CA Journal of Cultural Analytics

A Shared Task for the Digital Humanities Chapter 3: Description of Submitted Guidelines and Final Evaluation Results

Marcus Willand, Evelyn Gius, Nils Reiter

11.05.19

Article DOI: 10.22148/16.050

Journal ISSN: 2371-4549

Cite: Marcus Willand, Evelyn Gius, Nils Reiter, "A Shared Task for the Digital Humanities Chapter 3: Description of Submitted Guidelines and Final Evaluation Results," Journal of Cultural Analytics. November 4, 2019. doi: 10.22148/16.050

In this chapter, we give a descriptive overview of the annotation guidelines, their use of narrative level concepts, and the results of their quantitative evaluation. We will also connect some of the results to qualitative findings we uncovered during the workshop, although some references to the participants' objectives are conjecture. Finally, the chapter contains a reflection on the annotation and evaluation procedure.

General Observations

Since this shared task was targeted at diverse audiences, the submissions and the disciplinary backgrounds of their authors are as diverse as expected. Table 1 shows key properties of the research teams.

Guideline	I	II	III	IV	V	VI	VII	VIII
Group size	2	1	2 (S)	4	1	2 (S)	3	1 (S)
Disciplinary background	Computational linguistics	English literature/ Data science	English literature	Literary studies/Digital humanities	Digital humanities /Literary studies	English literature	Computational linguistics	Literary studies
Country	U.S.A.	Ireland	Germany	Germany	Germany	Germany	Sweden	Canada

Table 1. Properties of the guideline authors. Group size indicates the number of authors of a guideline, (S) indicates that the guideline has been developed within a seminar or lecture. The disciplinary background is based on self-designation of the participants.

The participating teams also differ greatly in terms of age, gender, group size, academic level, research field, and disciplinary affiliation. This diversity is reflected in the submitted guidelines. They differ strongly in shapes and sizes: they range from 1 to 50 pages length as well as from theoretical essays to practical how-to's (see Chapters 4-10 for the Guidelines).

While disciplinary differences come with diverging practical experiences and diverging genre knowledge about annotations and guidelines as such, there is no clear cut between computer scientists and humanists. Guidelines with authors from both areas aim at providing a mixture between conceptual definition and practical annotation instructions.

Guidelines and Levels

The definition of this first task given by us organizers left it to the participants to select a useful/correct/reasonable theoretical basis for their guideline. To give an overview of the main theoretical foundations of the guidelines, Table 2 shows which publications and Table 3 shows which concepts are referred to in which guideline and/or rationale. Please note that different guidelines may employ different understandings of the same narratologist or concept (e.g., focalisation). Two guidelines referring to the same narratologist are not necessarily compatible, neither are two guidelines referring to a similar set of concepts. Please also note that this summary is a descriptive one and it is not our intention to suggest that guidelines *should* contain references to theoretical research.

References to Research in Narratology [1]

[1]

Narratological publications	Guideline/Rationale							
	I	II	III	IV	V	VI	VII	VIII
Genette		x	x	x	x	x	x	
Jahn		x		x	x		x	x
Lahn/Meister				x	x			
Lämmert					x			
Mani								
Martínez/Scheffel				x				
Nelles			x (1997)		x (2005/1997)	x (1997)		
Neumann/Nünning					x			
Pier			x (2014)	x (2012)	x (2014)	x (2014)		
Rimmon-Kenan								
Romberg								
Ryan			x	x	x			
additional references		x	x	x	x	x	x	
own approach	x							x

Table 2. Narratological publications operationalized by the guidelines. The assignment is based on the references in the guidelines and/or rationale.[2]

As Table 2 shows, the most referenced publication was Genette (Narrative Discourse) with six guidelines referring to it. This is not surprising, since Genette introduced the concept of narrative levels and most other theorists relate to his work in some way. Other publications cited by three or more contributions that are thus comparably present in the guidelines are the introductory texts by Manfred Jahn (Narrative Levels) and John Pier (Narrative Levels), as well as Marie-Laure Ryan's account (Possible Worlds and others, see Endnote 2). While Jahn and Pier may have been chosen due to their introductory character—they don't develop anything new but summarize the most prominent existing approaches—Ryan is probably the most formalized approach among the suggested and may thus have been considered especially suited for the guideline development. The fact that no guideline refers to Mani (Computational Narratology) and Romberg (Narrative Technique of the First-Person Novel) is probably due to the rather

2

abstract description of computational narratology in the former and, on the contrary, the rather lengthy discussion and focus on first-person narrators in the latter. Seven out of eight contributions also cited additional research (cf. endnote 2). Guideline I developed their own approach. Even though the authors refer to certain concepts (cf. Table 3), there is no explicit reference to a theorist. Guideline I conceives of narrative as linguistic representation of a story and focuses on the identification of borders of narratives, introducing the notion of uninterrupted vs. embedded vs. interruptive narrative (referring also to analepsis and prolepsis). Guideline VIII gives its own definition of narratives and focuses on level changes that can be identified with a test ("Let me tell you a story").

