Early Modern Dutch Comedies and Farces in the Spotlight: Introducing EmDComF and its Emotion Framework

Florian Debaene[x][o], Kornee van der Haven[o], Veronique Hoste[x]
[x]LT^3, Language Technology & Translation Team, [o]Department of Literary Studies
[x]Groot-Brittanniëlaan 45, [o]Blandijnberg 2, 9000 Gent, Belgium
florian.debaene@ugent.be, cornelis.vanderhaven@ugent.be, veronique.hoste@ugent.be

Abstract

As computational drama studies are developing rapidly, the Dutch dramatic tradition is in need of centralisation still before it can benefit from state-of-the-art methodologies. This paper presents and evaluates EmDComF, a historical corpus of 466 both manually curated and automatically digitised early modern Dutch comedies and farces authored between 1650 and 1725, and describes the refinement of a historically motivated annotation framework exploring sentiment and emotions in these two dramatic subgenres. Originating from Lodewijk Meyer's philosophical writings on passions in the dramatic genre (±1670), published in *Naauwkeurig onderwys in de tooneel-poëzy* (Thorough instruction in the Poetics of Drama) by the literary society Nil Volentibus Arduum in 1765, a historical and genre-specific emotion framework is tested and operationalised for annotating emotions in the domain of early modern Dutch comedies and farces. Based on a frequency and cluster analysis of 782 annotated sentences by 2 expert annotators, the initial 38 emotion labels were restructured to a hierarchical label set of the 5 emotions *Hatred*, *Anxiety*, *Sadness*, *Joy* and *Desire*.

Keywords: Early Modern Dutch Theatre, Historical Drama, NLP, OCR, Emotion Analysis

1. Introduction

Drama as a literary genre has been gaining interest in the Natural Language Processing (NLP) research field in recent years. In 2019, the DraCor database (Fischer et al., 2019) established a standardized XML TEI encoding framework, allowing the dramatic tradition in Europe to be described structurally and language-independently. Thanks to digitizing and encoding initiatives of literature throughout Europe in previous decades, the dramatic genre has opened up to computational and comparative research on a European level. Predicting structure in plain text dramas for corpus expansion through encoding enrichment (Pagel et al., 2021), network analysis based on structural drama features (Botond and Bence, 2023; Santa María Fernández and Dabrowska, 2023), coreference resolution (Pagel and Reiter, 2021), emotion analysis (Schmidt et al., 2021a; Dennerlein et al., 2023) and authorial style development in writing tragedies and comedies (Cafiero and Gabay, 2023) have shown how drama is opening up to data-driven analysis and interpretation. In spite of this momentum for European drama, the Dutch dramatic tradition has not yet been object of such structural comparative research. Lacking standardised datasets first and encoding enrichment second, the Dutch dramatic tradition needs centralisation still before it can partake in riding the waves of computational drama analysis.

In this paper, our objective is to propel research on the Dutch dramatic tradition forward by focusing on early modern Dutch comedies and farces, spanning the period from 1650 to 1725. Comedies and farces, traditionally underexposed or deemed inferior in Dutch literary historiography on drama (te Winkel, 1924; Knuvelder, 1964; Erenstein et al., 1996), showcase the importance of desire and imagination in early modern consumption culture by staging characters who experience socially confirming situations or imaginary social expansions in a broad range of economical settings, recognisable to the early modern consumers in the audience (van Stipriaan, 1996; Porteman and Smits-Veldt, 2008). These types of plays, therefore, display cultural social conduct regarding possession and value assignment, revealing the moral, social and emotional dynamics of early modern consumption (Hinnant, 1995; Perry, 2003; Goldstein and Tigner, 2016; Ferket, 2021). We therefore aim to model how desire and its objects are staged in comedies and farces in the Low Countries, and by doing so individuate unexplored patterns in the theatrical representation of the early modern Dutch consumption culture.

First, we relate the creation of the EmDComF corpus, consisting of manually curated and automatically digitised early modern Dutch comedies and farces authored between 1650 and 1725 in txt format, and evaluate the implementation of the Transkribus Print M1 model for automatic corpus expansion (Section 2). Then, we elaborate how we refined a historically motivated emotion annotation framework for early modern Dutch comedies and farces through data-driven clustering algorithms (Section 3). In conclusion, we discuss our findings and discuss future work (Section 4).

OCR	Ground Truth (OCR + GOLD)		MANUAL	
Google Books	Google Books + CENETON	Google Books + DBNL	CENETON	DBNL
217	92	34	108	15

Table 1: Overview of the EmdComF corpus (n=466) and its subsets.

2. The EmDComF Corpus

2.1. Collecting Text Editions

Comedies and farces were productive dramatic subgenres in early modern Dutch society. In total, we collected 466 early modern comedies and farces written by 165 authors in the period from 1650 to 1725.

For the collection of the early modern Dutch comedies and farces, we made use of both open source editions in txt format from the databases Digitale Bibliotheek voor de Nederlandse Letteren (DBNL) and Census Nederlands Toneel (CENETON), and of scanned editions accessible on Google Books using OCR. In our dataset of 466 unique historical plays, we can individuate three subgroups according to their database provenance: there are 123 texts only available in manually curated form, 217 texts are only available in OCR form, and there are two ground truths (GTs) which contain 126 texts for which both manually curated and OCRed texts are available. With the gold-OCR pairs in both GTs, we are able to measure the quality of OCRed texts in the EmDComF corpus in general, as we can compare 126 OCRed texts to their manual references. The distinction between both GTs is maintained throughout this comparison, because their manual text editions correspond differently to OCRed editions due to differing markup implementations. In Table 1, we give an overview of the provenance of the texts collected in the EmdComF corpus.

Two full-text databases provided manually curated texts, roughly making up half of the dataset. DBNL provided 49 and CENETON provided 200 manually curated plays. The other 217 plays were obtained OCRing scans of printed plays made available by Google Books using the Transkribus Print M1 model. This model was chosen to perform OCR, as it is trained on more than 5,000,000 words in 16 languages, among which Dutch, English, French, German, Italian, Spanish and Latin which appear in varying degrees throughout the plays in the EmDComF corpus, from several print typologies, such as the roman and blackletter script, sometimes both used at the same time in the print editions of the plays in the corpus. Finally, Transkribus' Print M1 model digitises text with an acclaimed 2.20% Character Error Rate accuracy according to their website[1], which makes it an interesting model for multilingual text recognition on multiple historical and modern scripts.

