Abstract
In this paper, we present first work-in-progress annotation results of a project investigating computational methods of emotion analysis for historical German plays around 1800. We report on the development of an annotation scheme focusing on the annotation of emotions that are important from a literary studies perspective for this time span as well as on the annotation process we have developed. We annotate emotions expressed or attributed by characters of the plays in the written texts. The scheme consists of 13 hierarchically structured emotion concepts as well as the source (who experiences or attributes the emotion) and target (who or what is the emotion directed towards). We have conducted the annotation of five example plays of our corpus with two annotators per play and report on annotation distributions and agreement statistics. We were able to collect over 6,500 emotion annotations and identified a fair agreement for most concepts around a κ-value of 0.4. We discuss how we plan to improve annotator consistency and continue our work. The results also have implications for similar projects in the context of Digital Humanities.

2012 ACM Subject Classification Applied computing → Arts and humanities; Computing methodologies → Machine learning

Keywords and phrases Emotion, Annotation, Digital Humanities, Computational Literary Studies, German Drama, Sentiment Analysis, Emotion Analysis, Corpus

Digital Object Identifier 10.4230/OASIcs.LDK.2021.9

Funding This research is part of the project “Emotions in Drama” (Emotionen im Drama), is funded by the German Research Foundation (DFG) and part of the priority programme SPP 2207 Computational Literary Studies (CLS).

Acknowledgements We want to thank the following student annotators for their contributions to this project: Viola Hipler, Julia Jäger, Emma Ruß, and Leon Sautter.

1 Introduction

Emotions in dramatic texts are central for the dramaturgy, the characterization of characters, the intended effect on the reader as well as for the propagation of anthropological ideas. Emotions are a frequent and important subject in German literary studies of the 17th and 18th century. For example, literary scholars investigated the intended emotional effect [18, 39] or single emotions in plays of that time [2, 37]. We want to expand this hermeneutical research focused mostly on canonical texts. That is why we are applying computational emotion analysis on larger data sets of historical German plays around 1800. We are aiming at a more holistic view of emotion usage, progression and distribution in the plays of that time.

Computational emotion prediction in Natural Language Processing (NLP) describes the task of predicting the expressed emotion, predominantly in written text. Sentiment analysis, its neighbouring field, is focused on the prediction of the valence/polarity of text (if a text unit is rather positive or negative) while emotion prediction deals with more complex...
emotion categories like anger, joy, or surprise [16]. Both methods have gained a lot of interest in Digital Humanities (DH) and Computational Literary Studies (CLS) (cf. [9]) and are applied to analyze emotions and sentiment in historical plays [12, 17, 23, 25, 26, 27, 29, 40], novels [6, 12, 21], fairy tales [1, 12], political texts [38], or online forums [14, 35]. DH projects also explore more modern literary genres like fan fictions [8, 7], original creative works on the web [19], subtitles of movies [5, 42] or song lyrics [24]. From a methodological point of view, many of these projects employ lexicon and rule-based methods to perform the sentiment and emotion analysis [1, 12, 17, 19, 23, 24, 25, 38, 40] leading to the development of lexicon-based sentiment analysis tools specifically designed for the DH-community [33]. However, these methods are outperformed by modern machine learning approaches [16]. The reason for the application of lexicon-based methods is the lack of well-annotated corpora of the particular domains that are necessary to train machine learning algorithms [31]. Currently, however, many projects work towards closing this gap and create first corpora of emotion annotated literary texts to explore deep learning based emotion analysis [7, 8]. Annotation of emotions and sentiments can be a challenging task [13, 41]. The task has been shown as even more problematic for historical and poetic texts [1, 28, 30, 32, 36, 38]. While the application of large-scale crowd-sourcing is common for many text types in NLP (cf. [13]), researchers rather refer to expert-based annotation for historical and poetic texts because of the challenges in language and interpretation [1, 28, 30, 36, 32, 38]. Furthermore, due to the high level of subjectivity and complexity of these texts, agreement statistics among expert or common annotators are oftentimes rather low [1, 28, 30, 36, 32, 38] which poses challenges to creating a valid gold standard. Recent research explores the development of tools with gamification elements to improve upon these problems [34, 42].

