Article

Computational Approaches to Opera Libretti
An Experiment on DraCor Corpora

Luca Giovannini [1,3] [iD]
Daniil Skorinkin [2] [iD]

1. Institute for German Studies, University of Potsdam, Potsdam, Germany.
2. Digital Humanities Network, University of Potsdam, Potsdam, Germany.
3. Department of Linguistic and Literary Studies, University of Padova, Padua, Italy.

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 respective linguistic environments. On the other side, it explores the structural development of the genre across history, evaluating how its relationship with traditional genres changed in different periods. Results confirm known challenges of modelling dramatic texts while pointing out 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. As a reaction, the second half of the XX century actually saw the birth of a new discipline, librettology, which aimed at investigating operatic texts from a literary point of view.[2]

In this paper, we offer a contribution towards a 'digital librettology' by attempting an investigation of libretti[3] through computational techniques. In doing so, we strive to demonstrate how quantitative-oriented methods hold some promise for exploring the libretto as an autonomous dramatic form. In particular, we believe the now consolidated 'distant reading' paradigm (Moretti [2013]) might be useful for exploring the evolution of libretti over time.

In the following, we will try to discover whether operatic texts can be distinguished from contemporary comedies and tragedies according to elements other than the presence of a musical score. While looking for the existence of a peculiar 'genre signal' for libretti, we will also analyse the structural development of the genre across history, documenting how its relationship with the traditional genres changed in different periods. More specifically, we will work on a sizeable corpus of French- and German-language texts and explore which features best describe their libretti in structural terms, employing several computational methods such as dimensionality reduction algorithms, statistical significance tests and feature importance analysis within a classification procedure.

2. Related Literature

The 'digital turn' affecting the humanities in recent decades seems to have had a limited impact on opera studies. While there have been substantial efforts to build digital databases on operatic materials and performances,[4] the implementation of computational methods to investigate libretti appears still quite underdeveloped. Among existing literature, Muñoz-Lago et al. ([2020]) proposed a layered graphical visualisation of opera structures, working on texts by Pietro Metastasio, while Jeong and Yoo ([2022]) applied *k-means* clustering to confirm the validity of traditional periodisation frameworks. Furthermore, Bonora and Pompilio ([2021]) have worked on the automatic extraction of descriptions of operatic characters through lexical and syntactic patterns. There have also been some attempts to employ sentiment analysis on libretti, especially non-Western ones (e.g. Jin et al. [2022]). A sentiment analysis of arias on the basis of linked recitatives was also suggested 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 throughout the paper to designate 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 for Aristotle (*Poet.* 1450a, 10), the first integration of the two aspects in an art form (retrospectively) perceived as new happened in the early seventeenth-century Italian melodrama (cf. Leopold ([2003]) for a detailed discussion of the matter see also Louvat-Molozay ([2018])).
4. An inventory of these sources would go beyond the scope of this paper and is therefore not provided.

Comprehensive computational analyses of opera libretti, however, are still lacking, but such endeavours might 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 many methodologically diverse studies, ranging from stylometric approaches to topic modelling or networks analysis (see e.g. Algee-Hewitt (2017), Cuéllar (in press), Estill and Meneses (2018), Fischer et al. (2017a), and Lehmann and Padó (2022)).[5] This last technique, in particular, has proven useful for capturing structural patterns within large corpora and modelling literary concepts like plot or character systems (Fischer et al. 2017b; Trilcke et al. 2024).

3. Corpus Building

For our research, we employed the German- and French-language corpora from the DraCor project,[6] 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 (para)textual metadata. Crucially for our purposes, the DraCor markup contains a `textclass` element with a descriptive genre tag (such as 'Tragedy' or 'Comedy') and the genre's Wikidata entity, as in the following example:

```
1  <textClass>
2   <keywords>
3   <term type="genreTitle">Tragedy</term>
4   </keywords>
5   <classCode scheme="http://www.wikidata.org/entity/">Q80930</classCode>
6  </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*[7] is marked *both* as a tragedy and as a libretto). While this heterogeneity mostly stems from the corpora's different sources,[8] it also suggests how genre attribution could be all but unambiguous in different cultural contexts.

To ease our operationalisation, we decided to normalise the French genre column by having only one genre per play, i.e., marking all libretti with additional genre tags only as 'libretti'. This methodological choice, whose impact was anyway very limited,[9] had several reasons. On one side, since genre labels in the FreDraCor markup were derived from the *Théâtre Classique* labels through a combination of automated (re)generation

---

5. A good overview of the state of the field was offered by the *Workshop on Computational Drama Analysis: Achievements and Opportunities* (Cologne, 14-15 September 2022); contributions to the workshop are collected in Andresen and Reiter (2024).
6. See https://dracor.org/ger and https://dracor.org/fre.
7. See https://dracor.org/api/corpora/fre/play/chabanon-toison-d-or/tei.
8. 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).
9. Only 16 French texts out of 1560 have multi-genre labelling.

and manual tweaking,[10] we considered them not fully reliable. Furthermore, some additional uncertainty was due to the conventions of the genre itself: as Senici (2014, 38) points out, "[p]erhaps the weakest contribution to the genrification of opera [came] from the discursive space where genre is normally explicitly named, that is, the generic indicator on published librettos".

In other words, especially at the beginning of opera history, 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) than with the text's actual properties. Eventually, our first guess was that the intended usage of a libretto (as component of an operatic staging) would have been more 'distinctive' than its broader thematic alignment along the comic/tragic axis, and thus we choose to keep 'libretto' as the primary label.

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*:[11]

```
1  <standOff>
2  ...
3   <listRelation>
4   <relation name"="wikidata active"=https://dracor.org/entity/"ger000245
        passive"=http://www.wikidata.org/entity/"Q114640/>
5   </listRelation>
6  </standOff>
```

Scraping the XMLs 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 limited (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. Indeed, it has been 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 (Senici 2014, 38). A similar situation was therefore to be expected in other areas as well, where various translations of the Italian loanword *opera* (*Oper, opéra*) long coexisted with local, often quite diverse (sub)generic denominations.

To ensure that no possible libretto was neglected, we searched GerDraCor and FreDraCor for all plays that contained in their title or subtitle at least one of the German or French genre tags that the authoritative *New Grove Dictionary of Music and Musicians* associates,

---

10. See the README file at https://github.com/dracor-org/fredracor and https://github.com/dracor-org/theatre-classique/compare/dracor.
11. See https://dracor.org/api/corpora/ger/play/wagner-der-fliegende-hollaender/tei.

Genre	German sample	French sample
Tragedies	141	321
Comedies	197	824
Tragicomedies	8	83
Libretti (marked up)	55	58
Libretti (attributed)	28	32
Total	429	1318
Percentage of libretti	19.3%	6.8%

Table 1: Final French and German drama samples.

