# Inter-Rater Agreement and Usability: A Comparative Evaluation of Annotation Tools for Sentiment Annotation

Thomas Schmidt, Brigitte Winterl, Milena Maul, Alina Schark, Andrea Vlad
and Christian Wolff[1]

**Abstract:** We present the results of a comparative evaluation study of five annotation tools with 50 participants in the context of sentiment and emotion annotation of literary texts. Ten participants per tool annotated 50 speeches of the play *Emilia Galotti* by *G. E. Lessing*. We evaluate the tools via standard usability and user experience questionnaires, by measuring the time needed for the annotation, and via semi-structured interviews. Based on the results we formulate a recommendation. In addition, we discuss and compare the usability metrics and methods to develop best practices for tool selection in similar contexts. Furthermore, we also highlight the relationship between inter-rater agreement and usability metrics as well as the effect of the chosen tool on annotation behavior.

**Keywords:** Sentiment Annotation, Usability Engineering, Usability, Inter-rater agreement, Annotation, Sentiment Analysis, Emotion Analysis, Annotation Tools

## 1 Introduction

In recent years, computational methods of sentiment and emotion analysis have found their way into several areas of Digital Humanities (DH), most notable computational literary studies (cf. [KK18a]). The goal of this method is to analyze and predict sentiments and emotions in written text [Li16]. Concerning literary texts, recent research explores the application of sentiment analysis methods in fairy tales [ARS05], novels [KK11] and historic plays [Mo11, NB13, SB18, SBD18a] predominantly with rule-based prediction methods. However, when compared to human annotated gold standards, prediction accuracies are rather low [SB18, SBW19, KK18b]. Therefore, current studies strive to acquire large-scale sentiment- and emotion-annotated corpora that can be used for advanced machine learning purposes. However, annotators of studies for manual sentiment annotation in the context of literary texts report that this is a tedious, challenging and time-consuming task [SBD18b]. Furthermore, due to the subjective nature of literary texts, agreement among annotators tends to be rather low and at best moderate [SB18, KK18b] which also hinders the design of valuable corpora.

To facilitate and improve the sentiment annotation process for the annotators we want to highlight the role of the annotation tool. Depending on the specific task, researchers in DH can select from various annotation tools of different domains for manual annotation. In most studies, the selection of a specific tool seems arbitrary and reflections and expla-

[1] University of Regensburg, Media Informatics Group, Regensburg, Germany, firstname.lastname@ur.de

doi:10.18420/inf2019_ws12

nations about the selection process as well as about the usability and user experience of the used tools are often missing. Furthermore, systematic evaluation of annotation tools are rare [Bu12], mostly done heuristically via an expert analysis following usability guidelines [Ga04, SP05, Bu12] with basic usability tests with rather low sample sizes [HP15] or combining both approaches [DGS04]. At the same time, there are no specific recommendations which methods to employ from the plethora of quantitative and qualitative usability and user experience (UX)-metrics [AT13] for evaluating semantic annotation tasks. Furthermore, while usability and user experience are important aspects of an annotation tool, one of the most important metrics in the description of semantic corpora is the inter-annotator agreement. Additionally, oftentimes the quality of manual annotation tools is measured via task completion rates and the correctness of the annotations. However, when dealing with more subjective annotation types that do not have a definitive *right* or *wrong* annotation (as is often the case with semantic annotation) the usage of these metrics is not possible. Therefore, we also propose to integrate agreement metrics as substitute metric when evaluating annotation tools. Though it could sound counterintuitive that the annotation tool has an influence on the agreement among annotators when the annotation schema and the overall functionality is the same, we want to investigate if the tool might explain variance in agreement statistics or annotation behavior in general. We assume that usability and complexity of a tool can influence concentration and motivation of annotators and therefore influences the general annotation behavior.

