Giovannini, Luca, and Daniil Skorinkin (2023). "Computational approaches to opera libretti: An experiment on DraCor corpora". Preprint.
Submitted to the *Journal of Computational Literary Studies.*

Article

Computational approaches to opera libretti

An experiment on DraCor corpora

Luca Giovannini [1]
Daniil Skorinkin [2]

1. Institute for German Studies, University of Potsdam, Potsdam, Germany.
2. Digital Humanities Network, University of Potsdam, Potsdam, Germany.

Abstract. The paper offers a first computationally-informed look at German and French opera libretti by modelling them on the basis of their structural features. On one side, it strives to assess whether libretti – a relatively recent genre, born in the early 1600s – exhibit peculiar formal properties which set them apart from contemporary comedies and tragedies from the same linguistic environment. On the other side, it explores the structural development of the genre across history, evaluating how its relationship with the Aristotelian genres changed in different periods. Results confirm known challenges of modelling dramatic texts while pointing out to some structural patterns which help identify libretti, particularly non-comic ones.

1. Introduction

Antonio Salieri's one-act piece *Prima la musica e poi le parole* (1786) opens with a Poet and a Composer rushing to put together an opera within four days. According to the Composer, the task should be pretty easy: the score is ready, and now his collaborator should just adapt some words to it. The Poet protests:

> Questo è l'istesso, / che far l'abito, e poi / far l'uomo a cui s'adatti.
> ("That would be the same, as first designing a dress, and then creating a man who would fit it.")

The Composer, however, immediately retorts:

> Voi signori poeti, siete matti. / Amico, persuadetevi; chi mai / credete che dar voglia attenzione / alle vostre parole? / Musica in oggi, musica / ci vuole.
> ("You poets are crazy. My friend, be persuaded: who do you think would pay attention to your words? Music is what we need nowadays.")

By the end of the play, the Poet has begrudgingly come to accept this argument, and thus the piece's title ("First the music, and then the words") seems vindicated. Nonetheless, Salieri's elegant formula was not, by all means, the last word on the issue, nor was it the comment by his Salzburg colleague that "bey einer Oper muß schlechterdings die

Poesie der Musick gehorsame Tochter seyn" ("in an opera poetry must absolutely be the music's obedient daughter", qtd. in Kesting 2005: 21),

Indeed, the dispute on the relative weight and importance given to music and words within the symbiotic construction of operas both predates and goes on after Mozart and Salieri [1]. As Gorlée 1997: 237 points out, one could recognise among opera theorists and practitioners an ongoing confrontation between a "musicocentric" and a "logocentric" approach, with the first one being somehow prevalent throughout the centuries. It comes a little surprise, then, that in the second half of the XX century a new discipline, librettology, had arisen as a reaction, with the aim of establishing the libretto as an autonomous literary genre and investigating it from a literary point of view [2].

Studying an object as complex and multifaceted as the "libretto" (understood as a shorthand for "operatic texts" [3]) is an endeavour which might well profit from quantitative techniques, which seem well-equipped to trace its relation with other established dramatic genres, such as comedy and tragedy, and gauge its evolution across time.

Accordingly, this paper aims at offering a first computationally-informed investigation of libretti, with the main goal of discovering whether such texts from the same linguistic environment. At the same time, we will investigate the structural development of the genre across history, documenting how its relationship with the Aristotelian genres changed in different periods. More specifically, we will work on a sizable corpus of French- and German-language drama and explore which features best describe their libretti in structural terms, employing computational methods such as dimensionality reduction algorithms, statistical significance tests and feature importance analysis within a classification procedure.

2. Related literature

Opera studies as a whole seem to have been mostly unfazed by the digital turn the humanities experienced in the last decades, and it comes thus as little surprise that few studies on libretti and other operatic texts resorted to computational means. Among them, Muñoz-Lago et al. 2020 proposed a layered graphical visualisation of operas' structures, working on texts by Pietro Metastasio, while Jeong and Yoo 2022 applied k-means clustering to confirm the validity of traditional periodisation frameworks. There have also been some attempts to employ sentiment analysis on libretti, especially non-Western ones (Jin et al. 2022, Jeong 2021); a sentiment analysis of arias on the basis of linked recitatives was also proposed by Gervás and Torrente 2022.

1. The same topic is discussed, for example, in E. T. A. Hoffmann's short story *Der Dichter und der Komponist* (1813) and in Richard Strauss' last opera, *Capriccio* (1941).
2. Pioneers in this sense were, among others, Patrick J. Smith (*The Tenth Muse: A Historical Study of the Opera Libretto*, 1971) and Albert Gier (*Oper als Text: Romanistische Beiträge zur Libretto-Forschung*, 1986; *Das Libretto: Theorie und Geschichte einer musikoliterarischen Gattung*, 1998).
3. This is an admittedly rough moniker we employ here to designate all those modern dramatic texts where music plays a central role. Although dramatic forms had some sort of musical accompaniment since Antiquity, with music being one of the components of tragedy already in Aristotle (*Poet.* 1450a, 10), the first convincing integration of the two aspects in an art form (retrospectively) perceived as new was attested in the early seventeenth-century Italian melodrama.

Comprehensive computational analyses of opera libretti, however, are still lacking, but such an endeavour might certainly profit from the experience accumulated in the cognate field of quantitative drama analysis, where early approaches such as those by Markus 1970 or Reichert 1964 eventually opened the way for scores of studies with different methodologies, ranging from stylometry to topic modelling or networks analysis (see e.g. Cuéllar 2023, Lehmann and Padó 2022, Estill and Meneses 2018, Fischer et al. 2017a, Algee-Hewitt 2017) [4]. This last technique has proven particularly useful for capturing structural patterns within large corpora and modelling literary concepts like plot or characters' system (Fischer et al. 2017b, Trilcke et al. 2022).

3. Corpus building

For our research, we employed the German- and French-language corpora from the DraCor project [5], an open-access platform for hosting, accessing and analysing theatrical texts (Fischer et al. 2019). All plays in the DraCor collections are encoded in a semantically rich TEI-XML format, with specific annotation of character speech (which allows in turn to generate co-presence networks) and additional metadata on the texts and their authors.

Crucially for our purposes, the DraCor markup contains a `textclass` element with a descriptive genre tag ("Tragedy, "Comedy", etc.) and the genre's Wikidata entity, as in the following example:

```
<textClass>
 <keywords>
 <term type="genreTitle">Tragedy</term>
 </keywords>
 <classCode scheme="http://www.wikidata.org/entity/">Q80930</classCode>
</textClass>
```

The four Wikidata-linked genres currently present in the DraCor markup are *tragedy* (Q80930), *comedy* (Q40831), *tragicomedy* (Q192881), and *libretto* (Q131084); if no genre is given, the text class element is empty. It is important to note a major difference between GerDraCor and FreDracor: while the first corpus treats any genre label as exclusive, the second one allows "libretto" to coexist with other tags (for example, Chabanon's *Le Toison d'Or* [6] is marked *both* as a tragedy and as a libretto). While this heterogeneity likely stems from the corpora's different sources [7], it also underlines the blurred boundaries of genre attribution in different cultural contexts.

