

FAKULTÄT FÜR MATHEMATIK UND INFORMATIK

DIGITAL HUMANITIES

Module: Methods and Application in the Digital Humanities

Teacher: Dr.-Ing. Andreas Niekler

Leipzig, 31.03.2022

**TITEL DER ARBEIT**

|  |  |  |
| --- | --- | --- |
| Cecilia Graiff (3758332) | Richard Preußas | Fabian Strobel (3707349) |
| M. Sc. Digital Humanities | M. Sc. Digital Humanities | M. Sc. Digital Humanities |
| cg67qowo@studserv.uni-leipzig.de |  | fs92wybi@studserv.uni-leipzig.de |

**Introduction**

In the present work we aim to make use of the *DraCor[[1]](#footnote-2)* drama corpora by representing all of the italian and german plays currently present in the *ItaDraCor* and *GerDraCor* corpus respectively using a number of stylistic text features. We will then try to identify clusters of those plays, signifying a stylistic closeness among them. Finally, the goal is to compare these findings to those of traditional literary studies.

And with that we have already stated the main difference of our digital drama analysis approach in relation to the vast majority of recent research publications on the matter. If one takes a look at the papers linked on the dracor.org platform itself[[2]](#footnote-3) for example, they all seem to focus either on a specific subset of plays, say Russian five-act comedies (Wendell 2021) for instance, or select just a part of the dramatic text to analyze, like the stage directions in german drama (Trilke et al., 2020). Additionally, there is often an emphasis on using one specific distant reading method like social network analysis or topic modelling (Pavlova, Fischer 2018).

While there are obvious benefits to narrowing down either the number of texts to quantify or the methods to quantify and analyze them with, since a smaller subject matter can lend itself to a more detailed discussion of the chosen matter, such a decision also has its pitfalls.

As Estill (2019) lays out using the example of Shakespeare’s continuing dominance as a testing ground for all kinds of digital literary research methods, what we choose as the topic of our research can also reinforce predefined notions of the selected explanandum. It furthermore might influence following research to go in the same directions, as every publication offers an entry point for succeeding work. In other words: Most research is centered around a canonical set of texts (and perhaps some notions of what the place of these texts in literary history is).

We regard it as an important exercise to accompany that type of more specialized research with a macro-perspective approach that has the potential of finding relationships between texts that might have otherwise gone unnoticed or underappreciated. Therefore, as stated above, we will look at *all* the plays in the Italian and German corpus and compare them to each other based on a set of text features. We will use the k-means clustering algorithm to do that comparison. Then we will vary the feature input for k-means and analyze how that feature selection affects the outcome.

Ideally of course, we would use as many features as possible and apply them to as many texts as possible. But in order to not go beyond the possible scope of this work, we will use tf-idf data, POS data and a set of metadata on each play as their representing features. To be able to adequately interpret the results of the clustering, we will also limit ourselves to the Italian and German plays in the DraCor corpora collection.[[3]](#footnote-4)

Therein, as well as in the number of plays present in the mentioned corpora, lie some of the main limitations of our methodology, which are discussed in more detail in the following section.

**Methodology**

Data

At the core of the DraCor project is its documented API[[4]](#footnote-5), which offers scholars multiple easy ways to extract data for research purposes. It gives access to the raw textual data of the plays in its corpora, divided into spoken text, spoken text by character, stage directions, as well as metadata on the plays, characters, and the corpora themselves. It also features network and relational data for each play in various forms such as GraphML, GEXV and CSV. The corpus collection can easily be extended, since the only prerequisite is that the texts are TEI-encoded.

The data can be obtained either manually through the web interface or programmatically via API call. For our project, we use a script that makes the necessary API calls to access the metadata and text data, providing several options of varying that input (more on the specifics of the feature selection see the feature selection section below).

To give a quick summarized overview of our dataprocessing pipeline: First we make the API calls to obtain the raw text data and metadata for each text. That text data is then processed using natural language processing algorithms, so for each text we have tf-idf vectors and part of speech vectors alongside the metadata. All of those are stored in pandas dataframes. These data sets are subsequently passed on to the k-means clustering algorithm to find relationships between the texts in the corpus. Finally, the output of the clustering is used by visualization methods we have written to both evaluate the right value of k for k-means clustering (elbow, silhouette) and to finally use a scatterplot and an output of the top centroids to interpret the end result.

Before we delve deeper into our methodology, we shall also provide a quick overview of the corpora contents themselves.

