CA Journal of Cultural Analytics

December 15, 2021

On Organizing a Shared Task for the Digital Humanities – Conclusions and Future Paths

Evelyn Gius, Marcus Willand, and Nils Reiter

Evelyn Gius, Marcus Willand, Technical University of Darmstadt

Nils Reiter, University of Cologne

Dataverse DOI: https://doi.org/10.7910/DVN/2YQVM6

Article DOI: https://doi.org/10.22148/001c.30697

ABSTRACT

Shared tasks are a work format prevalent in the natural language processing and machine learning community. This special issue continues the reporting on the shared task SANTA (Systematic Analysis of Narrative levels Through Annotation), which has the development of annotation guidelines for narrative levels as its goal. Narrative levels, also known as embedded narrations, are omnipresent in many kinds of narrations, and one of the core concepts of narratology. In this introduction, we summarize the current state, report on the second annotation round in SANTA, draw some conclusions and, finally, derive some recommendations for future shared tasks in the digital humanities.

Introduction[1]

This is the second special issue on a shared task on the creation of guidelines for the annotation of narrative levels. We adopted the principle of shared tasks from natural language processing to digital humanities, bringing together two perspectives that typically are present in this interdisciplinary field: the information or computer science perspective and the humanities perspective.

The shared task took place in two rounds (see Figure 1 for an overview of the entire project). The first round was announced at various conferences in 2017 and started officially with the call for guideline submission in June 2018. In July 2018 the guidelines were used for annotations and in September 2018 we organized a workshop for the evaluation of the submitted guidelines. After this, the second round started and the guideline submitters got the opportunity to revise their guidelines and re-submit them by May 2019. A refined annotation pro-

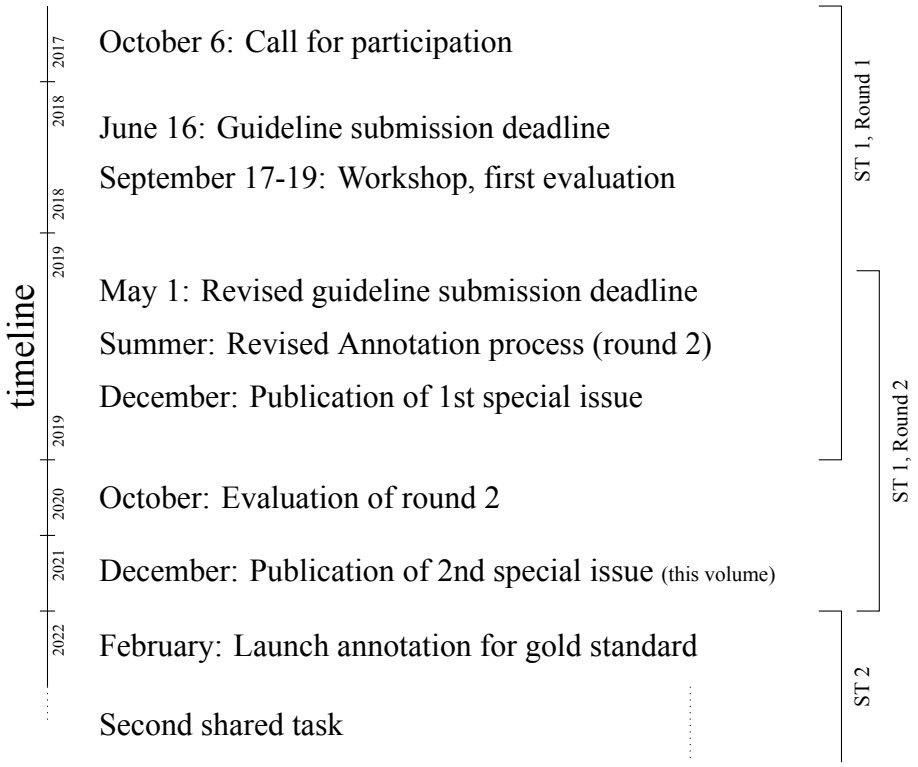

Figure 1: Schematic representation of the entire shared task endeavour

cess took place afterwards and was concluded in October 2019. Thereafter, we started the final evaluation of the guidelines, which is documented here. Even though the shared task is now concluded, this is not the end of our project: The outcomes of this shared task will be the starting point for a second shared task which will adhere to the typical shared task format in natural language processing. This second shared task will be dedicated to the automated recognition of narrative levels in literary texts and it will build on the guidelines created for the now concluded task.

In 2019 we published a first issue about the shared task on guidelines for narrative levels[2] consisting of the submitted guidelines, an in depth description of the overall approach and the evaluation setup as well as results of the first round. In this issue, we now present the description of the subsequent process, the revised guidelines and their evaluation together with a reflection of the whole undertaking of organizing a shared task for guideline creation as a shared task in the digital humanities. In *Revisions of the Shared Task Process*, we describe the

enhancement of the original process, consisting in an iteration of the process. Thereafter, we will present the evaluation of the revised guidelines, that is described both qualitatively (i.e., with regard to the changes that were made) and quantitatively (i.e., in terms of inter-annotator agreement and a discussion of the results), see *Analysis and Evaluation of Guideline Revisions*. The third part regards the insights we gained into the organisation of a shared task in the digital humanities that are intended to function as a contribution to a best practice for future shared tasks in this field, see *Humanities' Concepts and Shared Tasks: Lessons Learned*. Finally we summarize the most important points in terms of *Recommendations for Organizing Shared Tasks in the (Digital) Humanities*.

Revisions of the Shared Task Process

Initially, we had planned our shared task in three steps: (1) submission of the guidelines by the participants, (2) annotation, and (3) evaluation of the submitted guidelines. Thus in the beginning, our approach did not provide any iteration of the work on guidelines. This only came up during the evaluation workshop we had organized for the last step. There we discussed all guidelines thoroughly, also by comparing the different approaches and discussing which aspects could have made some guidelines better than others. This engagement resulted in the request by many participants to revise their guidelines.