To ease our operationalisation, however, we decided to normalise the French genre column by having only one genre per play. All libretti with additional genre tags were

4. A good overview on the state of the field was offered by the recent *Workshop on Computational Drama Analysis: Achievements and Opportunities* (Cologne, 14-15 September 2022).
5. https://dracor.org/ger; https://dracor.org/fre.
6. https://dracor.org/api/corpora/fre/play/chabanon-toison-d-or/tei.
7. GerDracor was (mainly) derived from the TextGrid repository, (https://textgridrep.de) while FreDraCor originates from the Théâtre Classique database (https://theatre-classique.fr/index.html).

thus marked only as libretti, assuming that their intended usage (as component of operatic staging) would have been more distinctive than their broader thematic alignment along the comic/tragic axis. This working hypothesis also tried to account for the fact that genre labels in the FreDraCor markup were seemingly auto-generated from the Théâtre Classique ones and sometimes followed non-transparent patterns in attributing multiple labels.

Furthermore, as Senici 2014: 38 points out, "[p]erhaps the weakest contribution to the genrification of opera comes from the discursive space where genre is normally explicitly named, that is, the generic indicator on published librettos"; especially at the beginning of the history of opera, the choice of (sub)titling a work "tragedy" or "comedy" was a deeply rhetorical one, and had more to do with the perception the author intended to convey (e.g. that of an elevated, serious work).

Another issue in the corpora was the large number of texts without any genre label. Again, we tried to address it by exploiting DraCor's Linked-Open-Data capabilities, since the plays' markup often contains a link to the Wikidata item of the work itself (the following example is from Wagner's *Der fliegende Holländer* [8])

```
<standOff>
 ...
  <listRelation>
  <relation name="wikidata" active="https://dracor.org/entity/ger000245"
     passive="http://www.wikidata.org/entity/Q114640"/>
  </listRelation>
</standOff>
```

Scraping the plays through the Python library `BeautifulSoup`, we recursively accessed all Wikidata items and checked if they contained the Wikidata property P136, designating "genre", and used it (after some manual disambiguation of the results) to assign a genre to unlabelled plays. Unfortunately, the information gain was negligible (18 new labels for German plays, 2 for French ones).

On another note, we had also to take into account that not all libretti in our corpora might have been properly marked as such, owing to the profusion of different terminology for designating operatic texts. It has been indeed noted that even in the cradle of opera, Italy, a "plethora of terms" for indicating such works "circulated freely" for decades before one label, *dramma per musica* ("music drama"), emerged in the Venice milieu and eventually became dominant (ibid.: 38). A similar situation was therefore to be expected also in other areas, where various translations of the Italian loanword *opera* (*Oper, opéra*) long coexisted with local, often quite diverse (sub)generic denominations.

To ensure no possible libretto was neglected, we therefore searched GerDraCor and FreDraCor for all plays which contained in their title or subtitle at least one of the German

8. https://dracor.org/api/corpora/ger/play/wagner-der-fliegende-hollaender/tei.

or French genre tags that the authoritative *New Grove Dictionary of Music and Musicians* associates, to various degrees, with opera [9]. After cleaning up the results manually and removing some false positives, we grouped the newly found libretti under the "libretti (attributed)" label. As a last step, we excluded all texts still without an assigned genre, which would have marred the visualisation without providing any added value for the interpretation. Table 1 shows the final composition of our two research samples.

Genre	German sample	French sample
Tragedies	140	312
Comedies	156	692
Tragicomedies	8	82
Libretti (marked up)	55	58
Libretti (attributed)	28	34
Total	387	1178
Percentage of libretti	21.4%	7.8%

Table 1: Final French and German drama samples.

4. Experiments

We first set out to explore the relation between libretti and the main historical genres (comedy and tragedy), with the goal of mapping them on a multidimensional space. An early, informal attempt at examining the interplay between different genres in an early version of GerDraCor corpus was already made by Trilcke et al. 2015. By looking at two significant network metrics, i.e. size and density, they argued that the "evolution [of drama] over two centuries shows [...] clearly the proximity of comedy and libretto and the persistent distance from the tragedy".

To further test this hypothesis and explore more thoroughly the topology of German and French drama, we decided to adapt a method recently proposed by Szemes and Vida 2022 to cluster dramatic genres according to content-independent properties. In our case, however, we were not strictly interested in a classification task, but we rather aimed at finding out which features (if any) set libretti apart from other major dramatic genres.

The method developed by Szemes and Vida relied on a number of measures, mostly related to network properties and speech distribution corpora, which were developed for the study and/or obtained from DraCor metadata. Based on these metrics, they carried out a supervised classification procedure on DraCor's German and Shakespeare [10] corpora to label comedies and tragedies using the Support Vector Machine (SVM) method. They found no "striking difference" between the two genres as structural fea-

9. Searched terms included: *ballet de cour, ballet-héroïque, burlesque, comédie-ballet, divertissement, drame lyrique, entrée, grand opéra, intermède, Lehrstück, Liederspiel, Märchenoper, masque, Monodrama, opéra-ballet, opéra bouffon, opéra comique, opéra-féerie, pantomime, pastorale-héroïque, Posse, Schuldrama, Schuloper, Singspiel, Spieloper, tragédie en musique, vaudeville, Zauberoper, Zeitoper* (see Brown et al. 2001 s.v.)
10. https://dracor.org/shake.

tures are concerned, but were nonetheless able to single out some properties which are highly predictive of generic alignment – thus "confirm[ing] the existence of a "genre fingerprint" that shapes the dramatic structure of tragedies and comedies" (10), and which authors may (un)consciously choose to adhere to or depart from.

Following a similar approach, we also took as starting point some of the features which are provided by the DraCor API [11] through the corpora's metadata tables. Out of the 41 features available we employed almost all the count-based ones, but, unlike Szemes and Vida we chose a leaner implementation and did not compute additional speech- or distribution-related measures, relying instead mostly on standard size- and network-related properties. We considered the following features:

- `num_of_segments`: number of subdivisions (scenes or acts) in the plays
- `num_of_speakers`: number of characters with at least one utterance
- `num_of_person groups`: number of characters marked as groups (e.g. "Women")
- `word_count_sp`: total word count in characters' utterances
- `word_count_stage`: total word count in stage directions
- `average_degree`: average number of nodes to which a node is connected
- `density`: ratio between the number of actual edges and the maximum number of possible edges
- `average_clustering`: average of the local clustering coefficients of all the vertices (cf. Watts and Strogatz 1998)
- `max_degree`: maximum number of nodes to which a node is connected
- `num_of_connected components`: number of independent subgraphs within the network
- `diameter`: the longest shortest path between two nodes
- `average_path_length`: average length of the shortest path that can be drawn between any two nodes

After collecting them, we calculated correlation coefficients to assess interdependence between variables. We did it separately for the two corpora since their features are not perfectly overlapping (e.g. FreDraCor does not contain encoding for the collective characters and thus no `num_person_group`). As the matrices (Figures 1 and 2) show, some features display an excessive correlation ($>\pm 0.75$) and therefore keeping both of them might have produced misleading results. In each pair or triplet of correlated features we chose to keep the ones easier to interpret and discard the other ones: this translated into dropping average path length, diameter, maximum degree, and number of connected components (while keeping density and number of speakers) for the German

11. https://dracor.org/doc/api.