The Italian corpus is the smaller one by far, featuring a total of just 139 plays, whereas the German corpus currently has 554 plays available. The Italian corpus however covers a significantly wider time span with 474 years (1449 - 1933), compared to the just 297-year span of the German plays (1650-1947). The histogram in Fig. 1 shows the distribution of plays over time with decade granularity, illustrating these differences between the two data sets. Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

*Figure 1: Distribution of plays per decade for German and Italian corpus*

As we can see, the density of the German corpus is much higher overall. It offers no information however about the time before 1650, and for the time prior to 1740 it features only plays by a single author, Andreas Gryphius. The Italian corpus is imbalanced as well, since it is heavily skewed towards the period around 1500 (Renaissance), the mid 1700s and around the year 1800. Past the early 1800s, only one play is present (‘Paola da Buti’ by Livio Cosci, 1933).

Another imbalance comes into play when we enquire about the distribution of authors in the corpora. As for the German corpus, the works of 199 unique authors are currently included, which computes an average of 2,78 plays per author. That number is quite similar to the 2,83 average plays per author for the Italian corpus (49 unique authors). Both corpora are skewed towards a small number of heavily represented authors, however. For the German corpus, unsurprisingly, Goethe’s works are most prominent (22 plays), followed by Scheerbart (20 plays) and Hofmannsthal (17 plays). Over half the authors (100) are represented by only a single play. The Italian corpus follows a similar trend. To get a sense of the most prominently featured authors see Figures 2 and 3, which show the 25 authors with the most plays in their respective corpus.

Chart, bar chart

Description automatically generated

Figure 2: Number of plays per author, Top 25 (GerDracor)

Chart, histogram

Description automatically generated

Figure 3: Number of plays per author, Top 25 (ItaDracor)

It is important to keep these things in mind when discussing and interpreting the clustering results, since the corpora compositions inevitably informs the possible conclusions one can draw from them.

Feature selection (main responsibility Fabian Strobel)

In order to cluster the plays we extract features regarding the metadata and the linguistic qualities of each text.

In regard to the metadata we utilize the pre-implemented metadata files provided by the DraCor API. We save their tables for metadata and use the columns *yearNormalized* that gives us information on the time the play was written in; *numOfSpeakers, numOfSpeakersFemale* and *numOfSpeakersMale* for information on the roles and their gender distribution and *wordCountText, wordCountSp* and *wordCountStage* to indicate the length of the play and its fraction of stage directions.

For the linguistic features we download the texts from the DraCor API. We enable to select for the full drama text or for the spoken text only, leaving out all the stage directions. We analyze the texts using *spacy* and its models *it\_core\_news\_lg* for Italian and *de\_core\_news\_sm* for German respectively. We gather relative counts of the POS-Tags in each play as a shallow representation of syntactic aspects of style. To cover the lexical aspects of style we compute TF-IDF over the vocabulary. We enabled a stopword filter using the *nltk* stopwordlists for German and Italian, however we found that this filter dramatically reduces the quality of our clustering and defaulted towards not filtering stopwords. We enabled lemmatization using *spacy* and found it always improving our clustering so we defaulted towards it.

We wanted to avoid TF-IDF being heavily tilted towards named entities like places or main characters that would in effect let plot design overshadow the true stilistic qualities of the plays. We achieved this goal by using a high cutoff number that makes the TF-IDF-count ignore all terms that do not appear in at least 20 (for DraCorGer) or 10 (for DraCorIta) plays. We manually inspected the highest TF-IDF values after the cutoff and found them to be sufficiently accounting for stilistic, not thematic words.

**Related Work**

Drama analysis has always played an important role in the realm of literary researches, even though the application of computational methods represents a relatively new approach. The study of related work therefore was not only useful to decide which approach would best suit our project, but also in motivating the use of quantitative methods for literary and drama analysis purposes. The poster *Katharsis – Ein Werkzeug für die quantitative Dramenanalyse*, written and presented by Burghardt et al. at the Forum CA3 2016 in Hamburg, argues that dramas present a structure that is apt to a quantitative analysis, due to the fact that they present a *dramatis personae*-inventary and a precise division in acts and scenes. This was also the case of this project, where the structure of drama allowed us e. g. to extract the spoken texts by removing the characters’ names, thus enabling an easier feature extraction and clustering.

Furthermore, the reading of the DraCor project documentation and the attending of the DraCor Workshop at the 8th annual convention of the DH in German speaking territories at the University of Potsdam gave us an insight into the possibility of computational and quantitative drama analysis methods, since the idea of a programmable corpus is *per se* an expression of a computational approach to literature.

If we restrict the field from the broad definition of “quantitative analysis” to machine learning, *Using Machine Learning for the Automated Classification of Stage Directions in TEI-Encoded Drama Corpora* by Frank Fisher and Daria Maximova provides an example for its application to drama analysis, even though its focus on stage directions marks a difference from our project.

However, the digital humanities approach to the field of quantitative text or drama analysis often suffers from the lack of a systematic nature; Therefore, it became important to read and carefully choose pertinent related work.

