CA Journal of Cultural Analytics

A Shared Task for the Digital Humanities Chapter 2: Evaluating Annotation Guidelines

Evelyn Gius, Nils Reiter, Marcus Willand

11.05.19

Article DOI: 10.22148/16.049

Journal ISSN: 2371-4549

Cite: Evelyn Gius, Nils Reiter, Marcus Willand, "A Shared Task for the Digital Humanities Chapter 2: Evaluating Annotation Guidelines," Journal of Cultural Analytics. November 4, 2019. doi: 10.22148/16.049

In this section, we will discuss our idea of guideline evaluation and the underlying considerations. Evaluating annotation guidelines in this way is a fairly new endeavor, and we have developed the evaluation setup from the ground up. Although we do not claim our choices to be universally valid or applicable, we believe that this approach to guideline evaluation is relevant for similar settings and can be adapted to projects that might have other preferences and priorities.

Preliminaries and Challenges

Our goal was to take into consideration requirements and principles from the humanities as well as from computational linguistics/natural language processing. Whichever evaluation method we would employ in the end, it needed to fulfill four basic requirements:

1. Establish a ranking: The method needs to be able to rank the guidelines. This ranking needs to be as clear as possible and avoid ties.
2. Be defined and explicit: The general design of shared tasks is a competition in which submissions are ranked according to an objective function. This

objective function needs to be defined in advance and as precisely as possible, in order for participants to know beforehand what they are getting into and so that it leaves little room for challenging this evaluation.
3. Be practical: The evaluation should be feasible to execute, within certain practical limitations. Concretely, we were aiming for an evaluation method that could be conducted within a two-day workshop.
4. Reflect our evaluation criteria: The evaluation method needs to reflect our evaluation standards, i.e., if a guideline contains aspects that are considered to be positive by the organizers, that guideline should be ranked higher than a guideline without these aspects. Defining positive/negative evaluation criteria was a decision that the organizers needed to make.

Those requirements are a consequence of aiming at creating annotation guidelines in a shared task. In shared tasks in natural language processing, the underlying intention is to reproduce the gold standard as closely as possible, which can then be measured in different ways, depending on the exact task (accuracy, f-score, MUC-score, …). But there is no "ground truth" conceivable for annotation guidelines. Even measuring inter-annotator agreement would not necessarily be that straightforward, since there may be cases in the data in which different textual readings are possible, stemming from a legitimate ambiguity of the text. In such cases, disagreements between annotators would not indicate a flaw in the guideline.

In addition to these general requirements that any evaluation method for a shared task needs to fulfill, there are several challenges related to the specific nature of this one:

As this shared task is an **interdisciplinary** endeavor, a heterogeneous set of participants was to be expected. The notion of annotation plays a different role in different disciplines, and a diverse set of best practices, rules, and traditions has been established in each field. In literary studies, for instance, annotation is typically understood as note-taking while reading. In computational linguistics, annotation is typically done in parallel, digitally, and with a high intersubjective agreement as the most important goal. The latter does not matter at all for annotation in (traditional) literary studies, as the disciplinary approach to text analysis is rather focused on a not necessarily reproducible overall meaning of a text.

Thus, participants have different previous experiences and expectations with regards to the annotation process. Still, the evaluation we conduct in this shared task needs to be valid and functional across the different disciplines and be beneficial for each participant's own discipline at the same time.

The **vagueness** of the source concepts provides another challenge. Narratology

represents a popular source for concepts in text-oriented DH, most likely because of its fundamental structuralist premises. Furthermore, narratologists and digital humanists agree on the idea that structural analyses expose interesting textual phenomena which remain hidden from purely content-related readings. However, as discussed above, the systematic application of narratological theory to texts also gives room for interpretation.

Given these considerations, we decided early on that the evaluation model needs to cover different perspectives. In particular, it should not ignore conceptual vagueness and complexity, but rather consider solution strategies for these problems.

Evaluation Model

Generally, the evaluation was conducted in **three different dimensions**: conceptual coverage, applicability, and usefulness. Figure 1 schematically shows where the evaluation dimensions are situated with respect to research activities in the digital humanities. It projects them onto the course of the entire work process, from narratological theory to guideline creation to annotated texts, and finally to the insights that could be drawn from applying the annotated texts to understand single literary texts or whole corpora.