We present first results of a collaborative project between computer scientists and literary scholars exploring computational emotion analysis on German plays around 1800. Our main corpus currently consists of over 200 plays of that time and we performed our first annotation study on five representative plays of this corpus. We report on annotation results and how we address the challenges of emotion annotation in this field. We developed annotation schemes and processes that are more directed towards the literary scholar’s perspective and goals than previous annotation schemes in NLP. Our experience with the annotation and overall results have implications for similar projects designing annotation schemes and performing emotion annotation in the context of CLS.

2 Annotation

In the following, we present the annotation scheme and process we developed. Please note that both the process as well as the scheme have been developed in an iterative process (cf. [22]) of pilot annotations on various scenes and plays of our corpus.

2.1 Annotation Scheme

We define emotion as a generic term for a character’s state of mind of distinguishable quality at a given time that is expressed, among other channels, through written language. We annotate emotions experienced by the characters and attributed to them as they are represented as text. Please note that we are interested in the “real” intention and meaning of the expressions of the characters in the context of the entire play. For example, in the case of an ironic expression we annotate the intention of the character in this specific context and not what the text would mean independent of the content and context of the play.
We started the annotation scheme with a list of categorical emotions collected from various established systems of psychology (e.g. [20]) which is rather common in emotion prediction in NLP (cf. [13, 16]). However, we realized that these emotion concepts are missing core emotion and affect ideas important to capture the concept of “emotion” in plays of that time. These are important for the research of literary scholars, however. Therefore, we deviate from established psychological concepts of emotion and integrate concepts such as love and friendship which are not regarded as emotions in many psychological definitions (cf. [15]) but important for this type of literature. We continued with some pilot annotations with a very large scheme containing various emotional concepts important throughout literary history. However, the sheer size and complexity hindered the annotation process. The set was filtered on the most important concepts for literary studies for this time and genre. The final annotation scheme for emotions consists of the 13 concepts mentioned below. In brackets we include the German original terms since we do annotate in German. We translated them to English to the best of our knowledge, but semantic details might get skewed.

- Emotions of affection (Emotionen der Zuneigung)
  - Desire (Lust) (−)
  - Love (Liebe) (+)
  - Friendship (Freundschaft) (+)
  - Adoration (Verehrung) (+)
- Emotions of joy (Emotionen der Freude)
  - Joy (Freude) (+)
  - Schadenfreude (+)
- Emotions of fear (Emotionen der Furcht)
  - Fear (Angst) (−)
  - Despair (Verzweiflung) (−)
- Emotions of suffering (Emotionen des Leids)
  - Suffering (Leid) (−)
  - Compassion (Mitleid) (−)
  - Anger (Ärger) (−)
- Other
  - Hate (Abscheu) (−)
  - Emotional movement (Emotionale Bewegheit)

We defined the set in a hierarchical order to deal with the imbalance problem or too few annotations in the later computational emotion prediction by mapping the emotions to the four main classes and two special types (hate, emotional movement). Emotional movement is used to annotate unspesific emotional arousal (that cannot be described with the other concepts) as well as astonishment. In the highest hierarchical order, emotions are represented by the two classes positive and negative (valence). We established a default valence for each emotion concept (marked as + and − in the above list) but annotators can also choose to deviate from this or mark an emotion as mixed via an attribute attached with each emotion annotation. Schadenfreude, although ambivalent, is assigned as positive per default since most of the time the emotion is perceived as positive by the experiencer in our texts. As with all emotions, annotators can however deviate from this assignment.

While this set of emotions is still not sufficient to fully capture emotional representations in the literature of that time, it is a compromise between the larger interest of literary scholars and the pragmatic limitations for the computational perspective as well as for the annotation process. Annotators annotate speeches (single utterances of a character separated
Towards a Corpus of Historical German Plays with Emotion Annotations

Figure 1 Illustration of an example annotation from Lessing’s Minna von Barnhelm (Act 1, Scene 3): First line is the German original, second line an English translation. The entire sentence is annotated with anger. “Ich” (I) is the source, “ihm” (him) the target of the emotion. “Just” and “Der Wirt” are the names of the specific characters.

