

A Unified Text Annotation Workflow for Diverse Goals

Janis Pagel, Nils Reiter, Ina Rösiger, Sarah Schulz

Institute for Natural Language Processing

University of Stuttgart

{janis.pagel,nils.reiter,ina.roesiger,sarah.schulz}@ims.uni-stuttgart.de

Abstract

In computational linguistics (CL), annotation is used with the goal of compiling data as the basis for machine learning approaches and automation. At the same time, in the Humanities scholars use annotation in the form of note-taking while reading texts. We claim that with the development of Digital Humanities (DH), annotation has become a method that can be utilized as a means to support interpretation and develop theories. In this paper, we show how these different annotation goals can be modeled in a unified workflow. We reflect on the components of this workflow and give examples for how annotation can contribute additional value in the context of DH projects.

1. Introduction

Annotation is a technique that we define very broadly as the process of enriching textual data with additional data. Our focus is on annotation as a *process and methodology*, and not on the created annotations as data objects or subject of analysis. We also focus on annotation tasks that have interpretative or associative aspects (i.e., are related to the explicit or implicit content of a text).¹

Annotation projects in computational linguistics (CL) have created a large volume of corpora annotated with linguistic notions (e.g., parts of speech, semantic roles, etc.). Furthermore, annotation projects in CL put most emphasis on consistent and agreeable decisions across annotators, as they are often used as (training/testing) data for machine learning methods.

In the Humanities, the individual is a recognized authority. Thus, annotations done in the Humanities do not necessarily follow the same inter-subjective paradigm. But even for the subjective, individual interpretation of, for instance, a literary text, annotation (e.g., adding notes to the margin) often plays a role, albeit sometimes an implicit one. Rendering this process explicitly has its benefits, as explicit annotations can support the interpretation by making it clearer and unambiguous.

In addition, a future perspective for the Humanities could be a more inter-subjective process of theory development. One approach to achieve this goal is the integration of the annotation methodology into Humanities research by applying theoretical notions to texts and iteratively sharpening these notions.

This paper compares annotation processes prevalent in CL with processes employed in the (Digital) Humanities. We argue that although the annotation processes serve different goals and set different priorities, they have much in common and can actually be integrated into a single conceptual model. In addition, we argue that annotation can be a very productive tool to improve theoretical definitions in the Humanities, which is a new way of using annotation.

¹Although adding structural markup to a text, as is done when creating editions in TEI/XML, is technically a very similar process, it is not related to the content and not interpretative.

2. Diverse Annotation Goals

Firstly, **exploratory annotation** offers to become familiar with a text (or another data object) in a semi-structured way. This way of annotating is the closest to long-lasting traditions of annotation in traditional Humanities (Bradley, 2008) where interesting ideas or important aspects that emerged while reading are noted down on the margin of a page. Bradley (2012) states that “this kind of annotation, indeed note-taking more generally, provides one of the bases for much scholarly research in the humanities. In this view note-taking fits into the activity of developing a personal interpretation of the materials the reader is interested in.” Thus, the goal of this kind of annotation is to end up with preliminary text knowledge that enables the scholar to formulate a more concrete research question or hypothesis. This question/hypothesis can later be addressed with a theoretical basis, while the initial reading is done without specific assumptions or questions.

Secondly, **conceptualizing annotation** aims at improving definitions of theoretical notions or pre-theoretic observations in need of explaining. Both are often described in secondary literature, but rarely defined in a way that they are applicable to new texts. Trying to apply them to texts through annotation is a way to improve their definitions as this process reveals differences in understanding. The core mechanism here is to identify instances of disagreement between different annotators and to refine the definitions until a sufficient agreement is reached.

Thirdly, **explicating annotation** aims at providing a formal representation of the textual basis for an interpretation hypothesis. While interpretation hypotheses (e.g., in literary studies) are typically based on textual evidence (at least partially), the text segments are not explicitly marked, and the argumentation path from text segments to the interpretation remains implicit. Explicating annotations make these steps explicit and formal. These annotations are not restricted to a single phenomenon, but cover all phenomena that are needed for an interpretation. In this setup, the main goal is not to create a single ‘true’ annotation, but different plausible ones that represent different readings of the text.