We first assess the quality of OCRed texts in the EmDComF corpus, before initiating further downstream content-wise NLP tasks such as sentiment and emotion detection or other profiling analyses. Doing this, we are able to evaluate which aspects of textual information are maintained or lost in the digitisation process using OCR.

2.2. Metrics

We use Character Error Rate ($CER = \frac{S+D+I}{C_{ref}}$) and Word Error Rate ($WER = \frac{S+D+I}{W_{ref}}$) to evaluate the performance of the digitisation process at the text level for gold-OCR pairs in both GTs. CER is the Levenshtein distance (Levenshtein et al., 1966) between predicted characters and their reference characters (C_{ref}), namely the minimal amount of substitutions (S), deletions (D) and insertions (I) needed to transform the OCRed characters into their reference characters, and WER is defined as the Levenshtein distance between predicted words and their reference words (W_{ref}) following the same logic (Neudecker et al., 2021). We report macro-averaged and micro-averaged CER and WER scores, with the former treating CER and WER scores equally regardless of text length and with the latter aggregating error rates cumulatively according to text length.

We complement CER and WER results with vectorisation similarity calculations, considering the averaged cosine similarity of the lexical and semantic vector representation of gold-OCR pairs in each GT on text level, to estimate the textual quality of the digitisation process. Lexical similarity is assessed through three perspectives on the combined vocabularies of the gold-OCR pairs per GT. First, lexical presence is modeled in gold-OCR pairs using a Bag-of-Words (BoW) representation of all gold and OCR word types per GT. Then, token frequency for each word type is captured through a count-based BoW representation. Finally, relative lexical significance is determined using a Term Frequency-Inverse Document Frequency (TF-IDF) representation, where token frequency per word type in each text is weighted based on the combined vocabulary frequency in its gold-OCR collection. Semantic similarity is modeled with a Doc2Vec representation, which considers context and the contextual meaning of words (Řehůřek and Sojka, 2010; Le and Mikolov, 2014).

[1] https://readcoop.eu/transkribus/public-models/

2.3. Cleaning & Preprocessing

Embedded in two separate databases, different markups were used for the original formatting of the manually curated plays. The manual editions from CENETON were automatically extracted from their html hyperlinks and converted to txt format. The manual editions from DBNL were downloaded in txt format. All manually curated editions downloaded in raw txt needed manual and semi-automatic (regex) cleaning to be able to form the ground truth for their respective raw OCR renderings to be compared to, since the manually curated editions incorporate textual noise superfluous and detrimental to this task and since the OCRed texts can only follow the scans of printed editions (Example 1).

1. -(==1==)(»pagina-aanduiding«) DE GEWAANDE ADVOCAAT, KLUCHTSPÉL.[2]
   - DE
   GEWAANDE
   ADV
   CCAAT
   KLUCHTSPÉL

We created two GTs, one for each manually curated subset, to get insight into the quality of the OCRed texts. The CENETON GT consists of 92 gold-OCR pairs and the DBNL GT consists of 34 gold-OCR pairs, which means that we OCRed print editions for which manually curated versions are available in the databases. We calculate the CER and WER values for all pairs in the GTs after different preprocessing steps, comparing the OCRed version of a text with its manually curated edition. Finally, we compare the lexical and semantic vector representations of the GTs after each preprocessing step to measure the textual similarity between gold and OCRed texts at text level. Doing this, we can make an informed estimation of the quality of the subset of 217 OCRed texts for which no manually curated text data are available.

Preprocessing steps undertaken to streamline raw gold-OCR pairs as much as possible, are:

1. Removing superfluous tabs and whitespaces.
2. Lowercasing and decoding diacritical marks.
3. Removing punctuation.
4. Automatic word segmentation and spellcheck using Symspellpy's edit distance[3] for out-of-vocabulary (OOV) OCRed tokens, based on the monogram and bigram frequency dictionary of all manually curated data.

Here follows a gold-OCR pairwise comparison per preprocessing step to illustrate the impact of preprocessing on CER and WER scores, where the first example is a gold sentence and the second an OCRed sentence. Hyphens separate the instances.

1) - DE GEWAANDE ADVOCAAT, KLUCHTSPÉL.
- DE GEWAANDE ADV CCAAT KLUCHTSPÉL
$CER: 12.12 \mid WER: 75.0$

2) - de gewaande advocaat, kluchtspel.
- de gewaande adv ccaat kluchtspel
$CER: 12.12 \mid WER: 75.0$

3) - de gewaande advocaat kluchtspel
- de gewaande adv ccaat kluchtspel
$CER: 6.5 \mid WER: 50.0$

4) - de gewaande advocaat kluchtspel
- de gewaande adv caat kluchtspel
$CER: 3.2 \mid WER: 50.0$

2.4. Results

2.4.1. CER and WER

The distribution of micro and macro-averaged CER and WER in the two databases throughout the preprocessing steps in Table 2 is telling for both how the manually curated editions functioned as references in the GTs and how well the OCR performed on the scanned editions.

Micro and macro-averaged CER and WER after preprocessing steps for the CENETON and DBNL GTs show opposite tendencies, with the CENETON GT obtaining better scores for micro-averaging and the DBNL GT obtaining better scores for macro-averaging. This indicates on the one hand that the DBNL GT had better scoring pairs in its collection (n=34) though cumulatively more errors were aggregated, whereas on the other hand CENETON had worse scoring pairs in its collection (n=92) though reaching lower accumulated error rates.

In general, though, the DBNL GT scores better than the CENETON GT, reaching lower CER and WER scores after preprocessing steps. This indicates that the DBNL gold-OCR pairs throughout correspond better textually, so that there are fewer out of gold strings in OCRed text and/or fewer out of OCRed print edition strings in manually curated text. Despite these preprocessing steps, both OCRed and gold texts in the GTs exhibit non-corresponding textual noise, which maintains Character Error Rate (CER) and Word Error Rate (WER) scores, including: OCR mistakes (wrongfully recognised and/or separated characters), structural deviations from the gold texts present in the OCRed texts, such as repeated titles, acts and scenes in headers and footers and (un)succesfull rendered Google Books vignets, and textual deviations from the OCRed texts indicating text structuring elements, occasional manual typos or word segmentation mistakes in the reference.