*Clustering with Sci-Kit Learn in Python*, published in *The Programming Historian* by Thomas Jurczyk, was one of the most meaningful readings. The author performs k-mean clustering on data and metadata regarding ancient Greek and Roman authors and abstracts from the journal *Religion*. His approach was an inspiration to this project, since he also adopts Elbow and Silhouette methods for the choice of cluster number.

*Analyzing Drama Networks with Machine Learning* by John R. Ladd, despite being an incomplete work, provides useful insights into the application of graph theory to drama network analysis. Like many other drama analysis case studies in the digital humanities, Ladd’s project focuses on Shakespeare’s plays.

On a more technical note, the [Python Data Science Handbook](http://shop.oreilly.com/product/0636920034919.do) by Jake VanderPlas was an useful guide to the clustering with Sci-Kit Learn. Moreover, several scientific articles, each regarding the specific tasks we had to perform, were taken into account for this project. These articles will be mentioned in the specific paragraphs.

The study of related work helped us realize that our project fills a gap in the previously conducted research. European dramas have been investigated with quantitative and more precisely also with machine learning methods, and k-means clustering has been performed on textual and literary data, but the application of k-means clustering to drama texts represent a new step in the direction of computational analysis of theatrical works.

**Clustering (general intro), Elbow and silhouette – draft**

Clustering algorithms belong to the realm of unsupervised machine learning models and, as their name itself suggests, aim to reach an optimal division of data points in groups (“clusters”) by analyzing their properties.

The algorithm chosen for this project was k-Means clustering. [explain why k-means].

This algorithm is based of an understanding of the “optimal cluster” as a cluster which center is the arithmetic mean of all its data points and where each point is closer to its own cluster center than to other cluster centers.

## Notable about this algorithm is the necessity to give a pre-determined number of clusters as input. The question of how to choose this number thus arises. In order to answer it, it was decided to perform two pre-analyses of the dataset: Elbow plot and silhouette plot. This idea mostly came from reading the article “Clustering with Sci-Kit Learn in Python” by Thomas Jurczyk.

**Elbow Plot**

The Elbow method is a heuristic technique which aims at choosing a number of plot from a data set based on the elbow of the curve in a representation in form of a graph.

In order to understand the process and analyze the results of elbow plots, it was necessary to familiarize with the concept of inertia, which in the above mentioned article by Jurczyk is defined as “the sum of squared distances of samples to their closest cluster center”. The inertia decreases with the number of clusters, and from a certain n onwards it decreases more slowly, thus forming a line plot which resembles an arm. The point of inflection in the curve, also known as the “elbow” of the arm-looking line plot, represents the ideal number of clusters. This happens because the elbow reflects the point of over-fitting: before the elbow the inertia gives significant information about the dataset and its variation, whereas after the elbow these values are not significant anymore and therefore a bigger cluster number is not necessary.

**Silhouette Plot**

# Another measure for finding the right amount of cluster is the silhouette method. The reading of related work, like Jurcyzk’s above mentioned article, but also the sci-kit learn guide on silhouette analysis and the article *Silhouette Coefficient: Validating clustering techniques* by Bhardwaj, published in Towards Data Science, provided us with useful insights into the topic.

# The sci-kit learn official use guide states that “silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually” (cite). This was measured with silhouette coefficients in a range between -1 and +1.

# The first step was to calculate these coefficients for our dataset. Once the coefficients had been computed, the second step was to visualize the obtained results as plots. A more accurate description of this process can be found in the code documentation.

# The outputs of silhouette calculus were then analyzed. Silhouette coefficients around +1 indicate that the sample is relatively far away from the neighbor clusters, whereas coefficient values near -1 indicate that the data point has been assigned to the wrong cluster.

**k-means clustering**

The sci-kit learn package in Python provides an implementation of k-means. In this case the reading of sci-kit learn’s documentation was fundamental to the further developing of the project. The parameters of the k-means algorithm in sci-kit learn allowed an experimentation with different inputs. Although an accurate documentation of the chosen parameters can be found in the source code of the project, some of the choices of parameters are worth mentioning. For example, random\_state accepts as values int, RandomState instance or None (which is also the default value). The sci-kit learn documentation states that “Whenever randomization is part of a Scikit-learn algorithm, a random\_state parameter may be provided to control the random number generator used.” The passed value should have an effect on the reproducibility of the results returned by the function. The most common values are integers between 1 and 42. It was therefore decided to confront the outputs for random\_state = None, random\_state = 1 and random\_state = 42.

**Visualization**

One of our goals for this project was to visualize the results. All visualizations have been done using the python libraries ‘matplotlib’ and ‘seaborn’, which has been built based on matplotlib. The main advantage of seaborn compared to plain matplotlib is that it has been designed to easily work with pandas dataframes, which is our main data structure.