Fourthly, **automation-oriented annotation** (cf. Hovy and Lavid, 2010; Pustejovsky and Stubbs, 2012) targets the compilation of consistently annotated data as training and

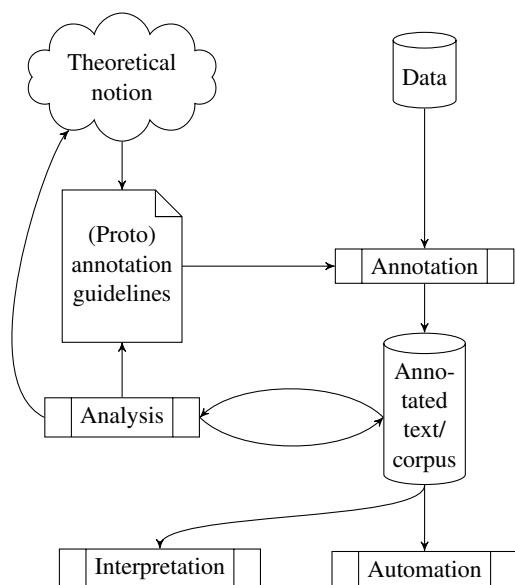


Figure 1: Annotation workflow schema. Arrows indicate (rough) temporal sequence.

testing material for automatic annotation tools. Consistency of the annotation is of utmost importance for the automation, because inconsistencies negatively impact the classification performance. Annotation projects that generate training/testing data put emphasis on high inter-annotator agreement.

These use cases for annotation methodology are not mutually exclusive. In fact, it is difficult not to at least touch on the different aspects of the other goals, even if one has a single goal in mind. Annotation to generate training/testing data, for instance, often discovers issues in the definitions and annotation schemata have to be refined, etc. This has an impact on the conceptual world, even if this impact is not considered or published within a single project.

3. A Unified Annotation Workflow

Figure 1 visualizes a model for an annotation workflow that encompasses annotations aimed at various goals. It describes both the annotation model prevalent in CL, annotation models originating in Humanities scholarship, and use cases that are new and specific to DH. The workflow does not imply that every annotation project employs every part of it, or that everything is done within a single project. Depending on the goal of the annotation, different areas receive more or less emphasis or are entirely ignored. Generally, the different annotation processes could also be seen as phases that new phenomena undergo until an inter-subjective understanding can be reached. Annotation guidelines established in one project can very well be continued or elaborated in the next.

One starting point is a **theoretical notion**. We use the term ‘notion’ here to include a variety of cases: The notion can be described/predicted based on a full-fledged theory (e.g., part of speech tags or narrative levels), but it can also be based on an observation in text data that needs to be explained or has been discussed in previous scholarly literature (e.g., similarities in the character representation in

adaptations of a literary piece). Theoretical notions are represented with a cloud to indicate that they often have ‘fuzzy edges’ and their application to textual data includes interpretation steps.

Theoretical notions interact with **data** in a complex way: Observations are made on data, even if quite indirectly or only transmitted through past scientific discourse. A concrete collection of data can never be chosen truly at random and thus assumes at least the broad limitation to a field of interest. The selection of data introduces a bias and restricts the space of possible findings. Canonization/standardization processes lead to a narrowed view, and make certain phenomena unobservable by design. This is irrespective of the exact state of the theoretical notion. Therefore, data selection must receive a big deal of attention and criteria for the selection need to be made explicit for users of the collection in order to make research transparent.

The actual annotation is (conceptually) always based on **annotation guidelines**. Initially, when a theoretical concept is first annotated, the guidelines might only be a fixation on a specific theoretical work (e.g., Genette (1980)) or a part of it (e.g., narrative levels). Iterations in the annotation workflow can lead to more and more elaborate annotation guidelines, that might even deviate from the theoretical concept. For the every-day business of the annotation, guidelines serve as a mediator between theoretical notions and the actual annotation of them. Ideally, this allows non-experts to do the annotations (e.g., student assistants or crowd workers). Annotation guidelines are often related to specific texts or corpora. When theoretical concepts are broken down for non-experts, they are often described in terms related to the corpus to be annotated; difficult, but irrelevant aspects might be ignored entirely. Limiting guidelines to certain aspects of a theory is reasonable in many projects, but makes guidelines less interchangeable.

