DraCor – Cecilia Graiff, Richard Prußas, Fabian Strobel
Exposé

Using stylometry to represent whole dramas, visualising their respective similarities and differences, comparing the visualisation to conceptualizations from comparative literature analysis like genre, epoch, authorship, plot etc.

The goal of our project is to perform a stylometric analysis of the plays present in 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.

Dracor provides data on plays in several different European languages such as French, Russian, Swedish, Greek, and more. Due to the fact that our group consists of German and Italian native speakers, our analyses will focus on the German and the Italian drama corpora, which contain 553 and 139 plays respectively.

Our aim in working with DraCor is more precisely to graphically visualise the dramatic works by representing their respective similarities and differences in style. Therefore, the performed analyses will be stylometric. We will then proceed to compare the thus obtained results with those of the traditional literary studies, which are normally based on concepts like genre, epoch, authorship, plot, etc., to see whether these corpora sufficiently represent the literary history of plays. The analysis of the results of the applied stylometric methods should thus provide information on the accuracy and reliability of such methods for drama analysis purposes. It also could give an insight into how accurate the grouping of sets of plays and authors into literary periods, a core task of literary studies, really is.

While we choose to look at both the German and Italian corpus from a macro-perspective – by finding a singular representation for each play and then cross-comparing all plays of those corpora to each other – we would then also like to take a closer look at a few selected cases or subsets that stand out in some way. Are there maybe plays/authors that stylistically go against the dominant contemporary style? Is there an author that produced vastly different plays? However, it is hard to predict at this stage of our project whether this step will indeed be needed, since it depends on our selection of textual properties to analyse. That selection will appear more clearly as the project proceeds.

We will take a pipeline approach to this project:

First we compile different representations using TF-IDF. We will later on only keep the most expressive of those representations. We will consider the full drama texts vs. only its spoken parts, normalised orthography vs. historic orthography, lemmatized texts vs. unlemmatized texts, versions without the dramatis personae vs. leaving them in, cleaning stopwords vs. leaving them in. In case there is extra time we also want to represent the TF-IDF of the

POS-Tags to include syntactic information in our modelling. Other additional features could be the average sentence length or the length of the whole drama.

In the next step of the pipeline we will cluster these representations with K-Means. An important property of this algorithm is that it searches for a predefined number of clusters. In the case of literary analysis, this could be useful when it comes to confirming or contradicting prior existing classifications. However, it could also be a downside, because the given number of clusters could be considered as a bias and therefore negatively affect the reliability of the results. This aspect will be taken into account while developing the project.

The results of the clustering will then be graphically displayed in the third step of the pipeline. We will use R-Shiny to visualise our findings; we will also consider additional metrics that can give us further insight into our data.

Once these methods have been applied, we will focus on interpreting and comparing the thus obtained results.

Timetable and Responsibilities:

Issue	**Responsibility**	**Deadline**
Representations	Fabian Strobel	04.02.2022
Clustering	Cecilia Graiff	21.02.2022
Visualisation	Richard Prußas	06.03.2022
Project Report	all	15.03.2022

Report, Git, Bibliography etc. will be handled by the group; each of us will focus on their respective responsibilities. The interpretation of our findings will be a group effort as well. Due to our respective linguistic and cultural background, the most constructive approach seems to divide the analyses: Fabian and Richard will take care of the analysis of the German corpus, whereas Cecilia will be in charge of analysing the Italian corpus. We will all help each other whenever this is needed.

Preliminary Bibliography:

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

Quantitative drama analysis/drametrics

Blessing, Andre / Bockwinkel, Peggy / Reiter, Nils / Willand, Marcus (2016): „Dramenwerkbank: Automatische Sprachverarbeitung zur Analyse von Figurenrede“, in: DHd 2016: Modellierung - Vernetzung - Visualisierung 281–284 http://dhd2016.de/boa.pdf

Krautter, B. (2018). Quantitative microanalysis? Different methods of digital drama analysis in comparison. Book of Abstracts, DH 2018 . Mexico-City, Mexico, pp. 225-228.

Romanska, M. (2015). Drametrics: what dramaturgs should learn from mathematicians. In Romanska, M. (ed.), The Routledge Companion to Dramaturgy. Routledge, pp. 472-481

Schmidt, T., Burghardt, M., Dennerlein, K. & Wolff, C. (2019). Katharsis - A Tool for Computational Drametrics. In: Book of Abstracts, Digital Humanities Conference 2019 (DH 2019). Utrecht, Netherlands.