[2] English: The Presumed Lawyer, Farce.

[3] https://github.com/mammothb/symspellpy

	CENETON (n=92)		DBNL (n=34)		COMBINED (n=126)	
Step	**CER^{M}**	**WER^{M}**	**CER^{M}**	**WER^{M}**	**CER^{M}**	**WER^{M}**
1	10.00	17.82	12.59	18.89	10.70	18.11
2	9.29	15.40	7.88	12.00	8.91	14.48
3	8.48	11.60	7.44	10.07	8.20	11.19
4	9.00	9.98	**7.11**	**8.39**	8.29	9.55
Step	**CER^{m}**	**WER^{m}**	**CER^{m}**	**WER^{m}**	**CER^{m}**	**WER^{m}**
1	9.76	17.26	13.13	19.86	10.67	17.96
2	9.09	14.89	8.10	12.36	8.82	14.21
3	**8.36**	11.50	7.65	10.40	**8.17**	11.20
4	8.85	**9.85**	7.31	8.69	8.43	**9.54**

Table 2: Macro-averaged^{M} and Micro-averaged^{m} CER and WER scores after preprocessing the CENETON GT, DBNL GT and the combined GTs.

	CENETON (n=92)				DBNL (n=34)			
vectorised	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
BoW	77.65	81.93	87.69	**90.12**	74.73	85.81	88.89	**91.50**
COUNT	99.26	99.38	99.31	**99.40**	99.41	99.53	99.52	**99.58**
TF-IDF	96.97	97.39	97.15	**97.91**	97.90	98.31	98.26	**98.73**
Doc2Vec	96.38	98.07	98.81	**98.92**	97.76	99.15	99.46	**99.51**

Table 3: Averaged cosine similarity scores of BoW, count-based, TF-IDF and Doc2Vec vector representations of the CENETON and DBNL GTs at text level after each preprocessing step.

Nonetheless, layout and string normalisation proves to textually align OCR outputs better in both GTs, as shown in Table 2, by removing superfluous tabs and spaces and punctuation and by lowercasing and ignoring accents. In this way, omitted or superfluously inserted accents or punctuation, or non-corresponding upper-cased or lower-cased characters in both gold and OCRed texts do not interfere with CER and WER scores. Using the previous preprocessing steps and a dictionary-based word segmentation and spellcheck algorithm for OOV OCRed tokens in step 4) from the Symspellpy library, we conclude that, despite the inevitable discrepancies caused by non-corresponding textual noise, OCRed text in this dataset on the average of 126 gold-OCR pairs reaches a correspondence of 91.5% on the character level and 90.5% on the word level.

2.4.2. Lexical and Semantic Vectorisation

The CER and WER scores are indicative of the performance of the Transkribus M1 model for digitising scanned print editions of early modern Dutch comedies and farces, and can be further supported by the averaged cosine similarity scores of the lexical and semantic vector representations of the GTs at the text level per preprocessing step to estimate textual quality after the digitisation process.

Comparing the lexical vectorisation of OCRed texts and their reference texts, we measure lexical differences between each text pair by modeling lexical presence on gold and OCR word types (BoW), lexical frequency on token frequency per word type (count-based BoW) and relative lexical significance (TF-IDF) based on the relative token frequency weighted on the gold-OCR combined vocabulary frequency per GT. This way, both the corresponding and deviating vocabulary items from OCRed texts are assessed from 3 lexical perspectives in the lexical similarity calculations. To perform these lexical vectorisations of the gold-OCR pairs, we used scikit-learn (Pedregosa et al., 2011). The semantical vectorisation of OCRed texts and their reference texts is performed by a Doc2Vec model that is trained on each gold-OCR collection per preprocessing step (5 models for CENETON GT, and 5 for DBNL GT), with each a vector size of 300 and a context window of 10 tokens. Gensim was used to obtain semantic Doc2Vec representations of gold-OCR pairs (Řehůřek and Sojka, 2010). The averaged similarity scores of the lexically and semantically vectorised gold-OCR pairs per preprocessing step on text level are found in Table 3.

The averaged cosine similarity scores of the lexical vector representations of the gold-OCR pairs per GT indicate that the BoW vectorisation, which models the vocabulary word types present in text pairs, closely correlates with the reported micro-averaged WER scores for both GTs after preprocessing step 3 and 4. For the DBNL GT after step 3 and 4, BoW scores 88.89 and 91.50 and WER scores 10.40 and 8.69; and for the CENETON GT, BoW scores 87.69 and 90.12 and WER scores 11.50 and 9.85. This correlation is to be expected, since a BoW representation assesses vocabulary presence at the word type level by creating new word types for OOV OCRed tokens and WER quantifies the percentage of these tokens that do not match their reference counterparts. Count-based BoW vector representations modeling the token frequency of vocabulary word types present in text pairs show an almost exact lexical frequency similarity between gold and OCRed texts, indicating that the distribution of terms is highly consistent across both gold-OCR collections, with minimal variation introduced by token frequencies of OCRed word types regardless of the preprocessing steps. TF-IDF vector representations demonstrate a high level of agreement regarding lexical importance in the GTs, yet exhibit a 2.09% deviation in the CENETON GT and a 1.27% deviation in the DBNL GT, underscoring the relative differences in which word types are important based on their token frequencies in correlation to their frequency distribution throughout all gold-OCR pairs per GT. Doc2Vec averaged cosine similarity scores reveal a high and incrementing semantic similarity between gold and OCRed text pairs after each preprocessing step in both GTs. This means that the semantic content and context captured by the Doc2Vec embeddings are becoming increasingly aligned through preprocessing.

Higher similarity scores are, again, generally reported for the DBNL GT, indicating that gold-OCR pairs in the DBNL GT lexically and semantically deviate less than pairs in the CENETON GT, as the former has less and the latter has more non-corresponding textual noise in its OCRed or gold texts on average. Finally, there is a tendency for the similarity scores to increase per preprocessing step in both GTs, with a slight deviation in the count-based and TF-IDF similarity scores for the CENETON GT and DBNL GT after preprocessing step 3 that removes punctuation. Therefore, we find that the proposed preprocessing steps generally lower the distance between the OCRed and gold texts in both GTs by effectively making their lexical and semantic similarities more explicit in all vector representations.