Our script ‘visualization.py’ instantiates an object of our custom visualization class, which lets us modify the text input (spoken text or full text, German or Italian corpus, with or without stopwords, to name a few). It then constructs one pandas dataframe for each of the feature domains tf-idf, POS and metadata, as well as one dataframe that holds all feature vectors at once. One can choose to visualize each of those four dataframes, so we are able to compare how the visualization outputs varies for different feature inputs.

There are four different types of visualizations implemented in our visualization class: metadata\_plot() for plotting the raw metadata (see Fig. 1-3), elbow\_plot() to draw an elbow plot, silhouette\_plot() to draw the silhouette plot alongside the scatterplot of the clustering (TODO: see Fig. if plot has been used in report) and cluster\_scatterplot() to draw just the scatterplot of the clustering (TODO: see Fig. if plot has been used in report).

DraCorGer

tf-idf clusters the German data into five big clusters with the fifth cluster representing different outliers to the main point cloud of the other clusters. POS has four clusters with less density and less outliers. If POS, metadata and tf-idf are combined, the density of the biggest part of our data gets bigger, while the outliers get more extreme. Also in this case the best clustering can be obtained counting five clusters, with the last cluster containing the outliers.

Our most prevalent centroids of those clusters are the three wordCount-features as well as the *yearNormalized*-feature. The non-metadata-centroids of our clustering are function words, the two most important being “ich” (“I”) and “der” (“the”). Sometimes the centroids contain modal verbs [“haben” (“have”), “werden” (“become”)] and one cluster has a dialect expression as centroid [“nich” (“not”)]. The POS centroids are PUNCT followed by NOUN, PRON and ADV.

Our five clusters are not distributed evenly, three clusters are big and close to equal among themselves (181, 145 and 143 plays), one is about half that size (73 plays) and one is very small (11 plays). The three big clusters contain the dense data point cloud of the data, the smallest cluster contains the outlier and the medium cluster contains the plays that are in between the main cloud and the outliers.

When looking at the three big clusters, they are by no means “clean” and easily interpretable: A lot of authors are spread across them, they cannot be easily pinpointed by time to any dramatic epoch. However, when looking at the unique authors in every cluster, there is a difference: Cluster 0 contains many unique authors associated with an expressionist style: Hasenclever, Barlach and Borchert. Cluster 2 contains unique authors associated with a romantic or neoclassical style: Rilke, Günderode and Kraus. It also contains the dramatic works by the expressionist Heym, but this even strengthens the distinction as Heym is known to be an expressionist who nevertheless alludes to a classical style. Cluster 3 is mixed between them, it contains works of expressionists (Lasker-Schüler, Sternheim) as well as romantic authors (Voß, Uhland, Droste-Hülshoff).

Thie two outlier clusters are the ones with considerably bigger text length. The smallest cluster contains all the extremely long plays, most notably Goethes *Faust II*, and the other cluster with the texts between the dense data cloud and the extreme outliers also has considerably higher average word count. These clusters cannot be interpreted from a stilometric standpoint, they are skewed by their metadata features.

**Analysis of Ita corpus**

This paragraph will now confront the output for a similar number of clusters (5) for tf-idf, POS, and their combination. Due to clarity purpose, it will not be possible to describe in detail every single cluster; This section of the project contains an overview of the outputs considered more important and useful to this project after a selection performed during the analysis phase.

Tf-idf for 5 clusters divides the data in clusters of various dimension: Cluster 0 contains 29 plays, Cluster 1 24, Cluster 2 64, Cluster 3 6, Cluster 4 16. Apart from Cluster 0 and Cluster 3, which in the visualization partially overlap, these clusters are quite separated from one another.

POS also presents clusters of various dimension, ranging from the 47 plays of Cluster 0 to the 13 plays of Cluster 4. Like tf-idf, also POS divides the data quite precisely, since the clusters do not overlap; However, the criteria of clustering will be further discussed, since in some cases authors who are considered to be a lot different in style are brought together and/or different plays of the same author are spread among the clusters. The neat division of clusters is *per se* no accuracy indicator.

The top 10 centroids are in the case of tf-idf some of the most common words of the Italian language, like “il”, “the”, or “essere”, “to be”; also some dialectal expressions (“el”, dialect for “il”) and Latin influences (“et”, Latin for “e”, “and”; quite used in old Italian) are present. For POS, the first centroid is PUNCT, followed by NOUN and VERB.

It is also worth mentioning that among the top 20 Centroids for all\_features also metadata features are present, like YearNormalized, or WordCountText; This is also by the German corpus the case.