Following ideas of aspect-based sentiment analysis [7] we also annotate the source (the character experiencing or attributing an annotated emotion) and the target (the instance an emotion is directed towards). Similar to the emotion set, the set for source and target was adjusted and developed throughout multiple pilot annotation iterations and deemed important for the literary studies perspective since the emotional interaction of the characters are the main aspects of these plays. Source and target consist of the following sub types and possible attributes:

- **Source**
  - Experiencer: characters of the play, the author, impersonal, unknown
  - Attributing instance: characters of the play, the author, impersonal, unknown

- **Target**
  - Character: characters of the play, impersonal, unknown
  - Non-Character: animal, state, event or object

**Impersonal** is a mark for addressing the general public while **unknown** points to characters that are not in the original character list of the play, which is the standard selection of which the annotators can select the characters of the play by their name. The annotation window is as variable as with the emotion annotation. Annotators mark each explicit mention of source and target in the annotated text. In certain cases, it is however possible that an emotion annotation consists of neither source nor target. Figure 1 illustrates the annotation of one example speech of our corpus consisting of an emotion annotation and an explicit annotation of source and target of this annotation.

2.2 Corpus

To start the annotation we decided to annotate five plays of different genres and authors of our main corpus. Plays are annotated in their entirety since we are interested in context and content dependent annotations that need thorough interpretation of the entire plot. While this poses challenges to later generalization processes on the computational side, this is in line with the focus of this project on literary criticism. One aspect to deal with this problem is to annotate plays that are representative concerning content and language of
clusters of the 200 plays corpus. Most plays are taken from the GerDracor corpus [3], one play was taken from a free repository.1 The following five plays have been annotated: *Minna von Barnhelm* (1767) by Lessing (comedy), *Kasperl der Mandolettikrämer* (1789) by Eberl (comedy), *Kabale und Liebe* (1784) by Schiller (tragedy), *Menschenhass und Reue* (1790) by Kotzebue (comedy), *Faust. Eine Tragödie* (1807) by Goethe (tragedy).

### 2.3 Annotation Process

Since the annotation of the plays is dependent on deeper knowledge of the language and the content of the plays (as we perform context-aware annotation), crowd-sourcing annotations was not a viable option. In similar projects, annotations are performed by experts and semi-experts with a specific training [1, 28, 30, 31, 38]. In our setting, each play was annotated independently from each other by two students of German literary studies who are compensated monetarily for the annotations and who are employed in the research project. For the annotation of this corpus, we employed three annotators; each play was annotated in different combinations of annotator pairs. The students were introduced to the annotation guidelines by a literary scholar during multiple annotation training sessions and they were offered support during the annotation process. The students participated in the pilot annotation studies to determine the annotation scheme, as well. They had access to an annotation guidelines document consisting of a description of the scheme and multiple examples. The annotation was performed with the tool CATMA [4] for which we created the annotation scheme as described. The annotators were assigned to a play and had a specific deadline to finish it. Depending on the length of the play, each annotator had one to two weeks time to finish the annotation. On average the entire annotation process was performed throughout multiple days of the set time frame and took around 8–12 hours concerning the absolute duration.

### 3 Annotation Results

We collected over 6,500 emotion annotations for the five plays. First, we look at annotation distributions among the main and sub categories as well as statistics of the annotation lengths via token statistics (see Table 1).

The most frequent annotated emotions are *suffering* (15%) *joy* (13%), *anger* (13%) and *love* (12%). Some emotions are annotated rather rarely in our study e.g. *desire* (1%), *friendship* (2%) and *Schadenfreude* (3%). The main categories themselves are annotated more equally, however with a dominance of more negative emotion categories like the emotions of *suffering* (33%). *Emotional movement* has been proven as an important annotation category (12%). Looking at the size of annotations, all categories have rather similar averages (around 25 tokens) with a large variance ranging from one word annotations to larger paragraphs consisting of over 300 tokens. The dominance of negative emotions is supported by the distribution of the highest hierarchical order valence: 54% of all emotion annotations were either per default *negative* or marked via an attribute as *negative* compared to 34% of *positive* assignments. The remaining annotations in the highest hierarchical order were *emotional movement* annotations (11%). The possibility to select *mixed* as attribute for annotations was rarely used. This attribute has been shown to be rather redundant in our scheme since annotators can assign multiple emotions with differing valence to one text unit.