We therefore iterated on these three steps. As a result, we were not only able to offer the participants the opportunity to revise their guidelines, but we were also able to establish a more controlled process of subsequent annotation. While this additional round seemed likely to improve the overall outcome, we did not have the resources to conduct the whole evaluation process again, since this would have involved another workshop for the qualitative evaluation of the guidelines. Therefore, we limited the evaluation to the quantitative approach in the second round of the shared task (hence, for the outcomes and discussions of this approach cf. section *Analysis and Evaluation of Guideline Revisions*).

Revision of the Submitted Guidelines

The revised guidelines are presented in this special issue. We would like to point out that in many cases the expertise of the guideline authors increased considerably between the first round and the second round for revision. This increase was due to a broader knowledge and understanding both of the phenomena in the guidelines and of the form of guidelines. In both areas, the guideline authors could build on insights from the evaluation. While some of this expertise is a general expertise in writing annotation guidelines, other is very specific to the phenomenon and can be gained best by iterating the process of guideline development based on annotation. The latter was achieved by introducing a second round for the guideline submission. For the former (the increased general expertise) the guideline authors could build on the discussions of their guideline during the workshop, on the consideration of the annotations that were made by others with their guideline as well as on insights they gained from reviewing other guidelines and deploying one of them for annotation. Besides the insights the guideline authors received during the workshop, we also provided them with the "How to Write an Annotation Guideline" depicted below that indicates suggestions on guideline creation (cf. Figure 2).

Revision of the Annotation Process

For the first round, an annotation process had been organized collaboratively in order to enable the calculation of inter-annotator agreement. For this purpose, each guideline was used three times for annotating the same texts. Every annotation was performed by one annotator of each of these groups: student assistants, the guideline author(s) and a 'foreign' participant from a competing team.[3] In this setting, the student assistants represent the only more or less controlled group since they were selected first and supervised during the annotation. Thus, the annotation process was probably not highly reliable in terms of stable annotation quality. In addition, each student annotator was asked to annotate according to two different guidelines. This proceeding could likely have led to

How to Write an Annotation Guideline

Some Recommendations for Guidelines from a Humanities Background

With the following we would like to give the authors of SANTA guidelines some suggestions which should lead to more understandable, applicable and comparable guidelines. The sections below describe important aspects of guidelines that should be separated.

Preliminaries

1. If your guideline is based on specific concepts or theories, specify them by referring to the concepts/theories and their authors.
2. Give definitions for the phenomena you are addressing. Demarcate the phenomena from each other explicitly. This also may help to facilitate a scholarly discussion about the concepts or other people's decisions about whether re-using your guideline or data that has been annotated according to it.

Annotation instructions

Defining the Annotation Span

3. Define the span of text an annotation typically covers, e.g., a sentence, word, paragraph or something different.
4. Define the borders of the annotations as exact as possible, e.g., specify whether to include/exclude interpunctuation, blanks at the beginning or end of a span etc.

Auxiliary indications

5. Give positive and, if possible, also negative **examples** for each phenomenon. Text examples might help as well as schematic illustrations do.
6. Name **markers** that indicate the presence of the phenomenon, if applicable. Think about syntactical, grammatical, semantic and other features that are typically connected to the phenomenon. E.g., specific words (as verbs with a specific semantic meaning, pronouns of a specific type etc.), tense, changes in mode or tense, preceding or subsequent phenomena etc.
7. Provide **tests** the annotators can perform in order to detect the phenomena. E.g., when replacing X with Y…; when paraphrasing it to Z…;

Organization of the Annotation Process

8. Provide an **overview** of the annotation categories (or overviews of subsets of related annotation categories)
9. If possible, organize the annotation routine from simple to complex phenomena
10. Where present, point out dependencies between phenomena (and consider them in 9)

Figure 2: The How To that was sent to the guideline authors for their revisions

some 'leaking' from one guideline to the other. The foreign annotators were not trained at all and also had to face the difficulty of unbiasedly deploying an annotation guideline for a phenomenon for which they had created a – typically rather different – guideline.

In fact, some of the most engaged discussions in our workshop evolved around the question, why some annotators seemed to have misunderstood the guideline they were working with. While we did not aim at clarifying the question whether the errors were caused by improper guideline deployment or by an improper guideline, we took the opportunity to also revise our annotation process for its use in the second round.

For the second round, we built a corpus that is similar to the corpus of the first round in terms of variety and frequency of narrative levels, but also of genre and authors/epochs[4]. It consists of 15 texts (cf. Table 5), two of which were also used in the first round. These texts have been re-used to ensure a certain level of continuity, because we otherwise would have run the risk of including only entirely new variants of the phenomenon.

Annotations in the second round were only made by student assistants inscribed in literary studies who also were trained beforehand.[5] Every guideline was annotated in CorefAnnotator[6] by two annotators and every annotator worked with only one guideline in order to avoid confusion by dealing with different annotation guidelines. To improve comparability between the annotations, we provided the annotators with generic annotation rules in addition to the specific guidelines (Figure 3).

We asked the guideline submitters to additionally provide two annotated texts (Heinrich von Kleist's Anecdote from the Last Prussian War, and a fragment of Theodor Storm's The Rider of the White Horse). Those were used during the training of the annotators for the purpose of comparing the annotations that the guideline authors considered to be correct with those made by the student assistants. Since cases of non-agreement can have two reasons – the student

assistant not applying the guideline correctly, or the guideline not covering the cases its authors wanted to cover – the annotations created by the guideline authors were only used as an additional heuristic approach during the annotators' training process.