POS and tf-idf together present a partial overlapping, but the clusters are still easily recognizable. The subdivision of authors is also worth mentioning. Cluster 0 mixes anonymous plays from the Renaissance era with some plays by Pietro Metastasio (18th century) and other Renaissance plays; Even one play by Giacomo Leopardi is present. The interpretation of these results is not clear at all, since the authors belong to the most different styles. Cluster 1 is marked by a relevant presence of Ariosto and Goldoni, two authors who no one would dare to pull close. Cluster 2 only presents one play by Torquato Tasso, and Cluster 3 also presents a curious mix of Machiavelli, Tasso, and Metastasio. Cluster 4 contains plays from the 17th century, apart from Guarini’s *La idropica*, which dates to the previous century.

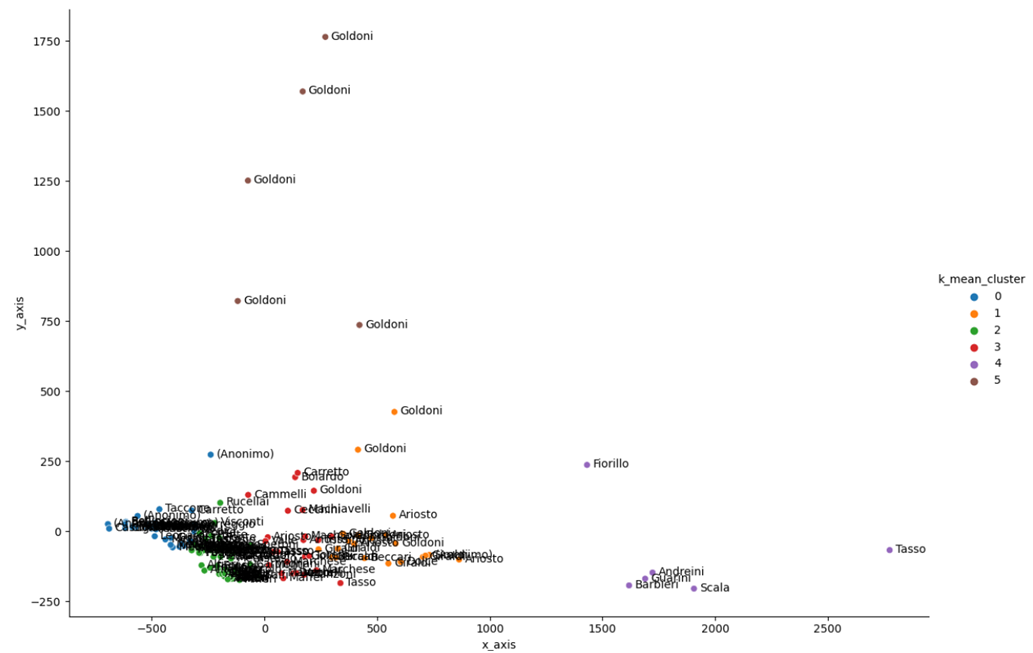
These fact that plays from very different eras are brought together in the same cluster was also made clear by the computation of the average value for the feature yearNormalized, which (unsurprisingly) confirmed this difference, whereas by the German corpus the yearNormalized values were quite similar.

This clustering clearly presents some aspects which are difficult to explain. In order to understand the logic of the clusters, we also calculated the average values for WordCount, one of the above mentioned metadata features among the top 20 centroids.

* + cluster 0: wordCount 4339,3
  + cluster 1: wordCount 19666,8
  + cluster 2: wordCount 49229,0 (only one element)
  + cluster 3: wordCount 11238,6
  + cluster 4: wordCount 38753,0

The WordCount values are also of great interest, because they show that in the case of this clustering the presence of very long plays has brought to a false analysis. In fact, cluster 4, which contains 5 plays, represents an outlier, and so does cluster 2, that only contains one (very long) play. Therefore, some normally very useful features like word count, POS tags, and tf-idf can also be invalidated from a data set which is too “extreme”, like in the case of very long texts.

One of the most interesting results regard Carlo Goldoni (Venezia, 1707 – Paris, 1793). For example, the tf-idf feature analysis of the spoken test (spoken\_min\_df=10\_cluster=6) shows a main cloud in which many Goldoni-scatter points are not included. These scatter points mostly come from Cluster 5, which contains *I Rusteghi*, *Il Campiello*, *Una delle ultime sere di Carnovale*, *Il servitore di due padroni* and *Le baruffe chiozzotte*. These comedies are all written in Venetian dialect (either the whole text or the most part of it), whereas other plays by Goldoni, like *La Locandiera* or *La bottega del caffè* (both in cluster 1) are written in Italian. This separation of Goldoni’s plays according to their language also recurs in other visualizations, such as the tf-idf feature analysis for the full text and the POS clusters. An exception by the POS-clusters is due to the fact that *Il servitore di due padroni* also presents many parts in Italian, alongside with characters who only speak Venetian dialect.



