizers or annotators are allowed

ment Challenge," J. Quiñonero-Candela et al. eds. *Machine Learning Challenges. Lecture Notes in Computer Science*, (Springer, 2006).

[35] Adam Kilgarriff and Joseph Rosenzweig, "Framework and Results for English Senseval," *Computers and the Humanities* 34, no. 1 (April 1, 2000): 15-48, https://doi.org/10.1023/A:1002693207386.

[36] Affect in Tweets, Multilingual Emoji Prediction, Irony Detection in English Tweets, Character Identification on Multiparty Dialogues, Counting Events and Participants within Highly Ambiguous Data covering a very long tail, Parsing Time Normalizations, Semantic Relation Extraction and Classification in Scientific Papers, Semantic Extraction from CybersecUrity REports using Natural Language Processing (SecureNLP), Hypernym Discovery, Capturing Discriminative Attributes, Machine Comprehension using Commonsense Knowledge, Argument Reasoning Comprehension Task. See http://alt.qcri.org/semeval2018/index.php?id=tasks.

[37] "Ethical Considerations in NLP Shared Tasks," *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing* (Valencia, Spain: Association for Computational Linguistics, 2017), 66-73, http://www.aclweb.org/anthology/W17-1608.

[38] Carla Parra Escartín, Wessel Reijers, Teresa Lynn, Joss Moorkens, Andy Way, and Chao-Hong Liu, "Ethical Considerations in NLP Shared Tasks," in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing* (Valencia, Spain: Association for Computational Linguistics, 2017), 66-73.

[39] Carla Parra Escartín, Wessel Reijers, Teresa Lynn, Joss Moorkens, Andy Way, and Chao-Hong Liu,"Ethical Considerations in NLP Shared Tasks," in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing* (Valencia, Spain: Association for Computational Linguistics, 2017), 68.

to participate and if they will publish the system results under agreed license.

Two Linked Shared Tasks

As the research practices, goals and not least communities in natural language processing and literary studies are clearly different, directly applying the shared task model known from natural language processing is not going to work. We therefore made several adjustments to the procedure. Our project consists of two shared tasks, and this volume appears after a milestone within the first one was reached: a guideline evaluation workshop with all participants. The two tasks have different goals, data sets, and target audiences, but both focus on the phenomenon of narrative levels. The goal of Task 1 is the generation of annotation guidelines which are then used to annotate a large corpus to be employed as training/testing data in Task 2. The second shared task is a 'regular' NLP shared task, i.e., its goal is to develop systems that automatically detect narrative levels.

Shared Task 1: Systematic Analysis of Narrative Texts through Annotation (SANTA)

In the first shared task, the challenges of conceptualizing and defining narrative levels, as well as manually applying them to texts, are in focus. The main task of the participants was to develop annotation guidelines for narrative levels. As discussed above, we did not specify an exact theoretical background to be used for the guidelines, but we pointed the participants to a bibliography for further readings. We also provided a "how to"-article on our web page explaining how annotation guidelines can be developed, which contained the same information as the above section on annotations.

To foster the development of generic guidelines that do not make a lot of assumptions on the text type in question, we decided early on that the guidelines should be tested on an unspecified corpus, but it was stated that it would contain literary texts of certain genres. Each participant thus had to write the guideline without knowing the exact texts it would be applied to in the end. To ensure comparability of the guidelines, however, there needed to be some homogeneity in the corpus. We thus decided to provide the participants with a development set that they could use when writing the guidelines. The texts in the final test set were similar to the ones in the development set. This setup is inspired by the distinction

between development, train and test data used in machine learning.[40]

Corpus considerations. The corpus was compiled to cover as many of the suggested level phenomena as possible. It is heterogeneous with respect to genre, publication date, and text length.[41] However, representativity (whatever that means for literature) was not a guiding principle. All texts were made available in both English and German, some being translations from a third language.

The maximum length for a text to be included in this corpus was 2000 words. Since the constraint might limit the use of narrative levels, we also included longer texts to avoid this bias. We made these available in a shortened form, omitting passages that do not affect the overall narrative level structure in a substantial manner, according to the level definitions we suggested on our web page and our own judgement. A set of 17 texts had been made available as a development corpus, to be used by the participants during guideline development. Table 3 shows the texts with some metadata. The actual annotation experiment was conducted on a set of eight texts that were previously unknown to the participants. This list can be found in Table 4. All texts are freely available and can be accessed through our GitHub repository.[42]

Creating Parallel Annotations. Measuring inter-annotator agreement is an established way of gaining insight into the intersubjective applicability of an annotation guideline. In order to measure inter-annotator agreement, the same text(s) need to be annotated by multiple people, using the same guideline.

To implement this in the shared task, we asked each participating group to annotate that same test corpus using someone else's guideline. In addition, a group of (paid) student assistants annotated with the same guideline. In this process, each guideline was used three times on the same set of texts (see Table 1 for an overview).

Label	Description
own	Annotations done by the guideline authors using their own guideline
foreign	Annotations done by the guideline authors using another guideline
student	Annotations done by a group of student assistants

Table 1. Overview of the annotations

Timeline. The full timeline of the various events of the shared task can be seen

[40] Cf. I. H. Witten and Eibe Frank, *Data Mining*, 2nd ed., *Practical Machine Learning Tools and Techniques* (Elsevier, 2005), 144ff.