The next step consisted of the discussion of the assigned guideline as well as of a first annotation of the two texts that also had been annotated by the guideline authors according to their guideline. Thereafter, the core training process started, combining phases of annotation, guideline discussion and comparison of annotation. The texts annotated by the guideline authors were used as an orientation during annotator training. Every annotation team was supervised by one of the shared task organizers. In an iterative process, we discussed issues that were pronounced by the annotators. We specified unclear parts of the guidelines if (and only if) we considered them to be deductible from the guideline. Respecting this condition, we avoided enhancing the guideline beyond the scope intended by the authors. We documented the annotation process in the training phase with regard to arising annotation challenges and specifications made for unclear cases. This built the basis for the final annotations. After the conclusion of the annotation process, we asked the annotators to compare the concrete tags they used for annotation and to adjust them if they were not conform. This was necessary for the identification of the categories for subsequent evaluation of the inter-annotator agreement and done without revising the annotations as such.

Revision of the Evaluation Process

For the second round, we also revised the evaluation process according to the enhanced design of the shared task. The major change was the restriction of the evaluation to the calculation of the inter-annotator agreement. From a conceptual point of view, we considered it not advisable to re-evaluate the guidelines within the same setting, since the second evaluation would have been heavily influenced by the insights and knowledge about the first version of the guidelines that were gained during the evaluation in the prior round. An additional,

Generic Annotation Rules

Ground Rules

1. Annotators annotate in groups of two people, who use the same guideline.
2. During the reading/annotation process, please do not interact with your counterpart (at least not about the annotation and/or text).
3. The first two texts you annotate are considered the training phase. This phase is done so that you get familiar with the guidelines and learn to interpret it correctly. After that (in the actual annotation phase), you need to make decisions on your own.
4. Read the guidelines and read the text you are going to annotate once before starting the annotation task.
5. If you have personal opinions about the guidelines and/or the text, and how you perceive narrative levels in the text: Please leave them out of the annotation. **Your annotation should *only* be based on the guidelines.**
6. If you come across cases that you are certain are not covered by the guideline, document them first. Then, make a decision that you feel is in line with the intention of the guidelines, and document this decision as well. Collect oddities or noteworthy observations during the annotation, and send them to us after you have finished your annotations. Cases that you find particularly difficult to annotate are of interest to us.
7. If you run into technical problems with the annotation tool, contact [email address]. Do not hesitate or feel bad, a certain amount of issues are to be expected. Help is fastest if you include a) what operating system you are using, b) what version of the tool you are running (bottom right of its windows) and c) the file that makes problems.
8. About the tool: [information about availability and 1-3 points about core functionalities]
9. The first text to be annotated during the training phase can be downloaded here: [link to online repository]

Figure 3: "Meta Annotation Guideline" provided as fundamental annotation rules

more pragmatic reason was that organizing another workshop for the evaluation of the guidelines was out of scope in terms of time and financial resources.

For the second round, the evaluation process was therefore focused on inter-annotator agreement, but we also reviewed the changes that were made in the revised version of the guidelines.

While the latter was a rather straightforward task (for the outcomes see section *Evaluation Results*), the calculation of the inter-annotator agreement posed a number of challenges we had to overcome. As for the metric to use, we stuck to Gamma,[7] the metric we already had used in the first round.[8]

Even though Gamma provides a solution for segmentation-oriented annotations, it still is not clear how it should be adapted best to our task. This concerns in particular the weighting of category discrepancies. As some guidelines used a large number of intermediate categories, this proved to be particularly important in our case. For the calculation of the inter-annotator agreement, it was necessary to take into account not only different categories in terms of type and quantity, but also considerable differences in the degree of its composition. While some of the guidelines defined the narrative level as their target category (i.e., the category to be measured for inter-annotator agreement), others combined the narrative level with the depth of embedding or the narrator of the level in a single category. This obviously affects the inter-annotator agreement, since identifying only the narrative level is a less complex task than identifying a narrative level together with an additional category as depth of embedding or narrator. For the evaluation this is problematic, because in the calculation of inter-annotator agreement for the latter case a corresponding identification of a narrative level will only be considered to be an agreement if also the additional categories are identified correspondingly. However, we had presented Gamma at the workshop at the end of the first round and therefore assumed that the participants were aware of the interdependency between the complex composition of an annotation category and its lower inter-annotator agreement rate. Nevertheless, we do not know to which extent this insight was taken into account for

the guideline revisions.

An additional problem was the standardization of the categories used by the annotators of the same guidelines. Even though we tried to be very restrictive on the labelling of categories and even asked the annotators to unify their labels after having made the annotations, in some cases we still were confronted with inconsistent labels. These inconsistencies have been unified with the help of the guidelines and, in some cases, after consultation with the annotators.

Further challenges were inconsistent specifications of subcategories and the handling of nested annotations. This issue could have been avoided by providing the annotators with predefined category sets and clear instructions on how to deal with nesting. The disadvantage of such a solution is that it requires an interpretation of the guidelines with regard to these points which is something we would consider a large intervention into the process.

Instead, we invested much effort in the following three adaptations:

- Inconsistent category labels were unified (e.g., for *narrative_level number="3"* vs. *narrative_level number=" 3"* the additional blank space in the latter was deleted; or for *A1A* vs. *1A1* the former was converted into the latter).
- In cases in which the target category for narrative level was a complex category (in the sense described above) or more than one category was specified as target category, we converted all the categories in question into a higher level category (e.g., we converted categories for the specification of the embedding type into the more general category narrative level).
- Different solutions for embedded levels were also unified (e.g., if a narrative level embedded in another narrative level was annotated continuously by one annotator and the other one chose a discontinuous annotation interrupted by the embedded level).

All in all, we had to adapt many category labels and found several issues where categories or annotations needed to be changed for inter-annotator agreement calculation.

Analysis and Evaluation of Guideline Revisions[9]

In the evaluation of the first round of annotations, we followed an evaluation scheme consisting of evaluations in three dimensions (conceptual coverage, applicability, usefulness).[10] While applicability could be (and was) measured in terms of inter-annotator agreement, the other dimensions were based on a questionnaire that the participants had to fill in during the workshop.