The actual **annotation process** then consists of reading texts, highlighting/selecting textual portions, and linking them to the categories defined in the guidelines. Sometimes, additional features of an instance of the notion are annotated. Depending on the annotation aims, annotations might be done in parallel, i.e., multiple annotators annotate the same text in parallel. This allows comparing the annotations directly in order to analyze potential shortcomings of the annotation guidelines. An additional parameter in the annotation process is that some annotations are based on linguistic units (e.g., phrases) that might be pre-annotated in the text. While annotation can in principle be done on paper, annotation tools can support the annotation process by proposing candidate annotations or sharing annotations digitally.

The immediate outcome of the annotation process is an **annotated corpus**. One obvious type of **analysis** is then to test certain hypotheses or assumptions against the newly created data. This type of analysis benefits a better understanding of the theory, for example in the form of more fine-grained theoretical notions. Analyzing actual data can also lead to finding evidence for or against certain theoretical claims. These results can then be used to re-fine the underlying theory. A different type of analyses is based on

the disagreements as produced by multiple annotators. The main goal of this type of analysis is to ensure that i) the annotation guidelines are sufficiently exact and well-defined and ii) they have been read, understood and followed by the annotators. A general mechanism is to manually inspect the annotations in which the annotators disagree, i.e., have made different annotation decisions. This can be done by the annotators themselves, or the annotators' supervisors. Quantitatively, the rate of disagreement can be expressed as the inter-annotator agreement (IAA) which is typically reported in the documentation accompanying a corpus release. While measuring IAA has quite a long tradition (Cohen, 1960), the discussion on how to exactly quantify IAA is still ongoing (Mathet et al., 2015). Measuring IAA quantitatively is especially important when comparing different annotation guidelines or annotated corpora, and can also serve as an upper bound for machine performance. If the goal of the annotation is to develop theoretical concepts, inspecting the actual disagreement made by the annotators is more insightful. Gius and Jacke (2017) propose to categorize disagreements in four categories, based on their causes: i) annotation mistakes, ii) annotation guideline shortcomings, iii) diverging assumptions and iv) 'real' ambiguities. Annotation mistakes can immediately be fixed, categories ii) and iii) require adaptation of the annotation guidelines. If disagreements of category iv) cannot be resolved by taking additional context into account, they remain annotated in the corpus.

Once an annotated corpus is available, two different subsequent steps are possible: Interpretation and automation. **Interpretation** of a text on a basis of annotations leads to an additional reading which is not established on vague observations, but on concrete annotations. Eventually, this will also lead to a more inter-subjective interpretation of texts and theories. We will not go into detail about the **automation** process, but it typically requires annotated data. One assumption made in CL is that the annotations are unambiguous, i.e., that all disagreements have been resolved. How true disagreements or unresolvable ambiguities can be handled with respect to the automation is not clear yet. Gius and Jacke (2017) suggest differently parameterized models for automatic prediction, at least for disagreement category iii). For example, applying a certain category might require a decision on a more basic related category. In a tool used for the automated detection of a certain notion, this parameter can be manually set to enforce a certain reading. However, they leave open the question how this could be realized for cases of disagreement stemming from a valid textual ambiguity. In the future, it would be beneficial if statistical methods could handle truly ambiguous data and if the annotations were not 'validated' to one gold version.

4. Exemplary Annotation Projects

We discuss several projects developed in the context of DH, in order to exemplify the different goals of annotation as well as showcase different paths that projects might take on our annotation workflow.

Exploratory An example for exploratory annotation is note-taking. One early project to support this for the DH

world is the Pliny project (Bradley, 2008). Pliny is a software released in 2009 to explore some of the new potential for annotation in the digital world. It is meant to support the traditional scholarship workflow (Bradley, 2008) by enabling the process of note-taking and the recording of initial reactions to a text with the goal of a subsequent phase in which a research question is developed. The developers give the example of a web page² where the user notes down observations they make while browsing the page. In our workflow, this phase of annotation corresponds to a pre-theoretical stage where `Data` triggers the `Annotation`. This can in a next step potentially result in the `Analysis of the Annotated text` which can lead to `annotation guidelines`. However, even though they claim that they move the "traditional way" of note-taking into the digital world, Pliny seems to lack acceptance in the DH scholarly world: there are few – if any – projects to be found that make use of the tool. However, this could also be an indication for an underdeveloped tradition of discussing methodology in the Humanities which results in a lack of publications of the process of annotation within specific projects.