References to Narratological Concepts

Narratological concept	Guideline/Rationale							
	I	II	III	IV	V	VI	VII	VIII
definition of "narrative"	x			x	x	x	x	x
narrative levels	x	x	x	x	x	x		x
discourse levels							x	
narrator (identity, relation to text)	x	x	x	x	x	x	x	x
narratee			x			x	x	
structure of the level change/embedding	x		x	x	x	x	x	
nature of boundary between levels (e. g. metalepsis)			x	x	x	x	x	
speaker (illocutionary boundary)			x	x	x		x	
change of world (ontological boundary)	x		x		x	x		
focalisation			x			x	x	
analepsis/prolepsis	x	x						
stream of consciousness/free indirect discourse		x						
extended/compressed time		x						

Table 3. Narratological concepts operationalized by the guidelines. The assignment is based on the explicit reference to them in the guidelines.

Table 3 gives a first impression of the concepts that were considered relevant by the guideline authors for the identification of narrative levels. The listed concepts can be divided into concepts connected directly to narrative levels (such as boundary and change related concepts) and concepts that typically co-occur with narrative levels (such as focalization or anachronies, i.e. analepses and prolepses). While directly connected concepts can be used for the operationalization of narrative levels, co-occurring concepts often appear within narrative levels and are interesting for further analysis. This intermingling of concepts is probably connected to theoretical openness as it is typical for the humanities and the vagueness of the narratological concepts.

Some discussed differences can be explained by the diverging research objectives of the participating teams; e.g., whether the narrative level annotation is supposedly used for narratological concept development (guideline IV), identifying narratological concepts other than levels in literary texts (as time, e.g. guideline II) or to recognize linguistic discourse levels (guideline VII).

Results of the Evaluation

Final Ranking

First of all, we would like to mention once again that both content (conceptual coverage, applicability, and usefulness) and method of the evaluation (questionnaire, IAA) arise from the specifics of a shared task in the humanities. The multidimensional approach allows for an evaluation of the guidelines irrespective of their disciplinary and research background, their aims, and their understanding of narratological concepts. As we have already pointed out, the knowledge before guideline writing and the aims of guideline application are crucial and lead to rather different guidelines. Evaluating such diverging guidelines in a fair manner requires the evaluation to be multi-dimensional and objective-agnostic. This is exactly what the three evaluation dimensions are supposed to capture without valuing one disciplinary paradigm over another.

Below, we present the final ranking for the overall evaluation and the three dimensions in tables 3 to 7. The tables show scores and standard deviation for each question in the questionnaire, grouped by dimension. The questions can be found in Chapter II: "Introduction to Annotation, Narrative Levels and Shared Tasks".

Guideline	**Dimension**			**Overall**
	Conceptual Coverage	Applicability	Usefulness	
Guideline V	14.14	12.09	**12.88**	**39.10**
Guideline II	11.17	11.89	12.57	35.63
Guideline VI	12.33	11.01	11.37	34.71
Guideline IV	**14.43**	7.71	11.26	33.40
Guideline VIII	8.10	**14.14**	9.12	31.36
Guideline VII	11.60	9.82	9.77	31.18
Guideline I	7.83	10.39	10.00	28.22
Guideline III	10.29	6.48	10.95	27.72

Table 4. Final results of the evaluation (overall scores). The highest score in each dimension is shown in bold.