*Cluster plot for just tf-idf features, spoken text, 6 clusters (spoken\_min\_df=10\_cluster=6).*

Nevertheless, the obtained clusters present some characteristics which are difficult to explain. As an example, it could easily be mentioned that a motivation for Tasso and Foscolo belonging to the same cluster is quite difficult to find, and the same goes for Ariosto and Leopardi. However, the presence of Ariosto and Giraldi Cinzio, both active in the first half of the 16th century and of similar style, or Metastasio and Da Ponte, both writers of the 18th century and moreover both authors of *librettis* of Mozart’s operas, certainly are significant and remarkable. Also the association of Ludovico Ariosto and Torquato Tasso, two different authors yet with some interesting common characteristics, representatives of Italian Renaissance and Mannerism, is of greatest interest.

Another interesting observation could be that in many cluster outputs the authors that are left outside the high-density area mostly are renowned for the uniqueness of their style, like the above mentioned Carlo Goldoni, Ugo Foscolo, and Torquato Tasso.

Despite being far from perfect, these clusters thus present many interesting features. A stylometric analysis of dramas certainly goes in the right direction; the difficulty stands in orientating and further developing the quantitative analysis by chosing the most meaningful features and parameters – also based on its results, as a “work in progress” analysis - in order to improve the qualitative analysis.

Overall Evaluation

In the beginning we were sceptic how TF-IDF would perform as a stilometric feature. Would it really capture stylistic features or just enforce thematic clustering? However, we found out that a high cutoff allows TF-IDF to capture the stylistic qualities of the plays over the thematic ones, mainly in accounting for the different distributions of very frequent words. This finding is reinforced by the fact that removing stopwords reduced the quality of our clusterings.

The POS features as well as the metadata features helped us to further improve the clustering. Measuring global TF-IDF meant that the number of lexical features is much higher than the number of syntactical or metadata features, however those features had significant impact.

We obtain the most expressive clusterings by regarding all those features together, this makes us confident in our chosen approach. Since we are the first to choose such an approach, there is no baseline that we can evaluate our findings against. However, the fact that the centroids for the clusters contain the *yearNormalized* metadata feature helps us evaluate our approach against the one of Literary Studies: They often “cluster together” different texts regarding to their epoch, claiming correlation between time and style, our clusterings come to the same conclusion.

Shortcomings / Additional Research

It would be interesting to have metadata on the speech acts (e.g. their count, their average length, the average count per person) available, however the DraCor API does not provide those numbers and we did not compute them.

?? The Italian drama corpus did cluster a lot better than the German one and the clusters were more expressive from a stylometric point of view. Probably our approach is better with smaller data sets and one could test this hypothesis with subsets of the German corpus.

Furthermore, one could think of formulating our approach the other way round: Could one write classifiers that reliably tell apart dramas with different stylistic qualities? However, the classes of such classifiers would need to be predefined.

Lessons learned  
   
Our approach to the whole project did not change significantly during the time of our work, however the approach was not detailed enough in hindsight. It would have been useful to communicate the data structures needed for all our steps in advance. We did not do so and had to refactor our code a lot because of that, the main point being the conversion of all our data in pandas dataframes.

We also struggled making our code reproducable amongst ourselves because we failed to communicate a fixed set of python-package versions and system requirements in advance.

It would have been useful to find a point where we can store all our features so that we can operate without being dependent on the DraCor API and the whole feature extraction algorithms rerunning every time we try out new visualizations. Our parameters and filters are applied at different stages of the code, this made it impossible to realize this utility.

LINK TO OUR CODE  
  
Main Responsibilities

|  |  |
| --- | --- |
| Feature Extraction | Fabian Strobel |
| Data Handling | Richard Prußas |
| Clustering | Cecilia Graiff |
| Elbow | Cecilia Graiff |
| Silhouette | Cecilia Graiff |
| Visualization | Richard Prußas |
| Analysis ITA | Cecilia Graiff |
| Analysis GER | Fabian Strobel |
| Report | all |
| Code / git | all |

**SOURCES**

Wilhelm, T., Burghardt, M., & Wolff, C. (2013) "To See or Not to See" - An Interactive Tool for the Visualization and Analysis of Shakespeare Plays. In *Kultur und Informatik: Visual Worlds & Interactive Spaces*. Glückstadt, Verlag Werner Hülsbusch, Franken-Wendelstorf, Regina and Lindinger, Elisabeth and Sieck, Jürgen, (eds.). Pp. 175-185. ISBN 978-3-86488-045-2.

Murrieta-Flores, P., Donaldson, C. & Gregory, I. (2017). GIS and literary history: Advancing digital humanities research through the spatial analysis of historical travel writing and topographical literature. Digital Humanities Quarterly. 11 (1).

Pagel, J. & Reiter, N. (2020). GerDraCor-Coref: A Coreference Corpus for Dramatic Texts in German. In *Proceedings of the 12th Language Resources and Evaluation Conference* (LREC), Marseille, France.