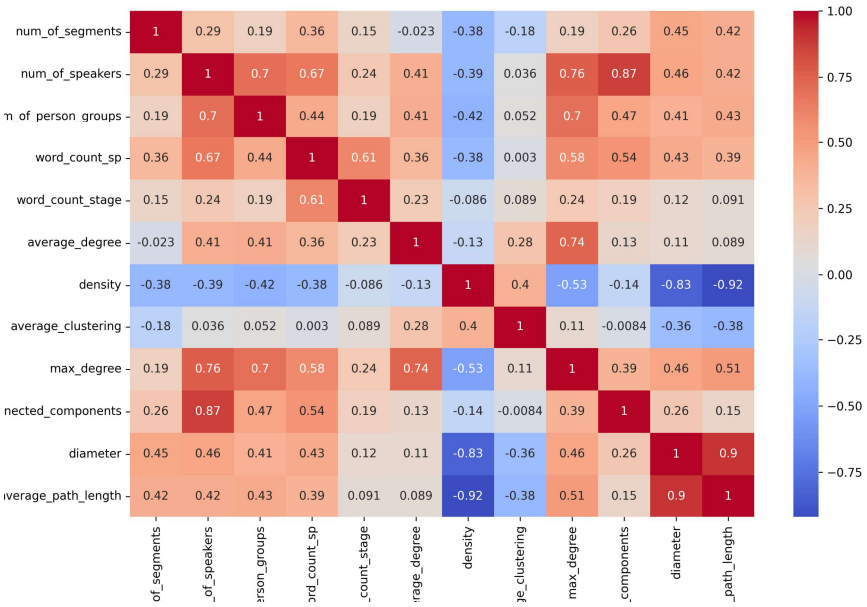

Figure 1: Correlation matrix for the German corpus.

corpus and dropping number of segments, average path length and maximum degree (while keeping word count of speeches, diameter, number of speakers and average degree) for the French one.

One of our initial assumptions was also that libretti – as a "new" genre with a relatively short history – would have shown some kind of structural evolution across the timespan covered by our corpora. In order to make this process visible in the following visualisations, we chose to subdivide the two samples according to smaller, roughly 50-year-long timeframes, starting from the year of the first recorded libretto [12] and trying to employ the same year spans in both corpora. The timeframes employed and their composition are outlined in Tables 2 and 3.

Timeframes	non-libretti	libretti (marked + attributed)
1626-1669	330	8
1670-1719	250	35
1720-1769	280	21
1770-1819	253	21
1820-1889	253	20

Table 2: Visualisation timeframes for our French corpus.

Aiming at a first exploratory analysis of our plays and their attributes, we set out by plotting them (now expressed as feature vectors) on a two-dimensional plane. To this aim, we tested various unsupervised (i.e. class-blind) techniques for dimension reduction

12. The earliest French libretto is from 1629 (*Maître Galimathias* by Claude de l'Estoile) and the earliest German one from 1770 (*Die Jagd* by Christian Felix Weiße), while the latest libretti available are dated respectively 1889 (*Thésée* by P. G.) and 1920 (*Doktor Faust* by Ferruccio Busoni).

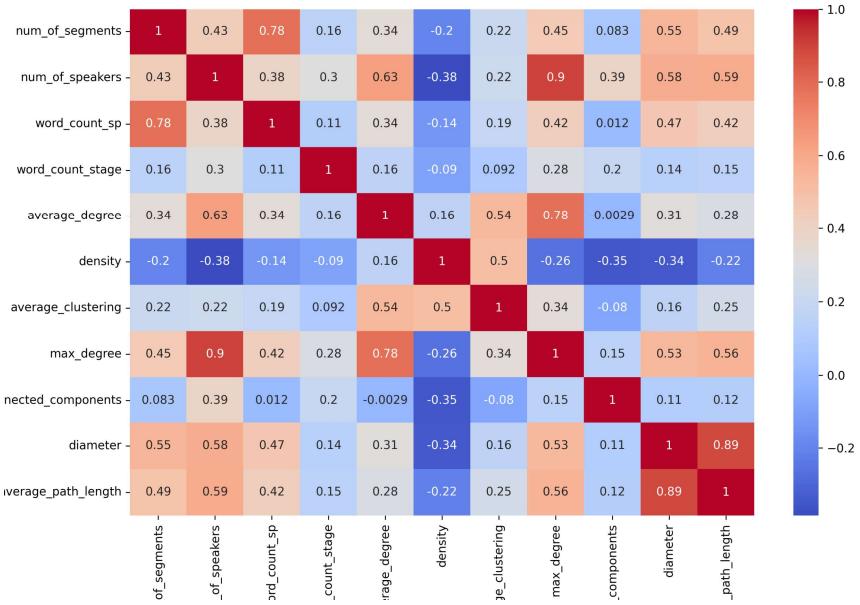

Figure 2: Correlation matrix for the French corpus.

Timeframes	non-libretti	libretti (marked + attributed)
1770-1819	156	25
1820-1869	113	41
1870-1920	158	17

Table 3: Visualisation timeframes for our German corpus.

such as Principal Component Analysis (PCA), T-distributed Stochastic Neighbour Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) (see Maaten et al. 2009, Burges 2010, Waggoner 2021 etc.). Standard PCA – a non-random linear mapping algorithm which tends to capture global structure better than local similarities, as t-SNE and UMAP do – was eventually chosen for plotting the various timeframes and thus implemented through the `sklearn` Pyhton library. This first attempt, however, led to somehow disappointing outputs, with both corpora displaying a substantial structural homogeneity between all four genres across most timeframes.

Eventually, one had to come to terms with the fact that, at least as our selection of formal variables was concerned, one cannot clearly detect a process of progressive "genrification" of libretti. While critical consensus tends to present the first two centuries of opera in terms of its "crystallization into a specific, identifiable genre" (Campana 2012: 206), the absence of meaningful clusters seemed to suggest that the novelty of libretti was difficult to capture through size- and network-related features only – especially if one kept envisioning the libretto as a *unitary* genre with clear-cut properties such as the traditional Aristotelian ones.

In a further attempt to refine our operationalisation, we decided to move away from this

perspective and tried instead to account for the libretti's heterogeneity by re-labelling them with respect to their proximity to the classic genres. Instead of using the traditional dichotomy *comic/tragic*, however, we chose to employ a binary *comic/non-comic* tagging, because qualitative analysis of genre descriptors (i.e. subtitles) in our sample showed how labels referring to comedy in an explicit (e.g. *Komödie für Musik*, *comédie-ballet*) or implicit manner (*divertissement*, *vaudeville*, *Posse*) were more frequent than the ones related to tragedy (e.g. *Trauerspiel*, *tragédie en musique*) or without a transparent reference (e.g. *opéra*, *Oper*) [13]. This led us to the conclusion that, at least as genre assignment was involved, comic-like libretti were somehow easier to identify and model as a group as against non-comic ones.

Accordingly, we semi-automatically extracted the new labels from the subtitles through keywords and run again the PCA algorithm, with the results being presented in Figures 3 and 4. While the model performed better than before, validating to some extent our refining of the libretti labels, it still failed to produce clear-cut clusterings of operatic texts; on top of that, as research by Szemes and Vida 2022 already showed, even the fundamental distinction between comic and tragic zones remained often blurred.

On the other hand, however, the choice of splitting our data in different timeframes represented a valuable improvement on previous attempts, insofar as it highlighted some topological idiosyncrasies which would have got lost in a catch-all visualisation. It is the case, for example, of the second French data frame (plays between 1670 and 1719), where a pattern is indeed visible: while comic libretti, as expected, mostly follow the structural model of comedies, non-comic libretti are clearly distinguishable from all other genres and build a definite, albeit sparse aggregation on the right side of the graph.

The relatively more pronounced clustering of French dramatic genres as against German ones, which the PCA plots show, could be explained by the corpora's different size and, even more, by their temporal coverage. While the French corpus mostly spreads over a period of normative aesthetics, where texts like d'Aubignac's *Pratique du théâtre* (1657) or Boileau's *Art poétique* (1674) set the rules for theatre-writing, the German corpus starts at a time in which Classicism was already losing ground and French theatrical conventions were being actively repudiated or deconstructed (cf. Lessing's *Hamburgische Dramaturgie*, 1767-1769) – leading to more pronounced interferences between genres which the graph seems to capture.