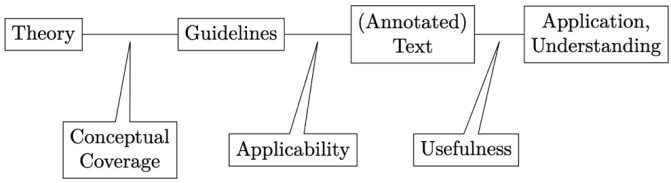

Figure 1: *The three evaluation dimensions connecting research areas in the digital humanities*

The dimension of **conceptual coverage** reflects how much of a theoretical basis is covered by an annotation guideline. If a guideline is explicitly based on a narratological theory, it might aim to fully implement every definition, rule, and exception of the theory. Another guideline based on the same theory might leave out some definitions or add others. This dimension is situated on the theoretical level, relating the guidelines to theory.

Applicability puts the guideline in relation to the text and reflects how well the guideline prepares annotators to do actual annotations, i. e., how well the guide-

line can be employed. A guideline's applicability may for instance be increased by thoughtful examples, a clear structure, and/or a careful use of terminology. The dimension of applicability also covers the achieved coherence and systematicity in the annotations.

Finally, the dimension of **usefulness** relates the annotated text to applications and understanding. "Applications", in this case, covers subsequent analysis steps as well as large scale analyses, while "understanding" refers to a hermeneutic interpretation of the text, that takes the annotations into account. Assuming corpora are annotated in accordance with the guideline (either manually or, in the case of large corpora, automatically), this dimension reflects how insightful they are, i.e., how "much" insight the annotations allow. Usefulness thus evaluates the insights gained by examining an annotated text or corpus.

The three dimensions allow a balanced evaluation of guidelines with diverse disciplinary and research backgrounds, aims, and understanding of narratological concepts. Focusing on only one of the dimensions will diminish the score in at least one other: A guideline addressing narratological theory exclusively might achieve a high score in the first dimension, but will be penalized in the second dimension, as mere theory is not very applicable. Optimizing for applicability could lead to guidelines that specify everything or nothing as narrative level, thus not being very useful. Finally, the blind optimization on usefulness will lead to guidelines that are unrelated to narratological theory. Thus, the challenge that this shared task poses to the guideline authors is to strike a balance between the three dimensions.

Arguably, an annotation guideline does not generally need to cover all three dimensions in order to be a useful guideline for a certain purpose. Guidelines that are detached or totally unrelated to a theoretical concept, for instance, could still address a relevant issue. Likewise, it is not always necessary to look at applications and aim, i.e., at the usefulness of a guideline. As guidelines and/or annotations are also an excellent tool for text analysis, their creation might be a sufficient research goal in its own right.

Implementation

A Multi-Purpose Questionnaire

In order to implement the three-dimensional evaluation model, we associated each dimension with a number of specific questions to be answered for each guideline. The questions represent different aspects of each dimension and

should be answerable directly for a guideline. Section 4 lists each question with a brief description.

The questions were made available to the participants before they submitted their guidelines. In the evaluation, they were used in two ways: Firstly, they provided a guide for qualitative evaluation. By following the online questionnaire we distributed during the workshop and discussing each question for each guideline, we assured that the same criteria were employed in the judgement of each guideline and that the same aspects were covered in the discussion. This is important to ensure both fairness and coherence in the evaluation. The discussion was quite extensive and thus difficult to document, but all teams described it as very helpful. The discussion gave rise to a number of guideline improvements, which will be documented in the second volume of this special issue.

Secondly, the questions were answered quantitatively. Each question was evaluated on a 4-point Likert scale, i.e., participants were asked to assign points for each guideline in each question with more points reflecting the more favorable choice. Thus, if guideline A has higher score than guideline B, it is considered the better guideline.

Our evaluation defined four questions for the dimensions of conceptual coverage and usefulness, and two questions for the dimension of applicability. In order to weigh the dimensions equally, two more scores regarding applicability were provided through the inter-annotator agreement score (see below), scaled to lie between one and four points. In the end, each guideline was given four scores in each dimension, which were added up, first by dimension and then to a total score. Each team evaluated all the other guidelines, leading to seven judgements and thus scores per question per guideline.

Questionnaire

Conceptual Coverage

1. Is the narrative level concept explicitly described?

Explanation: Narrative levels can be described or defined. This depends on the narratology used; some of them are structuralist, others are post-structuralist. Regardless of the mode, is the description/definition understandable and clear?

- 1: I did not understand what the guideline describes as "narrative level".
- 4: I fully understood the concept described in the guideline.

2. Is the narrative level concept based on existing concepts?

Explanation: The level concepts can be self-designed, oriented on existing narratologies or copied from an existing level definition

- 1: The theory relation of the used level concept is not clear.
- 4: It is clearly mentioned whether the level concept is made up or (partially) based on a theory.