Conceptual Coverage

Guideline	Rank	Question			
		Q1	Q2	Q3	Q4
Guideline I	8	3.00 ± 1.10	1.00 ± 0.00	1.00 ± 0.00	2.83 ± 0.98
Guideline II	5	2.67 ± 1.03	3.00 ± 1.10	2.83 ± 0.75	2.67 ± 1.21
Guideline III	6	2.43 ± 0.98	2.86 ± 1.07	2.71 ± 1.38	2.29 ± 0.76
Guideline IV	1	3.71 ± 0.49	4.00 ± 0.00	3.71 ± 0.49	3.00 ± 0.82
Guideline V	2	3.71 ± 0.49	4.00 ± 0.00	3.43 ± 0.53	3.00 ± 1.00
Guideline VI	3	3.33 ± 0.82	3.33 ± 0.82	2.67 ± 0.52	3.00 ± 0.89
Guideline VII	4	2.71 ± 1.11	3.33 ± 0.82	2.71 ± 0.95	2.83 ± 1.33
Guideline VIII	7	2.29 ± 1.11	1.00 ± 0.00	1.67 ± 1.21	3.14 ± 0.69

Table 5. Evaluation results for conceptual coverage (dimension 1). The table shows mean and standard deviation for each question. The questions are shown in Chapter 2 (see above).

The following reflections about the final results cover the three dimensions one after another. The results shown in Table 4 (conceptual coverage) are based on four questions. Guideline IV achieved the top position in this dimension. It is one of the guidelines that focuses on an in-depth description of the used narratological categories and definitions, which seems to be reflected by the positive evaluation of its conceptual coverage. On the opposite end of the spectrum is guideline I, which was ranked lowest. This coincides with the self-description and primary research interest of its authors, which is a "computational understanding of sto-

ries" (see guideline I, Chapter 4 of this issue). Their focus on future automation plans led to a guideline without ties to a theory and subsequently to a poor rating in this category. Guideline I also includes dialogical text genres (as scripts of TV shows and the transcripts of court cases) which caused some confusion for those who expected a narration to be necessarily narrated by a narrator. It is possible that the large number of narratologically versed participants penalized a deviation from established narratological consensus.

Applicability

Guideline	**Rank**	**Questions**		**Inter-Annotator Agreement**	
		Q1	Q2	Evaluation score (counted twice)	Gamma
Guideline I	5	2.60 ± 0.55	2.67 ± 0.52	2.56	0.18
Guideline II	3	3.17 ± 0.41	2.17 ± 0.98	3.28	0.24
Guideline III	8	2.57 ± 1.13	1.43 ± 0.53	1.24	0.07
Guideline IV	7	3.57 ± 0.53	2.14 ± 0.38	1.00	0.05
Guideline V	2	3.00 ± 1.15	2.29 ± 0.95	3.40	0.25
Guideline VI	4	3.17 ± 1.17	2.00 ± 0.89	2.92	0.21
Guideline VII	6	2.33 ± 0.52	1.17 ± 0.41	3.16	0.23
Guideline VIII	1	2.71 ± 1.38	3.43 ± 0.79	4.00	0.30

Table 6. Evaluation results for applicability (dimension 2). For the two questions, the table shows mean and standard deviation. For the inter-annotator agreement,

the table shows the scores used for the ranking (in the interval) as well as raw gamma scores. The questions are shown in Chapter 2 (see above).

The applicability score is based on the inter-annotator agreement and two questions from the questionnaire (cf. Table 5). The first of the two questions queried how well an expert in narratology would be able to apply the guideline to a narrative text; the other question asked the same for laypersons. Guideline VIII is the overall winner in this dimension while guideline III scored the lowest points. For an interpretation of the results, it is worth looking at the scores for individual questions and the IAA separately. The fact that Guideline VIII was rated best raises the question of a relation between the guideline's relatively simple level concept and its applicability. The detailed results reveal that the first place is partly due to the guideline having the highest inter-annotator agreement score. It seems obvious to deduce that simplicity is related positively to applicability. However, this is not completely supported by the answers in the questionnaire for this dimension. Guideline VIII gained only a lower midfield position in the question about expert applicability, but it was considered to be the most applicable guideline for laypersons (note that this judgement was cast before the IAA scores were revealed). So, in practice, a basic level concept seems to be applicable with great congruence, but the results of the questionnaire suggest that simplicity is understood to come with restrictions for experts. We attribute this to the assumed incapacity of the guideline to do justice to the complexity of narrative levels.

Complexity and applicability thus seem to correlate negatively. However, the comparison to other guidelines raises doubts about the derivability of such a general rule. Guideline V, a guideline with a relatively complex level concept and the overall winner of the evaluation, achieved second rank not only in the first dimension of conceptual coverage, but also in the applicability dimension. In this dimension, the result is based on another second rank in the IAA and a third/fourth rank in the questions. Thus, the guidelines with the two best results (VIII and V) in the applicability dimension are very different in nature and there seems to be no direct correlation between guideline complexity and applicability.