To further investigate this aspect, we decided to examine the individual features behind them through different methods; at this stage, we chose to leave out tragicomedies (because of their highly irregular distribution and globally low number) and focused instead on comparing the two main genres (i.e. comedies and tragedies) against libretti as a whole. We set out by measuring statistical significance in the features' differences to assess how such parameters performed in telling apart libretti and non-libretti. After applying the Shapiro-Wilk test for checking the normality of distributions, which as ex-

13. The German sample has 39 clearly "comic" libretti and 41 unassigned libretti, while the French one has 45 clearly "comic" libretti, 18 clearly "tragic" libretti, and 27 unassigned libretti.

Figure 3: Evolution of French drama, 1626-1889, visualised through PCA.

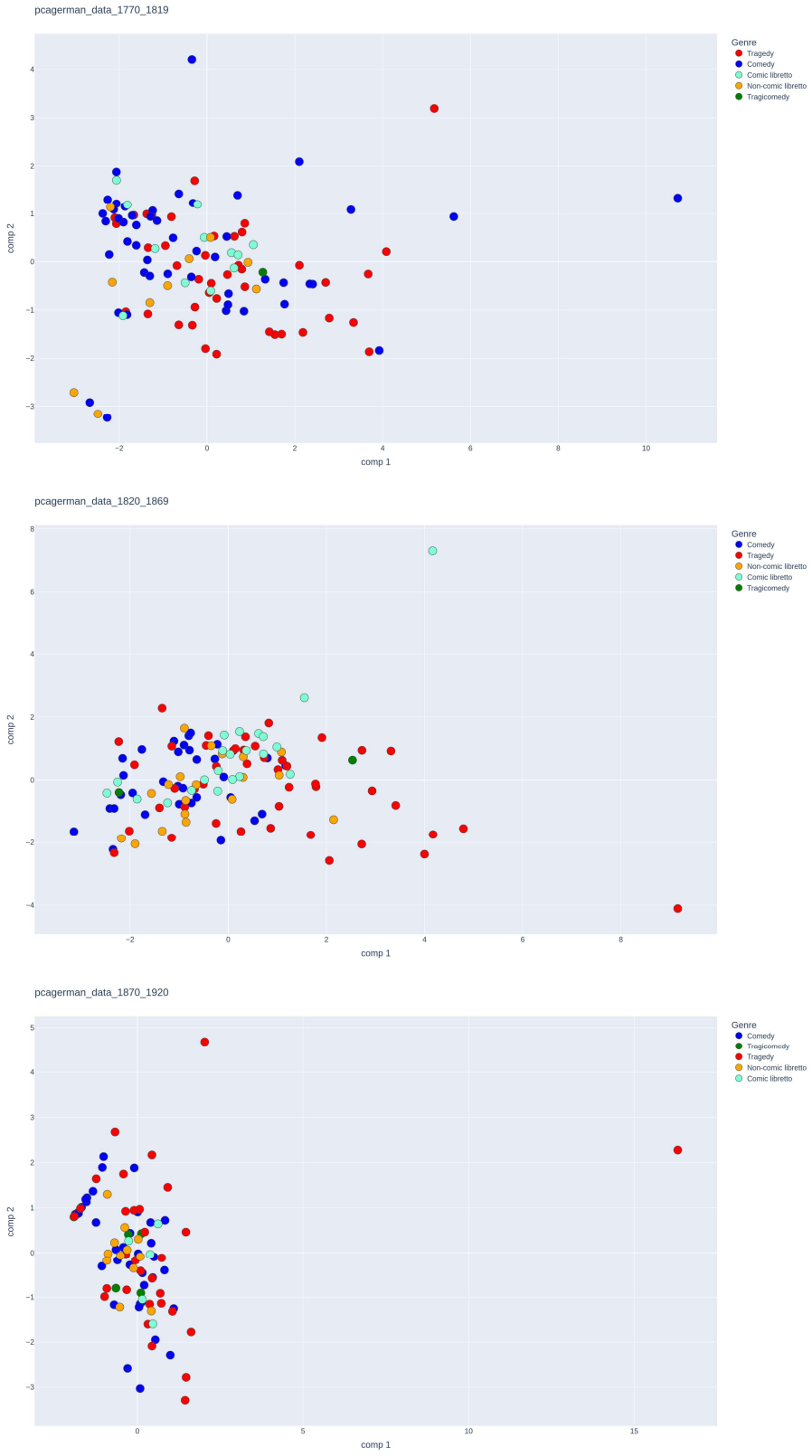

Figure 4: Evolution of German drama, 1770-1920, visualised through PCA.

Features	German sample	French sample
num_of_segments	0.42	*
num_of_speakers	0.3	**4.3e-08**
num_of_person_groups	**1.34e-06**	n/a
word_count_sp	**7e-09**	**4.53e-19**
word_count_stage	0.13	**1.12e-19**
average_degree	**0.03**	0.16
density	0.91	**6.3e-09**
average_clustering	0.23	0.82
num_connected_components	*	**1.16e-16**
diameter	*	**0.009**

Table 4: Statistical significance of features in the two samples according to the Wilcoxon Rank Sum test. The values marked with an asterisk were not computed, since the corresponding features were dropped beforehand because of high correlation.

Measures	German sample	French sample
overall accuracy	84.3%	93.1%
precision for class *Non-libretto*	85.4%	94.1%
recall for class *Non-libretto*	96.6%	98.7%
f1 for class *Non-libretto*	90.7%	96.3%
precision for class *Libretto*	75.6%	68.2%
recall for class *Libretto*	38.7%	31.1%
f1 for class *Libretto*	51.2%	42.7%

Table 5: Values for the best performing random forest classifier.

pected were almost always not normal, we implemented the non-parametric Wilcoxon Rank Sum test to check if the differences between the distributions were substantial (Table 4).

As an additional benchmark, we also run a binary random forest classifier tasked to differentiate between libretti and non-libretti; the number of estimators was optimised through iterative hyperparameter tuning based on five-fold cross-validation. While high overall accuracies were expected on such an imbalanced dataset, the binary classifier also showed mediocre performance in identifying libretti (i.e. the minority class) in both corpora (see Table 5).

Our intention, however, was not finding a method to efficiently automate libretti classification, but rather ascertaining whether the most relevant features for the classification task (presented in Figure 5) were the same whose variance was statistically significant according to our tests.

These two pipelines (statistical significance tests and classifier) broadly agreed in showing that three features (*num_of_person_groups*, *word_count_sp*, *average_degree* [14]) were helpful for distinguishing German libretti from non-libretti, while six of them proved

14. The classifier actually considered the average degree slightly less important than the word count for stage, but taking into account the confidence interval and the result of the statistical significance test we preferred to use this parameter instead of the other one.