Rzepka, A., Williams, P., & Royston, J. (2017). The Social Network of Early English Drama: A Digital Humanities Lesson Plan, in *The Emerging Learning Design Journal*: Vol. 5 : Iss. 1 , Article 4.

Ladd, J. R. (2019). *Analyzing Drama Networks with Machine Learning*

Burghardt, M., Dennerlein, K., Schmidt, T., Mühlenfeld, J., & Wolff, C. (2016). *Katharsis – Ein Werkzeug für die quantitative Dramenanalyse.* CLARIN-D Forum CA3, 7.-8. Juni 2016, Hamburg.

Wendell, I. (2021). A Statistical Analysis of Genre Dynamics: Evolution of the Russian Five-Act Comedy in Verse in the Eighteenth and Nineteenth Centuries. *UCLA*. ProQuest ID: Wendell\_ucla\_0031D\_19638. Merritt ID: ark:/13030/m51c7rt0

<https://escholarship.org/uc/item/9rr5k9p7>

Peer Trilcke, Christopher Kittel, Nils Reiter, Daria Maximova, Frank Fischer: [Opening the Stage: A Quantitative Look at Stage Directions in German Drama.](https://dh2020.adho.org/wp-content/uploads/2020/07/337_OpeningtheStageAQuantitativeLookatStageDirectionsinGermanDrama.html) In: DH2020: »carrefours/intersections«. 22–24 July 2020. Book of Abstracts. University of Ottawa.

<https://dh2020.adho.org/wp-content/uploads/2020/07/337_OpeningtheStageAQuantitativeLookatStageDirectionsinGermanDrama.html>

Irina Pavlova, Frank Fischer: [Topic Modeling 200 Years of Russian Drama.](https://eadh2018.exordo.com/files/papers/158/final_draft/Pavlova___Fischer_-_Topic_Modeling_-_EADH_conference.pdf) EADH2018: »Data in Digital Humanities«. 7–9 December 2018. National University of Ireland, Galway.

<https://eadh2018.exordo.com/files/papers/158/final_draft/Pavlova___Fischer_-_Topic_Modeling_-_EADH_conference.pdf>

**BIBLIOGRAPHY NOTES**

* **Bibliography tips by Niekler**

The Stylometry of Film Dialogue: Pros and Pitfalls, in DHQ: Digital Humanities Quarterly 2020, Volume 14, Number 4

[*http://www.digitalhumanities.org/dhq/vol/14/4/000498/000498.html*](http://www.digitalhumanities.org/dhq/vol/14/4/000498/000498.html)

Mainly sentiment analysis, not really useful for us; nevertheless, some mentions of stylometry as a form of preprocessing could be useful (e.g. page 3

Wilhelm, Thomas, Burghardt, Manuel and Wolff, Christian (2013) "To See or Not to See" - An Interactive Tool for the Visualization and Analysis of Shakespeare Plays. In: Franken-Wendelstorf, Regina and Lindinger, Elisabeth and Sieck, Jürgen, (eds.) Kultur und Informatik: Visual Worlds & Interactive Spaces. Verlag Werner Hülsbusch, Glückstadt, pp. 175-185. ISBN 978-3-86488-045-2.

[*https://epub.uni-regensburg.de/28417/1/KuI\_2013\_VisualShakespeare.pdf*](https://epub.uni-regensburg.de/28417/1/KuI_2013_VisualShakespeare.pdf)

This paper is not really useful to us. The authors used XML and TEI annotation ==> technological methods quite different from the ones we chose. Maybe we could mention this paper in comparison to DraCor (if we want to cite similar projects/tools).

Maybe we could briefly mention that they focused on annotated plays with the tags provided by TEI (e.g. speech, dramatis personae, gender, stage directions, etc), because nearly every approach to drama texts consists of extrapolating these elements (and ours is no exception). However, I do not know if this idea makes any sense at all ;)

Pagel, Janis, Reiter, Niels (2020) GerDraCor-Coref: A Coreference Corpus for Dramatic Texts in German. In Proceedings of the 12th Language Resources and Evaluation Conference (LREC), Marseille, France.

[*http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.7.pdf*](http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.7.pdf)

Contains a lot of statistics, does not really concern us.

Murrieta-Flores, P., Donaldson, C. & Gregory, I. (2017). GIS and literary history: Advancing digital humanities research through the spatial analysis of historical travel writing and topographical literature. Digital Humanities Quarterly. 11 (1).

[*https://chesterrep.openrepository.com/handle/10034/620256*](https://chesterrep.openrepository.com/handle/10034/620256)

I did not read this paper carefully, because it is quite long and does not really concern us. It contains spatial analysis methods and I do not think it can be useful to us.