3. How comprehensive is the guideline with respect to aspects of the theory? Does it omit something?

Explanation: If the guideline is based on a theory or multiple theories, does it include the whole theory or only parts of it? Are there reasons mentioned why aspects are in-/excluded?

- 1: The guideline does not clearly state the extension of its dependence on theory/ies.
- 4: The guideline unambiguously states the scope of its theory-dependance.

4. How adequately is the narrative level concept implemented by this guideline in respect to narrative levels?

Explanation: Narratologies differ in their complexity. Firstly, you have to decide whether complexity or simplicity (in relation to x) is desirable, then you have to answer:

- 1: The guideline is too simple or too complex for narrative levels and thus not adequate.
- 4: The guideline's complexity is adequate.

Applicability

1. How easy is it to apply the guideline for researchers with a narratological background?

Explanation: The question asks for an assessment of the ease of use of the guideline for an annotator with some narratological background. Indicators can be: Complexity of the concepts, length of the guideline, clarity of examples, clear structure, difficulty of finding special cases, etc.

- 1: Even as a narratology expert, I needed to read the guideline multiple times and/or read additional literature.
- 4: The guideline is very easy to apply, and I always knew what to do.

2. How easy is it to apply the guideline for researchers without a narratological background?

Explanation: The question asks for an assessment of the ease of use of the guideline if we assume an annotator who doesn't have a narratological background (e.g., an undergraduate student). Indicators can be: Complexity of the concepts, length of the guideline, use of terminology, clarity of examples, reference to examples only by citation, clear structure, difficulty of finding special cases, etc.

- 1: Non-experts have no chance to use this guideline.
- 4: The guideline is very easy to apply, and non-experts can use them straight away.

3./4. Inter-annotator agreement: gamma scores (see below)

Usefulness

1. Thought experiment: Assuming that the narrative levels defined in the annotation guideline can be detected automatically on a huge corpus. How helpful are these narrative levels for an interesting corpus analysis?

Explanation: This question focuses on the relevance of the narrative level annotations for textual analysis of large amounts of texts, e.g., for the analysis of developments over time with regard to narrative levels or a classification of texts with regards to genre, based on narrative levels.

- 1: The narrative levels annotations are irrelevant for corpus analysis.
- 4: The annotations provide interesting data for corpus analysis.

2. How helpful are they as an input layer for subsequent corpus or single text analysis steps (that depend on narrative levels)?

Explanation: The analysis of some other textual phenomena depends on narrative levels, e.g., chronology should be analyzed within each narrative level before analyzing it for the whole text. This question asks whether the analysis of such phenomena is possible or even better when based on the narrative level annotations.

- 1: The usage of the narrative levels annotations makes no difference for subsequent analyses.
- 4: Subsequent analyses are possible only because of the narrative level annotations.

3. Do you gain new insights about narrative levels in texts by applying the foreign guideline, compared to the application of your own guideline?

Explanation: In most cases annotating a text in accordance to a guideline changes the evaluation of textual phenomena in the text, e.g., the quality (or quantity) of

narrative levels in the text.

- 1: It doesn't make a difference—I get no additional insights with the foreign guideline.
- 4: I gain a lot of new insights about narrative levels in texts based on this guideline.

4. Does the application of this guideline influence your interpretation of a text?

Explanation: Interpretations are normally based on the analysis of a text and thus on the observation of the presence (or absence) of certain textual phenomena. Therefore, the application of the guidelines may result in annotations that are relevant for your interpretation, e.g. the detection of a narrative level of a certain type may influence your interpretation of the reliability of a narrator.

- 1: My interpretation is independent from the annotations based on the guideline.
- 4: My interpretation is based primarily on the annotations based on the guideline.

Measuring Inter-Annotator Agreement

In this shared task, we employed the metric γ (gamma) as developed by Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier.[1] Its final score combines observed disagreement with chance disagreement (γ is thus calculated using *dis*agreements, while most metrics are calculated using agreements). This is done in order to be able to compare evaluation schemes with different complexities and to avoid favouring more simple schemes (if the scheme is simpler, chance agreement is higher). Gamma is thus calculated as shown in equation 1, with δ_0 and δ_e for the observed and expected disagreement respectively.

$$\gamma = 1 - \frac{\delta_o}{\delta_e} \qquad (1)$$

Chance Disagreement δ_e. For calculating the chance disagreement, Gamma takes the real annotations provided by an annotator, splits the text at a random point, and permutes the two parts. This is done repeatedly, until the disagreement in the permuted "text" approximates the real disagreement (in the entire population) with high confidence (above 95%). Based on these annotations, the

[1] "The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment," *Computational Linguistics* 41, no. 3 (2015): 437-79.

chance disagreement can be calculated in the same way as observed disagreement.