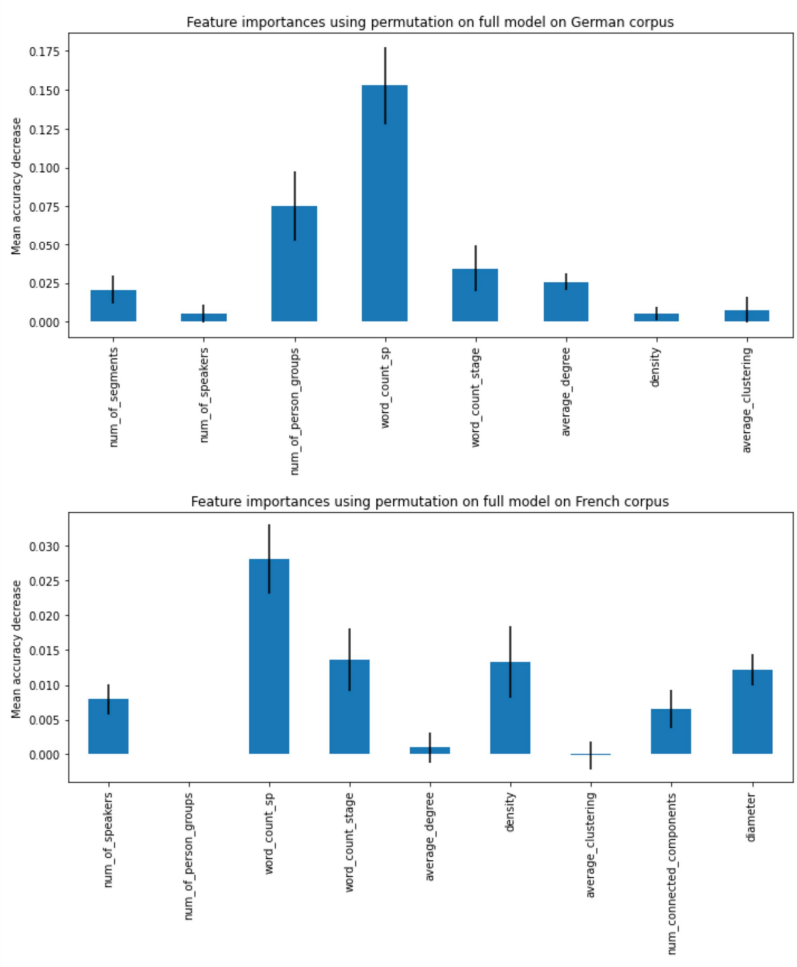

Figure 5: Relative features importance according to the random forest classifier. The barplot indicates which features, if left out, lead to the biggest accuracy decrease.

useful for sorting out the French data (*num_of_speakers*, *word_count_sp*, *word_count_stage*, *density*, *num_connected_components*, *diameter*).

At this point, we charted some of the most interpretable features as scatterplots [15], using them as hermeneutical tools for discovering which traits were distinctive for libretti at different stages of history. In order to better capture the granularity of the process, we switched again to a four-class visualisation (comedies/tragedies/comic libretti/non-comic libretti) and plotted each play individually; we also applied a local regression algorithm (LOWESS, cf. Cleveland 1979) to draw a smooth curve between the data points and help visualise the distances between the genres and their evolution.

While the unequal distribution of texts across the investigated timespan suggests caution in speaking of *longue-durée* evolutionary phenomena, trends emerging from the pictures give a glimpse into some long-lasting relations between genres. One of the patterns seemingly common to both German and French data, for instance, is the relative independence of non-comic libretti from the other genres in terms of several structural features.

To start with, such phenomenon can be observed when comparing the two most discriminative features for the mapping of German data, i.e. the total word count for utterances and the number of collective entities (Figure 6). The two curves show indeed how non-comic libretti (the yellow line) chart quite an autonomous path, while comic libretti often adhere more to the structural model set by comedies, as seen especially in the `word_count_sp` graph.

Such pattern, which the sparse nature of the German data space makes somehow less evident, emerges with more force in the French data. Figure 7, for example, shows scatterplots for the network-related features `num_of_speakers` and `density`, which are often found as inversely correlated (the higher the number of the characters in a play, the lower the chance they interact with all the other ones). While all genres broadly follow this pattern, comic libretti tend to structurally resemble comedies in having somehow tighter plots, with fewer characters which are highly interconnected, while non-comic libretti display an unusual dissimilarity from any other genres.

A supplementary confirmation of the non-comic libretti's peculiar status comes from another classification experiment we conducted, where we asked the random forest algorithm to sort plays into our four genres. As the confusion matrices in Figure 8 and Figure 9 illustrate, the classifier struggled to distinguish comic libretti and comedies more often than in separating non-comic libretti from tragedies.

Some prominent spikes which are especially visible in the non-comic libretti curves, coupled with the clustering evidence emerging from the PCA, concur in underscoring the particular relevance of the second half of the XVII century – something one cannot explain only on account of the higher number of texts available in that timeframe. Critics indeed consider it as a pivotal moment in the diffusion of opera beyond Italy and

15. We removed outliers which were more than three standard deviations away from the mean.

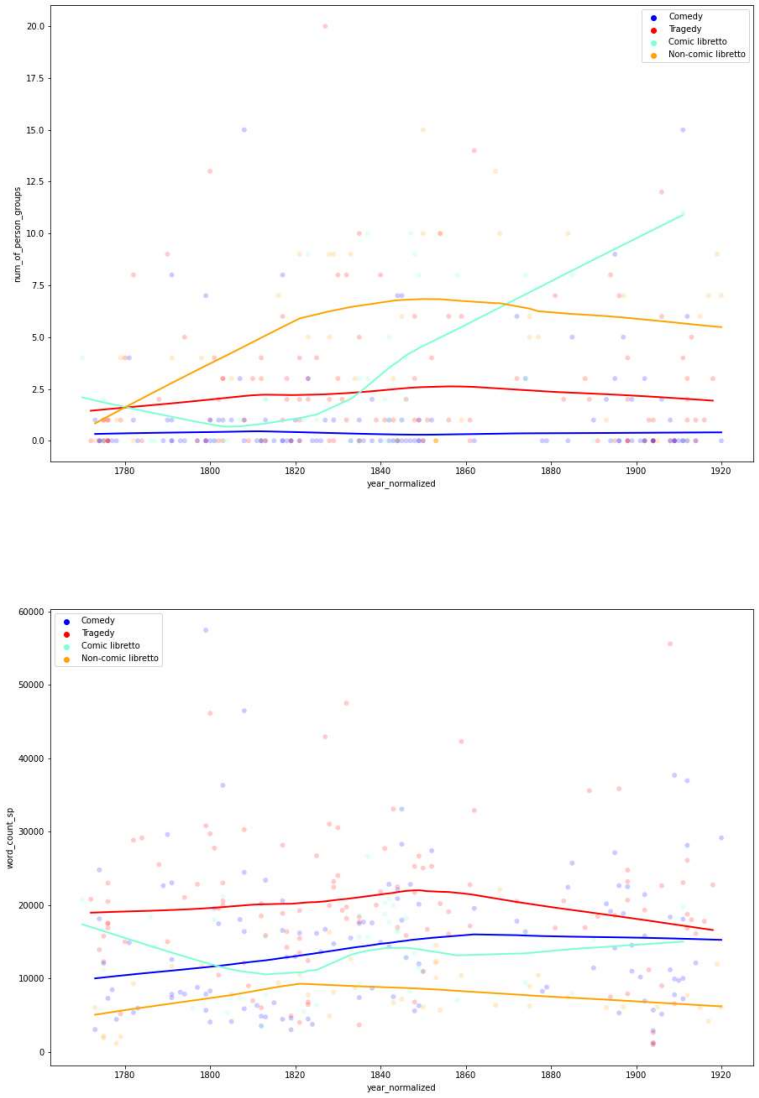

Figure 6: Evolution of selected features in German data: *word_count_sp* (below) and *num_of_groups* (above). The numbers in the graph indicate the number of plays in this timeframe.