In the following, we present results in the context of sentiment and emotion annotation for German historic plays. The study followed a between-subject design with 50 participants and five different tools. Ten participants per tool were presented with the same annotation tasks and multiple usability and UX-metrics as well as annotation and agreement statistics were gathered and compared. The research goals (RG) of this study are:

(RG1) To identify the most user-friendly tool for this specific annotation task and compare the tools to each other.
(RG2) To compare and analyze usability and UX-metrics and discuss which are most fitting in the context of sentiment annotation tasks in DH.
(RG3) To examine if those usability and UX-metrics relate to annotation behavior and agreement statistics.
(RG4) To examine if the annotation behavior and agreement statistics are influenced by the tool chosen.

## 2 Methods

### 2.1 Annotation Material, Scheme, and Process

As material for the sentiment and emotion annotation, we used historic German plays, more specifically the play *Emilia Galotti* by the German playwright *Gotthold Ephraim Lessing*. Plays have been in the focus of sentiment analysis in computational literary

studies [NB13, SB18] and specifically for Lessing annotation studies have already been carried out [SBD18b].

All participants annotated the first 53 speeches of the play. A speech is a single utterance of a character separated by utterances of other characters and can consist of one or multiple sentences. All speeches were presented with the name of the character and the speech in the correct order. First, annotators had to annotate the sentiment, if the speech is rather positive, neutral or negative (we also refer to this concept as *polarity*). We instructed the annotators to annotate the sentiment that they feel is the most adequate for the given speech. In a second step, annotators could choose up to eight emotion classes (e.g. anger, sadness, surprise) they feel are present in the specific speech. Annotators could select no emotion or multiple ones. The entire annotation scheme and process was similar to studies by [SBD18b]. The annotation scheme and process were set up for every tool in a way that participants were able to perform the task with minimal effort and did not have to deal with any settings.

Before the start of the annotation task, a moderator explained the annotation process and the tool for every participant. The first three annotations were done together with the moderator and served as training. We do not include those three annotations into our analysis, thus only 50 speech annotations are used. The annotators were instructed to work in the pace they prefer and inform the annotator when the annotation was finished.. We did not include techniques like "thinking-aloud" since we wanted to measure the time needed for the task and methods like this may skew usability performance metrics. After the annotation participants had to fill out questionnaires and we conducted a semi-structured interview.

We chose a between-subject design so every annotator must annotate the same speeches and there is no influence concerning the individual annotation difficulty of the speeches. However, the individual characteristics of the annotators certainly have an influence on all metrics, so if the annotators of a specific tool have specific characters, data might get skewed. Nevertheless, we try to control this annotator-specific influence by gathering rather large sample sizes with 10 participants per tool.

## 2.2 Tool Selection

To approach the tool selection for our study systematically, we first collected a list of annotation tools by researching the web and contacting experts. The list consists of 29 tools and is available online[2]. From this list, we included five different tools into our study. First, we decided to include all tools that have been used in similar projects about sentiment annotation of literary texts. Therefore, we chose *Microsoft Word*, which has been used by [SBD18b]. The usage of tools like *Word* or *Excel* for annotation tasks is not uncommon in DH (e.g. [DBW17, SBD18b]) since those tools are well-known and

[2] https://docs.google.com/spreadsheets/d/1PygbNWEiNEY8QzqjPbiU9KfHDvc_SAazGAtKM1uK5gM/edit?usp=sharing

adaptable. To perform the sentiment and emotion annotation in *Word* every speech is presented with tables. Participants can mark their selection for a sentiment or emotion in a table (cf. [SBD18b]). Another tool that has been used for sentiment annotation of literary text by [KK18b] is *WebAnno*[3] [Yi13]. *WebAnno* is a web-based annotation tool that has become popular for semantic annotation in DH [Pe14]. We also included the tool *Sentimentator*[4] [ÖK18] since it has been used for sentiment annotation of a similar text sort: subtitles of movies. The *Sentimentator* is a web tool designed specifically for the context of sentiment and emotion annotation integrating gamification concepts. In addition, we also included two other tools often used in DH research: *CATMA*[5] (cf. [Bö15]) is one of the most popular tools for annotation in DH being used in research and education as well[6]. As last tool we also included *eMargin*[7] [KG12], which is a very adaptable online collaborative annotation tool usable for more general private annotations but also for research [AP17]. Overall, we selected a reasonable mixture of general as well as specialized tools to analyze the impact of the tool selection. Note that the annotation is presented and performed in a rather different way in every tool. However, we configured the setting of each tool in a way that the annotations are comparable. We will refer to these differences in more detail when discussing the results.