But there are still interpretable results. As the first question refers to annotators *with* a narratological background, it is not surprising that the highest score was reached by guideline IV (the winning guideline of the conceptual coverage dimension) followed by Guideline V. The fact that both guidelines with best results in conceptual coverage scored only average points for layperson application (question two) might be explained by the consideration that laypersons can benefit from clear and explicit conceptual level description, but also run the risk of being overwhelmed by complexity.

The low rank of guideline III[3] might be explained by a combination of factors. The guideline covers a broad range of narratological concepts. At the same time, it neither defines these narratological concepts in depth nor does it give examples on how to apply its annotation categories. The mere description of annotation categories seems to lead to difficulties in their application (see the very low inter-annotator agreement). Conversely, it can be said that applicable guidelines should have to demonstrate their categories by way of example. This is what the top two guidelines do in great detail. Even though for the most part they only achieved average results in the questionnaire, they were elevated in the *applicability* dimension by relatively high inter-annotator agreement scores.

[3] Guideline III is not printed in this volume, as its authors withdrew their submission.

Usefulness

Guideline	Rank	Question			
		Q1	Q2	Q3	Q4
Guideline I	6	3.17 ± 0.75	3.00 ± 0.71	2.17 ± 0.75	1.67 ± 0.52
Guideline II	2	3.50 ± 0.55	3.40 ± 0.89	3.00 ± 0.89	2.67 ± 0.52
Guideline III	5	3.33 ± 0.82	3.17 ± 0.98	2.29 ± 0.76	2.17 ± 0.75
Guideline IV	4	3.29 ± 0.76	2.83 ± 0.98	2.71 ± 0.76	2.43 ± 0.98
Guideline V	1	3.50 ± 0.55	3.40 ± 0.55	3.14 ± 0.69	2.83 ± 0.75
Guideline VI	3	3.40 ± 0.55	3.40 ± 0.89	2.17 ± 0.75	2.40 ± 0.55
Guideline VII	7	3.00 ± 0.63	2.60 ± 0.55	2.17 ± 0.98	2.00 ± 0.89
Guideline VIII	8	3.20 ± 0.84	2.80 ± 1.10	1.29 ± 0.76	1.83 ± 0.75

Table 7. Evaluation results for usefulness (dimension 3). The table shows mean and standard deviation for each question. The questions are shown in Chapter 2 (see above).

The top rank of guideline V in usefulness is most likely due to 1) the multifold examples of literary texts that illustrate the use of the elements to be annotated and 2) the very clear description of the research objectives of the guideline authors. This is also the case for guideline II, which ranked second in this dimension: Guideline II states that it is "designed for annotating analepsis, prolepsis, stream-of-consciousness, free indirect discourse, and narrative levels, with facility also for annotating instances of extended or compressed time, and for encoding the identity of the narrator" (see guideline II, Chapter 5). Since usefulness is the category that addresses a variety of possible cases in which the annotated texts might

allow further research, this result is quite interesting. It shows that expressing a specific application area in the guideline allows the evaluators to picture opportunities for use more clearly, but on the other hand might be more restricting compared to a guideline that does not express a specific application area.

Finally, some remarks on guideline V, the **overall winner guideline** in the first shared task: Guideline V was ranked second best in conceptual coverage and applicability (as well as inter-annotator agreement); it also reached the top position in usefulness, which in combination makes it the overall winner with some distance to the second rank. Both the quantitative results from the evaluation as well as the qualitative results from the discussion suggest that guideline V defines narrative levels in quite some detail and with particular precision: The guideline distinguishes the narrative level concept from "narrative acts" and refers to other narratological concepts (such as narrator) in a way that is helpful for identifying levels. Abstract examples in the form of diagrams and tables are given, illustrating systematically how narrative levels are to be understood. Furthermore, concrete text examples are annotated with these concepts. Last but not least, a workflow for the practical annotation is given as well as a very clear description of the aims of the annotation of each concept.

Observations on the Evaluation Process

As this is the first time such a shared task takes place, we think it is important to end this introductory chapter with an assessment of the evaluation process and the shared task as a whole. This includes our version of measuring inter-annotator agreement, the questionnaire and—the very heart of our evaluation—the three dimensional model.