to various degrees, with opera[12]. After manually cleaning up the results and removing some false positives, we grouped the newly found libretti under the 'libretti (attributed)' label. As a last step, we excluded all texts that still had no 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.

4. Experiments

In our investigation, we deliberately refrained from formulating a rigid initial hypothesis on how a libretto would have looked like, i.e., about its distinctive features. Conversely, we took inspiration from standard practices in cultural analytics research, preferring instead to follow the exploratory data analysis paradigm (Tukey 1977), trying to explore the structure of libretti as embodied in clusters, networks, and relations between features (Manovich 2020, 252). To do so, we organised our research as a series of concatenated 'experiments' aimed at exploiting the epistemological potential of feature analysis to its fullest extent. More concretely, we used three distinct strategies to explore our corpora: we plotted them on a Cartesian plane according to their structural features (subsection 4.1), we deployed some statistical tests to assess the features' usefulness in distinguishing libretti from comedies and tragedies (subsection 4.2), and we created scatterplots to measure the evolution of individual metrics (subsection 4.3).

All three approaches, which allowed us to explore more thoroughly the topology of German and French drama, were based on a method recently proposed by Szemes and Vida (2024) to cluster dramatic genres according to content-independent, form-related properties. In our case, however, we did not aim to perform a classification task, but were rather interested in finding out which features (if any) set libretti apart from comedies and tragedies.

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 through the API. Based on these metrics, they carried out a supervised classification procedure on DraCor's German and Shakespeare[13] corpora to label comedies and tragedies using the Support Vector

---

12. 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.).
13. See https://dracor.org/shake.

Machine (SVM) method. They found no "striking difference" between the two genres insofar as structural features were concerned, but were nonetheless able to single out some properties which are highly predictive of generic alignment. According to their results, it is indeed possible to single out "the existence of a 'genre fingerprint' that shapes the dramatic structure of tragedies and comedies" (Szemes and Vida 2024, 10) and which authors may (un)consciously choose to adhere to or depart from.

Following a similar approach, we also took as starting point the array of features provided by the DraCor API, which can be obtained in tabular format through a direct API call.[14] Out of the 41 metrics available, we employed almost all the numerical ones, which are mostly related to the plays' size or character networks. For pragmatic reasons, we did not follow Szemes and Vida in computing additional speech-related or distributional measures. 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. 'Soldiers');[15]

- `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;[16]

- `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 path between any pair of nodes;

- `average_path_length`: average length of the shortest path that can be drawn between any two nodes.

4.1 Visualisation

As a first step in our exploratory analysis, we plotted the plays (i.e., their feature vectors) on a two-dimensional plane. To this aim, we tested various unsupervised (i.e., class-blind) techniques for dimension reduction such as Principal Component Analysis (PCA), T-distributed Stochastic Neighbour Embedding (t-SNE) and Uniform Manifold

---

14. To ensure full reproducibility of our results, we followed the procedure laid down by Börner et al. (2023b) and used the Docker technology to create a container hosting a snapshot of the DraCor platform as it was at the time of data extraction. We then run the API calls for metadata (cf. https://dracor.org/doc/api/) directly within this container, which can be reconstructed in Docker by using the Github commit numbers *82c90cd0fef330a5547d8d86058d9238e46effad* for GerDraCor and *d5bf3be983650e04e5a17c558501fea437438468* for FreDraCor.
15. This feature is available only for GerDraCor.
16. Cf. Watts and Strogatz (1998).

Approximation and Projection (UMAP) (see Burges (2010), Maaten et al. (2009), and 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 implemented through the `Scikit-learn` Python 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, it seemed not possible to confidently 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 however 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 ones (Szemes and Vida 2024, 16).

Consequently, 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 main genres. Instead of using the traditional *comic/tragic* dichotomy, 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*).[17] 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 than non-comic ones.

Accordingly, we semi-automatically extracted the new labels from the subtitles through keywords and ran the PCA algorithm again, with results being presented in Figure 1 and Figure 2. While the model performed better than before, validating to some extent our refining of libretti labels, it still failed to produce clear-cut clusterings of operatic texts; on top of that, as research by Szemes and Vida (2024) 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 timeframe (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

---

17. 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.
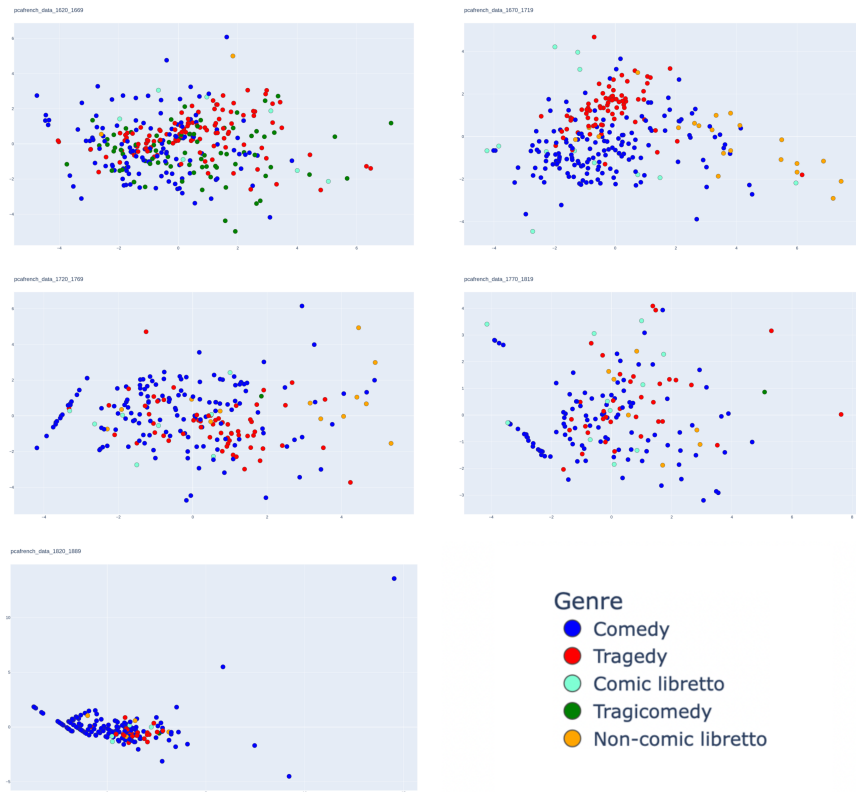
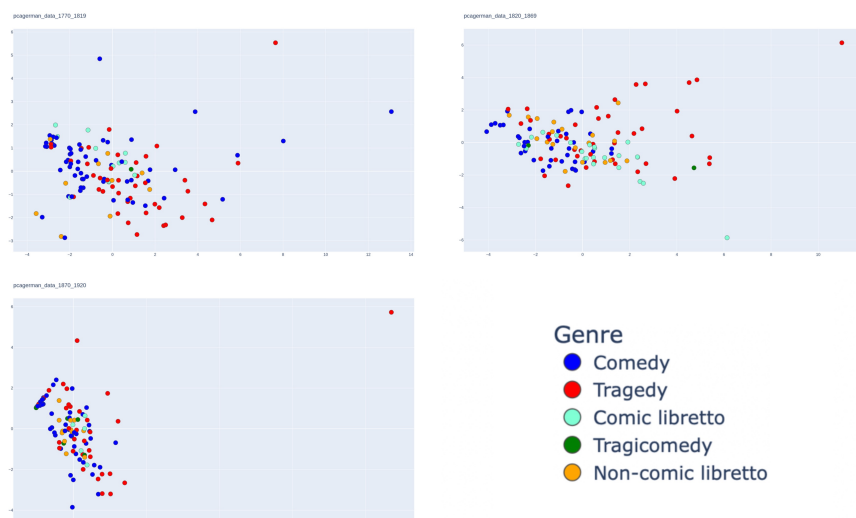**Figure 1:** Evolution of French drama, 1626-1889, visualised through PCA.