### 2.3 Usability and User Experience Methods Used in the Experiment

Usability professionals often differ between more subjective self-reported data and more objective performance metrics [AT13]. To get a holistic view on the comparison of the tools, we included both types of metrics. We gathered usability and UX-metrics that are rather established. One such metric to operationalize the performance is the time needed to complete a task (also called *time on task*, [AT13]). We measure the time needed for the entire annotation. The lower this metric the more efficient the tool.

The first questionnaire we employed is the *System Usability Scale* (SUS; [Br96]). The SUS is an established and validated instrument to measure usability [BKM09]. The SUS consists of 10 statements concerning the subjective overall usability of a tool. Users can agree upon these statements on a 5-point-Likert-scale. Via calculation recommendations, a tool can achieve up to 100 points for a "perfect" overall usability. To operationalize the concept of UX we use a short version of the *User Experience Questionnaire* (UEQ-S; [SHT17]), which is also an established questionnaire in usability engineering. The short version consists of eight semantic differentials like *boring-exciting*. Participants can mark their tendency towards an attribute on a 5-point scale. Another short questionnaire that gathered attention in usability engineering in recent years is the *NASA Task Load Index* (*NASA-TLX*; [HS88]). This questionnaire allows the assessment of the perceived workload of a task on multiple dimensions like mental, physical or temporal demand.

[3] https://webanno.github.io/webanno/
[4] https://github.com/Helsinki-NLP/sentimentator
[5] https://catma.de/
[6] For information about projects using CATMA visit: https://catma.de/documentation/affiliates/
[7] https://emargin.bcu.ac.uk/

For six dimensions, participants can rate the demand on a scale from 1 (very low) to 10 (very high). By adding up all eight values, we gather an overall value for the workload of the annotation task with a specific tool. This way of calculation was recommended and validated by [HS88]. We propose that this questionnaire is fitting for the evaluation of annotation tools in our context since it has been shown that the overall effort is high and the task challenging [SBD18b]. Therefore, tools that lower this effort can be regarded as better for this task.

Besides these standard questionnaires, we also integrated specific questions on the annotation task. Participants rated if they understood what they had to do during the annotation, how difficult the task was perceived and how confident they are about their annotations. Participants answered via 5-point Likert scales. [SBD18b] were able to gather further insights using a similar questionnaire. In addition, we also analyze metrics that are usually analyzed in annotation projects like the annotation distributions for every annotation layer and the inter-rater agreement per tool. Finally, participants also completed a general demographic questionnaire.

After completing all questionnaires, we also conducted a semi-structured interview with all participants about all positive and negative aspects they noticed. Due to length constraints, we will not report the results of those interviews in detail but integrate them when discussing the results in section 4.

### 2.4 Participants in the Study

Our sample consists of 50 participants, 10 for each tool with 26 female and 24 male participants. The youngest participant was 17 years old, the oldest 55 years ($M$=25.7); however, the majority of the participants were in the age group from 20-35 years ($n$=48). Most of the participants were students ($n$=28) or employed ($n$=20). We purposefully only chose non-experts in the context of literary studies and Lessing since this is the annotator group we want to focus on in further research.

## 3 Results

In the following, we first present descriptive statistics for all the metrics. Furthermore, we perform significance tests with the tool as independent variable and usability metrics and annotation distributions as dependent variables examining if there is a significant effect of the chosen tool. The significance level is chosen as $p$<.05.