Inter-Annotator Agreement

Firstly, the role and calculation of the inter-annotator agreement should be reflected. The agreement scores are based on a comparatively low number of annotations done by annotators with no systematic training. Therefore, their level of expertise varied: The student annotators were basically untrained (although some had experience in other annotation tasks) and had virtually no knowledge of narratological theory. The foreign annotators naturally were trained on their own guidelines and may have had problems to disengage themselves from them. Both issues applied to all guidelines. Nevertheless, the participants stated differ-

ent degrees of satisfaction with the foreign annotations based on their guidelines in the discussion. Some participants did not feel adequately understood by the annotators, which was especially the case with narratologically complex guidelines. Therefore, this problem may be at least partially caused by the interdisciplinarity of the shared task.

In addition, it was observed that in the case of guideline IV the annotations made by student annotators had a higher agreement with the guideline authors than the annotations made by the other groups. Therefore, the question arose whether we should have taken into account the expected disciplinary competences when distributing the guidelines for foreign annotation among the participants. We do not believe that this is a viable way to go, but we believe that a profound revision of the mutual annotation model between participating teams ("foreign annotation") is worth considering. Ultimately, this approach was born out of the need to obtain as many annotations as possible for each guideline to get data for the IAA. Given appropriate funding, it makes sense not to calculate the agreement on the basis of mutual guideline annotations, but to have it done exclusively by "external", similarly trained annotators. In fact, this is what we will do in a second annotation round, the results of which will be published in the second volume of this special issue.

A further observation worth mentioning was found looking at the inter-annotator agreement scores for guideline VII, where the foreign and student annotations were more similar to each other than to the annotations by the guideline authors. This also points to the different disciplinary competences, in this case to the strong linguistic influence of this guideline. Since our annotators were aware of the narratological focus of most other guidelines, they might have translated the guideline into a narratological perspective in the same way the co-annotating participant(s) did.

Questionnaire

We also want to highlight several issues that have been raised about the questionnaire itself during the workshop discussion: Filling in questions in the conceptual coverage dimension requires a broad narratological knowledge, which not all participants possessed. The two questions in the applicability dimension were intended to test the comprehensibility of the guideline for experts and non-experts, but as participant groups were homogeneous with respect to their expertise, one question could only be filled in with a grain of salt. As we have seen with guideline VIII, there is a clear mismatch between measuring applicability by us-

ing inter-annotator agreement, and by predicting applicability in a questionnaire. This is not surprising per se, but the magnitude of the mismatch is. Furthermore, it has been mentioned that clarity is such a relevant feature of guidelines that it should have been explicitly evaluated.

Some literary scholars voiced concerns about presenting the results of our complex evaluation in mere numbers. The dissatisfaction, though, did not refer to the results of the inter-annotator agreement of their own guideline, but was rather based on a general methodological skepticism, which was not shared to this extent by the participants with an affinity for automation. In the usefulness dimension, participants found it difficult to assign scores to guidelines they had not been working with intensively. Without this practical knowledge, the answers can only be conjecture. Nevertheless, despite the difficulties in filling in the questionnaire, the numeric scores are relatively homogeneous across the groups (low standard deviation, see above).

The Three Dimensional Evaluation Model

Lastly, we would like to reflect on the nucleus of the whole process, the three dimensional evaluation model. Since each evaluation dimension favors characteristics that may be related to the disciplinary origin of the guideline authors and thus may lead to biases, the combination of the three dimensions was arranged in such a way that they cancel out those biases. For example, guideline IV, whose authors all have a background in literary studies, achieved the first position in the dimension of *conceptual coverage*, a midfield position in *usefulness* and the second to last position in *applicability*. Guideline I, written by researchers in natural language processing, was ranked last in *conceptual coverage*, but received average scores in the other two dimensions. This gives us reason to believe that disciplinary advantages and disadvantages are indeed offset by our evaluation approach. The fact that guideline IV and VIII reached inverted ranks in dimensions one and two also indicates that the evaluation dimensions neutralize disciplinary advantages. The guideline that was ranked highest overall received high scores in all dimensions, but was ranked first only in one dimension. This suggests that to succeed in general, one needs to strike a compromise between the dimensions. This is the effect we were aiming for when designing the evaluation scheme.

As an outlook, for our readers both the quantitative results and the distribution of the narratological concepts might serve another purpose: They provide a structure and categorization of the submitted guidelines. Researchers and scholars who are interested in narrative levels and/or their annotation for whatever pur-

pose can browse through the rationals, guidelines and short reviews, all published in the following. Those instructive documents as well as our introduction allow for an informed decision on any guideline or the combination of multiple guidelines, both to be used as a starting point for original, new work.