Figure 2: Evolution of German drama, 1770-1920, visualised through PCA.

Features	German sample	French sample
numOfSegments	0.62	0.2
numOfSpeakers	0.09	**8.8379471044e-10**
numOfPersonGroups	**8.304284707413e-08**	nan
wordCountSp	**2.2689986819869e-07**	**7.983046e-14**
wordCountStage	0.045007599370324045	**1.1e-19**
averageDegree	**0.0072945052710442024**	**0.04833803602286693**
density	0.58	**1.7687716293e-10**
averageClustering	0.41	0.47
maxDegree	**0.03540074483912862**	**6.9572145605217e-07**
numConnectedComponents	0.39	**1.7e-18**
diameter	0.19	**0.0004643100289270638**
averagePathLength	0.35	**0.0001684502205165606**

Table 2: Statistical significance (p-value) of features in the two samples according to the Wilcoxon Rank Sum test. Numbers in bold are the ones below the significance threshold (α = 0.05) we use to deem the difference statistically significant.

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.

4.2 Statistical Testing and Classifier

While the PCA yielded some first insights into the composition of our corpora, it also pushed us to develop a stronger understanding of the individual features composing our textual vectors. To this aim, we tried an alternative approach, which consisted in measuring the statistical significance of the features' differences and thus assessing how such parameters performed in telling libretti and non-libretti apart.[18] After applying the Shapiro-Wilk test for checking the normality of distributions – which as expected 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 2).

As an additional benchmark, we 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. At the beginning of the procedure, 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_of_person_groups`).

As the matrices (Figure 3 and Figure 4) show, some features displayed a strong correlation (>±0.75), and keeping both of them might make it difficult to understand which features actually contributed the most to the classifier's output. Following the example of Szemes and Vida (2024, 6–7), we therefore chose to keep within each pair or triplet of correlated features the ones easier to interpret, while discarding the other

---

18. From here onwards, we removed tragicomedies from our sample because of their irregular chronological distribution and globally low number.
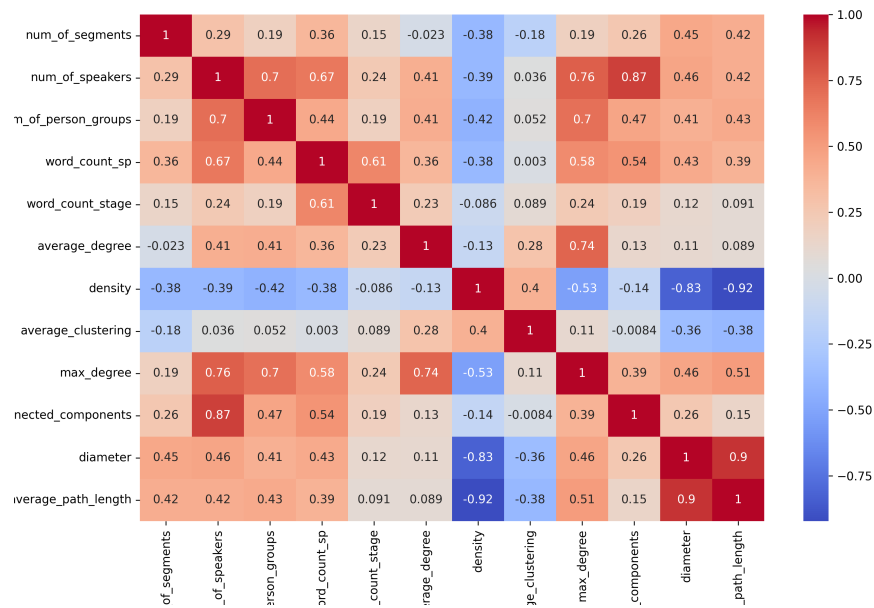
Figure 3: Correlation matrix for the German corpus.

Measures	Sample	Baseline (random)	Baseline (most frequent class)
overall accuracy	84.3%	48.7%	76.7%
precision for class *Non-libretto*	85.4%	75.7%	76.7%
recall for class *Non-libretto*	96.6%	48.9%	100%
f1 for class *Non-libretto*	90.7%	59.4%	86.8%
precision for class *Libretto*	75.6%	22.3%	-
recall for class *Libretto*	38.7%	48.2%	0%
f1 for class *Libretto*	51.2%	30.4%	-

Table 3: Values for the best performing random forest classifier, German sample.

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 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.

After running the model, results were unsatisfying: 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 3 and Table 4). Our intention, however, was not to find a method to efficiently automate libretti classification, but rather to ascertain whether the most relevant features for the classification task (presented in Figure 5) were the same ones whose variance was statistically significant according to our tests.

Crossing the results of the two pipelines (statistical significance tests and classifier), one could eventually argue that four features (*num_of_person_groups*, *word_count_sp*, *word_count_stage*, *average_degree*) were particularly helpful for distinguishing German libretti from non-libretti, while six of them proved useful for sorting out the French data
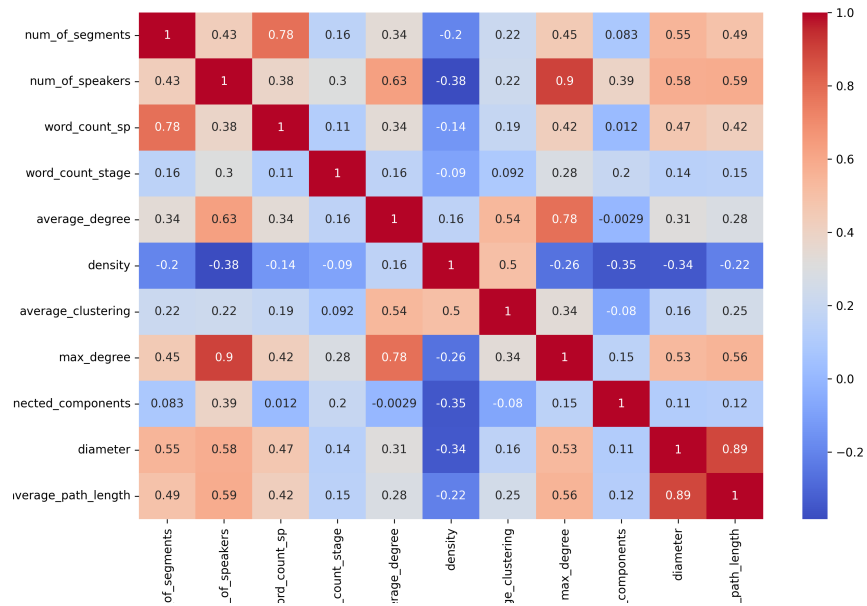
Figure 4: Correlation matrix for the French corpus.