Speech lenght statistics

Ilsemann, H. (2005). Some statistical observations on speech lengths in Shakespeare’s plays. Shakespeare Jahrbuch, 141: 158–68.

Ilsemann, H. (2008). More statistical observations on speech lengths in Shakespeare’s plays. Literary and Linguistic Computing, 23(4): 397-407.

Visualization

Wilhelm, T., Burghardt, M., and Wolff, C. (2013). “To See or Not to See” - An Interactive Tool for the Visualization and Analysis of Shakespeare Plays. In R. Franken-Wendelstorf, E. Lindinger, and J. Sieck (Eds.), Kultur und Informatik: Visual Worlds & Interactive Spaces . Glückstadt: Verlag Werner Hülsbusch, pp. 175–185.

R

Eder, M., Rybicki, J. and Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. R Journal 8(1): 107-121. https://journal.r-project.org/archive/2016/RJ-2016-007/index.html

Methods

Kent K. Chang and Simon DeDeo (2020): Divergence and the Complexity of Difference in Text and Culture. In: Journal of Cultural Analytics 4.11 (2020): 1-36. doi: 10.22148/001c.17585

Jurczyk, T.(2021). Clustering with Scikit-Learn in Python. Programming Historian, (10):

The goal of our project is to perform a comparative analysis of the plays present in 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 relation data for each play in various forms such as GraphML, GEXV and CSV.

Dracor provides data on plays in several different European languages such as French, Russian, Swedish, Greek and more. Due to the fact that our group consists of german and italian native speakers, our analyses will focus on the German and the Italian drama corpora, which contain 553 and 139 plays respectively.

Our aim in working with DraCor is more precisely to graphically visualize the dramatic works by representing their respective similarities and differences in style and theme. Therefore, the performed analyses will mostly be stylometric. We will then proceed to compare the thus obtained results with those of the traditional literary studies, which are normally based on concepts like genre, epoch, authorship, plot, etc. to see whether these corpora sufficiently represent the literary history of plays. The analysis of the results of the applied stylometric methods should thus provide information on the accuracy and reliability of such methods for drama analysis purposes. It also could give an insight into how accurate the grouping of sets of plays and authors into literary periods, a core task of literary studies, really is.

While we choose to look at both the German and Italian corpus from a macro-perspective - by finding a singular representation for each play and then cross-comparing all plays of those corpora to each other - we would then also like to take a closer look at a few select cases that stand out in some way. Are there maybe plays/authors that stylistically go against the dominant contemporary style? Is there an author that produced vastly different plays?

We shouldn't fail to mention that a further division of the corpora into subcorpora grouped by epoch, author, etc., might turn out to be necessary in order to perform a more precise analysis. However, it is hard to predict at this stage of our project whether this step will indeed be needed, since it depends on our selection of textual properties to analyze. That selection will appear more clearly as the project proceeds.

In order to perform the stylometric analysis of our two selected corpora, each of us will approach our subject using a different distant reading methodology.

Fabian will analyze the texts using Tf-idf, …,...,... [insert description Fabian].

Cecilia, on the other hand, has decided to perform a k-means clustering of the given corpora. Useful sources for designing the approach were an article by T. Jurczyk on The Programming Historian and the Python Data Science Handbook by Jake Van der Plas.

[Clustering algorithms aim for an optimal division or discrete labeling of groups of points according to the properties of data. Among those, the k-means clustering algorithm is implemented in SciKit-Learn, a widely applied and well-documented machine learning framework in Python.]

An important property of this algorithm is that it searches for a predefined number of clusters. In the case of literary analysis, this could be useful when it comes to confirming or contradicting prior existing classifications. However, it could also be a downside, because the given number of clusters could be considered as a bias and therefore negatively affect the reliability of the results. This aspect will be taken into account while developing the project.

The results of the clustering will then be graphically displayed.

Richard will perform an LDA topic modelling of the plays in the german and the italian corpus to find thematic similarities and differences among the plays over time.

He will visualize all our findings and integrate them in a small web application created using R-Shiny, which will be permanently hosted online for others to explore.

Once these methods have been applied, we will focus on interpreting and comparing the thus obtained results. Due to our respective linguistic and cultural background, the most constructive approach seems to be to divide the analyses: Fabian and Richard will take care of the analysis of the German corpus, whereas Cecilia will be in charge of analyzing the Italian corpus.