[41] *Genres*: anecdote, fable, folktale, literary fairy tale, novel, novella, narration, short story. *Publication date*: the majority of the texts were written in the 19th and 20th century. *Text length*: 2000 words maximum.

[42] https://github.com/SharedTasksInTheDH

in Table 2.

Date	Event
October 6, 2017	Call for Participation
June 16, 2018	Submission of the guidelines
June 25, 2018	Submission of the annotations on test corpus, using own guideline
July 6, 2018	Submission of the annotations, using foreign guideline
September 17-19, 2018	Workshop

Table 2. Timeline

Workshop. As a milestone in the first shared task, all participants were invited to a workshop that took place in Hamburg, Germany. All but one team were physically present. The three-day event was structured as follows: The goal of the first day was for all participants to gain a better understanding of the other guidelines. This was realized in the form of brief presentations and a discussion to identify commonalities and differences. On the second day, the guidelines were evaluated in detail. To this end, a questionnaire was first presented and discussed. All questions could be answered in the form of a four point Likert scale. Each team was then asked to fill out the questionnaire (in digital form) for every guideline except their own. In addition, they were asked to keep notes on why they assigned which scores. We will cover the evaluation details in Chapter 3 of this volume. On the last day, the organizers presented the evaluation results as well as the inter-annotator agreement scores, and the entire group discussed the results and next steps.

Outlook: Shared Task 2—Automatic Detection of Narrative Levels

The second shared task can be considered a "regular" NLP shared task, and is thus intended to primarily attract researchers in natural language processing. It is envisaged to take place in the summer of 2021.The annotated corpus will be split into development, training, and testing data sets, and will be made available to the participants at certain points in time. The final evaluation will then require participants to submit their automatic predictions to the organizers, who in turn will compare the predictions to the manual annotations of the test set. This shared task is planned to be organized with the SemEval community to attract a large enough number of participants. The participants are not required to be familiar with or experienced in literary studies, narratology, or digital humanities, as the task and its difficulties are encoded in the annotations. The result of the second shared task will be a comparison of automatic systems that detect narrative levels.

Preparations. After having completed the first, guideline-oriented shared task, the organizers will conduct an annotation phase. The goal of the annotation

phase is to provide an annotated corpus which is large enough to allow for methodological experiments, including machine learning.

This annotation phase will be executed using the best performing guideline of the first shared task as a starting point. It can be expected, however, that it will need updating during the annotation phase, as new phenomena are expected to arise. The final version of the guideline will be made available along with the annotated data for the second shared task.

Title (orig.)	Author	Title (en)	Genre	Year	Language(orig.)	Comment
	Aesop	The Wolf and the Lamb	fable	600 BCE		
Rosen-Alfen	Andersen, Hans-Christian	The Elf of the Rose	folktale	1839	dk	
Kjærestefolkene [toppen og bolden]	Andersen, Hans Christian	The Top and Ball	folktale	1862	dk	
Se una notte d'inverno un viaggiatore	Calvino, Italo	If on a Winter's Night a Traveller	novel	1979	it	Shortened
Мститель	Čechov, Anton Pavlovič	An Avenger	short story	1887	ru	
The Child's Story	Dickens, Charles	The Child's Story	short story	1852	en	
Die drei Federn	Grimm, Brüder	Feathers	folktale	1819	de	
Das wohlfeile Mittagessen	Hebel, Johann Peter	The Cheap Meal	anecdote	1804	de	
Der geheilte Patient	Hebel, Johann Peter	The Cured Patient	anecdote	1811	de	
Hills Like White Elephants	Hemingway, Ernest	Hills Like White Elephants	short story	1920	en	
How the Leopard got his Spots	Kipling, Rudyard	How the Leopard got his Spots	short story	1901	en	
Beyond the Pale	Kipling, Rudyard	Beyond the Pale	short story	1888	en	
Unwahrscheinliche Wahrhaftigkeiten	Kleist, Heinrich von	Improbable Veracities	anecdote	1810	de	
	Lagerlöf, Selma	Among the Climbing Roses	narration	1894	sv	
The Cask of Amontillado	Poe, Edgar Allen	The Cask of Amontillado	short story	1846	en	
Frankenstein or The Modern Prometheus	Shelley, Mary	Frankenstein or The Modern Prometheus	novel	1818	en	shortened
A Haunted House	Woolf, Virginia	A Haunted House	short story	1921	en	

Table 3. Overview of development corpus.

Title (orig.)	Author	Title (en)	Genre	Year	Language	Comment
Lenz	Büchner, Georg	Lenz	novella	1839	de	shortened
Выигрышный билет	Čechov, Anton Pavlovič	The Lottery Ticket	short story	1887	ru	
The Gift of the Magi	Henry, O.	The Gift of the Magi	short story	1905	en	
Kleine Fabel	Kafka, Franz	A Little Fable	fable	1831	de	
Der blonde Eckbert	Tieck, Ludwig	The White Egbert	literary fairy tale	1797	de	shortened
Der Schimmelreiter	Storm, Theodor	The Rider of the White Horse	novella	1888	de	shortened
Anekdote aus dem letzten preußischen Kriege	Kleist, Heinrich von	Anecdote from the Last Prussian War	anecdote	1810	de	
Herr Arnes penningar	Lagerlöf, Selma	The Treasure	narration	1904	sv	shortened

Table 4. Overview of the test corpus.