Introduction draft

The DraCor corpus offers lots of dramatic works, however computational analysis has mostly focussed on distant-reading methods for single plays. In this paper we explore methods to compare all oft he plays of italian and german corpus respectively, in regard to their stylometric features. We use TF-IDF, POS data and metadata to create those features and analyze them with k-means clustering. The results are compared to the findings of literary studies.

Data section (copied from Exposee)

DraCor, an open platform for European drama corpora.

At the core of the DraCor project is its documented API, 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.

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.

All those features are aggregated in a *pandas* dataframe for the clustering.

**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 divison 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, which performs k-mean clustering on data and metadata regarding ancient Greek and Roman authors and abstracts from the journal *Religion*.

**Elbow Plot**

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”.

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.

* **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