Measures	Sample	Baseline (random)	Baseline (most frequent class)
overall accuracy	93.1%	49.7%	92.7%
precision for class *Non-libretto*	94.1%	93.3%	92.7%
recall for class *Non-libretto*	98.7%	49.3%	100%
f1 for class *Non-libretto*	96.3%	64.5%	96.2%
precision for class *Libretto*	68.2%	0.08%	-
recall for class *Libretto*	31.1%	55.6%	0%
f1 for class *Libretto*	42.7%	13.8%	-

Table 4: Values for the best performing random forest classifier, French sample.

(*num_of_speakers*, *word_count_sp*, *word_count_stage*, *density*, *diameter*, *num_connected_components*).

4.3 Evolutionary Trends

To offer a qualitative interpretation of these features and better understand their importance in distinguishing libretti at different stages of history, we decided to take a closer look by charting some of them as scatterplots.[19] In doing so, we followed the approach already informally proposed by Trilcke et al. (2015), who tracked the progression of two significant network metrics (size and density) in an early version of the GerDraCor corpus and found out a "proximity of comedy and libretto and [a] persistent distance from the tragedy". In our case, we switched again to a four-class visualisation (*comedies/tragedies/comic libretti/non-comic libretti*) in order to better capture the granularity of the process. We plotted each play individually and applied a local regression algorithm (LOWESS, Cleveland (1979)) to draw a smooth curve between the data points and help visualise the distances between the genres and their evolution.

19. We removed outliers which were more than three standard deviations away from the mean.
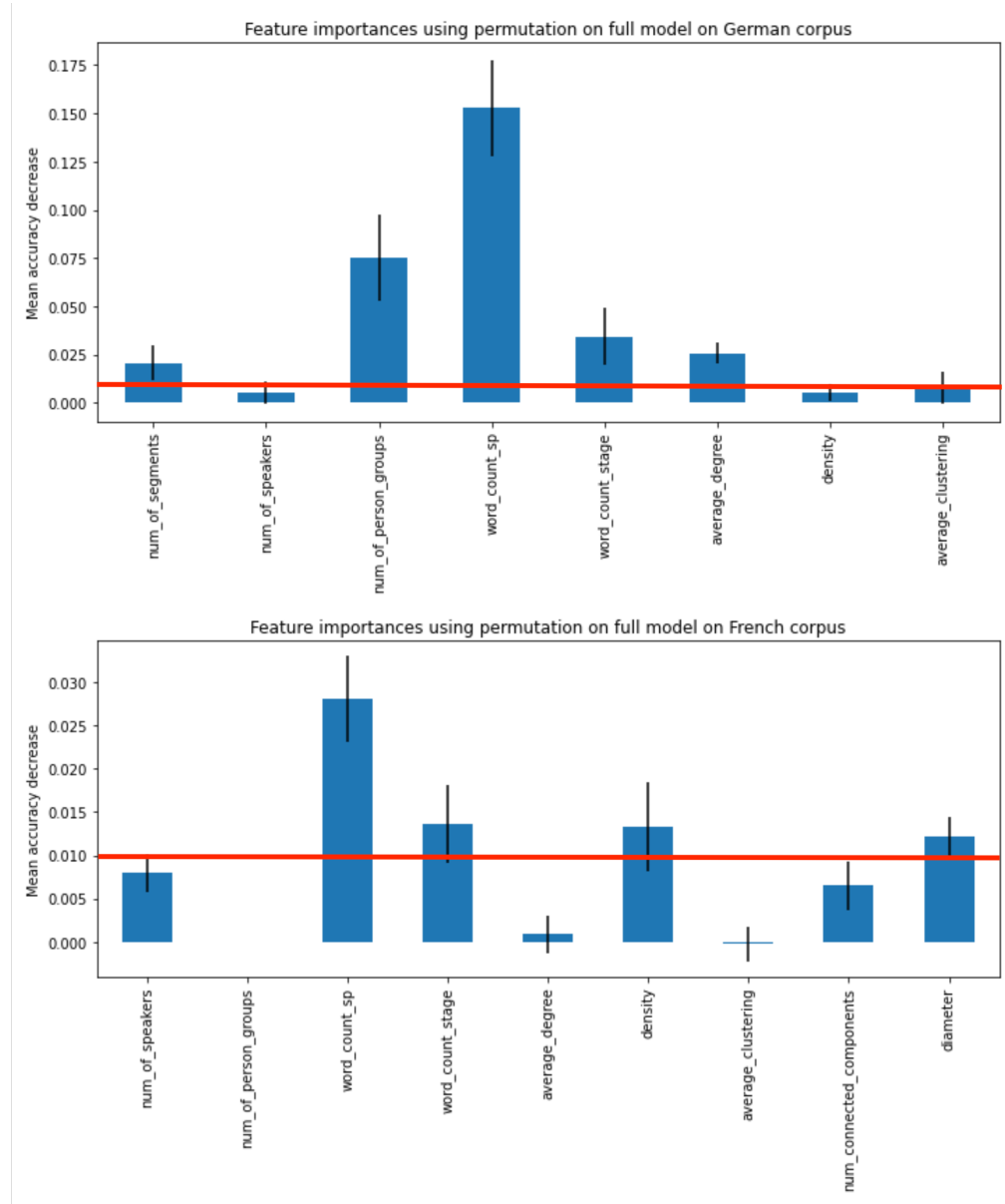
Figure 5: Relative importance of features according to the random forest classifier. The barplot indicates which features, if left out, lead to the biggest accuracy decrease. We deemed relevant features which cause an accuracy decrease equal to or superior than 0.01 (the red line), with confidence intervals taken into account.

Although the unequal distribution of texts across the investigated time span suggests caution in speaking of *longue-durée* evolutionary phenomena, trends emerging from the pictures give a glimpse into some long-lasting relations between genres. While aware of the limitation of such a hermeneutical tool in DH practice, as outlined by Moretti and Sobchuk (2019), we believe trendlines might still help to recognise patterns seemingly common to both German and French data. Most notably, one recognises the relative independence of non-comic libretti from the other genres in terms of several structural features.

To begin with, such a 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 a 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 highly interconnected characters, while non-comic libretti display an unusual dissimilarity from any other genre.