As the evaluation of the second round required a rather deep insight into other participants' guidelines, it was hard to realize without being in a shared workshop room. In the second evaluation, we therefore focused on the inter-annotator agreement, again using Gamma as a metric. We calculated inter-annotator agreement across the annotations produced by two student annotators for each guideline respectively.[11]

In the following, we will begin with looking at changes that the guideline authors made from the first to the second round, then we will present a comparative evaluation of the guidelines in terms of inter-annotator agreement, and finally discuss these new results.

Changes

In general, most of the changes between the original and the revised guidelines concern aspects such as structure, inclusion of (more) examples and explication of annotation routines. However, there are also some changes with regard to notions of narrative levels. Though the underlying theoretical works on which the notions build have not been changed by any guideline in terms of references (cf. Table 1).

Narratological Publication	**Guideline no.** I	II	IV	V	VI	VII	VIII
Genette		x	x	x	x	x	
Jahn		x	x	x		x	x
Lahn & Meister			x	x			
Lämmert				x			
Martinez & Scheffel			x				
Nelles (a: 1997), (b: 2005)				x[a,b]	[a]		
Neumann & Nünning				x			
Pier (c: 2012), (d: 2014)			x[c]	x[d]	x[d]		
Ryan			x	x			
Additional references		x	x	x	x	x	
Own approach	x						x

Table 1: References to narratological works in the guidelines

Nevertheless, there were some changes with regard to the concepts connected to narrative levels included in the revised guideline (cf. Table 2). Interestingly, the changes all consist in the addition of concepts, while no concepts were removed during revision.

Guideline 1 (J. Eisenberg, M. Finlayson; FIU, Miami; Eisenberg since 2019: Artie, Los Angeles) In the revision of guideline 1, the authors did not change anything with respect to the content except for an addition of a closer specification of the terms *narrator and type of narrative*.

Guideline 2 (E. Kearns; NUI, Galway) In its previous version, guideline 2 consisted of a list of selected XML elements and attribute explanations with many sample annotations of literary text snippets. In the revision, it was transformed into a revised annotation guideline on the basis of the literary concepts of *narrative levels* and *anachronism*. The guideline revision received a new structure and was supplemented by an additional chapter containing definitions and explanations needed for the annotation, as well as hints for self-checking procedures with sample annotations to guarantee a stable annotation process. In this revision, two specific improvements were made to the guideline: Firstly, the

Narratological Concept	**Guideline no.** I	II	IV	V	VI	VII	VIII
Definition of "narrative"	x		x	x	x	x	x
Narrative levels	x	x	x	x	x	**xx**	x
Discourse levels					**xx**	x	
Narrator (identity, relation to text)	x	x	x	x	x	x	x
Narratee			**xx**		x	x	
Structure of the level change/embedding	x		x	x	x	x	
Nature of level boundary (e.g., metalepsis)			x	x	x	x	
Speaker (illocutionary boundary)			x	x	**xx**	x	
Change of the world (ontological boundary)	x		**xx**	x	x		
Focalisation					x	x	
Analepsis/prolepsis	x	x	**xx**		**xx**		
Stream of consciousness/free indirect discourse		x	**xx**				
Extended/compressed time		x					

*Table 2: Narratological concepts operationalized by the guidelines. The assignment is based on the explicit reference to them in the guidelines. The references that were added to the revised guidelines are represented with **xx**.*

authors added further explanation of annotation spans and their possible beginning and ending in the middle of a sentence independent of the sentence structure and of the given punctuation. Secondly, they included helpful indicators that facilitate the detection of a narrative level change.

Guideline 3 was withdrawn from the shared task after the workshop. To avoid confusion, we have left the numbering as it was before.

Guideline 4 (N. Ketschik, B. Krautter, S. Murr, Y. Zimmermann; Stuttgart University) Guideline 4 is characterized by a broader approach to the annotation task by extending explanations that were already given in the previous version as well as adding new definitions and strengthening several relevant concepts for the detection of narrative levels. At the beginning of the guideline, the authors provide a new introductory overview dealing with the relatedness of the annotation guidelines to applications in (digital) literary studies. Furthermore, they enriched the introduction by strengthening the position of the narrator. In

the revised guideline, they added a further distinction of different types of distant narrators. In addition, they integrated a more precise distinction between the theoretical conception, the recognition of narrative levels and the actual application of the annotation guideline. A further extension was the explanation of different subfunctions of narrative levels (explicative, actional, thematic). Altogether, the revision provides a more detailed explanation of the concept of narrative level changes accompanied by a detailed description and sample annotations.

Guideline 5 (F. Barth; Stuttgart University, since 2020: University of Göttingen / State and University Library, Göttingen) Revised guideline 5 is three times the length of its previous version. The guideline was extended by a preface that sets out the aim of the guideline followed by an expanded introduction to the topic in general. Furthermore, the author provides more detailed definitions of narrative levels and narrative acts supplemented by clarifying diagrams. The addition of the chapter *Annotation Instructions and Examples* to the guideline offers a detailed guide through the annotation steps by providing extensive sample annotations to each sub point of the guideline. This extension helps reduce the complexity of the described annotation task.

Guideline 6 (M. Bauer, M. Lahrsow; Universities of Tübingen and Osnabrück) In the second version of the guideline 6, the structure of each individual chapter was enhanced by the addition of sub-items, namely, *span, borders, what does not belong here?* and *frequent markers & test*. Additionally, a further paragraph on annotation routine was added with guiding questions that suggest a structured approach to the annotation process. The final part of the guidelines is a graphical representation of the annotation categories in the form of a tree structure (overview of the annotation categories), which shows the dependencies of the annotations and the guiding questions as a graphical round off of the guidelines.