A more recent project supporting exploratory annotations is the 3DH project³, which concentrates on the visualization and exploration of Humanities data from a DH perspective in form of exploratory free annotations (Kleymann et al., 2018). This aids the goal of sharpening a research question.

For this kind of annotation, IAA is not important because it predominantly serves the aim of developing an understanding of important concepts and potential departure points for a research project.

Conceptualizing As an example for conceptualizing annotation, we want to cite Moretti (2013). He describes the departure from the definition of "character-space" by Woloch and Woloch (2003). The operationalization of this literary theory by approximating it as the textual space that a character occupies, more concretely how many words a character speaks in a dramatic text, strengthens the underlying theory by leading "back from theories, through data, to the empirical world." (Moretti, 2013, p. 4). He deems this crucial for literary theories because it makes "some concepts 'actual' in the strong sense of the word." (Moretti, 2013, p.4). In our workflow, this project has a strong focus on the formalization of a `Theoretical notion`, thus the translation from the concept of character-space into the space of actual text portion. The annotation itself is trivial, however the annotated text is then used as a basis for `Interpretation`.

A more thorough attempt at using annotation to develop theoretical concepts has been made by (Bögel et al., 2015). The goal of the project heureCLÉA is to annotate time-related narrative phenomena in literary texts. The published guidelines⁴ are already more specific than the underlying

²From the Proceedings of the Old Bailey site: <http://www.oldbaileyonline.org>

³<http://threedh.net/3dh/>

⁴<http://heureclea.de/wp-content/uploads/2016/11/guidelinesV2.pdf>

theory, as they define how to deal with, e.g., hypothetical prolepses. This process of refining theoretical notion through annotation can also be conducted as a shared task (cf. Reiter et al. (2017) for a focus on embedded narratives).

Potentially, the confrontation of the theory with an inter-subjective understanding will lead to implications for this theory. For this kind of annotation, IAA builds a basis to discuss predefined theoretical concepts on an inter-subjective basis. Thus, IAA is a measurement that can provide information on how specified a theory is and how objectively it allows the definition of indicators to verify it.

Another example for conceptualizing annotation is coreference annotation. Annotation of coreference is well established in CL and supported by already existing theoretical notions and guidelines (Pradhan et al., 2007; Dipper and Zinsmeister, 2009; Riester and Baumann, 2017). However, application of these guidelines on ‘new’ text types reveals the need to improve the guidelines further. A concrete example is the DH project ‘QuaDramA’⁵. First insights of the continuing work on the guidelines have been published in Rösiger et al. (2018). QuaDramA focuses on dramatic texts, and on gathering information about characters in particular. The project complies with the workflow as follows: Existing annotation guidelines were adopted and the annotation process initiated. After the first texts were annotated, the circle of analyzing the results was entered, meaning that the guidelines were adapted towards the data and specific problems and new texts were either annotated with the adopted guidelines or the existing annotated texts were revised in order to adopt them to the new version of guidelines as well. This is the conceptualizing step, since the new guidelines reflect new insights, which were gained from looking at concrete coreference phenomena. Finally, a single text might also be interpreted based on the given annotations. A possible case in the setting of coreference and dramas might be to come to a different interpretation of a play based on agreeing on a different reference for an ambiguously mentioned character. Depending on the reference of that character, the plot might be seen in a new light and require diverging interpretations.

Explicating An example for an explicating annotation project in an early stage is the work presented in (Nantke and Schlupkothén, 2018). The authors focus on the annotation of intertextual references, in order to formalize possible interpretations of a text. Only a subset of the proposed formalizations are actually textual annotations in the narrow sense – others are relations between textual annotations, or between textual annotations and (digital representations of) historical context. On a technical level, the annotations as well as the relations are represented using semantic web technologies. It is important to realize that these annotations do not cover a single phenomenon. Instead, they may include a large number of “basic annotations” for various phenomena. Given the complexity of these annotations, a large scale annotation project seems difficult to realize – annotations of this kind are mainly produced for

a single text. This makes the inter-subjective agreement less important. With respect to the workflow presented in Figure 1, explicating annotations employ *theoretical notions* as the basic inventory of textual evidence (if possible using *annotation guidelines*), without aiming to improve on them. Instead, projects such as these take the right path using *Annotation* which results in an *annotated text*, followed by an interpretation or a justification of the interpretation using the annotations.