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 categories. 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 patterns especially visible in the non-comic libretti curves, coupled with the clustering evidence emerging from the PCA, underscore the particular relevance of the second half of the XVII century – a fact that cannot be explained only by the higher number of texts available in that timeframe. Indeed, critics consider it as a pivotal moment in the diffusion of opera beyond Italy and 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., mostly 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 corpus plays written between 1670 and 1719. Closer inspection of some of these operas, such as *Cadmus et Hermione* (1673) or *Persée* (1682),[20] confirms the aforementioned quantitative findings: these works feature large ensemble casts whose characters often appear together, thus resulting in well-connected social networks.

---

20. See https://dracor.org/fre/quinault-cadmus-hermione; https://dracor.org/fre/quinault-persee.
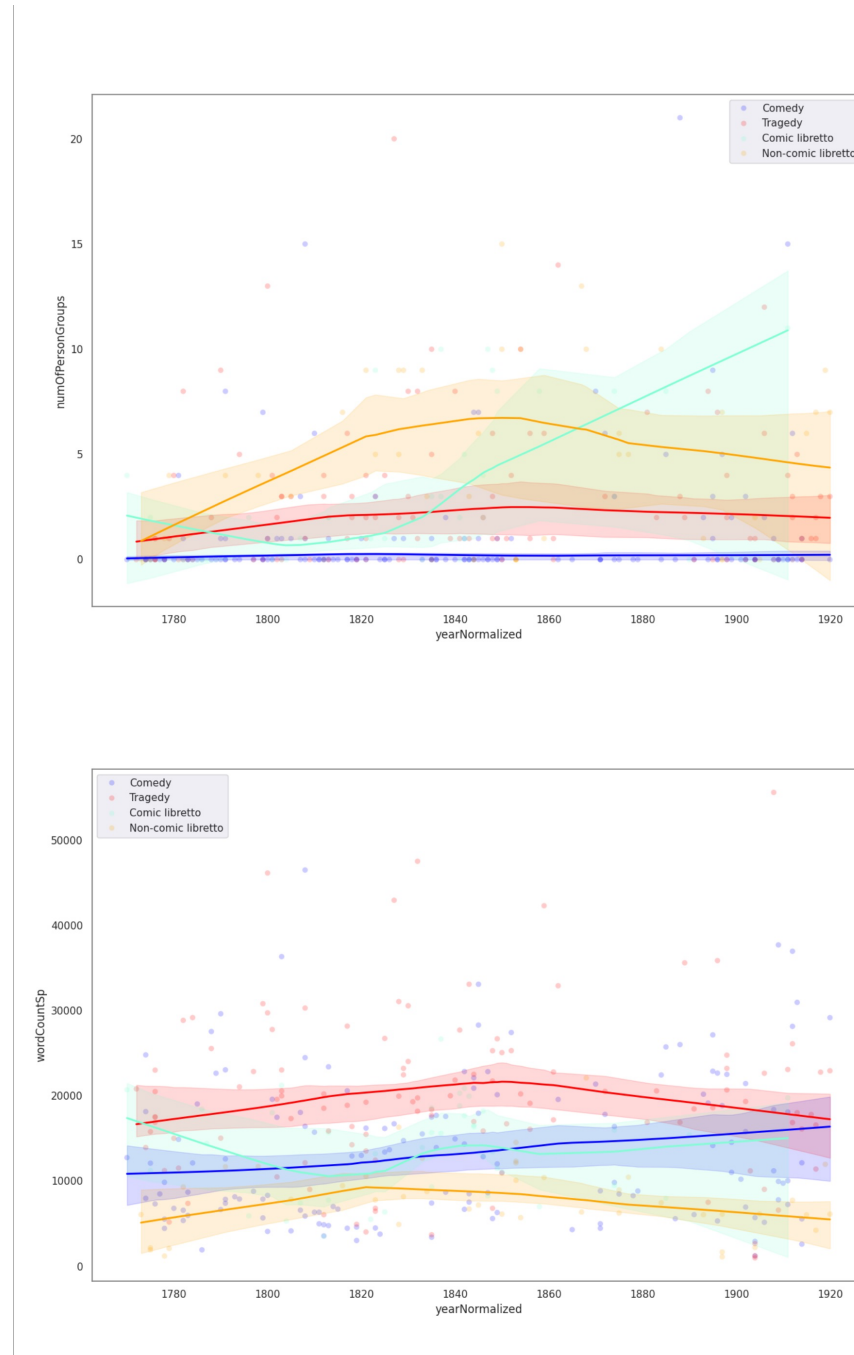
Figure 6: Evolution of selected features in German data: *word_count_sp* (below) and *num_of_groups* (above). The numbers in the graph indicate the number of plays in this timeframe.
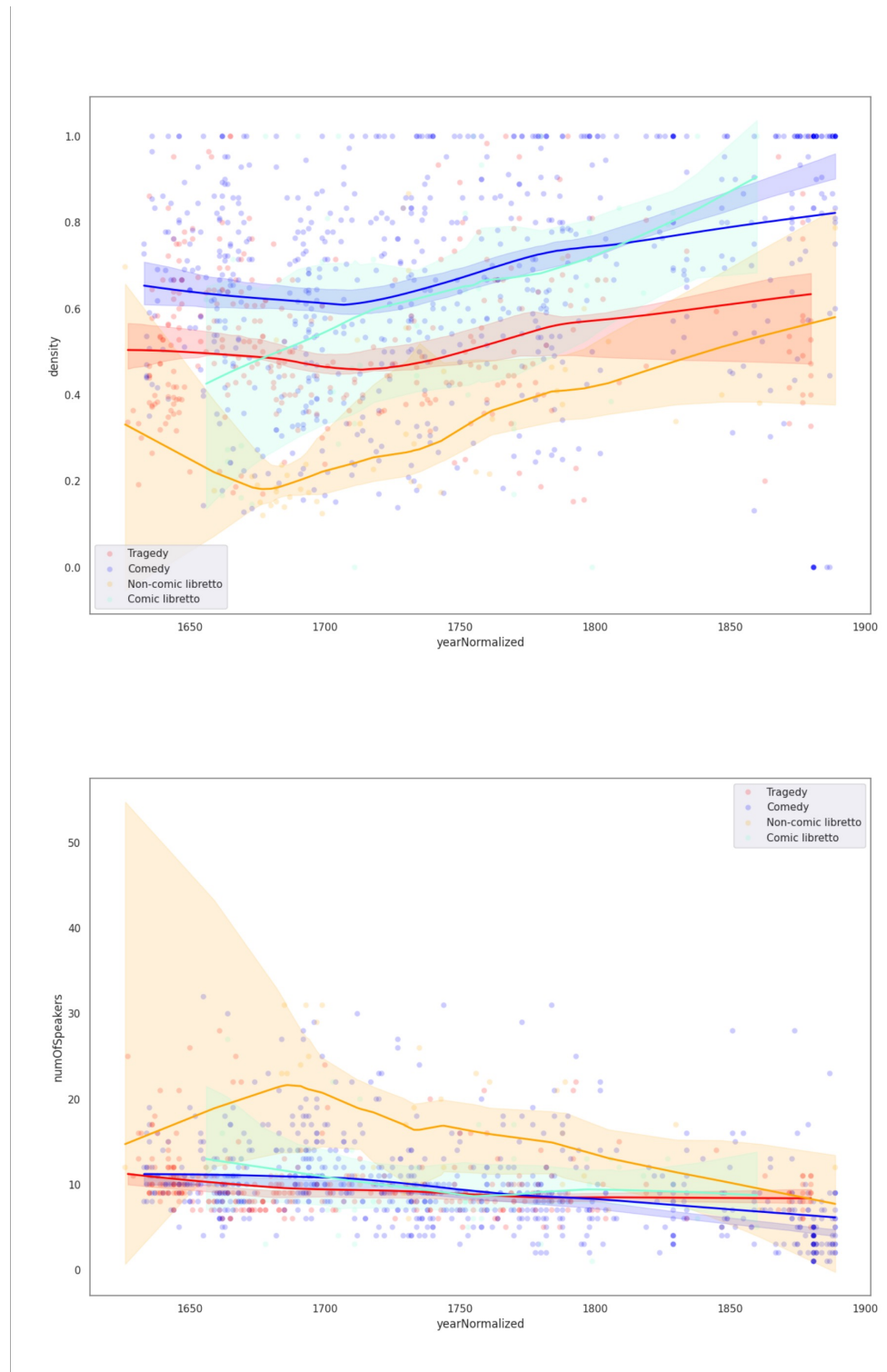
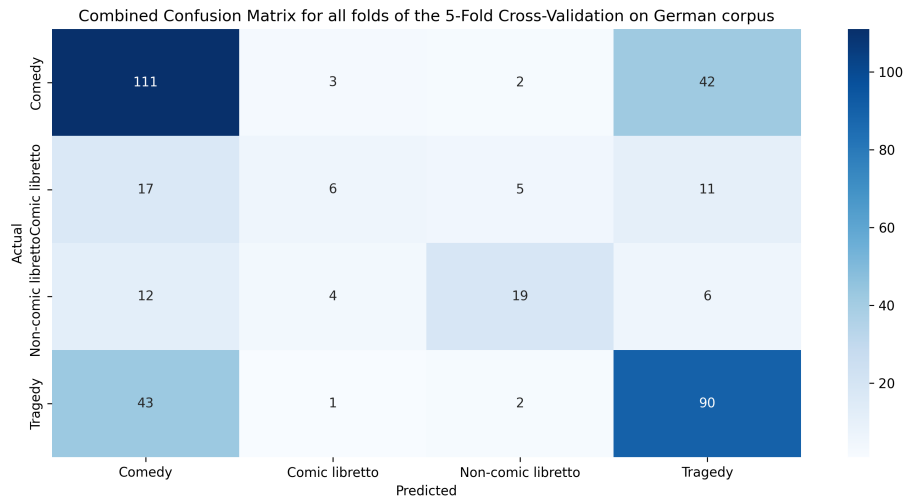**Figure 7:** Evolution of selected features in French data (I): *density* (above) and *num_speaker* (below).