Based on the averaged vectorised comparison of 126 gold-OCR pairs, we conclude that the OCRed texts can be expected to be qualitative enough for further textual analysis despite persisting CER and WER scores averaging around 8.5% and 9.5% respectively due to non-corresponding textual noise, since they capture very similar amounts of lexical and semantic information to their manually curated counterparts. By analogy, this means that the subset of 217 uniquely OCRed texts should convey similar amounts of lexical and semantic information on average after the proposed preprocessing steps, which makes them valuable assets to this dataset. This also suggests the automatic digitisation of early modern Dutch comedies and farces using Transkribus M1 model to be a worthwhile corpus expansion method.

At last, we contend that the presence of textual noise previously identified in both OCRed and gold texts should not necessarily undermine the overall textual quality of the OCRed texts in the EmDComF corpus. In future work, our focus will be on testing the usability of these OCR data through textual analysis. This will best illustrate the real impact of the observed average 2.09% deviation in the relatively significant content words in the OCRed CENETON vocabulary and of the 1.27% deviation in the OCRed DBNL vocabulary on the one hand, and the impact of the semantic deviations of 1.08% for OCRed CENETON texts and 0.49% for OCRed DBNL texts on the other hand. These deviations might eventually be deemed negligible within this corpus, which could have important implications for the automatic corpus expansion of other historical dramatic traditions for which no manually curated but scanned editions are available. Nonetheless, we plan to explore additional OCR post-correction or language normalisation techniques to further process deviations in OCRed texts of the EmDComF corpus.

3. Historical Emotions in EmDComF

The EmDComF corpus consists of 466 early modern Dutch comedies and farces, with OCRed texts that we have demonstrated to display very high lexical and semantic similarities to the manually curated editions on average. Now that the textual quality of both types of text editions has been put into perspective, we proceed to the content-wise emotion analysis of these text data as early modern Dutch comedies and farces have been suggested to particularly display moral, social and emotional dynamics of early modern Dutch consumption culture (Hinnant, 1995; Perry, 2003; Goldstein and Tigner, 2016; Ferket, 2021). After a discussion on Emotion Analysis in historical drama, we describe the refinement of a data-driven genre-specific emotion annotation framework to be implemented in a Machine Learning (ML) approach. With this, we aim to create expert systems capable of automatically detecting emotion within the EmDComF corpus by fine-tuning pre-trained LLMs (large language models) on historical and modern Dutch, such as GysBERT (Manjavacas Arevalo and Fonteyn, 2022) and BERTje (de Vries et al., 2019) respectively, based on the manual annotations of sentiment and emotion in the corpus.

3.1. Emotion Analysis in Historical Drama

Sentiment Analysis (SA) is defined by Liu (2020) as the field of study that analyses people's opinions, sentiments, appraisals, attitudes, and emotions toward entities and their attributes expressed in written text. As a popular application from the

Emotion	A-a	B-a	A-r	B-r	K	F1
Affection	11	13	1.99	2.96	0.66	0.67
Ambition	37	27	6.70	6.15	0.61	0.63
Anger	43	49	7.79	11.16	0.82	0.83
Audacity	4	31	0.72	7.06	0.22	0.23
Aversion	50	53	9.06	12.07	0.76	0.78
Compassion	0	1	0.00	0.23	0.00	0.00
Confidence	81	7	14.67	1.59	0.12	0.14
Consternation	26	15	4.71	3.42	0.48	0.49
Courage	3	2	0.54	0.46	0.40	0.40
Cowardice	0	1	0.00	0.23	0.00	0.00
Curiosity	20	6	3.62	1.37	0.38	0.38
Desperation	8	8	1.45	1.82	0.75	0.75
Devotion	18	8	3.26	1.82	0.61	0.62
Enjoyment	1	6	0.18	1.37	0.28	0.29
Favor	8	9	1.45	2.05	0.94	0.94
Fear	9	9	1.63	2.05	0.66	0.67
Friendship	8	8	1.45	1.82	0.87	0.88
Gratitude	5	6	0.91	1.37	0.54	0.55
Hatred	3	5	0.54	1.14	0.75	0.75
Hope	14	4	2.54	0.91	0.33	0.33
Indecision	1	0	0.18	0.00	0.00	0.00
Indignation	66	41	11.96	9.34	0.61	0.64
Joy	17	9	3.08	2.05	0.45	0.46
Love	23	27	4.17	6.15	0.88	0.88
Peace of mind	2	4	0.36	0.91	0.67	0.67
Pity	1	1	0.18	0.23	1.00	1.00
Pride	17	9	3.08	2.05	0.69	0.69
Regret	12	11	2.17	2.51	0.78	0.78
Remorse	1	1	0.18	0.23	1.00	1.00
Sadness	20	12	3.62	2.73	0.62	0.63
Satisfaction	25	26	4.53	5.92	0.82	0.82
Shame	1	1	0.18	0.23	1.00	1.00
Uneasiness	15	26	2.72	5.92	0.58	0.59
Vindictiveness	2	3	0.36	0.68	0.80	0.80

Table 4: Absolute frequency (a), relative frequency (r), Cohen's Kappa (K), and F1-score for the 34 emotion labels annotated by annotator A and annotator B in the annotation test set of 782 sentences.

NLP domain, SA is nowadays often being used to identify positive, neutral, negative or mixed sentiments expressed in product reviews or social media posts, as well as the targets of these sentiments (Liu, 2020). Emotion Analysis (EA), a subdomain of SA, deals with the more complex task of identifying different emotion classes like joy and sadness in texts, instead of the aforementioned sentiment polarity (Kim and Klinger, 2019; Rebora, 2023). In the last decade, SA and EA have been increasingly applied at the intersection of NLP and Digital Humanities (DH) in the field of Computational Literary Studies, as literary research is often concerned with understanding sentiments and emotions that organise and orient narratives throughout literary genres since the emergence of literary traditions (Hogan, 2011).