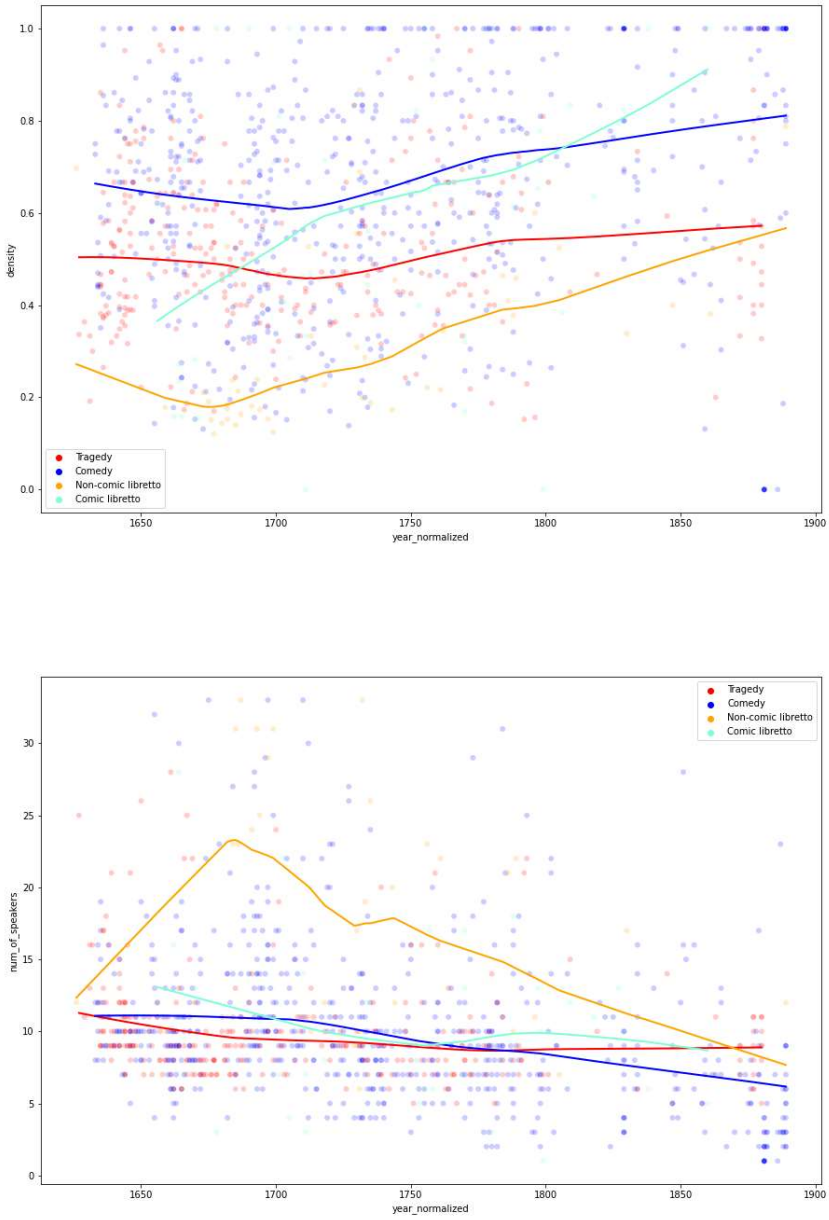

Figure 7: Evolution of selected features in French data (**I**): *density* (above) and *num_speaker* (below).

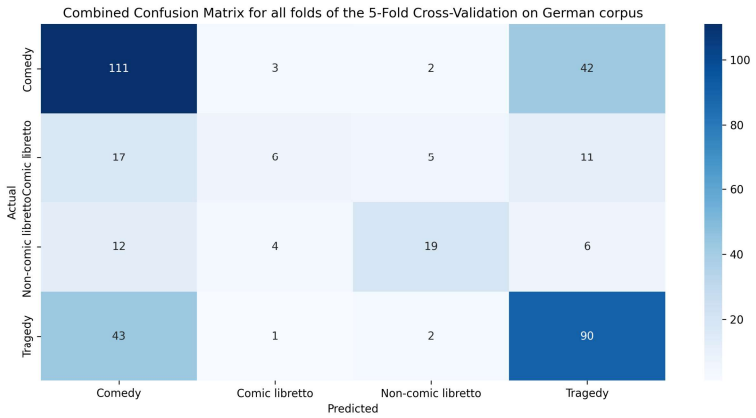

Figure 8: Confusion matrix for a four-class classifier trained and tested on the German sample.

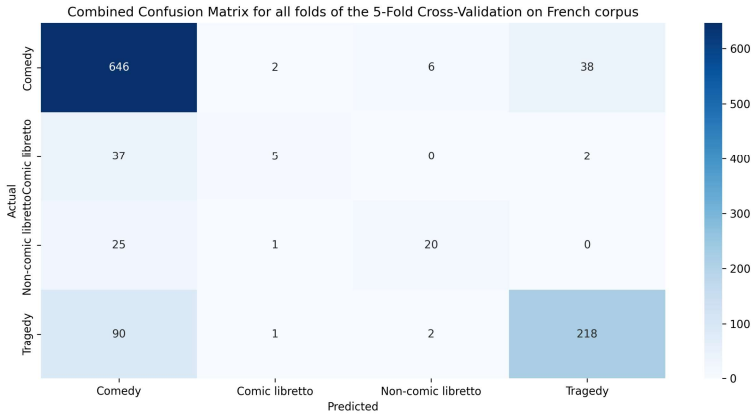

Figure 9: Confusion matrix for a four-class classifier trained and tested on the French sample.

into France, where new hybrid forms of theatre, music, and dance (such as the *comédie-ballet*, best exemplified by Molière's *Le Bourgeois gentilhomme*) were soon joined by truly "operatic" (i.e. completely sung) genres such as the *tragédie lyrique*.

The merit of popularising this last mode of expression, which has been considered "the definitive form [of French opera], capable of rivalling the spoken theatre" (Norman 2009: 17), is to be shared between composer Jean-Baptiste Lully and librettist Philippe Quinault, whose operas account indeed for almost one third of the plays written between 1670 and 1719. Closer inspection of some of these operas, such as *Cadmus et Hermione* (1673) [16] or *Persée* (1682) [17], confirms the aforementioned quantitative findings: they indeed feature large ensemble casts whose characters have frequent interactions with each other, thus resulting in well-connected social networks.

Due to higher text availability, French scatterplots are also useful for empirically verifying traditional assumptions on opera by literary critics and musicologist. This is the case, for example, of the relation between diegetic and non-diegetic elements within a dramatic text. As Ulrich Weisstein has argued in his seminal essay *The Libretto as Literature*, for example, "music lacks the speed and verbal dexterity of language, [and] fewer words are needed in opera than would be required in a play of comparable length"; therefore, "librettos are usually shorter than the texts of ordinary dramas, and often to the point of embarrassing the listener or reader" (Weisstein 1961: 19). On the other side, one would expect operatic texts to have a greater share of stage directions, due to the necessity of setting the stage for musical numbers or dances (something along the lines of "Enter five dancers, dressed as knights..."). As Figure 10 illustrates, this trends seems indeed confirmed in both kinds of libretti: comic and non-comic operatic texts follow similar paths in having less interaction between characters and (sizeably) more stage directions.