---

1 [http://lithes.uni-graz.at/maezene/eberl_mandolettikraemer.html](http://lithes.uni-graz.at/maezene/eberl_mandolettikraemer.html)
Table 1 Distribution of emotions and corresponding main categories. First, the sub emotions are listed followed by the summed results of the main categories in bold. Percentages are rounded.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>absolute</th>
<th>%</th>
<th>avg. tokens</th>
<th>min tokens</th>
<th>max tokens</th>
<th>std. tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desire</td>
<td>50</td>
<td>1</td>
<td>23.22</td>
<td>4</td>
<td>83</td>
<td>16.49</td>
</tr>
<tr>
<td>Love</td>
<td>783</td>
<td>12</td>
<td>26.16</td>
<td>1</td>
<td>326</td>
<td>33.67</td>
</tr>
<tr>
<td>Friendship</td>
<td>127</td>
<td>2</td>
<td>22</td>
<td>1</td>
<td>120</td>
<td>18.66</td>
</tr>
<tr>
<td>Adoration</td>
<td>306</td>
<td>5</td>
<td>19.63</td>
<td>1</td>
<td>96</td>
<td>16.36</td>
</tr>
<tr>
<td><strong>Emotions of affection</strong></td>
<td>1,266</td>
<td>19</td>
<td>24.05</td>
<td>1</td>
<td>326</td>
<td>28.61</td>
</tr>
<tr>
<td>Joy</td>
<td>850</td>
<td>13</td>
<td>22.78</td>
<td>1</td>
<td>223</td>
<td>24.3</td>
</tr>
<tr>
<td>Schadenfreude</td>
<td>201</td>
<td>3</td>
<td>25.02</td>
<td>1</td>
<td>121</td>
<td>21.89</td>
</tr>
<tr>
<td><strong>Emotions of joy</strong></td>
<td>1,051</td>
<td>16</td>
<td>23.21</td>
<td>1</td>
<td>223</td>
<td>23.86</td>
</tr>
<tr>
<td>Fear</td>
<td>424</td>
<td>6</td>
<td>16.87</td>
<td>1</td>
<td>173</td>
<td>17.45</td>
</tr>
<tr>
<td>Despair</td>
<td>282</td>
<td>4</td>
<td>30.78</td>
<td>1</td>
<td>206</td>
<td>30.15</td>
</tr>
<tr>
<td><strong>Emotions of fear</strong></td>
<td>706</td>
<td>11</td>
<td>22.42</td>
<td>1</td>
<td>206</td>
<td>24.32</td>
</tr>
<tr>
<td>Suffering</td>
<td>998</td>
<td>15</td>
<td>26.12</td>
<td>1</td>
<td>302</td>
<td>28.91</td>
</tr>
<tr>
<td>Compassion</td>
<td>318</td>
<td>5</td>
<td>21.61</td>
<td>1</td>
<td>156</td>
<td>21.87</td>
</tr>
<tr>
<td>Anger</td>
<td>880</td>
<td>13</td>
<td>22.14</td>
<td>1</td>
<td>261</td>
<td>24.35</td>
</tr>
<tr>
<td><strong>Emotions of suffering</strong></td>
<td>2,196</td>
<td>33</td>
<td>23.87</td>
<td>1</td>
<td>302</td>
<td>26.27</td>
</tr>
<tr>
<td>Hat</td>
<td>614</td>
<td>9</td>
<td>25.05</td>
<td>1</td>
<td>167</td>
<td>26.19</td>
</tr>
<tr>
<td><strong>Emotional movement</strong></td>
<td>763</td>
<td>12</td>
<td>24.4</td>
<td>1</td>
<td>313</td>
<td>32.74</td>
</tr>
</tbody>
</table>

Table 2 Agreement statistics per play for the overall valence, the main emotion class and the sub emotions respectively for the text unit of speeches. $\kappa$ refers to Cohen’s $\kappa$ while $\%$ is the proportion of agreed upon speeches among all speeches.