Automation-oriented The last type of annotation that we want to discuss is the automation-oriented one that is prevalent in computational linguistics. As shown in Figure 1, the purpose of the annotation hereby is to enable automation, i.e. provide data for the (often statistical) algorithms to learn from, or in rule-based approaches, to function as evaluation data.

One prominent example for annotations that are used as input to a fully automated approach is the annotation of parts of speech (pos). Parts of speech is one of the CL task that is best suited as an example for the automation-oriented annotation, as it is a task that is conceptionally clear, which can be seen in the extremely high inter-annotator agreement which is reported for this task. The recent GRAIN corpus (Schweitzer et al., 2018), for example, contains annotations by three annotators for German radio interviews, which comprises rather complex and spontaneous speech. In their paper, they state a pair-wise Cohens κ of 0.97, which is generally described as almost perfect agreement. The fact that the annotation can be consistently performed by humans is a necessary requirement for the development of automatic tools. As a consequence, pos tagging has been one of the first CL tasks for which the performance of automatic tools has reached a satisfactory level, with an accuracy of over 97 percent (cf. Manning (2011)), and is now considered an almost solved task, at least for standard text.

Pos tagging has also been applied to texts from the DH domain, e.g. historical text, where of course the performance of off-the-shelf tools is not satisfactory. However, Schulz and Kuhn (2016) have shown that, for Middle High German, a small amount of annotated data (e.g. around 200 sentences) can already lead to reasonable results of automatic systems.

5. Discussion and Conclusions

We show that annotation can not only function as a means to create training material for machine learning approaches. Annotation as a process can function as a tool to develop a focused understanding of relevant concepts that can be found in texts as well as an instrument for the specification and verification of theoretical or pre-theoretical concepts. This is especially fruitful for disciplines such as literary studies where concepts often stay underspecified in the scholarly discourse which complicates an inter-subjective exchange. Generally, the annotation of non-standard (from the point of view of CL) texts can help uncovering new phenomena which call for an adaptation or extension of assumptions. E.g., assuming the existence of a ‘ground truth’ – a single annotation that is correct – potentially needs to be relaxed for literary texts concepts, because reading and interpreting a text can allow for different and yet correct

⁵<https://quadrana.github.io>

readings. It remains a challenge for machine learning methods how to deal with these ‘real’ ambiguities with respect to training and evaluation of automatic systems.

Another consideration that these different types of annotations trigger is the choice of annotation tool: Annotation tools developed in CL (e.g., WebAnno (Yimam et al., 2013) or MMAX2 (Müller and Strube, 2006)) naturally incorporate standards used in CL. They typically include a method to compare annotations, but the actual annotation categories and schemes need to be defined in advance. Annotation tools used for exploratory annotation have been developed, but they work quite differently: The tool developed in the 3DH project (Kleymann et al., 2018) allows marking arbitrary text spans and offers much more functionality on interacting with these text spans (e.g., grouping and/or visualizing them). Explicating annotations would contain a lot of formal relations that are not directly text-related. For these, a generic ontology development tool such as Protégé (Musen, 2015) might be well suited. In any case, the relation between functionality offered by the tool and the goal of the annotation process is still an under-researched area. We have noticed that there are almost no documented DH projects that document the use of annotation as a means to explore new texts or sharpen research questions. Not surprisingly, automation-oriented annotations are not difficult to find.

In summary, we have described a workflow for annotations performed in the DH. The workflow aims to be as open and flexible as possible, in order to account for the different possible perspectives and fields coming together in the DH, while at the same time focusing on and requiring steps that should be necessarily shared by all annotation undertakings. We define four major goals that the different branches of DH might pursue: Exploratory, conceptualizing, explicating, and automation-oriented goals. We discuss the purpose and differences of each goal on a general level, followed by an examination of concrete projects in the DH following one of these goals. This examination also showcases the use of the workflow in different settings, emphasizing its flexibility. We believe that our workflow is generally applicable for all the kinds of DH goals and hope that in the future more projects will make use of annotation in order to view old questions of the humanities in a new perspective.