### 3.1 Overall time and time per annotation

Table 1 illustrates the average time needed (in seconds) and the standard deviation for every tool. We will mark important results in the table as bold.

| | Measure | *WebAnno* | *Word* | *CATMA* | *eMargin* | *Sentimentator* |
|---|---|---|---|---|---|---|
| **Time** | M | 1530 | 1243 | 1491 | **1621** | **946** |
| | Sd | 395.12 | 289.15 | 296.63 | 597.5 | 193.49 |

Tab. 1: Descriptive statistics - Time and time per annotation

Using *eMargin* resulted in the highest duration for the annotation. Here, the annotation took around 29 minutes taking on average half a minute to perform an annotation, while the annotation was performed fastest with *Word* (20 minutes) and *Sentimentator* (15 minutes). Furthermore, the variance for *eMargin* is much larger, while the duration is rather stable among participants when using *Word* or *Sentimentator*. Using the tool as an independent variable we conducted a one-way between subjects ANOVA to show that there is a significant difference between the tools ($F(4, 45) = 5.18$, $p$=.002).

### 3.2 Questionnaires

Table 2 summarizes the results of the questionnaire-based metrics: SUS, UEQ-S and NASA-TLX. Note that the ranges are different: The SUS-value can range from 0 (very bad usability) to 100 (very good usability), the UEQ-S from 8 (very bad UX) to 40 (very good UX) and the NASA-TLX from 1 (very low workload) to 60 (very high workload).

| | Measure | *WebAnno* | *Word* | *CATMA* | *eMargin* | *Sentimentator* |
|---|---|---|---|---|---|---|
| **SUS** | M | **42.5** | **82.25** | 69 | 56.75 | 76.5 |
| | Sd | 13.94 | 9.46 | 20.21 | 18.37 | 23.4 |
| **UEQ-S** | M | 21.9 | 26.1 | 24.5 | **21.8** | **30.6** |
| | Sd | 8.61 | 2.92 | 4.5 | 7.12 | 3.57 |
| **NASA-TLX** | M | **37.4** | **23.1** | 37.1 | 29.5 | 26.6 |
| | Sd | 8.92 | 5.78 | 6.64 | 4.43 | 5.76 |

Tab. 2: Descriptive statistics - SUS, UEQ-S, NASA-TLX

Considering the SUS metric, *Word* achieves the highest score with 82.25, which can be regarded as "good" usability according to [BKM09]. All other tools achieve values that can be regarded as "OK" except for *WebAnno* which would be regarded as "poor". *Sentimentator* is rated highest considering the UX (M = 30.6). The results for NASA-TLX are rather similar to the SUS, the subjective workload is regarded the lowest for *Word* (M = 23.1) and the highest for *WebAnno* (M = 37.4). One way ANOVAs for every metric show that the effect of the tool on the individual metric is significant: SUS ($F(4, 45) = 8.08$, $p$<.000), UEQ-S ($F(4,45) = 9.64$, $p$=.008), NASA-TLX ($F(4,45) = 9.64$, $p$<.000).

We also integrated three questions about the overall understanding of the annotation task, the perceived easiness and the certainty of the annotations on a 5-point Likert scale. The higher the value the higher the understanding/certainty and the lower the perceived difficulty. Table 3 summarizes the descriptive statistics:

| | Measure | *WebAnno* | *Word* | *CATMA* | *eMargin* | *Sentimentator* |
|---|---|---|---|---|---|---|
| **Understanding** | M | **4.1** | 4.5 | **4.7** | 4.6 | 4.6 |
| | Sd | 0.74 | 0.52 | 0.48 | 0.52 | 0.52 |
| **Easiness** | M | 2.8 | **2.9** | **1.8** | 2.7 | 2.8 |
| | Sd | 1.14 | 0.88 | 1.23 | 0.68 | 0.92 |
| **Certainty** | M | **2.3** | **3.1** | 2.4 | 2.7 | 3.0 |
| | Sd | 1.25 | 1.29 | 0.97 | 1.16 | 0.82 |

Tab. 3: Descriptive statistics - Understanding, easiness, certainty

The average values considering all three metrics are very similar, especially for the overall understanding. For the easiness *Word* is identified as the tool with the highest perceived task easiness (M = 2.9). Furthermore, participants felt the most certain about their annotations with *Word* (M = 3.1) and *Sentimentator* (M=3.0). However, one-way ANOVAS showed that the differences among the tools for those items is not significant.

