**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 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 [TODO] 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.

Another argument in favour of the reasonability of a quantitative analysis of dramas could be the fact the so-called “Shakespeare problem”, that is the question of the disputed attribution of some Shakesperean works, together with the analysis of Shakesperean text reuse, was addressed by many DH-researchers, as papers like *Digital Humanities’ Shakespeare Problem* by Laura Estill demonstrate.

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. (eventually add more “technical” readings?)

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).

Notes on Ita clustering

Telesilla = tragedia pastorale

Cluster 0

Bellincioni: egloga ovvero pasturale (1493)

: ripresentazione di Pavia (1492)

Cascina: Alfea reverente (1639)

Cosci: Paola da Buti (1933)

Da Correggio: fabula de Cefalo (1487)

Dal Carretto: comedia de Timon greco (1497)

De Medici: Rappresentazione di San Giovanni e Paolo (1491)

Gigli: la dirindina (1715)

Guidi: Endimione (1691)

Leopardi: Telesilla (1819)

Metastasio: Achille in Sciro (1736)

: Alessandro nell’Indie (1730)

: il re pastore (1751)

: Ipermestra (1744)

: Issipile (1732)

: l’eroe cinese (1752)

: l’impresario delle Canarie (1724)

: Romolo ed Ersilia (1765)

Mariano Muzi: La rappresentazione del vitello sagginato (1449)

Poliziano: Orfeo (1480)

**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.

Gao, J, Nyhan, J. Visualising The Digital Humanities Community: A Comparison Study Between Citation Network And Social Network

<https://discovery.ucl.ac.uk/id/eprint/10051991/1/Gao_dh2018.pdf>

I actually found this paper a little confusing... Could someone have a second look at it? I actually would not cite it (not even in the introduction part), but I may be wrong.

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)