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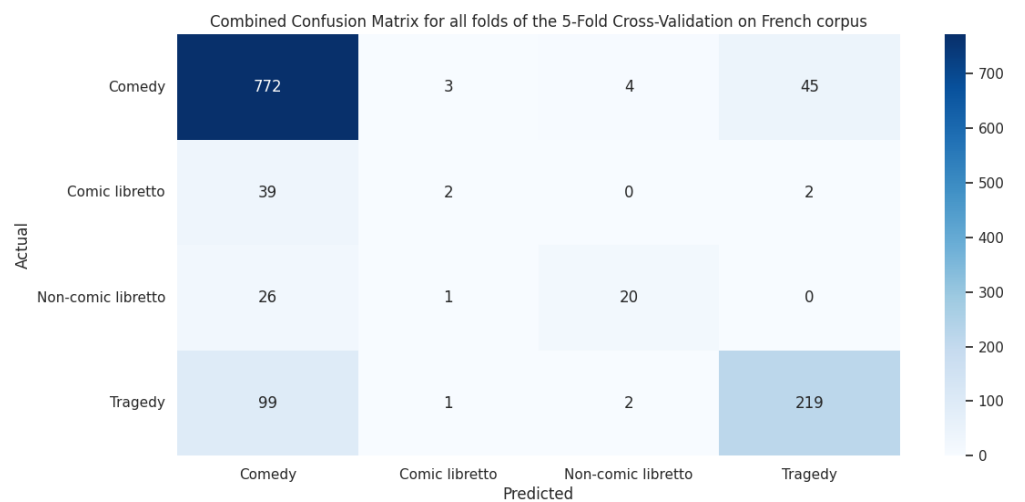

Figure 9: Confusion matrix for a four-class classifier trained and tested on the French sample.

Due to higher text availability, French scatterplots are also useful for empirically verifying traditional assumptions on opera by literary critics and musicologists. 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, these trends seem indeed confirmed in both kinds of libretti: comic and non-comic operatic texts follow similar paths in having less character speech and (sizeably) more stage directions.[21]

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 into the dynamics of the age, the results obtained in others might be radically changed by any corpora overhaul, be it in terms of corpus enlargement[22] or markup refining.

The limited text availability also forced us to focus on only two cultural milieus (German and French) and to 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[23] contains so far only a handful of libretti (mostly early *melodrammi* by 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[24] to perform more encompassing analyses, while possibly enlarging ItaDraCor as well (through the onboarding procedure described in Börner et al. (2023a); see also Giovannini et al. (2023)).