5. Conclusion

As in many CLS projects, a major limitation of this investigation was represented by the relatively small size of the corpora employed. By relying on DraCor, one of the largest scholarly databases of dramatic texts available, we tried to collect a sample big enough to draw meaningful conclusions, but the results were sometimes mixed. While some timeframes were populated enough to give some actual insight on the dynamics of the time, the results obtained in others might be radically changed by any corpora overhaul, be it in terms of corpus enlargement or markup refining.

Reduced text availability also forced us to focus on only two cultural milieus (German and French) and forsake any further comparative attempt. The absence of texts from Italy, one of opera's major playfields, is particularly lamentable, but since DraCor's Italian corpus [18] contains so far only a handful of libretti (mostly early *melodrammi* by

16. https://dracor.org/fre/quinault-cadmus-hermione.
17. https://dracor.org/fre/quinault-persee.
18. https://dracor.org/ita.

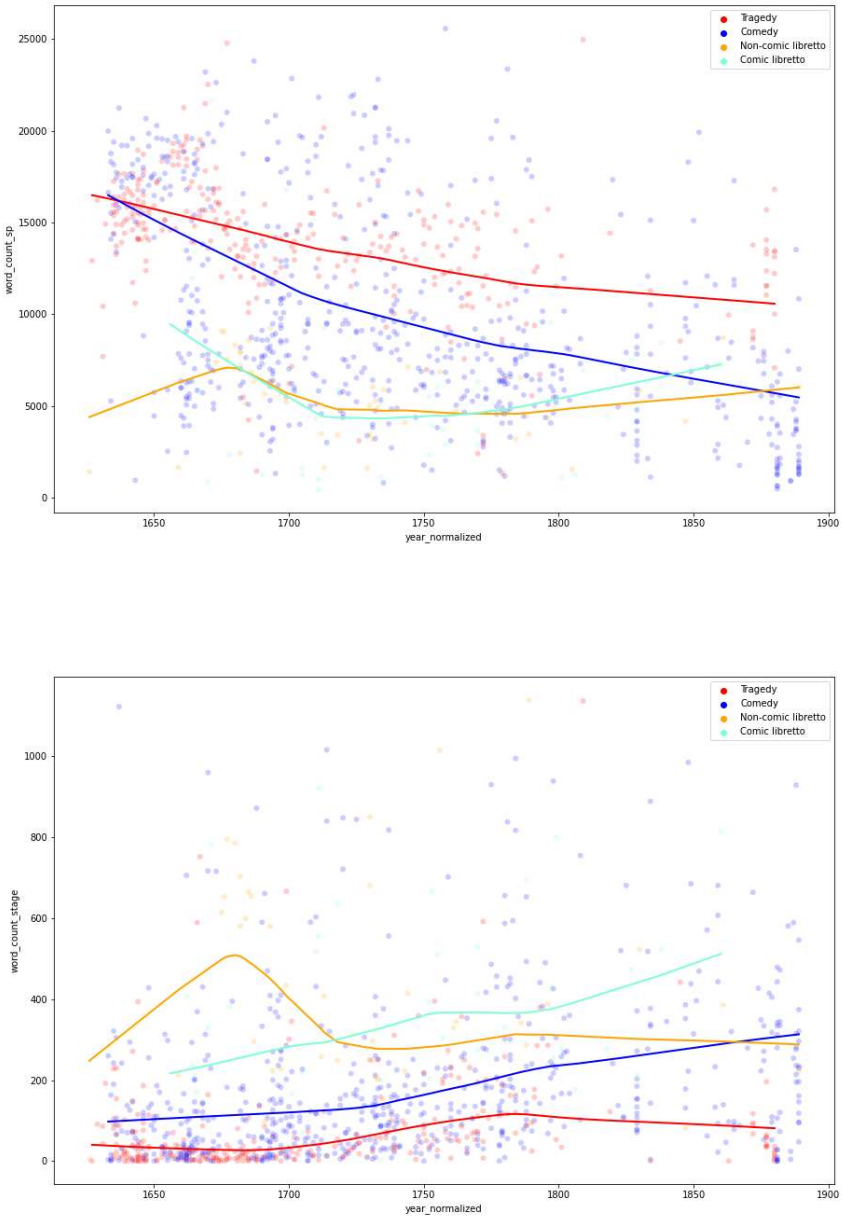

Figure 10: Evolution of selected features in French data (II): *word_count_sp* and *word_count_stage*.

Figure 11: Principal Components Analysis for German and French libretti, with centroids (X).

Metastasio), adding them would have not dramatically improved the quality of our findings while posing several challenges in the implementation. Future studies, however, might exploit the wealth of freely accessible Italian libretti [19] to perform more encompassing analyses, while possibly enlarging ItaDraCor as well (through the onboarding procedure described in Börner et al. 2023).

Eventually, much more work is needed in order to achieve a satisfying diachronic picture of opera in its relationship with other dramatic genres. Nonetheless, our first computational foray into operatic texts yielded some insights on libretti as part of the wider dramatic system. On one side, our mapping of operatic texts clearly showed how "[o]pera spread throughout Europe both as an 'Italian' product and also *as a 'native' musical theatre*" (Campana 2012: 207, emphasis added), with its local iterations possessing idiosyncratic features which make them stand apart. Specifically, the findings revealed how the two types of libretti often display different behaviours, with comic libretti mainly aligning with comedies but non-comic libretti manifesting a definite distance from both tragedy and the other genres.

Furthermore, analysis of the different timeframes through PCA clusterings and feature lineplots also suggested that in German data it is more difficult to discriminate effectively between the two kinds of libretti, while such distance is more substantial within the French space. This difference is even clearer if one plots through PCA only German and French operatic texts (see Figure 11).

On the other side, our clustering attempts on a selection of purely formal (size- and network-based) features failed to identify libretti as a genre possessing a strong degree

19. Many libretti with simple or no markup, mostly in HTML or PDF format, are available in online databases such as `https://opera-guide.ch`, `https://opera.stanford.edu`, `https://librettidopera.it`, `https://www.operalib.eu`, etc.

of formal independence. Such outcomes actually play into the established critical narrative which sees opera as a Protean art form, whose generic essence is continuously contested:

> Opera's identity as a genre [...] relies from the start on mixing, as a contamination of music and theatre, music and word, singing and acting, showing and telling. Thus, it defines itself historically and systematically as a hybrid, challenging at the outset the foundational law of genre discourse. [...] Even more so than literary genres, the hybridity of opera, and its dialogue with the demands of production and performance, contains the possibility of genre being disrupted. (Campana 2012: 205)

Nonetheless, our operationalisation did show that one could identify, to some extent, traits which are clearly distinctive of comic and non-comic libretti, and that such traits do not always align with the ones characterising comedies and tragedies – thus pointing to a complex relationship of imitation/departure from the spoken theatre models.

Ultimately, this paper showcases once again the complexity of modelling the relationship between different dramatic genres – and more generally, the whole concept of dramatic text – on account of formal features. While fine-tuning and enlarging our model with additional features might actually help to achieve better accuracy in classification, we believe nonetheless that major performance optimisations, able to mirror the clustering effectiveness and explanatory power of other techniques such as e.g. topic modelling [20], could be achieved only through a more radical rethinking of operationalisation patterns – an ambitious task for future CLS scholarship.

20. See Schöch 2017's successful attempt on French Classical and Enlightenment theatre, based on texts now contained in FreDraCor.

6. Data availability

Data and scripts employed can be found here: [REPOSITORY REMOVED].

7. Acknowledgements

The authors would like to thank Artjoms Šeļa, Henny Sluyter-Gäthje, and Peer Trilcke for their helpful suggestions.