Guideline 7 (A. Ek, A. Kasaty, M. Wirén; Universities of Stockholm and Gothenburg) The second version of the guideline 7 underwent a drastic shortening from

a total of 51 pages in the first version (23 pages guideline text + 28 pages appendix) to 19 pages in the revision in total. The authors removed more generic information on guiding principles that underlay the guideline generation process in favour of focussing more on narrative concepts. The chapters were restructured in a way that previous short subchapters became more detailed independent chapters with further explanations and definitions (e.g., the addition of three new chapters: one on Narrator's discourse, a second on Characters' discourse and a third on Embeddings of narrative transmission levels). Furthermore, with chapter 2 an own section is provided to indicate all the changes that were made in the revised version of the guideline.

Guideline 8 (A. Hammond; University of Toronto) Guideline 8 received a new content structure by integrating the theoretical introduction into the first chapter and reserving the second chapter for sample annotations only. Additionally, the author completed the theoretical introduction by formulating the aim of the annotation of narrative levels as well as by referring to a definition of narrative itself. Furthermore, in the revision of the guideline, he adds two new distinguishable attributes and provides two additional charts with sample annotations for narrative levels and degrees of narrative. Another difference to the first version of the guideline is the extension of the terminology of level attribute and open attribute by the addition of more detailed explanations.

Evaluation Results

Table 3 consists of a list of both the inter-annotator agreement results from the first round as well as those from the second round. In both cases, the scores were averaged over the different texts. According to this, Guideline 8 achieved the highest agreement in the first round, while Guideline 7 outperformed all other guidelines in the second round. In total, the picture looks relatively symmetric: Three out of seven guidelines could improve their inter-annotator agreement score, for three further guidelines the scores are below their previous performance, and one achieved exactly the same result as in the previous round. However, two of the three guidelines that could improve their scores, recorded rather

	Round 1	Round 2		Delta
Guideline no.	**Mean**	**Mean**	**Dev**	
I	0.18	0.31	0.42	0.13
II	0.24	0.03	0.22	-0.21
IV	0.06	0.15	0.37	0.10
V	0.25	0.16	0.29	-0.09
VI	0.21	0.21	0.44	0.00
VII	0.23	**0.46**	0.28	0.23
VIII	**0.30**	0.23	0.42	-0.07

Table 3: Inter-annotator agreement scores

small improvements meanwhile the third increased its score considerably. Two of the guidelines for which the score has fallen show a very small, one a larger decrease.

For at least one of the guidelines with decreasing scores, the explanation for this development is quite simple: The decrease of 0.21 for Guideline 2 is due to the fact that in several texts (but not all) one annotator did not annotate any level so that no agreement could be measured. The reduced scores therefore reflect not only the quality of the guideline but also the quality of the annotators' performance.

On the positive side, we noted that the best agreement that was recorded in the second round is substantially higher than the best agreement in the first round (0.46 compared to 0.3). Guideline 7 achieves this score with the lowest standard deviation (cf. column "dev"), showing that these results are consistent across the different texts. While these are not entirely satisfying results, we note that some improvement was achieved in this shared task. However, the task of defining narrative levels for annotation remains a difficult one. Connecting the changes in inter-annotator agreement directly to the changes in the guidelines is difficult. The fact that Guideline 1, which was almost not changed at all, gained 0.13 points in inter-annotator agreement score, points out that there are factors influencing the annotation quality besides the guidelines themselves. In

Text	Mean	Dev	Min	Max	Best Guideline
Buechner	0.25	0.55	-0.31	1.00	VIII
Carroll	0.08	0.26	-0.15	0.53	VI
Salsbury	0.33	0.23	-0.12	0.61	IV
Mansfield	0.13	0.32	-0.22	0.68	VI
Twain	0.48	0.33	0.19	1.00	IV
Boccaccio	0.52	0.51	-0.15	1.00	II
Twain	0.07	0.25	-0.17	0.30	IV
Bierce	-0.03	0.25	-0.32	0.40	I
Melville	0.37	0.43	-0.13	1.00	I
Kafka	0.29	0.34	-0.05	0.77	V
Anderson	0.09	0.25	-0.10	0.45	V
Wilde	0.14	0.24	-0.15	0.60	VII
Bierce	0.30	0.45	-0.30	0.94	I

Table 4: The Inter-annotator agreement scores for each annotated text. "Best Guideline" shows the guideline that achieved the highest inter-annotator agreement-score

addition to the individually varying care and proficiency of the annotators, we see technical complexities with the annotation tool as well as an individually different supervision style by the organizers (replicating double-blind studies where the organizer supervising the annotators does not know who wrote the guideline cannot realistically be conducted in this setting).

The assembled material (guidelines, annotations, texts) allows a multitude of interesting analysis and observations, which we can only scratch the surface of here. Table 4 shows the achieved inter-annotator agreement per text, i.e., across the different guidelines. Looking at the texts, those authored by Carroll, Bierce and Anderson seem to be the most difficult to annotate, while those by Boccaccio and Twain seem the easiest. The fact that for each text at least one guideline achieved an agreement less than if annotated by chance (i.e., the expected agreement) underlines the difficulty of the task. Looking at the mean of the scores, we see better than chance agreement for almost all of the texts. We also observe that for many texts, at least one guideline was able to reach a high agreement. However, there is not one guideline providing for a good agreement

for a considerable portion of texts. It seems that the guidelines have specialized in particular phenomena and the main challenge is to achieve high agreement on different texts.

Humanities' Concepts and Shared Tasks: Lessons Learned

In the introduction to the first volume of this shared task-special issue we have claimed iteration as an important element for annotation and the process of guideline creation.[12] We generally believe in iteration as a principle that allows significant progress for both computational and human workflows. Therefore, we asked the participants to revise their guidelines. Not all did this with fundamental changes, but as we discussed above, most participating researchers share our opinion.

While shared tasks in natural language processing built on the idea of a competition, the participants of this shared task had other reasons to take part than winning, e.g., to learn about narratological concepts. In unison, they have expressed that they benefited a lot from the overall experience (gaining insights, knowledge, practices, etc.) and in the end, the ranking of the guidelines was only of secondary importance. This overall impression may be influenced by the fact that this shared task was attended mainly by humanities scholars, what we will discuss below.