<table>
<thead>
<tr>
<th>Drama</th>
<th>Valence ($\kappa$)</th>
<th>Valence (%)</th>
<th>Class ($\kappa$)</th>
<th>Class (%)</th>
<th>Emotion ($\kappa$)</th>
<th>Emotion (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faust</td>
<td>0.44</td>
<td>67.853</td>
<td>0.345</td>
<td>59.399</td>
<td>0.342</td>
<td>58.064</td>
</tr>
<tr>
<td>Kabale und Liebe</td>
<td>0.382</td>
<td>58.908</td>
<td>0.325</td>
<td>50.313</td>
<td>0.312</td>
<td>47.992</td>
</tr>
<tr>
<td>Menschenhass und Reue</td>
<td>0.402</td>
<td>75.28</td>
<td>0.347</td>
<td>72.331</td>
<td>0.347</td>
<td>71.91</td>
</tr>
<tr>
<td>Minna von Barnhelm</td>
<td>0.406</td>
<td>74.619</td>
<td>0.377</td>
<td>72.752</td>
<td>0.356</td>
<td>71.23</td>
</tr>
<tr>
<td>Kasperl’ der Mandolettkrämer</td>
<td>0.42</td>
<td>70.83</td>
<td>0.344</td>
<td>65.34</td>
<td>0.312</td>
<td>62.72</td>
</tr>
<tr>
<td>Overall</td>
<td>0.41</td>
<td>69.498</td>
<td>0.3476</td>
<td>64.027</td>
<td>0.333</td>
<td>62.383</td>
</tr>
</tbody>
</table>
Table 3 Source and target distributions. The sub categories are listed followed by the summed results of the main categories in bold. The percentages of the sub groups refer to the main class.

<table>
<thead>
<tr>
<th>Annotation Type</th>
<th>absolute</th>
<th>%</th>
<th>avg. tokens</th>
<th>min tokens</th>
<th>max tokens</th>
<th>std. tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiencer</td>
<td>6,573</td>
<td>97</td>
<td>1.06</td>
<td>1</td>
<td>7</td>
<td>0.33</td>
</tr>
<tr>
<td>Attributing Instance</td>
<td>187</td>
<td>3</td>
<td>1.05</td>
<td>1</td>
<td>3</td>
<td>0.27</td>
</tr>
<tr>
<td>Source</td>
<td>6,760</td>
<td>50</td>
<td>1.06</td>
<td>1</td>
<td>7</td>
<td>0.33</td>
</tr>
<tr>
<td>Character</td>
<td>5,336</td>
<td>79</td>
<td>1.28</td>
<td>1</td>
<td>14</td>
<td>0.82</td>
</tr>
<tr>
<td>Non-Character</td>
<td>1,390</td>
<td>21</td>
<td>3.97</td>
<td>1</td>
<td>26</td>
<td>3.68</td>
</tr>
<tr>
<td>Target</td>
<td>6,726</td>
<td>50</td>
<td>1.84</td>
<td>1</td>
<td>26</td>
<td>2.13</td>
</tr>
</tbody>
</table>

Since the annotations are performed on variable text lengths, we decided on the following heuristic to calculate agreement among annotators: We focus on the speech and stage directions as central structural units of plays. They can consist of one word to multiple sentences. For every annotator we assign the specific emotion that is annotated the most (in total token count) for one speech. Thus, if multiple emotions are annotated, we assign the emotion that is annotated the most. We decided for this heuristic in order to be able to apply the traditional agreement metric Cohen’s $\kappa$ and get a first overview of agreement among annotators. We explore possibilities for more fitting fuzzy agreement metrics in future work. If no emotion was annotated the unit is marked as none. None is regarded as additional annotation class in this concept. Table 2 illustrates the agreements. The $\kappa$-value according to Cohen’s $\kappa$ is shown as well as the percentage wise agreement. We identified mostly moderate agreement for the valence according to [11] (0.41-0.6) and fair agreement for the main emotion category and the sub emotions (0.21-0.4). Due to the higher number of classes the agreement gets lower for the sub emotions.

We also gathered over 12,000 source and target annotations (see Table 3). Both classes are annotated to an equal extent. For sources, characters are mostly marked as experiencer of emotions (97%) and rarely as the ones attributing emotions to other characters (3%). Targets of emotions are mostly characters (79%). For the sub groups of these classes, the following findings could be made. Sources, being it experiencer or attributing instances, are for the most part one character (94%) or multiple characters (2%). The attributes for unknown and impersonal sources are rarely used (2%). If a character is chosen as a target, the distribution is similar with one to multiple characters being the most frequent annotation (89%) compared to unknown (7%) and impersonal (4%). If the target is a non-character, the attribute assigned most frequently is event (61%) followed by state (19%), objects (16%) and animals (4%). Regarding the annotation lengths, source and target annotations are mostly one word annotations like pronouns or character names which points towards token based prediction mechanisms in later computational approaches to predict source and target.