### 3.3 Annotation metrics

Considering annotation metrics we first present results about the annotation distributions. Table 4 shows the distribution for the first annotation task: the polarity. Note that for several tools annotators missed out annotations, so this is another class next to negative, neutral and positive.

| | *WebAnno* | *Word* | *CATMA* | *eMargin* | *Sentimentator* | *Overall* |
|---|---|---|---|---|---|---|
| **No annotation** | 1 | 0 | 54 | 3 | 0 | 58 |
| **negative** | 135 | 133 | 128 | 122 | 124 | 642 |
| **neutral** | 167 | 169 | 175 | 231 | 230 | 972 |
| **positive** | 197 | 198 | 143 | 144 | 146 | 828 |

Tab. 4: Annotation distributions (Polarity)

In general, most annotations were neutral (39%) and positive (33%). We made two interesting findings analyzing tool-specific differences between the distributions. First, *CATMA* has the highest number of missing annotations (11%). Second, while the distributions between *WebAnno* and *Word* on the one hand and between *eMargin* and *Sentimentator* on the other hand are very similar, the distributions between those groups are rather different. Annotators tend to choose most of the times neutral annotations with *Sentimentator* and *eMargin* (56%) while they choose most of the time positive annotations when using *Word* and *WebAnno* (39%) and neutral only second most (33%). Calculating a Chi-square test of independence, we found a significant effect of the chosen tool on the distributions because of those differences ($\chi^2(12) = 238.26$, $p<.000$).

For the emotion annotation we do not want to go in depth about distributions of every of the eight classes but focus solely on the general number of emotion annotations. Note that annotators could chose 0-8 emotions per annotation. Table 5 shows the average

number of emotion annotations per tool:

| | Measure | *WebAnno* | *Word* | *CATMA* | *eMargin* | *Sentimentator* |
|---|---|---|---|---|---|---|
| **Number of annotations** | M | 1.01 | **1.36** | **.81** | .82 | 1.1 |
| | Sd | 0.15 | 0.95 | 0.57 | 0.71 | 0.88 |

Tab. 5: Descriptive statistics - Number of emotion annotations

On average, annotators tended to annotate the most emotions with *Word* while the participants with *CATMA* and *eMargin* tended to avoid emotion annotations. This effect of the tool on the number of emotion annotations is significant ($F(4, 2495) = 51.32$, $p<.000$).

To analyze the agreement among annotators we use two metrics: Krippendorff's α is an established metric to measure inter-annotator agreement recommended for annotations with more than two annotators and proven to be stable [AVL14]. We also report the average agreement among all annotators in percent by calculating the agreement for all annotator pairs and dividing it by the number of annotator pairs (table 6). We only report the results about the polarity annotations (negative, neutral, positive).

| | *WebAnno* | *Word* | *CATMA* | *eMargin* | *Sentimentator* |
|---|---|---|---|---|---|
| **Krippendorff's α** | 0.23 | **0.35** | **0.17** | 0.27 | 0.28 |
| **Average agreement in percent** | 48.8% | **57.2%** | **40.4%** | 53.2% | 54% |

Tab. 6: Agreement metrics (Polarity)

Overall, the agreement among annotators is rather low. According to [LK77], agreement levels for the majority of the tools are regarded as fair agreement (0.2-0.4). Participants using *CATMA* show a poor agreement (<0.2) mostly because users oftentimes forgot to annotate a speech at all when using this tool (table 4). The agreement levels are close to each other; however, participants using *Word* have the highest agreement. The low to fair agreement levels among annotators are in line with other research in the context of sentiment annotation of literary texts [SBD18b, [KK18b].

# 4 Discussion

## 4.1 Tool Evaluation

*Word* and *Sentimentator* are consistently rated the highest, they are the easiest and fastest to use and lead to higher perceived annotation certainty. We assume that there are multiple reasons for this result. Both tools are similar concerning the annotation in that it is done rather easy by clicking one button or marking one field without interaction with the text. For the other tools, it is necessary to first mark text and then select the annotation. The annotators reported in the interviews that this is a very tedious task since some tools take several seconds time after selecting the text before one can assign the annotation.