Rzepka, Adam; Williams, Pierce; and Royston, Jennifer (2017) "The Social Network of Early English Drama: A Digital Humanities Lesson Plan," The Emerging Learning Design Journal: Vol. 5 : Iss. 1 , Article 4.

[*https://digitalcommons.montclair.edu/eldj/vol5/iss1/4*](https://digitalcommons.montclair.edu/eldj/vol5/iss1/4)

I really do not see anything useful here.

Henning, Urs: Dramenanalyse mit DraCor

<https://web2-unterricht.ch/uncategorized/dramenanalyse-mit-dracor/>

It focuses on distant reading and other methods which are not really useful to us; However, we could cite it when writing about APIs and in the general presentation of DraCor.

Ladd, John R. (2019). Analyzing Drama Networks with Machine Learning

[*https://jrladd.com/ach.html*](https://jrladd.com/ach.html)

This paper is actually interesting and provides explanations and interpretations regarding k-means clustering and clustering in general. At the beginning of the page stands “Work in progress, please do not cite or circulate”; However, the link is quite old and I would cite it (Niekler sent it to us, so I think he will be ok with us citing it).

Manuel Burghardt, Katrin Dennerlein, Thomas Schmidt, Johanna Mühlenfeld & Christian Wolff (2016). Katharsis – Ein Werkzeug für die quantitative Dramenanalyse. CLARIN-D Forum CA3, 7.-8. Juni 2016, Hamburg.

<https://dhregensburg.files.wordpress.com/2016/06/2016_katharsis-ca3-abstract.pdf>

<https://dhregensburg.wordpress.com/2016/06/06/katharsis-ein-werkzeug-fuer-die-quantitative-dramenanalyse/>

I would cite this article as an example of text analysis of dramas. However, it does not involve clustering or stylometric methods similar to ours.

Estill, Laura. 2019. "Digital Humanities’ Shakespeare Problem" *Humanities* 8, no. 1: 45. https://doi.org/10.3390/h8010045

<https://www.mdpi.com/2076-0787/8/1/45/html>

I would cite this article as an example of drama analysis and more broadly of dh approaches to literature. I would also generally mention the Shakespeare problem as an example of wide application of dh methods to literature and dramas, because many articles I read regard this problem.

Wendell, I. (2021). A Statistical Analysis of Genre Dynamics: Evolution of the Russian Five-Act Comedy in Verse in the Eighteenth and Nineteenth Centuries. *UCLA*. ProQuest ID: Wendell\_ucla\_0031D\_19638. Merritt ID: ark:/13030/m51c7rt0

<https://escholarship.org/uc/item/9rr5k9p7>

Peer Trilcke, Christopher Kittel, Nils Reiter, Daria Maximova, Frank Fischer: [Opening the Stage: A Quantitative Look at Stage Directions in German Drama.](https://dh2020.adho.org/wp-content/uploads/2020/07/337_OpeningtheStageAQuantitativeLookatStageDirectionsinGermanDrama.html) In: DH2020: »carrefours/intersections«. 22–24 July 2020. Book of Abstracts. University of Ottawa.

<https://dh2020.adho.org/wp-content/uploads/2020/07/337_OpeningtheStageAQuantitativeLookatStageDirectionsinGermanDrama.html>

Irina Pavlova, Frank Fischer: [Topic Modeling 200 Years of Russian Drama.](https://eadh2018.exordo.com/files/papers/158/final_draft/Pavlova___Fischer_-_Topic_Modeling_-_EADH_conference.pdf) EADH2018: »Data in Digital Humanities«. 7–9 December 2018. National University of Ireland, Galway.

<https://eadh2018.exordo.com/files/papers/158/final_draft/Pavlova___Fischer_-_Topic_Modeling_-_EADH_conference.pdf>

* **DraCor Workshop Bibliography**

<https://lehkost.github.io/slides/2022-03-08-potsdam-dhd/index.html>

<https://www.fu-berlin.de/sites/dhc/programme/termine/dh-gespraech-sose-22-2.html>

Introduction to DraCor (Prof. Fischer is one of its creators).

I would definitely cite his works.

API Dokumentation: <https://dracor.org/doc/api>

https://dh-abstracts.library.cmu.edu/works/9656

1. For an introduction on the DraCor project see: Fischer, Frank, et al. (2019). Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. In: Proceedings of DH2019: "Complexities", Utrecht University, [doi:10.5281/zenodo.4284002](https://doi.org/10.5281/zenodo.4284002). [↑](#footnote-ref-2)
2. dracor.org/doc/research [↑](#footnote-ref-3)
3. That is simply for the practical reason that our group consists of Italian and German native speakers. [↑](#footnote-ref-4)
4. See: https://dracor.org/doc/api [↑](#footnote-ref-5)