4 Discussion

The annotated corpus will be made publicly available and is currently in the process of preparation.

To validate findings of the annotation analysis, we discussed our results with the annotators after the annotation. The extension of our scheme beyond established categories of psychology has been well received by annotators and we recommend this for similar projects. Concepts such as love, suffering and emotional movement are important parts of literature of that time and genre and have been annotated in large numbers. However, other concepts such as desire or adoration were rarely annotated. We are discussing the need for these concepts
since any complexity reduction of the scheme is beneficial for annotation speed, consistency and the later prediction. Please note that the annotation of emotions is highly influenced by the plays chosen to be annotated. Concepts such as desire and adoration are more important for earlier periods which we will investigate in the future and which will likely lead to the collection of more annotations. Looking at the main categories, the distribution becomes more equal. Negative categories are more frequent, although the majority of our chosen corpus consists of comedies. This is in line with previous annotation results in similar contexts [1, 28, 30] showing that negativity is an integral part of the narrative of most plays. The genre assignment comedy just points towards a positive ending, the play itself still consists of conflicts and disputes up until the end. Annotators rarely used the annotation of the attribute mixed for emotion. This attribute is redundant in our scheme and will be discarded. Considering source and target, we identified that annotators mostly annotate them as characters and not as non-characters which is quite intuitive in light of the content of the plays which are driven by emotional interactions of characters. We will reflect upon the question if differentiated sub classes for non-characters make sense if this main class is annotated rather rarely. The variable annotation lengths have also been perceived rather positively by the annotators and we also recommend the application of this idea for similar projects in a literary studies context. Emotions were annotated in variable sizes concerning number of words and sentences. This resembles the reality of the emotion expressions in these plays and is also in line with the general annotation behavior of literary scholars. Forcing annotations for a concrete window size would be challenging for decision processes during the annotation and would prolong and complicate the process. We plan to apply heuristics to map annotations on structural units and perform speech, sentence, n-gram and token based multi-label emotion prediction in our computational approaches.

The current agreement results indicate fair to moderate agreement. This is mostly in line with results of projects with similar text types [1, 28, 30, 32, 36, 38] since the material is more subjective and challenging to interpret. Our approach to perform context-sensitive annotation reinforces this aspect. In future work, we plan to explore sentence and token based agreements but also agreements of source and target annotations to get a better overview of the annotation problems. We also see potential in fuzzy agreement scores to represent the agreement in our variable and complex setting in a more fitting way [10] since our heuristic certainly leads to further disagreement in certain instances. Furthermore, we argue that we will reach higher agreements the more experience the annotators gain. To support this process and to find a way to deal with the disagreements among the annotators, we decided to add a subsequent post-annotation phase after the first two independent annotations by the students. This post-annotation phase is performed under the guidance of a literary scholar expert annotator who discusses the annotation with the students and creates a consensus annotation during these sessions. Although this might increase the annotation duration, it will improve the understanding of all annotators and might lead to more consistent annotations. Kajava et al. [5] argue that $\kappa$-values of 0.6 are acceptable for multi-label emotion annotations to validate the consistency of a scheme. The consensus annotation will also be the material we use to train and evaluate computational emotion analysis based on machine learning. We will adjust the annotation scheme and continue the annotations in the described way.

References


Eric T. Nalisnick and Henry S. Baird. Character-to-Character Sentiment Analysis in Shakespeare’s Plays. In Proceedings of the 51st Annual Meeting of the Association for Computa-
Towards a Corpus of Historical German Plays with Emotion Annotations


Thomas Schmidt, Manuel Burghardt, Katrin Dennerlein, and Christian Wolff. Sentiment annotation for lessing’s plays: Towards a language resource for sentiment analysis on german literary texts. In Thierry Declerck and John P. McCrae, editors, 2nd Conference on Language,