In this section, we would like to summarize our central insights from this endeavour. This reflection is also addressed to potential future organizers of such shared tasks. Since this is the first shared task with the goal of guideline creation, there are currently no best practices and/or established workflows to follow. We will focus on the aspects we consider fundamental for the organization of shared tasks in the interdisciplinary field of digital humanities. Those aspects are particularly (i) the role of theories and concepts, (ii) the issue of guideline creation, (iii) annotation, (iv) the challenge of evaluation and (v) participants. All five aspects depend on the humanities perspective that we introduced into

the shared task idea. The most important insights are summarized in our final recommendations for the organization of shared tasks in the (digital) humanities.

About Theories and Concepts in DH-Shared Tasks

Our shared task was about the identification of narrative levels. We chose them since, *prima vista*, narrative levels are a textual phenomenon that can be theorized at a very low level of complexity - especially in comparison to other narratological categories such as focalization or distance. Thus, defining narrative levels *in theory* is a manageable task, but when it comes to the application, these seemingly clear categories show their fuzziness and inconsistencies. Like probably every other concept that is used in the humanities, the detection of its instances in texts is a partial act of interpretation. This interpretive act gets even more challenging when discussing it with other researchers, notably when they come from different research fields with different research interests.

Looking at this from the perspective of theoretical modelling, they are (in application even more than in theory) not to be understood as autonomous concepts, but to be related to many other narratological ones. As we did not per se specify the theoretical background for the narrative level concept, the guideline authors contextualized narrative levels not only with regard to the narrator (which seems obvious) and the narratee, but also to narration, focalisation and narrative time. This relation, however, goes in both directions: Other narratological concepts may be used to define narrative levels, or vice versa. In terms of operationalization, one can operationalize narrative levels using narratee, narrator etc. as more basic concepts which then need to be defined as well. But of course, one can also operationalize narratee, narrator etc. using narrative level as a more basic concept. All of this is not to be understood as a general criticism of known narratologies, but rather as an indication of their meaningful – because intertwined – arrangement of concepts. To examine one concept always also means putting another one to the test, which makes the isolated annotation of a single concept challenging.

About Guideline Creation

In the second round of the shared task, the contributors had the option to modify their guideline based on the insights they had gained during and after the workshop. Thereby, a rather long list of revisions developed: The lengthy guideline 7 was drastically shortened, guidelines 2, 5, 7 and 8 have put more emphasis on defining their concepts, and sample annotations were added to guidelines 2, 4, 5 and 8. All these features were identified at the workshop as characteristics of good guidelines. It can therefore be assumed that some standardization has taken place, which is mainly due to the intensive examination and discussion of all submitted guidelines at large.

Based on this insight, we think that future endeavours following these footsteps should regulate guideline creation more strictly. Our initial decision was to regulate as little as possible, both formally and in terms of content, in order to facilitate typical, rather open humanities approaches to text analysis. We did not make any specifications regarding future automation, the theoretical conception of narrative levels (as mentioned above), the formal design of annotation tags, etc. Although we still believe that participants should not be put in too tight a corset, we are also convinced that specifications at the level of the formal formulation of guidelines are advantageous.

For instance, it is crucial to ask participants to specify one "target tag", i.e., the concept that is central in their guideline and that will be used for their evaluation. The fact that most guidelines also included additional concepts led to the additional and rather tricky task to decide which of the concepts should be taken into account for evaluation. Furthermore, concentrating on only one concept can help humanities scholars to draw up a guideline, since for many approaches in the humanities neither the drawing up of guidelines nor the limitation of analytical processes to only one concept are typical practices.

Another aspect that should be clarified is whether the guidelines address experts or laypersons as intended 'users' (i.e., *annotators*). As we have shown,[13]

both variants have advantages and disadvantages. Since the selection of the addressees has an impact on the guideline, in future shared tasks it should be defined from the outset for which user groups and application scenarios the guideline is to be prepared for. Guideline developers must take into account the type of expertise of the annotators working with their guidelines.

One conclusion we can draw from our observations is that the applicable guidelines should underpin their categories with examples. Guidelines that followed this path achieved high inter-annotator scores.

About Annotation

By comparing the inter-annotator agreement of the first and second round, we have identified an interesting development. As we already argued,[14] achieved inter-annotator agreement is not the only aspect to take into account for measuring guideline quality. In addition, even the raw inter-annotator agreement is influenced by factors that prohibit a direct inference from inter-annotator agreement-improvement to an improvement of the guideline as a whole. One of these factors is the "quality" of the annotators, i.e., their care for details, their text understanding and their precision. While this is hard to formalize, some annotators are better at understanding the guideline authors' intentions, while others simply tend to forget things during their annotation work. In the second round, we trained the annotators with two texts that previously had been annotated by the guideline authors themselves, which is a time consuming process. Nevertheless, we believe that in a future shared task, this training process should be extended and potentially supplemented by more supervision during the annotation process. It is an open question whether to include the guideline authors during annotator training and/or during the annotation. On the one hand, they have authority on the interpretation and use of their own guideline, but on the other hand, their involvement might "contaminate" the annotators with reasonings that are not explained by the guideline, leading away from an evaluation of the guidelines as they were originally submitted. More control over the annotator training goes hand in hand with the question of pre-requirements for

the annotators (e.g., as discussed above, the question of laypersons vs. expert annotators).