In computational literary research adopting SA and EA in historical drama, Leemans et al. (2017) aimed to trace historical changes in emotion expression and in the embodiment of emotions in a corpus of 29 historical Dutch theatre plays from between 1600 to 1800. To this end, the first lexicons and emotion classification schemes for early modern Dutch were created by annotating 27,993 sentences with 38 historically accurate emotion labels, body part labels, bodily process labels, emotional action labels and body sensation labels (van der Zwaan et al., 2015; Leemans et al., 2017). Using a combination of dictionary-based approaches, this first historical emotion classification methodology for early modern Dutch drama reached a 10% precision and 60% recall on the test set.

In historical German drama, state-of-the-art methodologies have been applied for sentiment and emotion classification using transformer-based language models (Schmidt et al., 2021a; Dennerlein et al., 2023). Anchoring their hierarchical emotion annotation scheme in a German literary studies perspective to annotate 13 sub-emotions coming from 6 main emotion classes expressed or attributed to characters (Schmidt et al., 2021b; Dennerlein et al., 2022), Schmidt et al. (2021a) acquired 13,264 annotations from 11 historical German plays and Dennerlein et al. (2023) acquired 11,939 annotations from 17 historical German plays. Both studies evaluated multiple transformer-based ML approaches to classify text sequences with single emotion labels from their emotion framework. Schmidt et al. (2021a) separately report polarity classification accuracy and F1-score up to 90% for their 2 polarity classes (positive/negative), 75% accuracy and F1-score for main emotion class classification and 66% accuracy and F1-score for the 13 sub-emotion classification after fine-tuning on an annotation subset filtered on disagreeing annotations. Dennerlein et al. (2023) report an accuracy of 73% for the 14 sub-emotion classification (a neutral category was added) in cross-validation from which the 6 main emotion classifications and 4 polarity classifications are derived, after fine-tuning on a similarly filtered annotation subset. With this performance, Dennerlein et al. (2023) succeeded in detecting emotional differences between historical German comedies and tragedies.

3.2. Operationalising Historical Emotions

To be able to detect emotions that are historically relevant to the comedies and farces from the EmDComF corpus, we operationalised the historical emotion framework for early modern theatre composed by Lodewijk Meyer around 1670 for emotion annotation. Meyer defined emotions, or passions, as abnormal motions of the heart caused by the notions of good or evil and perceived by the soul (Steenbakkers, 1999). This vision on emotions being caused by individual moralistic judgment about

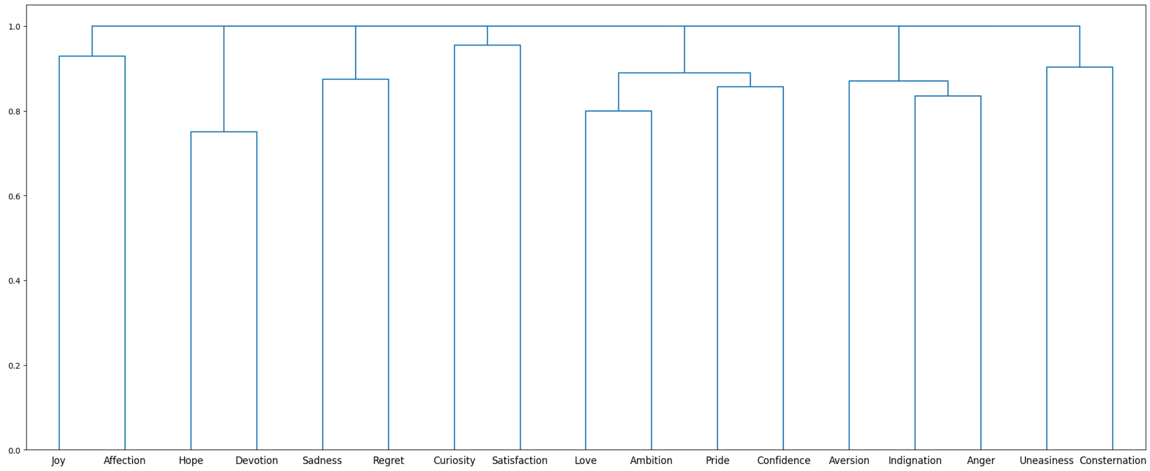

Figure 1: Annotator A's emotion clusters with weighted-linking filtered on infrequent emotions.

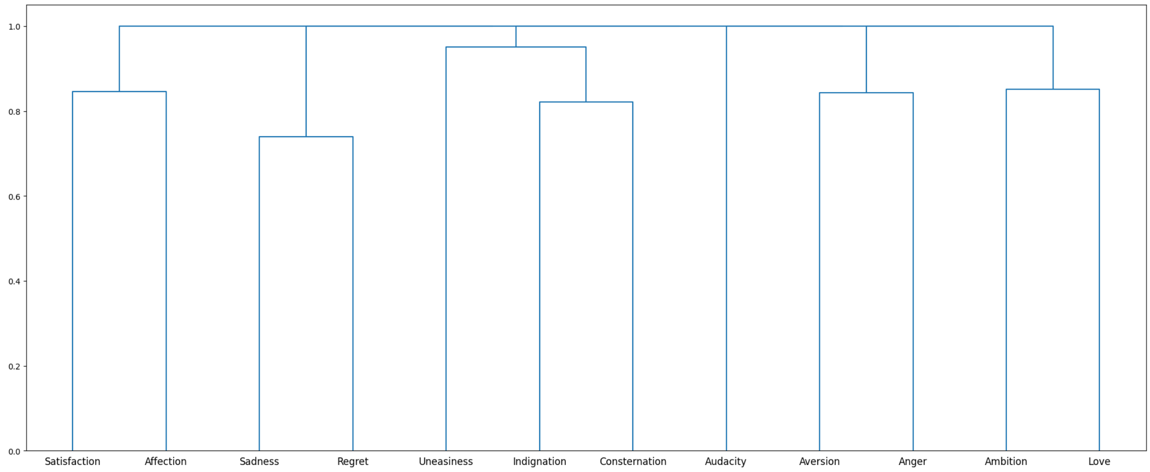

Figure 2: Annotator B's emotion clusters with weighted-linking filtered on infrequent emotions.

good or evil was rooted in contemporary philosophical and literary debates on ethics and human nature. In the instructive work on theatre poetics *Naauwkeurig onderwys in de tooneel-poëzy* (Thorough instruction in the Poetics of Drama) published by literary society Nil Volentibus Arduum in 1765 (Harmsen, 1989; Steenbakkers, 1999), Meyer's moralistic and individualistic conceptualisation of emotion was authoritative in early modern Dutch theatre writing. His description of 38 emotions in the domain-specific context of the EmDComF corpus therefore validates our approach to adopting this emotion annotation framework.