- Bögel, T., Gertz, M., Gius, E., Jacke, J., Meister, J. C., Petris, M., and Strötgen, J. (2015). Collaborative Text Annotation Meets Machine Learning: heureCLÉA, a Digital Heuristic of Narrative. *DHCommons*, 1.
- Bradley, J. (2008). Thinking about interpretation: Pliny and scholarship in the humanities. *Literary and Linguistic Computing*, 23(3):263–279.
- Bradley, J. (2012). Towards a richer sense of digital annotation: Moving beyond a “media” orientation of the annotation of digital objects. *Digital Humanities Quarterly*, 6(2).
- Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.
- Dipper, S. and Zinsmeister, H. (2009). Annotating dis-

- course anaphora. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, ACL-IJCNLP, pages 166–169, Singapore.
- Genette, G. (1980). *Narrative Discourse – An Essay in Method*. Cornell University Press, Ithaca, New York. Translated by Jane E. Lewin.
- Gius, E. and Jacke, J. (2017). The hermeneutic profit of annotation: On preventing and fostering disagreement in literary analysis. *International Journal of Humanities and Arts Computing*, 11(2):233–254.
- Hovy, E. and Lavid, J. (2010). Towards a ‘science’ of corpus annotation: A new methodological challenge for corpus linguistics. *International Journal of Translation Studies*, 22(1):13–36.
- Kleymann, R., Meister, J. C., and Stange, J.-E. (2018). Perspektiven kritischer Interfaces für die Digital Humanities im 3DH-Projekt. In *Book of Abstracts of DHd 2018*, Cologne, Germany, February.
- Manning, C. D. (2011). Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part I*, CICLing’11, pages 171–189, Berlin, Heidelberg. Springer-Verlag.
- Mathet, Y., Widlöcher, A., and Métivier, J.-P. (2015). The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479.
- Moretti, F. (2013). “Operationalizing”: or, the function of measurement in modern literary theory.
- Müller, C. and Strube, M. (2006). Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, et al., editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany.
- Musen, M. (2015). The Protégé project: A look back and a look forward. *AI Matters*, 1(4), June.
- Nantke, J. and Schlupkothen, F. (2018). Zwischen Polysemie und Formalisierung: Mehrstufige Modellierung komplexer intertextueller Relationen als Annäherung an ein ‚literarisches‘ Semantic Web. In *Proceedings of DHd*.
- Pradhan, S. S., Ramshaw, L., Weischedel, R., Macbride, J., and Micciulla, L. (2007). Unrestricted coreference: Identifying entities and events. In *International Conference on Semantic Computing*.
- Pustejovsky, J. and Stubbs, A. (2012). *Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications*. O’Reilly Media, Sebastopol, Boston, Farnham.
- Reiter, N., Gius, E., Strötgen, J., and Willand, M. (2017). A Shared Task for a Shared Goal - Systematic Annotation of Literary Texts. In *Digital Humanities 2017: Conference Abstracts*, Montreal, Canada, August.
- Riester, A. and Baumann, S. (2017). The RefLex Scheme – Annotation guidelines. SinSpeC. Working papers of the SFB 732 Vol. 14, University of Stuttgart.
- Rösiger, I., Schulz, S., and Reiter, N. (2018). Towards Coreference for Literary Text: Analyzing Domain-

- Specific Phenomena. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, Santa Fe, USA.
- Schulz, S. and Kuhn, J. (2016). Learning from within? comparing pos tagging approaches for historical text. In Nicoletta Calzolari, et al., editors, *LREC*. European Language Resources Association (ELRA).
- Schweitzer, K., Eckart, K., Gärtner, M., Faleńska, A., Rieger, A., Rösiger, I., Schweitzer, A., Stehwien, S., and Kuhn, J. (2018). German radio interviews: The GRAIN release of the SFB732 Silver Standard Collection. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, LREC 2018.
- Woloch, A. and Woloch, P. (2003). *The One Vs. the Many: Minor Characters and the Space of the Protagonist in the Novel*. ACLS Humanities E-Book. Princeton University Press.
- Yimam, S. M., Gurevych, I., Eckart de Castilho, R., and Biemann, C. (2013). Webanno: A flexible, web-based and visually supported system for distributed annotations. In Miriam Butt et al., editors, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 1–6. Association for Computational Linguistics.