Furthermore, it is far easier to forget annotations and it also leads annotators to annotate fewer emotions in general. The positive results for *Sentimentator* are not surprising since this tool has been designed specifically for sentiment annotation. We assumed that it is a disadvantage that annotators can only see one speech for the annotation while with other tools multiple speeches are visible thus allowing to integrate the context. However, annotators did not report this as a major issue. At first sight, the good rating for *Word* may appear as counterintuitive since *Word* is not specifically designed for annotation. However, the fact that the overall UI of *Word* is well-known might be a significant advantage, especially when dealing with a sample of non-experts in the field of annotation. In addition, this software has been developed and optimized with respect to usability and UX for decades. Therefore, for some use cases in DH, one might recommend to rely on a standard tool with a good usability / UX baseline like this. Comparing *Word* with *Sentimentator*, we only identified that *Word* seems to produce a lower workload while the *Sentimentator* has a higher rated UX. Note that we do not make any general assumptions about the usability and UX of the tools in any other context since usability results are very task depending. Changing the type of annotation, e.g. to linguistic annotations, other tools might very well be more valuable. For example, *Sentimentator* and *Word* are rarely adaptable to other types of annotations. Also, it is necessary to mention that we did not regard any functionality for administrators. Apparent disadvantages for a tool like *Word* are missing automatic exports and user management functions, which are not part of this study.

### 4.2 Analysis and Comparison of Usability and UX metrics

Comparing the usage of our questionnaires, we did not identify noticeable differences between the perceived usability (SUS), the user experience (UEQ-S) and workload (NASA-TLX). Overall, we would regard one questionnaire as sufficient and only recommend the usage of multiple ones if a more detailed analysis is necessary. Regarding user feedback, we realized that while participants were focused on basic usability issues and problems with the interpretation of the annotation process itself, UX-specific aspects were rarely mentioned. Since annotation in DH is oftentimes embedded in a (complex) working context, usability seems to be more important, thus we recommend surveys like the SUS and NASA-TLX. However, regarding methods like crowdsourcing, UX and joy of use might become more critical. The short questionnaire derived by Schmidt et al. (2018) did not lead to more insights except for the measuring of the subjective certainty. Nevertheless, this aspect is analyzed via agreement statistics anyway. Gathering qualitative data via interviews was helpful to find explanations for apparent problems and to gather recommendations on possible tool and process improvements. However, we realized that the feedback became rather repetitive after five participants per tool. If the task is to only decide upon multiple tools, one is capable doing so solely by gathering survey data. If the goal is indeed to decide between multiple tools and furthermore resources are rare, we recommend designing usability tests with few questionnaires. However, if the goal is to design a new tool we recommend acquiring fewer participants but integrating more qualitative methods like interviews.

### 4.3 The Relationship of the Tool and Annotation Behavior

We did find an effect of tool selection on annotation behavior and agreement metrics. However, this effect seems to be rather small. First, participants using the more usable tools *Word* and *Sentimentator* report higher levels of certainty. Second, as already outlined in section 4.2, the higher usability of the annotation process for *Word* and *Sentimentator* lead to more emotion annotations while for other tools some speeches were entirely missed out. Third, if we rank the tools according to their average ratings in the questionnaires as well as with the agreement statistics, we get a very similar order (see table 7) thus proving the influence of the tool on annotation behavior and inter-rater agreement.

| Ordered by SUS | Ordered by UEQ-S | Ordered by NASA-TLX | Ordered by Krippendorff's α |
|---|---|---|---|
| Word | Sentimentator | Word | Word |
| Sentimentator | Word | Sentimentator | Sentimentator |
| CATMA | CATMA | eMargin | eMargin |
| eMargin | WebAnno | CATMA | WebAnno |
| WebAnno | eMargin | WebAnno | CATMA |

