Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama

Christof Schöch (University of Würzburg)

To appear in: *Digital Humanities Quarterly*
http://digitalhumanities.org/dhq/

Abstract

The concept of literary genre is a highly complex one: not only are different genres frequently defined on several, but not necessarily the same levels of description, but consideration of genres as cognitive, social, or scholarly constructs with a rich history further complicate the matter. This contribution focuses on thematic aspects of genre with a quantitative approach, namely Topic Modeling. Topic Modeling has proven to be useful to discover thematic patterns and trends in large collections of texts, with a view to class or browse them on the basis of their dominant themes. It has rarely if ever, however, been applied to collections of dramatic texts.

In this contribution, Topic Modeling is used to analyse a collection of French Drama of the Classical Age and the Enlightenment. The general aim of this contribution is to discover what types of semantic coherence topics show in this collection, whether different dramatic subgenres have distinctive dominant topics and plot-related topic patterns, and inversely, to what extent clustering methods based on topic scores per play produce groupings of texts which agree with more traditional genre distinctions. This contribution shows that interesting topic patterns can be detected which provide new insights into the thematic, internal structure of a genre such as drama as well as into the history of French drama of the Classical Age and the Enlightenment.

Introduction

The concept of literary genre is considered highly complex for several reasons. First, genres can be defined and described on a number of levels of description, such as plot, theme, personnel, structure, and style, with style in turn concerning a range of aspects, such as spelling, lexicon, morphology, semantics, and syntax as well as rythm, imagery, or complexity (for an overview, see herrmann2015). Second, related genres are frequently defined on several different levels, such as length, form or theme: for example, subgenres of narrative prose fiction include the short story, the epistolary novel, and the libertine novel, creating overlap if not contradictions. Finally, consideration of genres as cognitive, social, and of course scholarly constructs with a rich history further complicate the matter, raising the question what part formal features, on the one hand, and tradition and construction, on the other, play in the perception of literary genres (see hempfer1973, schaeffer1989).

The wider context of this contribution is the junior research group on Computational Literary Genre Stylistics (CLiGS). There, we consider it of interest to analyse genres (and especially subgenres of drama and the novel) on a wide range of levels of description, stylistic as well as structural and thematic, and ranging from the use of function words all the way to plot structures. A particular focus lies on how these various levels correlate and interact, and how they evolve over time in a given generic subsystem. The present contribution is one brick in that building, laying the focus on thematic aspects of genre and using Topic Modeling. This technique has proven to be useful to discover thematic patterns and trends in large collections of texts. Here it is applied, as has rarely

been done so far, to collections of dramatic texts. *[Note: schmidt2014 has used topic modeling for the analysis of screenplays of TV shows, which may be considered a genre related to the more traditional dramatic texts considered here.]*

In this contribution, Topic Modeling is used to analyse a collection (further described below) of French Drama of the Classical Age and the Enlightenment. This is a well-researched domain in French literary studies (for an account of the genre's formal poetics, see, for example, scherer2001, and for a thorough overview, mazouer2010). The general aim of this contribution is to assess whether Topic Modeling can be a useful, quantitative complement to established, qualitative methods of analysis of French dramatic texts from this period. More specifically, the contribution would like to discover: (1) what types of semantic coherence topics show in this collection, knowing that topic modeling applied to fictional texts usually yields less abstract topics than when applied to non-fictional texts; (2) whether different dramatic subgenres have distinctive dominant topics, and if yes, whether those topics are expected or suprising, specific or vague, and abstract or concrete; (3) whether dramatic subgenres have distinctive plot-related topic patterns, and if yes, whether they concern the generically distinctive topics or others; (4) and finally, to what extent clustering methods based on topic scores produce groupings of texts which agree with more traditional genre distinctions.

Hypotheses

Based on what we know about the history of French drama on the one hand, and the basic principles of topic modeling on the other hand, we can formulate a number of hypotheses or questions concerning the relations between topics and dramatic genres and subgenres.

First of all, dramatic subgenres such as comedy or tragedy being in part defined on the basis of their themes, and topic modeling bringing to the fore the hidden thematic structure of text collections, it can be expected that topic modeling applied to a collection of dramatic texts from a small range of different subgenres (comedies, tragedies and tragicomedies, in the present case) should bring out relatively strong genre-related topic patterns in the data.

More specifically, and because there is only a small number of subgenres in the collection, there should be a relatively large proportion of topics which are clearly distinctive of one of the subgenres involved. It will be interesting to note which of the topics will be most distinctive of comedies and tragedies: for instance, will they be clearly thematic topics, or show some other type of coherence? Will they be topics which we would expect to be characteristic of tragedies and comedies written in the seventeenth and eighteenth centuries (such as royalty vs. bourgoisie, honor vs. love, etc.), or will they be unexpected? It will also be of interest to investigate the topic-wise position of tragicomedy, which may either turn out to mix topics of both comedy and tragedy in a specific manner, or may also contain distinctive topics of its own.