Our concept of annotation is largely based on the one established in linguistics and intensively used in computational linguistics. However, transferring this workflow to concepts from literary studies leads to certain challenges, both practical and conceptual in nature. This annotation workflow is mostly used for local phenomena that require a limited, and clearly defined amount of context such as a sentence for the task of part of speech annotation. Narrative levels, in contrast, can differ considerably in length, which in turn requires varying amounts of contextual information. This poses challenges for the annotation tool in use (e.g., text selection works best with smaller segments and worse with longer ones), requires profound text understanding/reading proficiency by the annotators and generates parameters regarding the evaluation that are hard to define (e.g., what level of disagreement is still acceptable?). Thus, transferring the "regular" annotation workflow to long segments requires careful adaptation, because some of the assumptions holding for linguistic annotations are not fulfilled.

About the Evaluation

Competitions and competition-like events require clearly defined evaluation methods in order to be fair and transparent. While our evaluation strategy (a questionnaire covering the three dimensions mentioned above) was not without problems, we made several encouraging observations during the process, regarding the adequacy of this evaluation method:

1. The answers to the questionnaire made full use of the range of possible points. From this, we conclude that the questions were suitable to identify positive and negative aspects of individual guidelines and to distinguish between the guidelines.

2. The standard deviation across evaluators is small. The evaluators there-

fore agreed on how a guideline should be evaluated.

3. During the workshop, we observed that the qualitative plenary discussion of the guidelines led to assessments that were reflected by the results of the quantitative evaluation in the questionnaires. As mentioned above, the guidelines which were lauded for being highly differentiated on a theoretical level were the ones that received the most points in the dimension of conceptual coverage, etc. Thus, the quantitative results of a standardized questionnaire with a fixed scale not only seem to map the entire evaluation with the three dimensions in a sufficiently differentiated way. The correspondence of results also leads us to the conjecture that the subjective evaluation of vague concepts in the humanities can be expressed in figures even in the context of an interdisciplinary field of research as the digital humanities.

Finally, the creation of guidelines depends highly on disciplinary backgrounds. In this regard, two major research objectives could be discerned in our shared task: 1) Deepening the understanding of the narrative level concept (e.g., for one's own research) and 2) generating annotated data for automatisation purposes. Therefore, guidelines can be determined both by the theoretical (in this case: narratological) knowledge prior to guideline writing as well as by the aims of guideline application. Evaluating such diverging guidelines in a fair manner requires the evaluation to be multi-dimensional and objective-agnostic.

Given this, we have to acknowledge that our goal of providing an explicit and clearly defined evaluation function was met, but in our implementation the evaluation was not reproducible. In addition, it might not have been entirely fair nor restricted to the guidelines as such (among other, because at voting time, guideline authors were well known to each other). Furthermore, eliciting unbiased answers via questionnaires is a complex problem on its own, and our questionnaire would have benefited from involving social science expertise. Regardless of these remaining challenges, we are quite optimistic that the procedural framework of a shared task can be employed to tackle other issues. For this, adaptions

of the framework are could be made, e.g., by modifying the evaluation dimensions (or the number or weighting thereof) or by prioritizing quantitative or qualitative evaluation methods.

About Participants of DH-Shared Tasks

At the start of our endeavour, it was not clear at all whether it would attract enough participants – or even no participants at all. Through presentations at various conferences, however, we observed quickly that embedding in narratives is something that many researchers can relate to and that annotating them is perceived to be a challenge. It was thus important to find these researchers or to give them the chance to find the shared task. Future shared task organizers should also think about the incentives for participating: These include extrinsic motivation, such as the winner's reputation, but also intrinsic motivation, like the prospect of automating one's own research approach, or the pure excitement of participating in a new kind of academic experience.

Here, also issues connected to ambition and reputation have to be considered. While shared tasks provide a new, potentially productive framework for literary, and more generally for humanities scholars, their innovation also bears its risk. These risks are connected to the process and the output of shared tasks, which are not yet established in the (digital) humanities. As for the process, competition is a central property to the workflow of shared tasks. In natural language processing, winning a shared task comes with mentionable symbolic gratification whereas in the (digital) humanities competitions have still to be established as a acknowledged form of research. Also the output of shared tasks differs clearly from the typical output in literary studies. In the case of our shared task, the output of a long research process was an annotation guideline which also in has been developed and written jointly many cases. This diverges both in content and form from the publications specific for literary studies. These differences from the established workflow and publication formats in the humanities can negatively affect the motivation of those who are concerned about promoting their reputation in publications. On the other hand, the shared tasks

as formats of research are in line with the development towards short-term publication routines, availability, visibility, linkability, processability, etc., as well as with the shift towards publishing also intermediate steps of research in the humanities. Therefore, we expect the reputation issue to fade away in the course of the next years.

Moreover, there is a kind of side effect of the competitive format that can considered an additional benefit of shared tasks. The main reason for doing a quantitative evaluation in our shared task is that we were aiming for a ranking, and needed the possibility to proclaim a winner. As this was also discussed with the participants during the workshop, we learned that the opportunity to win was not their motivation to participate (which can be attributed to the high number of humanities scholars participating and their first contact with shared tasks). Thus, potential organizers of future shared tasks need to bear in mind how important the competitive aspect will be for their workflow.

Recommendations for Organizing Shared Tasks in the (Digital) Humanities

For potential future shared task organizers, the transferability of our shared task workflow to other tasks might be of interest. Within the digital humanities there is of course an extreme variety of research questions, of which certainly not all are equally well suited to pursue in a sequence of two related shared tasks. Some tasks are better addressed without a prior shared task for guideline creation, as known from the MUC/CoNLL tradition. At the same time, digital humanities questions origin foremost in the humanities and rely on – in quotation marks – "soft concepts". This led us to design a shared task workflow in which the underlying concepts are previously problematized in a first shared task. Since there is nothing to be said against replacing the text analysis category "narrative level" with, e.g., historiographic categories (such as "power" or "discourse") or categories from the spectrum of qualitative methods of the social and cultural sciences, we have no doubt that our shared task workflow can be transferred

to, e.g., digital musicology, digital art history, digital religious studies etc. The caveat of this transfer, however, is the certainly quite complex adaptation of the evaluation workflow to other research questions. It would probably be appropriate to evaluate the submissions in a quantitative and qualitative manner, but – as mentioned above – this depends on the underlying research objects.