In the annotation study conducted to get insight in the emotionality of the EmDComF corpus, we made use of Meyer's initial 38 emotion labels to annotate emotions expressed or attributed to characters in sentences. Sentiment was annotated on sentence level, using a positive, neutral or negative label. Per sentence, only one sentiment but multiple emotions could be annotated if this was necessary. NLTK Punkt sentence segmentation (Bird et al., 2009) was used to create the sentences, as this greedy sentence splitting method seemed most fit for this task instead of relying on regular expressions. Guided by the description of these 38 emotion categories as summarised by Harmsen (1989) during the annotations, two expert annotators independently annotated emotions and sentiment in one act of a comedy from the corpus, consisting of 782 sentences in authentic early modern Dutch. In these sentences, annotator A annotated 552 emotions, and annotator B annotated 439 emotions using 34 of the 38 emotion labels with varying frequencies. This is due to the fact that delineating emotions in these sentences is at times an interpretative task, which is why the annotators often disagreed in their annotations. In Table 4, we give an overview of the annotation study from the perspective of both annotators per annotated emotion label, the absolute and relative frequencies, and Cohen's kappa (Cohen, 1960) and F1-scores to determine the Inter-Annotator Agreement (IAA).

Throughout the 34 annotated emotion labels, emotion agreement is moderate as the mean Kappa score is 0.59 (0.4 < k < 0.6) and F1-score is 0.60 , whereas sentiment agreement is substantial with a mean Kappa score of 0.75 (0.6 < k < 0.8) and F1-score of 0.85. Nevertheless, class imbalances due to different emotion frequency annotations per annotator were created by this fine-grained annotation set, resulting in a few emotion labels with non-existing or perfect IAA scores. For the emotions *Compassion*, *Cowardice* and *Indecision*, IAA scores are 0.00 as these single-time annotated labels were not used by the other annotator, and IAA for the emotions *Pity*, *Remorse* and *Shame* is theoretically perfect as the single time that these emotions occurred in the annotation set, they were annotated. For example in sentence 656 "Zou hy zich zo verstooren?"[4], annotator A labeled *Indecision* and B labeled *Uneasiness*; in sentence 73 "Ik kon immers 't arme maag're beest zo niet in de open lucht laaten staan."[5], both annotators labeled *Pity*. More frequently occurring emotions like *Ambition*, *Anger*, *Aversion*, *Indignation*, *Love* and *Satisfaction* report more meaningful substantial IAA scores as these were throughout consistently annotated by both annotators. Finally, some emotions like *Audacity*, *Confidence*, *Curiosity*, *Enjoyment* and *Hope* were not annotated with consistent frequency by both annotators, meaning that often the other annotator did not label that sentence or used another emotion label having interpreted it differently. For example in sentence 714 "Zeer wel, daar wil ik wel van snoepen."[6], annotator A labeled *Joy* and B labeled *Enjoyment*.

3.2.1. Clustering Emotion Annotations

Annotating 38 emotion labels remains a challenge as some emotion labels are hard to be distinguished from one another and seem open to interpretation, even though moderate IAA was reached on annotating 34 of the initial 38 emotions found in the annotation test. To operationalise this fine-grained label set and to establish an emotionality framework that fits the EmDComF corpus best, we apply methodological solutions for emotion class imbalances in fine-grained emotional frameworks, illustrated by similar research analysing emotion in another domain. De Bruyne et al. (2019) created a domain-specific emotion set of 5 emotion labels by clustering the annotations from an annotation study labeling 25 emotions in modern Dutch tweets. Their approach increased efficiency in the annotation process by hierarchically structuring the emotion framework, which means that similar emotion labels were grouped together under a broader label that captures a shared emotional essence. Therefore, as hierarchical emotion frameworks have also shown to be effective in research detecting emotion in historical German drama (Schmidt et al., 2021a; Dennerlein et al., 2023), we adopt the clustering methodology to hierarchically structure the 34 annotated emotion labels in the annotations by both annotators as proposed by De Bruyne et al. (2019).

To cluster the annotated emotion labels per annotator, each emotion label is transformed into a vector, resulting in 34 782-dimensional vectors as 782 sentences were annotated. Emotion presence in these vectors is binarised per dimension, with 1 indicating emotion presence and with 0 indicating its absence. These binarised vectors are then fed to hierarchical clustering algorithms performed with SciPy (Virtanen et al., 2020). SciPy's weighted linkage method (WPGMA: $d(u,v) = \frac{\text{dist}(s,v)+\text{dist}(t,v)}{2}$)[7] resulted in the most intuitive dendrograms (emotion label clusters), as it iteratively merges clusters based on the average distances between them, ultimately forming a hierarchical structure that reflects those average linkage distances, and are therefore the only clustering results we report. We applied the weighted linkage method on both annotation sets and on the annotation sets filtered on infrequent emotion labels occurring less than 10 times per set, showcasing the consistent cluster intervals based on the average linkage distances. Figures 1 and 2 show the weighted-linking dendrograms based on both annotators' filtered annotation set.

Concluding, 7 hierarchical clusters result from annotator A's and 5 from annotator B's annotations. Both dendrograms acknowledge main classes for *Joy*, *Sadness*, *Hatred*, *Anxiety* and *Desire*; with annotator A's dendrogram distinguishing another two main classes for *Hope-Devotion* and *Curiosity-Satisfaction*. Based on the emotion frequency and weighted cluster results of the 782 annotated sentences and a historical literary studies perspective on the dramatics and emotions in early modern Dutch comedies and farces, we propose the hierarchical emotion set of 5 labels that we will continue to use in future annotations in Table 5. Based on the dendrograms, we merged *Fear*, *Desperation* and *Cowardice*; *Consternation*, *Uneasiness* and *Indignation*; *Regret* and *Remorse*; *Joy* and *Enjoyment*; *Affection*, *Friendship*, *Compassion* and *Gratitude*; *Satisfaction*, *Peace of mind* and *Relief*; *Devotion* and *Favour*; and finally *Pride*, *Confidence*, *Audacity* and *Courage*. We left the under-represented emotions of *Vindictiveness*, *Indecision*, *Shame* and *Pity* extant to be annotated in future work to de-

[4] English: Would he be so upset?