Another aspect of the relations between topic and genre concerns plot. We know that on a very fundamental level, comedies and tragedies from the period studied here are distinguished by their typical plot patterns: tragedies tend to have a final act in which a lot of the protagonists are defeated or die, with conflictual power-relations and violent crime dominating. Comedies tend to have a final act leading up to one or several marriages, that is with a triumph of socially-accepted love relationships and happiness. If, as can be expected, there are topics related to such themes or motives, we should see a pattern across textual progression showing an increased importance of such topics towards the end of tragedies and comedies, respectively.

Finally, it is possible to invert the perspective from *a priori* categories and their distinctive characteristics to a data-driven, entirely unsupervised grouping of texts into (potentially genre-related) clusters. If topics turn out to be strongly distinctive of genre, then it can also be expected that clustering based on scores of topic proportions per play should result in groupings strongly related to genre. However, it remains to be seen whether such groupings confirm traditional genre-

related divisions or diverge from them.

Data

The data used in this study comes from the Théâtre classique collection maintained by Paul Fièvre (fievre2007). This continually-growing, freely available collection of French dramatic texts currently contains 750 plays published between 1610 and 1810, thus covering the Classical Age and the Enlightenment. The majority of the texts are based on early editions made available as digital facsimiles by the French national library (BnF, Bibliothèque nationale de France). The quality of the transcriptions is relatively good without always reaching the consistent quality expected of more formally edited scholarly editions. However, the plays contain detailed structural markup applied in accordance with the Guidelines of the Text Encoding Initiative (TEI P4; for a general introduction, see burnard2014. For example, the plays' structure with respect to act and scene divisions as well as speeches, speaker names, stage directions and many other phenomena are all carefully encoded. In addition, detailed metadata has been added to the texts relating, for instance, to their historical genre label (e.g. comédie héroique, tragédie, or opéra-ballet) as well as the type of thematic and regional inspiration (e.g. French history, Roman mythology, or Spanish mores). A large part of this information can fruitfully be used when applying topic modeling to this text collection.