How-to: Organizing a Shared Task in the (Digital) Humanities

Our findings and principles about the entire process can be summarized in these six points:

1. Be clear about your goals: In a competitive shared task, a clear objective function is needed. In a non-competitive shared task, create opportunities for in-depth discussions (e.g., workshops or online forums). Mixing these goals is possible, but raises a lot of challenges.

2. Make it clear to the participants that it is assumed that they follow an iterative guideline development scheme before submitting the guidelines to the shared task (i.e., that their guidelines undergo multiple tests before submission). If possible, offer technical and/or intellectual support in this phase.

3. Specify a guideline template. The template should include a clear requirement that every guideline needs to define the (same) target category label. Depending on the evaluation, one might allow for multiple target category labels.

4. Make it clear to the participants how the annotation and evaluation process work, especially with respect to the expertise of annotators and the evaluation system (and metric, if applicable).

5. Control the annotation process

    (a) Define annotator training (if possible, in multiple stages). Guideline

authors need to provide material for this, e.g., annotated texts to be used as reference.

(b) Define the annotation environment/tool beforehand to avoid too many moving variables. The guideline template, for instance, could contain a section on technical aspects of the annotation, such that guideline authors can refer to it and its jargon.

6. Aim at a limited, but fixed time frame: Academic projects tend to take longer than anticipated, but for participants and organizers alike, a limited time frame is easier to handle. Fixed conference and/or publication dates are a reasonable way to enforce this.

Title (orig.)	Author	Title (en)	Genre	Year	Language (orig.)	Comment
Anekdote aus dem letzten preußischen Kriege	Kleist, Heinrich von	Anecdote from the Last Prussian War	anecdote	1810	de	Annotated by guideline-submitters for training purpose
Der Schimmel-reiter	Storm, Theodor	The Rider of the White Horse	novella	1888	de	Annotated by guideline-submitters for training purpose
Lenz	Büchner, Georg	Lenz	novella	1839	de	shortened
Alice's Ad-ventures in Wonderland	Carroll, Lewis	Alice's Adventures in Wonderland	novel	1866	en	shortened
Perdition	Salsbury, Richard	Perdition	short story	1998	en	
Bliss	Mansfield, Katherine	Bliss	short story	1918	en	
Luck	Twain, Mark	Luck	short story	1891	en	
Federigo degli Alberighi	Boccaccio, Giovanni	Federigo's Falcon	novella	~1349-1353	it	
Extracts from Adam's Diary	Twain, Mark	Extracts from Adam's Diary	short story	1904	en	
The Moonlit Road	Bierce, Ambrose	The Moonlit Road	short story	1907	en	
Bartleby the Scrivener	Melville, Herman	Bartleby the Scrivener	short story	1856	en	shortened
Gespräch mit dem Beter	Kafka, Franz	Conversation With the Supplicant	novella	1909	de	
Svinedrengen	Andersen, Hans Christian	The Swineherd	fairy tale	1846	dk	
The Remarkable Rocket	Wilde, Oscar	The Remarkable Rocket	fairy tale	1888	en	
An Occurrence at Owl Creek Bridge	Bierce, Ambrose	An Occurrence at Owl Creek Bridge	short story	1891	en	

Table 5: Overview of corpus for second round

Notes

[1]We would like to thank Alina Klein and Janis von Keitz for their assistance with typesetting of all contributions in this special issue.

[2]Evelyn Gius, Nils Reiter, and Marcus Willand, "A Shared Task for the Digital Humanities," *Journal of Cultural Analytics*, 2019,

[3]For the annotation process cf. Gius, Reiter, and Willand.

[4]For a description of the corpus of round 1 Nils Reiter, Marcus Willand, and Evelyn Gius, "A Shared Task for

the Digital Humanities Chapter 1: Introduction to Annotation, Narrative Levels and Shared Tasks," *Journal of Cultural Analytics*, 2019, https://doi.org/10.22148/16.048

[5]We would like to thank our fourteen assistants for their help: Clemens Baiker, Rabia Ferahkaya, Jaqueline Haas, Marie-Angelina Hartstock, Yumeng Hu, Karoline Huber, Antje Jörgensen, David Klein, Maximilian Müller, Franziska Putz, Malte Schmid, Achim Schmid, Mira Schwarzer and Annika Spiegel.

[6]Nils Reiter, "CorefAnnotator : a new annotation tool for entity references" [in en], Publisher: Universität Stuttgart, 2018, https://doi.org/10.18419/OPUS-10144.

[7]Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier, "The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment" [in en], *Computational Linguistics* 41, no. 3 (2015): 437–79, https://doi.org/10.1162/COLI_a_00227.

[8]For a more detailed discussion of the suitability of Gamma cf. Gius, Reiter, and Willand, "A Shared Task for the Digital Humanities"

[9]This section was co-authored by Svenja Guhr, TU Darmstadt.

[10]For the evaluation of the first round cf. Evelyn Gius, Nils Reiter, and Marcus Willand, "A Shared Task for the Digital Humanities Chapter 2: Evaluating Annotation Guidelines," *Journal of Cultural Analytics*, 2019, https://doi.org/10.22148/16.049.

[11]In total, 14 annotators were involved in the annotations, as each pair of annotators only annotated according to a single guideline. Annotators were trained using two texts on which the guideline authors provided reference annotations.

[12]Cf. Reiter, Willand, and Gius, "A Shared Task for the Digital Humanities Chapter 1"

[13]Marcus Willand, Evelyn Gius, and Nils Reiter, "A Shared Task for the Digital Humanities Chapter 3: Description of Submitted Guidelines and Final Evaluation Results," *Journal of Cultural Analytics*, 2019, https://doi.org/10.22148/16.050.

[14]Gius, Reiter, and Willand, "A Shared Task for the Digital Humanities Chapter 2."