[5] English: After all, I couldn't leave the poor, skinny animal out in the open like that.

[6] English: Very well, I would like to snack on that.

[7] Weighted Pair Group Method with Arithmetic Mean: $d(u,v)$: distance between clusters u and v. $dist(s,v)$: distance from s to v, $dist(t,v)$: distance from t to v. The formula averages these distances to compute $d(u,v)$.

Label	A-a	B-a	A-r	B-r
Hatred	**98**	**18%**	**110**	**25%**
aversion	50	9%	53	12%
anger	43	8%	49	11%
vindictiveness	2	0%	3	1%
(hatred)	3	1%	5	1%
Anxiety	**126**	**23%**	**101**	**23%**
fear	17	3%	18	4%
indecision	1	0%	0	0%
shame	1	0%	1	0%
consternation	107	19%	82	19%
Sadness	**34**	**6%**	**25**	**6%**
(sadness)	20	4%	12	3%
regret	13	2%	12	3%
pity	1	0%	1	0%
Joy	**69**	**13%**	**73**	**17%**
(joy)	18	3%	15	3%
affection	24	4%	28	6%
satisfaction	27	5%	30	7%
Desire	**225**	**41%**	**130**	**30%**
devotion	26	5%	17	4%
love	23	4%	27	6%
ambition	37	7%	27	6%
pride	105	19%	49	11%
hope	14	3%	4	1%
curiosity	20	4%	6	1%

Table 5: Absolute (a) and relative (r) emotion frequency distribution in the hierarchical label set of 5 emotions and 20 sub-emotions based on the annotations of annotator A and B.

cide if they have their own place in this framework or should be merged. We finally remark that the resulting hierarchical emotion framework of 5 labels seems to correspond quite closely to some modern day emotion classifications, namely with the 5 labels of *Joy*, *Sadness*, *Anger*, *Nervousness* and *Love* established by De Bruyne et al. (2019) or with 4 of the 6 basic emotions linked to universal facial expressions of *Joy*, *Anger*, *Fear* and *Sadness* by Ekman (1992). Other than these emotion frameworks, we explicitly maintain more fine-grained emotion sub-classifications as we hope they will eventually be useful in an Aspect-based Sentiment Analysis (ABSA) methodology to find the objects of the detected emotions in early modern Dutch comedies and farces, with specific regard to the different sub-emotions that were clustered under the label of *Desire*.

4. Conclusion & Future Work

In this paper, we presented and evaluated the EmDComF corpus of 466 early modern Dutch comedies and farces written between 1650 and 1725 in txt format, of both manually curated and OCRed text editions. The quality of OCRed texts in the corpus was measured using CER and WER on 126 gold-OCR text pairs, which resulted in a micro-averaged CER score of 8.43 and a WER score of 9.54 after preprocessing. Finally, we calculated the lexical and semantic vectorisation similarity of 126 gold and OCR texts on text level to further estimate the textual quality of the OCRed texts. These results indicated high lexical and semantic similarities between OCRed texts and their manually curated edition on average, which generally increased after preprocessing. Based on these average CER and WER scores and average lexical and semantic vectorisation similarity scores, we can expect the subset of 217 uniquely OCRed plays in the EmDComF corpus to be similarly qualitative in line with said averages.

Having framed the digitisation performance and the lexical and semantic quality of OCRed plays in the EmDComF corpus, we then related how we refined a historical emotion annotation framework for emotion analysis in early modern Dutch comedies and farces. Lodewijk Meyer's philosophical and literary work on emotions in early modern Dutch theatre provided us the framework of 38 emotion labels. An annotation study on 782 sentences labeling these 38 emotions was conducted by two expert annotators independently. Their annotations indicated a dense emotionality spectrum in the sentences, as 34 emotions had been used in the annotations with a mean Kappa score of 0.59, indicating moderate annotation agreement. Expectedly, annotation sparsity was evident for some emotions using this fine-grained framework. To operationalise the initial emotion framework, clustering algorithms were performed on the annotated emotion labels to establish a hierarchical emotion label set. The refined emotion annotation framework we propose consists of 5 hierarchical emotions *Hatred*, *Anxiety*, *Sadness*, *Joy* and *Desire* with 20 possible sub-emotions based on the clusterisation and the annotation frequencies. Recalculating IAA on the 5 main emotion labels, annotation agreement is now substantial with a mean Kappa score of 0.68 and F1-score of 0.72. *Hatred* has almost perfect agreement with a 0.85 Kappa score; *Joy*, *Sadness* and *Anxiety* have substantial agreement with Kappa scores of 0.79, 0.68 and 0.64 respectively; *Desire* is the hardest category to agree on, having moderate agreement with a 0.43 Kappa score.

Our future plans involve expanding the annotation of plays using this refined emotion framework. These annotations will then serve as training data to fine-tune LLMs, allowing for the creation of expert systems capable of automatically detecting emotion in early modern Dutch comedies and farces. Additionally, we plan to integrate this approach into an ABSA methodology to automatically link identified emotions with their respective objects in the EmDComF corpus, aiming to establish an emotional object typology specific for these types of historical plays with specific regard to expressions of *Desire*.

5. Acknowledgements

This research was carried out with the support of the Research Foundation – Flanders (FWO) under grant G032123N.

6. Bibliographical References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Szemes Botond and Vida Bence. 2023. Tragic and Comical Networks. Clustering Dramatic Genres According to Structural Properties. (arXiv:2302.08258).

Florian Cafiero and Simon Gabay. 2023. Dating the Stylistic Turn: The Strength of the Auctorial Signal in Early Modern French Plays. In *Computational Humanities Research*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Luna De Bruyne, Orphée De Clercq, and Véronique Hoste. 2019. Towards an empirically grounded framework for emotion analysis. In *HUSO 2019: The Fifth International Conference on Human and Social Analytics*, pages 11–16. IARIA, International Academy, Research, and Industry Association.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT Model. (arXiv:1912.09582).