Observed Disagreement δ_0. Calculating the observed disagreement is based on an alignment and the pairwise comparison of the annotated segments. The alignment encodes which annotation of annotator 1 corresponds to which annotation of annotator 2 and is created in such a way that the overall inter-annotator agreement is maximal, i.e., all possible alignments are considered. For each possible alignment, the algorithm calculates an averaged observed disagreement by comparing the aligned segments.

For two aligned segments, Gamma considers both the positional and categorial disagreement. The *positional disagreement* expresses how different the position of two aligned segments is and is calculated as shown in equation 2. The functions *end(x)* and *start(x)* refer to the start and end position of the annotated segments, which is measured in token positions.

$$d_{pos}(u,v) = \left(\frac{|start(u) - start(v)| + |end(u) - end(v)|}{(end(u) - start(u)) + (end(v) - start(v))}\right)^2 \qquad (2)$$

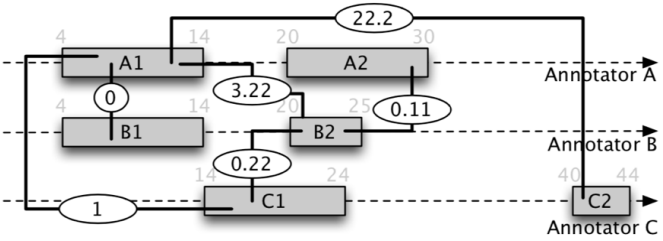

Figure 2: *Example calculations of positional disagreement. Grey numbers show index positions, numbers in white oval shapes show the calculated disagreement (Mathet, Widlöcher, and Métivier, 451).*

In equation 2, the enumerator represents the difference between the starting and end positions of the two annotations, while the denominator incorporates the length of the respective annotations. Figure 2 shows several example situations and the resulting positional disagreement score. As can be seen, numbers between zero and one indicate some overlap; if d_{pos} >1, the two annotations do not overlap. There is no upper limit on the positional disagreement. If the annotations differ widely in their position (e.g., are placed at the beginning and end of the text), they get a d_{pos} -value that is roughly as high as the text is long.

Incorporating *categorical disagreement* (d_{cat}) allows to evaluate whether different annotation categories have been selected. If, for example, annotator 1 has assigned category A, while annotator 2 has assigned category B, a category disagreement is noted. Using a matrix, the difference between each pair of categories can be weighted by assigning a number between zero and one. Thus, a user of gamma can express that using category A instead of B is less severe than using A instead of C. There is, however, no way to automatically determine the severity of categorical disagreement. Instead, it is a preference that the user of Gamma has to provide.

In our evaluation, categories play a minor role and have thus been treated as equally distant: If a guideline specifies multiple categories, all pairs have been assigned a distance of one. Features' values attached to the annotation have been suffixed to the category name, so that differences between features are treated in the same way as category disagreement.

Finally, equation 3 shows how the two sub metrics are combined, using α and β to express weighting (in our setting, both are set to one: $\alpha = \beta = 1$).

$$d_{combi}(u, v) = \alpha d_{pos}(u, v) + \beta d_{cat}(u, v) \qquad (3)$$

To measure Gamma, we employed an implementation provided by the developers on their web page. The way expected disagreement is calculated here leads to issues when annotations are sparse. If a single annotation covers the entire text, which is entirely plausible for narrative level annotation, there is no way to split the text and reshuffle the annotations. To circumvent this, we calculated Gamma individually for each text and on all eight texts concatenated together. The latter score was then used for the final ranking.

Integration of the Evaluation Scores

The final score for each guideline was calculated as follows:

1. For each of the ten questions, the arithmetic mean over all answers is calculated. This results in ten values, distributed over three dimensions: four questions/values in the first dimension, two questions/values in the second and four questions/values in the third.

2. The Gamma scores are scaled to the interval of [1;4] and added twice as the scores of "virtual questions" in the second dimension. This results in four values per dimension, each in the interval [4;16].

3. In each dimension, all four (mean) values are added up.

This results in one score for each dimension, so guidelines can be ranked accordingly. As an overall score, we calculated the sum of the scores in all dimensions.