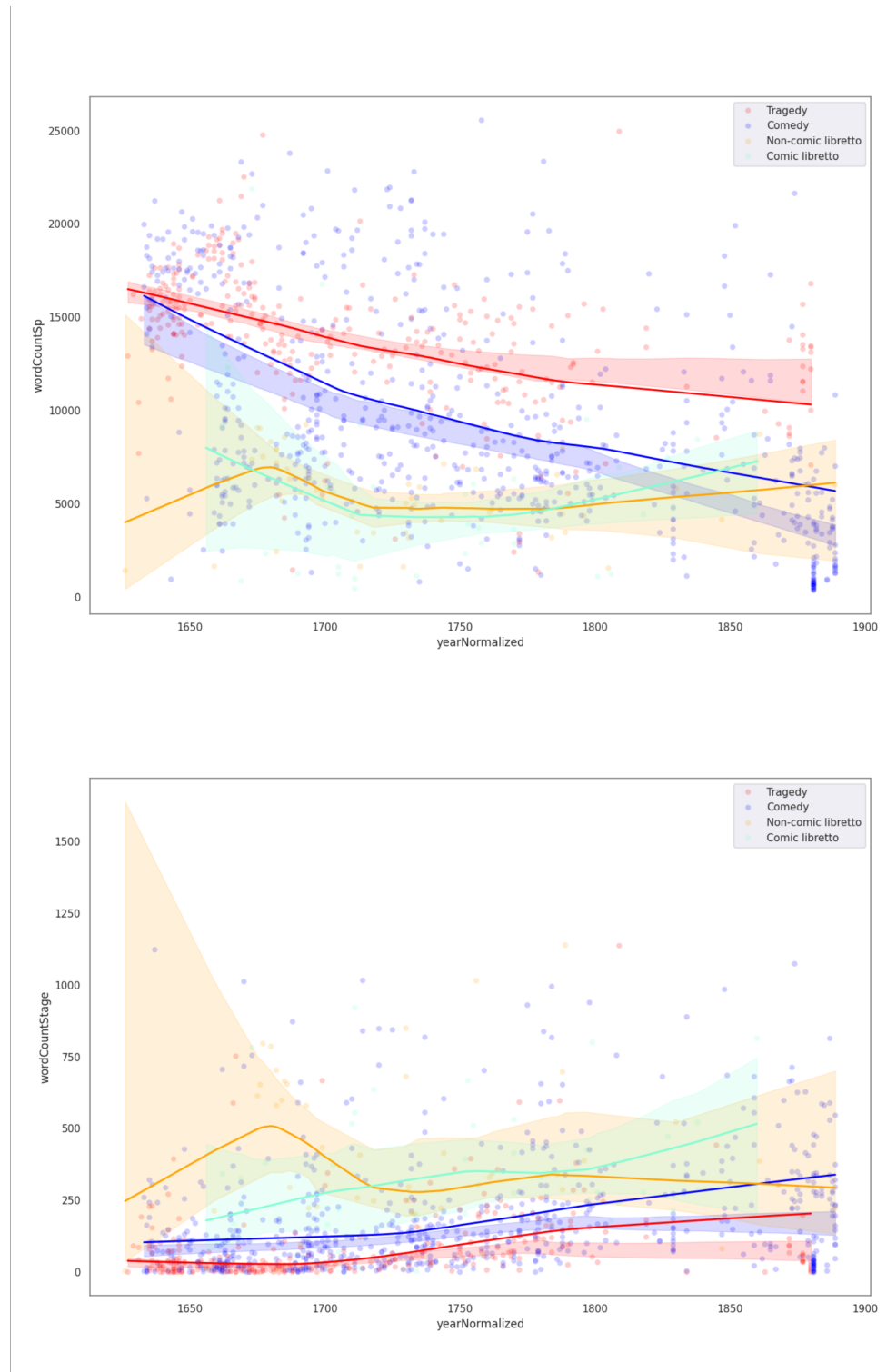
Eventually, much more work is needed in order to achieve a satisfying diachronic picture of opera in its relationship with other dramatic genres. Nonetheless, our first computational foray into operatic texts yielded some insights on libretti as part of the wider dramatic system. On one side, it was possible to single out idiosyncratic features defining the genre's local iterations in the French- and German-speaking lands. Furthermore, taking into account the libretti's generic alignment showed how the two

---

21. Coincidentally, the general increase in the share of stage directions for all genres – except for non-comic libretto – supports Trilcke et al. (2020)'s argument about a 'tendency to epification' in the history of drama even beyond their original case study (GerDraCor).
22. For example, about 100 new plays have been added to GerDraCor since the paper's first submission, and while we do not expect the overall results to change much, we plan to rerun our algorithms and publish updated results in the project's Github repository.
23. See https://dracor.org/ita.
24. 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.

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).

main types (*comic/non-comic*) often displayed different behaviours as far as structural features are concerned.

More generally, the analysis of the different timeframes through PCA clusterings and feature lineplots also suggested that it is more difficult to discriminate effectively between the two kinds of libretti in German data, while such distance is more substantial within the French dramatic 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 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 highlights once again the complexity of modelling the relationship between different dramatic genres – and, more generally, the concept of dramatic text itself – in terms of structural features alone. A possible line of improvement, of course,

would have been to refine our metrics, pragmatically chosen due to their direct availability through the DraCor API, in order to address different dimensions of drama (e.g. plot, dialogue, characters, etc.) that might be relevant for the identification of libretti subtypes.[25] Broadly speaking, however, we have to acknowledge that the clustering effectiveness and explanatory power of this methodology does not match up (yet) to those of other content-based CLS techniques, such as topic modelling, which have been applied to similar corpora in the past (see e.g. Schöch (2017) on French Classical and Enlightenment plays).

Nonetheless, we argue that there is still room for improvement, and that structure-oriented methods could eventually yield even more productive results if coupled with a better operationalisation of the investigated concept. This could also be achieved, for example, by referring to the key components of opera as singled out by Gier (2000),[26] or by expanding the metrics pool through existing or newly-concocted measures.[27] Still, our experiment has shown how even a limited set of structural features can help to gain insights into a notably complex literary form such as the libretto. From this starting point, modelling of operatic texts can be further enhanced and refined to achieve a more nuanced understanding of libretto morphology and its relation with the major dramatic genres.

6. Data Availability

Data and scripts employed can be found here: https://github.com/DanilSko/opera (DOI: https://doi.org/10.5281/zenodo.8356601).

7. Acknowledgements

The authors would like to thank Artjoms Šeļa, Henny Sluyter-Gäthje, and Peer Trilcke for their helpful suggestions. We are also thankful to the anonymous reviewers and the conference participants for their insightful comments.

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), 751–782. 10.1353/nlh.2017.0038.

---

25. For an experiment in this sense see Giovannini (2024).
26. The five components identified by Gier (2000, 14) are brevity, a discontinuous time structure (arias slow the dramatic progression, recitatives speed it up), the independence of parts, a contrast-based structure (opposition as a central theme), and a "primacy of the perceivable" (meaning is conveyed at multiple levels due to the genre's intrinsic multimediality). Cf. also Overbeck (2011, 16–17).
27. An additional metric could be created, for example, by computing the proportion between prose and verse, which might capture well the libretti's bipartite structure (recitatives vs. arias).

Andresen, Melanie and Nils Reiter, eds. (2024). *Computational Drama Analysis. Reflecting Methods and Interpretations*. De Gruyter. `10.1515/9783111071824`.

Bonora, Paolo and Angelo Pompilio (2021). "Estrazione automatica delle caratteristiche del personaggio d'opera attraverso pattern lessico-sintattici". In: *Umanistica Digitale* 10, 193–210. `10.6092/issn.2532-8816/12426`.

Börner, Ingo, Frank Fischer, Luca Giovannini, Christopher Lu, Carsten Milling, Daniil Skorinkin, Henny Sluyter-Gäthje, and Peer Trilcke (2023a). "Onboard onto DraCor: Prototyping Workflows to Homogenize Drama Corpora for an Open Infrastructure". In: *DHd 2023 Conference Abstracts*. `10.5281/zenodo.7711513`.