Katrin Dennerlein, Thomas Schmidt, and Christian Wolff. 2022. Figurenemotionen in deutschsprachigen Dramen annotieren. Zenodo.

Katrin Dennerlein, Thomas Schmidt, and Christian Wolff. 2023. Computational emotion classification for genre corpora of German tragedies and comedies from 17th to early 19th century. *Digital Scholarship in the Humanities*, 38(4):1466–1481.

Paul Ekman. 1992. An argument for basic emotions. volume 6, pages 169–200, United Kingdom. Taylor & Francis.

Robert L. Erenstein, Dirk Coigneau, Robert van Gaal, Flor Peeters, Herman Pleij, Karel Porteman, Jaak van Schoor, and Mieke B. Smits-Veldt. 1996. *Een Theatergeschiedenis Der Nederlanden. Tien Eeuwen Drama En Theater in Nederland En Vlaanderen*. Amsterdam University Press.

Johanna Ferket. 2021. *Hekelen Met Humor: Maatschappijkritiek in Het Zeventiende-Eeuwse Komische Toneel in de Nederlanden*. Uitgeverij Verloren.

Frank Fischer, Ingo Börner, Mathias Göbel, Angelika Hechtl, Christopher Kittel, Carsten Milling, and Peer Trilcke. 2019. Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. Zenodo.

David B. Goldstein and Amy L. Tigner. 2016. *Culinary Shakespeare: Staging Food and Drink in Early Modern England*. Penn State Press.

Antonius Johannes Engbert Harmsen. 1989. *Onderwys in de tooneel-poëzy: de opvattingen over toneel van het Kunstgenootschap Nil Volentibus Arduum*. Ordeman.

Charles H. Hinnant. 1995. Pleasure and the Political Economy of Consumption in Restoration Comedy. *Restoration: Studies in English Literary Culture, 1660-1700*, 19(2):77–87.

Patrick Colm Hogan. 2011. *Affective Narratology: The Emotional Structure of Stories*. University of Nebraska Press.

Evgeny Kim and Roman Klinger. 2019. A Survey on Sentiment and Emotion Analysis for Computational Literary Studies. *Zeitschrift für digitale Geisteswissenschaften*.

Gerard Petrus Maria Knuvelder. 1964. *Handboek Tot de Geschiedenis Der Nederlandse Letterkunde*. Malmberg, Den Bosch.

Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. (arXiv:1405.4053).

Inger Leemans, Janneke M. van der Zwaan, Isa Maks, Erika Kuijpers, and Kristine Steenbergh. 2017. Mining Embodied Emotions: A Comparative Analysis of Sentiment and Emotion in Dutch Texts, 1600-1800. *Digital Humanities Quarterly*, 11(4).

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

Bing Liu. 2020. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*, 2 edition. Studies in Natural Language Processing. Cambridge University Press, Cambridge.

Enrique Manjavacas Arevalo and Lauren Fonteyn. 2022. Non-parametric word sense disambiguation for historical languages. In *Proceedings of* *the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 123–134, Taipei, Taiwan. Association for Computational Linguistics.

Clemens Neudecker, Konstantin Baierer, Mike Gerber, Christian Clausner, Apostolos Antonacopoulos, and Stefan Pletschacher. 2021. A survey of ocr evaluation tools and metrics. In *Proceedings of the 6th International Workshop on Historical Document Imaging and Processing*, HIP '21, page 13–18, New York, NY, USA. Association for Computing Machinery.

Janis Pagel and Nils Reiter. 2021. DramaCoref: A Hybrid Coreference Resolution System for German Theater Plays. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 36–46, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Janis Pagel, Nidhi Sihag, and Nils Reiter. 2021. Predicting Structural Elements in German Drama. In *Proceedings of the Second Conference on Computational Humanities Research*, volume 1613, page 0073.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Curtis Perry. 2003. Commerce, Community, and Nostalgia in The Comedy of Errors. In Linda Woodbridge, editor, *Money and the Age of Shakespeare*, pages 39–51. Palgrave Macmillan US, New York.

K. Porteman and Mieke B. Smits-Veldt. 2008. *Een nieuw vaderland voor de muzen: geschiedenis van de Nederlandse literatuur, 1560-1700*. Bakker.

Simone Rebora. 2023. Sentiment Analysis in Literary Studies. A Critical Survey. *Digital Humanities Quarterly*, 017(2).

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Luis Rei and Dunja Mladenić. 2023. Detecting Fine-Grained Emotions in Literature. volume 13, page 7502. Multidisciplinary Digital Publishing Institute.

María Teresa Santa María Fernández and Monika Dabrowska. 2023. Análisis comparativo del Coro como personaje en tres tragedias griega y tres dramas españoles del Corpus DraCor. *Neophilologus*, 107(3):389–412.

Thomas Schmidt, Katrin Dennerlein, and Christian Wolff. 2021a. Emotion Classification in German Plays with Transformer-based Language Models Pretrained on Historical and Contemporary Language. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 67–79, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.

Thomas Schmidt, Katrin Dennerlein, and Christian Wolff. 2021b. Towards a Corpus of Historical German Plays with Emotion Annotations. In *3rd Conference on Language, Data and Knowledge (LDK 2021)*, volume 93 of *Open Access Series in Informatics (OASIcs)*, pages 9:1–9:11, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Piet Steenbakkers. 1999. The Passions according to Lodewijk Meyer: Between Descartes and Spinoza. In *Desire and Affect: Spinoza as Psychologist ; Papers Presented at the Third Jerusalem Conference (Ethica III)*, pages 193–210.

J. te Winkel. 1924. *De Ontwikkelingsgang Der Nederlandsche Letterkunde, Deel IV. Geschiedenis der Nederlandsche letterkunde van de Republiek der Vereenigde Nederlanden (2)*. De erven F. Bohn, Haarlem.

Janneke M. van der Zwaan, Inger Leemans, Erika Kuijpers, and Isa Maks. 2015. HEEM, a complex model for mining emotions in historical text. In *2015 IEEE 11th International Conference on E-Science*, pages 22–30. IEEE.

René van Stipriaan. 1996. *Leugens en vermaak: Boccaccio's novellen in de kluchtcultuur van de Nederlandse renaissance*. Amsterdam University Press.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.