Tab. 7: Tools ordered by different metrics

At first glance, these results are counterintuitive since the annotation task and scheme stay the same. Rating the sentiment of a speech, in theory, has nothing to do with the chosen tool. In a more general perspective, though, tools shape the way we think and act to a certain degree. We hypothesize that the less user friendly the annotation process is designed the lower the motivation and concentration which leads to rather arbitrary annotation. However, if the tool is designed in a way the annotator can solely focus on the annotation task, the annotations among annotators become more similar and stable, thus annotators are more capable to choose the "objectively" correct annotation. These results prove that especially when dealing with vague annotations that are open to interpretation, the usability and UX of a tool is very important and researchers should take this into account when selecting the annotation tools. In the future, we want to conduct further large-scale studies to gather more insights about the effect of the tool on the annotation behavior.

## Bibliography

[AP17] Arévalo, Enrique Manjavacas; Petré, Peter: Enabling Annotation of Historical Corpora in an Asynchronous Collaborative Environment. In: Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage. ACM, S. 9–14, 2017.

[ARS05] Alm, Cecilia Ovesdotter; Roth, Dan; Sproat, Richard: Emotions from text: machine learning for text-based emotion prediction. In: Proceedings of the conference on human-language technology and empirical methods in natural language processing. Association for Computational Linguistics, S. 579–586, 2005.

[AT13] Albert, William; Tullis, Thomas: Measuring the user experience: collecting, analyzing, and presenting usability metrics. Newnes, 2013.

[AVL14] Antoine, Jean-Yves; Villaneau, Jeanne; Lefeuvre, Anaïs: Weighted Krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation. In: EACL 2014. S. 10–p, 2014.

[BKM09] Bangor, Aaron; Kortum, Philip; Miller, James: Determining what individual SUS scores mean: Adding an adjective rating scale. Journal of usability studies, 4(3):114–123, 2009.

[Bö15] Bögel, Thomas; Gertz, Michael; Gius, Evelyn; Jacke, Janina; Meister, Jan Christoph; Petris, Marco; Strötgen, Jannik; Cordell, Ryan; Baillot, Anne: Collaborative Text Annotation Meets Machine Learning: heureCLÉA, a Digital Heuristic of Narrative. DHCommons journal, 1, 2015.

[Br96] Brooke, John et al.: SUS-A quick and dirty usability scale. Usability evaluation in industry, 189(194):4–7, 1996.

[Bu12] Burghardt, Manuel: Usability recommendations for annotation tools. In: Proceedings of the Sixth Linguistic Annotation Workshop. Association for Computational Linguistics, S. 104–112, 2012.

[DBW17] Döhling, Lars; Burghardt, Manuel; Wolff, Christian: PaLaFra–Entwicklung einer Annotationsumgebung für ein diachrones Korpus spätlateinischer und altfranzösischer Texte. In: Book of Abstracts, DHd 2017. 2017.

[DGS04] Dipper, Stefanie; Götze, Michael; Stede, Manfred: Simple annotation tools for complex annotation tasks: an evaluation. In: Proceedings of the LREC Workshop on XML-based richly annotated corpora. S. 54–62, 2004.

[Ga04] Garg, Saurabh; Martinovski, Bilyana; Robinson, Susan; Stephan, Jens; Tetreault, Joel; Traum, David R: Evaluation of transcription and annotation tools for a multi-modal, multi-party dialogue corpus. Bericht, UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY CA INST FOR CREATIVE . . . , 2004.

[Ha06] Hart, Sandra G: NASA-task load index (NASA-TLX); 20 years later. In: Proceedings of the human factors and ergonomics society annual meeting. Jgg. 50. Sage publications Sage CA: Los Angeles, CA, S. 904–908, 2006.

[HP15] Hoff, Karoline; Preminger, Michael: Usability testing of an annotation tool in a cultural heritage context. In: Research Conference on Metadata and Semantics Research. Springer, S. 237–248, 2015.