8. Author contributions

Luca Giovannini: Conceptualisation, Formal Analysis, Methodology, Writing

Daniil Skorinkin: Software, Visualisation, Methodology, Formal Analysis

References

Algee-Hewitt, Mark (2017). "Distributed character: Quantitative models of the English stage, 1550–1900". In: *New Literary History* 48.4, pp. 751–782. DOI: 10.1353/nlh.2017.0038.

Börner, Ingo, Frank Fischer, Luca Giovannini, Christopher Lu, Carsten Milling, Daniil Skorinkin, Henny Sluyter-Gäthje, and Peer Trilcke (2023). "Onboard onto DraCor: Prototyping Workflows to Homogenize Drama Corpora for an Open Infrastructure". In: *Dhd 2023 Conference Abstracts*. Accepted. University of Luxemburg and Trier, Luxemburg/Germany.

Brown, Howard Mayer, Ellen Rosand, Reinhard Strohm, Michel Noiray, Roger Parker, Arnold Whittall, Roger Savage, and Barry Millington (2001). "Opera (i)". In: *Grove Music Online*. Oxford University Press. DOI: 10.1093/gmo/9781561592630.article.40726.

Burges, Christopher J. C. (2010). *Dimension Reduction: A Guided Tour*. Boston and Delft: now.

Campana, Alessandra (2012). "Genre and poetics". In: *The Cambridge Companion to Opera Studies*. Ed. by Nicholas Till. Cambridge: Cambridge University Press, pp. 202–224. DOI: 10.1017/CCO9781139024976.013.

Cleveland, William S. (1979). "Robust Locally Weighted Regression and Smoothing Scatterplots". In: *Journal of the American Statistical Association* 74.368, pp. 829–836. DOI: 10.1080/01621459.1979.10481038.

Cuéllar, Álvaro (2023). "Stylometry and Spanish Golden Age Theatre: An Evaluation of Authorship Attribution in a Control Group of Undisputed Plays". In: *Digital Stylistics in Romance Studies and Beyond*. Ed. by Robert Hesselbach, José Calvo Tello, Ulrike Henny-Krahmer, Christof Schöch, and Daniel Schlör. Forthcoming. Heidelberg: Heidelberg University Press.

Estill, Laura and Luis Meneses (2018). "Is Falstaff Falstaff? Is Prince Hal Henry V?: Topic Modeling Shakespeare's Plays". In: *Digital Studies/Le champ numérique* 8.1. DOI: 10.16995/dscn.295.

Fischer, Frank, 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". In: *Proceedings of DH2019: "Complexities"*. University of Utrecht, The Netherlands. DOI: 10.5281/zenodo.4284002.

Fischer, Frank, Gilles Dazord, Mathias Göbel, Christopher Kittel, and Peer Trilcke (2017a). "Le drame comme réseau de relations : une application de l'analyse automatisée pour l'histoire littéraire du théâtre". In: *Revue d'historiographie du théâtre*. URL: https://hal.science/hal-01811799.

Fischer, Frank, Mathias Göbel, Dario Kampkaspar, Christopher Kittel, and Peer Trilcke (2017b). "Network Dynamics, Plot Analysis: Approaching the Progressive Structuration of Literary Texts". In: *Proceedings of DH2017: "Access/Accès"*. URL: https://dh2017.adho.org/abstracts/071/071.pdf. McGill University, Montreal, Canada.

Gervás, Pablo and A. Torrente (2022). "Emotional Interpretation of Opera Seria: Impact of Specifics of Drama Structure". In: *14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. URL: http://nil.fdi.ucm.es/sites/default/files/KDIR2022-opera-CRC.pdf.

Gorlée, Dinda (1997). "Intercode Translation: Words and Music in Opera". In: *Target. International Journal of Translation Studies* 9.2, pp. 235–270. DOI: 10.1075/target.9.2.03gor.

Jeong, Harim (2021). "Study on sentiment analysis for Opera". In: *Proceedings of APICIST 2021*, pp. 89–91.

Jeong, Harim and Joo Hun Yoo (2022). "Opera Clustering: K-means on librettos datasets". In: *Journal of Internet Computing and Services* 23.2, pp. 45–52. DOI: 10.7472/jksii.2022.23.2.45.

Jin, Cong, Zhen Song, Jiaqi Xu, and Huiyue Gao (2022). "Attention-Based Bi-DLSTM for Sentiment Analysis of Beijing Opera Lyrics". In: *Wireless Communications and Mobile Computing*. DOI: doi.org/10.1155/2022/1167462.

Kesting, Hanjo (2005). *Der Musick gehorsame Tochter: Mozart und seine Librettisten*. Göttingen: Wallstein.

Lehmann, Jörg and Sebastian Padó (2022). "Classification of comedies and tragedies written in Calderón de la Barca's Comedias Nuevas". In: *Zeitschrift für Digitale Geisteswissenschaft* 7. URL: 10.17175/2022_012.

Maaten, Laurens van der, Eric O. Postma, and Jaap van den Herik (2009). *Dimensionality Reduction: A Comparative Review*. Preprint. URL: https://lvdmaaten.github.io/publications/papers/TR_Dimensionality_Reduction_Review_2009.pdf.

Markus, Solomon (1970). *Poetica matematică*. Bucharest: Academiei.

Muñoz-Lago, Paula, Nicola Usula, Emilia Parada-Cabaleiro, and Álvaro Torrente (2020). "Visualising the Structure of 18th Century Operas: A Multidisciplinary Data Science

Approach". In: *24th International Conference on Information Visualisation*, pp. 530–536. DOI: 10.1109/IV51561.2020.00091.

Norman, Buford (2009). *Quinault, librettiste de Lully: le poète des grâces*. trans. by Thomas Vernet and Jean Duron. Wavre: Mardaga.

Reichert, Waltraud (1964). "Kybernetische Methoden der Dramenforschung". In: *Grundlagenstudien aus Kybernetik und Geisteswissenschaften* 5.3-4, pp. 115–120.

Schöch, Christof (2017). "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama". In: *Digital Humanities Quarterly* 11.2.

Senici, Emanuele (2014). "Genre". In: *The Oxford Handbook of Opera*. Ed. by Helen M. Greenwald. Oxford University Press. DOI: 10.1093/oxfordhb/9780195335538.013.002.

Szemes, Botond and Bence Vida (2022). "Tragic and Comical Networks: Clustering Dramatic Genres According to Structural Properties". In: *Workshop on Computational Drama Analysis: Achievement and Opportunities*. Forthcoming. University of Cologne, Germany.

Trilcke, Peer, Frank Fischer, Mathias Göbel, and Dario Kampkaspar (July 2015). *Comedy vs. Tragedy: Network Values by Genre*. URL: https://dlina.github.io/Network-Values-by-Genre.

Trilcke, Peer, Evgeniya Ustinova, Frank Fischer, Carsten Milling, and Ingo Börner (2022). "Detecting Small Worlds in a Corpus of Thousands of Theater Plays: A DraCor Study in Comparative Literary Network Analysis". In: *Workshop on Computational Drama Analysis: Achievement and Opportunities*. Forthcoming. University of Cologne, Germany.

Waggoner, Philip D. (2021). *Modern Dimension Reduction*. Cambridge: Cambridge University Press.

Watts, Duncan and Steven Strogatz (1998). "Collective dynamics of 'small-world' networks". In: *Nature* 393, pp. 440–442. DOI: 10.1038/30918.

Weisstein, Ulrich (1961). "The Libretto as Literature". In: *Books Abroad* 35.1, pp. 16–22. DOI: 10.2307/40115290.