Börner, Ingo, Peer Trilcke, Carsten Milling, Frank Fischer, and Henny Sluyter-Gäthje (2023b). "Dockerizing DraCor – A Container-Based Approach to Reproducibility in Computational Literary Studies". In: *DH2023 Book of Abstracts*, 293–295. `10.5281/zenodo.8107836`.

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. `10.1093/gmo/9781561592630.article.40726`.

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

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

Cleveland, William S. (1979). "Robust Locally Weighted Regression and Smoothing Scatterplots". In: *Journal of the American Statistical Association* 74 (368), 829–836. `10.1080/01621459.1979.10481038`.

Cuéllar, Álvaro (in press). "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. Preprint available. Heidelberg University Press. `https://www.academia.edu/53262871/Stylometry_and_Spanish_Golden_Age_Theatre_An_Evaluation_of_Authorship_Attribution_in_a_Control_Group_of_One_Hundred_Undisputed_Plays` (visited on 05/15/2024).

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). `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: *DH2019 Book of Abstracts*. `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*. `https://hal.science/hal-01811799` (visited on 05/15/2024).

Fischer, Frank, Mathias Göbel, Dario Kampkaspar, Christopher Kittel, and Peer Trilcke (2017b). "Network Dynamics, Plot Analysis: Approaching the Progressive Structuration of Literary Texts". In: *DH2017 Book of Abstracts*. `https://dh2017.adho.org/abstracts/071/071.pdf` (visited on 05/15/2024).

Gervás, Pablo and Álvaro 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*. `http://nil.fdi.ucm.es/sites/default/files/KDIR2022-opera-CRC.pdf` (visited on 05/15/2024).

Gier, Albert (2000). *Das Libretto: Theorie und Geschichte einer musikoliterarischen Gattung*. Insel.

Giovannini, Luca (2024). "Quantitative Ansätze zur Untersuchung des frühneuzeitlichen Dramas". In: *DHd2024 Book of Abstracts*. Universität Passau. `10.5281/zenodo.10698266`.

Giovannini, Luca, Ingo Börner, Frank Fischer, Carsten Milling, Daniil Skorinkin, and Peer Trilcke (2023). "Quali prospettive per ItaDraCor? Risorse e strumenti per la codifica di testi drammatici in lingua italiana". In: *AIUCD2023 Book of Abstracts*. Università di Siena, 108–111. `https://iris.unive.it/retrieve/0f226d38-e332-418b-9b14-d5558d1a0d9d/AIUCD2023.pdf` (visited on 05/29/2024).

Gorlée, Dinda (1997). "Intercode Translation: Words and Music in Opera". In: *Target. International Journal of Translation Studies* 9 (2), 235–270. `10.1075/target.9.2.03gor`.

Jeong, Harim and Joo Hun Yoo (2022). "Opera Clustering: K-means on Librettos Datasets". In: *Journal of Internet Computing and Services* 23 (2), 45–52. `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*. `10.1155/2022/1167462`.

Kesting, Hanjo (2005). *Der Musick gehorsame Tochter: Mozart und seine Librettisten*. 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. `10.17175/2022_012`.

Leopold, Silke (2003). "Die Anfänge von Oper und die Probleme der Gattung". In: *Journal of the Seventeenth Century Music* 9 (1). `https://sscm-jscm.org/v9/no1/leopold.html` (visited on 05/15/2024).

Louvat-Molozay, Bénédicte (2018). "Théâtre et musique". In: *Le théâtre au miroir des langues: France, Italie, Espagne XVIe–XVIIe siècles*. Ed. by Marc Vuillermoz Véronique Lochert and Enrica Zanin. Droz.

Maaten, Laurens van der, Eric O. Postma, and Jaap van den Herik (2009). *Dimensionality Reduction: A Comparative Review*. Preprint. `https://lvdmaaten.github.io/publications/papers/TR_Dimensionality_Reduction_Review_2009.pdf` (visited on 05/15/2024).

Manovich, Lev (2020). *Cultural Analytics*. The MIT Press.

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

Moretti, Franco (2013). *Distant Reading*. Verso.

Moretti, Franco and Oleg Sobchuk (2019). "Hidden in Plain Sight: Data Visualization in the Humanities". In: *New Left Review* (118), 86–115. `https://newleftreview.org/issues/ii118/articles/franco-moretti-oleg-sobchuk-hidden-in-plain-sight` (visited on 05/15/2024).

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*, 530–536. `10.1109/IV51561.2020.00091`.

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

Overbeck, Anja (2011). *Italienisch im Opernlibretto. Quantitative und qualitative Studien zu Lexik, Syntax und Stil*. De Gruyter. `10.1515/9783110258349`.

Reichert, Waltraud (1964). "Kybernetische Methoden der Dramenforschung". In: *Grundlagenstudien aus Kybernetik und Geisteswissenschaften* 5 (3-4), 115–120.

Schöch, Christof (2017). "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama". In: *Digital Humanities Quarterly* 11 (2). `http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html` (visited on 05/15/2024).

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

Szemes, Botond and Bence Vida (2024). "Tragic and Comical Networks: Clustering Dramatic Genres According to Structural Properties". In: *Computational Drama Analysis. Reflecting Methods and Interpretations*. Ed. by Melanie Andresen and Nils Reiter. De Gruyter. `10.1515/9783111071824-009`.

Trilcke, Peer, Frank Fischer, Mathias Göbel, and Dario Kampkaspar (2015). *Comedy vs. Tragedy: Network Values by Genre*. `https://dlina.github.io/Network-Values-by-Genre` (visited on 05/15/2024).

Trilcke, Peer, Christopher Kittel, Nils Reiter, Daria Maximova, and Frank Fischer (2020). "Opening the Stage – A Quantitative Look at Stage Directions in German Drama." In: *DH2020 Book of Abstracts*. `https://dh2020.adho.org/wp-content/uploads/2020/07/337_OpeningtheStageAQuantitativeLookatStageDirectionsinGermanDrama.html` (visited on 05/15/2024).

Trilcke, Peer, Evgeniya Ustinova, Frank Fischer, Carsten Milling, and Ingo Börner (2024). "Detecting Small Worlds in a Corpus of Thousands of Theater Plays: A DraCor Study in Comparative Literary Network Analysis". In: *Computational Drama Analysis. Reflecting Methods and Interpretations*. Ed. by Melanie Andresen and Nils Reiter. De Gruyter. `10.1515/9783111071824-002`.

Tukey, John (1977). *Exploratory Data Analysis*. Pearson.

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

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

Weisstein, Ulrich (1961). "The Libretto as Literature". In: *Books Abroad* 35 (1), 16–22. `10.2307/40115290`.