[HS88] Hart, Sandra G; Staveland, Lowell E: Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: Advances in psychology, Jgg. 52, S. 139–183. Elsevier, 1988.

[KG12] Kehoe, Andrew; Gee, Matt: eMargin: A collaborative text annotation tool. In: Proceedings 33rd Conference on International Computer Archive of Modern and Medieval English (ICAME 33). 2012.

[KK11] Kakkonen, Tuomo; Kakkonen, Gordana Galic: SentiProfiler: Creating comparable visual profiles of sentimental content in texts. In: Proceedings of the Workshop on

Language Technologies for Digital Humanities and Cultural Heritage. S. 62–69, 2011.

[KK18a] Kim, Evgeny; Klinger, Roman: A Survey on Sentiment and Emotion Analysis for Computational Literary Studies. arXiv preprint arXiv:1808.03137, 2018.

[KK18b] Kim, Evgeny; Klinger, Roman: Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In: Proceedings of the 27th International Conference on Computational Linguistics. S. 1345–1359, 2018.

[Li16] Liu, Bing: Sentiment analysis: Mining opinions, sentiments, and emotions. Cambridge University Press, 2016.

[LK77] Landis, J Richard; Koch, Gary G: The measurement of observer agreement for categorical data. biometrics, S. 159–174, 1977.

[Mo11] Mohammad, Saif: From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. Association for Computational Linguistics, S. 105–114, 2011.

[NB13] Nalisnick, Eric T; Baird, Henry S: Character-to-character sentiment analysis in Shakespeare's plays. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Jgg. 2, S. 479–483, 2013.

[ÖK18] Öhman, Emily; Kajava, Kaisla: Sentimentator: Gamifying Fine-Grained Sentiment Annotation. In: DHN. S. 98–110, 2018.

[Pe14] Pedersen, Bolette S; Nimb, Sanni; Olsen, Sussi; Søgaard, Anders; Sørensen, Nicolai: Semantic Annotation of the Danish CLARIN Reference Corpus. In: Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation. Citeseer, S. 25–29, 2014.

[SB18] Schmidt, Thomas; Burghardt, Manuel: An Evaluation of Lexicon-based Sentiment Analysis Techniques for the Plays of Gotthold Ephraim Lessing. In: Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. Association for Computational Linguistics, S. 139–149, 2018.

[SBD18a] Schmidt, Thomas; Burghardt, Manuel; Dennerlein, Katrin: „Kann man denn auch nicht lachend sehr ernsthaft sein?" – Zum Einsatz von Sentiment Analyse-Verfahren für die quantitative Untersuchung von Lessings Dramen. In (Vogeler, Georg, Hrsg.): Book of Abstracts, DHd 2018. Cologne, Germany, S. 244–249, 2018.

[SBD18b] Schmidt, Thomas; Burghardt, Manuel; Dennerlein, Katrin: Sentiment Annotation of Historic German Plays: An Empirical Study on Annotation Behavior. In (Kübler, Sandra; Zinsmeister, Heike, Hrsg.): Proceedings of the Workshop for Annotation in Digital Humantities (annDH). Sofia, Bulgaria, S. 47–52, August 2018.

[SBW19] Schmidt, Thomas; Burghardt, Manuel; Wolff, Christian: Toward Multimodal Sentiment Analysis of Historic Plays: A Case Study with Text and Audio for Lessing's Emilia Galotti. In: 4th Conference of the Association of Digital Humanities in the Nordic Countries (DHN 2019). 2019.

[SHT17] Schrepp, Martin; Hinderks, Andreas; Thomaschewski, Jörg: Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S). IJIMAI, 4(6):103–

108, 2017.

[SP05] Sazedj, Peyman; Pinto, H Sofia: Time to evaluate: Targeting annotation tools. Proc. Of Knowledge Markup and Semantic Annotation at ISWC, 2005, 2005.

[Yi13] Yimam, Seid Muhie; Gurevych, Iryna; de Castilho, Richard Eckart; Biemann, Chris: WebAnno: A flexible, web-based and visually supported system for distributed annotations. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations. S. 1–6, 2013.