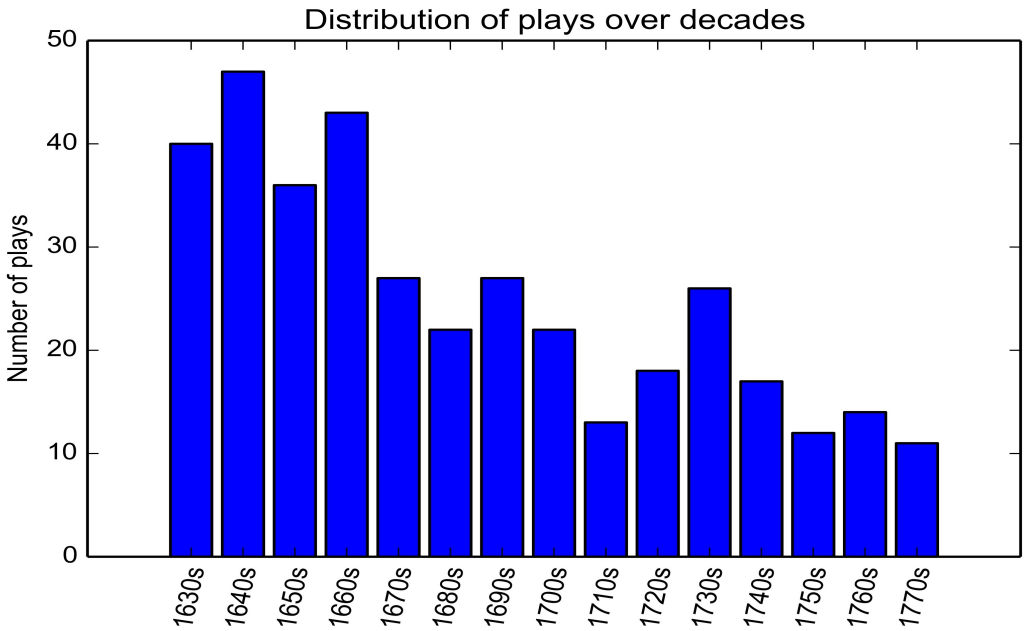

Figure 1 Distribution of plays

The collection of plays used for the research presented here is a subset of the Théâtre classique collection, defined by the following criteria: Only plays from the period between 1630 and 1780 were included, because the collection contains only a comparatively small number of plays for the decades before and after these dates. Plays have only been included if they have between three and five acts, effectively excluding a certain number of one-act plays, in order for the plays to have a similar length and a comparable plot structure. Finally, all the plays included belong to one of the following subgenres: comedy, tragedy or tragi-comedy. For most of the other relevant genres, only a relatively small number of examples is available. These criteria resulted in a collection of 375 plays (154 comedies, 181 tragedies, and 40 tragicomedies in verse or in prose), corresponding to approximately 5.4 million word tokens or 29.8 MB of plain text. Their distribution over the time period covered by this study is shown in Figure 1. As can be seen, the coverage is not entirely balanced, with more plays from the seventeenth than from the eighteenth century, but there is a mimimum number of 10 plays per decade. This collection is probably close to the lower bound of the amount of text necessary for useful topic modeling results which relies on a certain amount of data to play out its strengths. Also, it should be noted that this collection corresponds neither to a

random nor to a representative and/or balanced sample of all dramatic works produced during the period in question, which are more varied in length and more diverse in genre, and whose total number can only be estimated. As an indication, one may consider the registry of known plays from the period 1620 to 1720 maintained by the Théâtre classique project, which currently contains 1914 plays for a 100-year period.

Method

In the following, two aspects of the method used here are described. First, a brief explanation of topic modeling will introduce readers unfamiliar with the procedure to some of its most basic assumptions. However, topic modeling itself is just one step in a larger workflow involving both preprocessing and post-processing. Therefore, the more general processing workflow is also briefly described, with a focus on some of the decisions that need to be made with respect to several parameters used in the procedure.

Topic Modeling

Topic Modeling is an unsupervised method used to discover latent semantic structure in large collections of texts (for an introduction, see blei2011). In practice, individual words with the highest scores in a given topic are assumed to be semantically related words. This does not mean they must all belong to a common abstract theme (such as justice or biology). Especially in literary texts, it is also common for the shared semantic basis of words in a topic to be a particular setting (such as interiors or natural landscape), a narrative motive (such as horse-riding or reading and writing letters), or a social group of characters (such as noblemen or family members). However, the basis of similarity can also be some other aspect, such as the fact that all words are character names, or that all words come from a certain register (such as colloquial words) or from a distinct language (such as Latin terms in otherwise French text). This fact somewhat qualifies the general assumption that the topics with the highest scores in a given text represent that text′s major themes. Also, literary texts do not necessarily treat their dominant themes explicitly: unlike an essay or a research paper, a novel can be about social injustice, or a poem about death, without using these specific terms explicitly, showing them through concrete examples rather than explaining them through conceptual discussion.

On a slighly more technical level, a topic is a probability distribution over word frequencies; in turn, each text is characterized by a probability distribution over topics. Topic modeling is an entirely unsupervised method which discovers the latent semantic structure of a text collection without using lexical or semantic resources such as electronic dictionaries. This means that topic modeling is not only language-independent, but also independent of external resources with potential built-in biases. Rather, topic modeling is based on assumptions about language first developed in distributional semantics, whose basic tenet is that the meaning of a word depends on the words in whose context it appears. As John R. Firth famously put it in 1957, ‘a word is characterized by the company it keeps’. In line with this idea, the highest-ranked words in a topic are those words which frequently occur together in a collection of documents. A second, related assumption of topic modeling is a specific view of how the writing process is envisioned. In this view, text is generated from several groups of semantically related terms which are chosen, in different proportions for each text, when the text is written. Topic modeling reconstructs, based on the resulting text alone, which words must have been in which group and which probability they had of being selected in the writing process (see steyvers2007). Because this is a model with a very high number of unknown variables, it cannot be solved deterministically. Rather, it is solved with an initial random or arbitrary distribution of values which is then iteratively improved until a certain level of convergence between the texts predicted by the model and the actual texts in the collection analysed is reached. This also means that the results of topic modeling a given collection, with identical parameters, may not yield exactly identical results every time the technique is applied, although

generally speaking, it is a rather robust technique.

The most commonly used implementation of Topic Modeling uses an algorithm called Latent Dirichlet Allocation (blei2003), but there are several precursors (such as Non-Negative Matrix Factorization) and an increasing number of alternative algorithms. Also, besides the most commonly used tool, MALLET (mccallum2002), which is written in Java, several other tools are available, such as gensim (rehurek2010) and lda (riddell2014b), both written in Python. Several extensions to topic modeling have been proposed, such as hierarchical topic modeling (blei2004) and supervised or labeled topic modeling (ramage2009), for which, however, no ready-to-use implementation is currently available to the application-oriented research community. Due to the availability of relevant tools and tutorials (e.g. graham2012 or riddell2014a), and because it answers to the wish of many scholars in the humanities to gain a semantic access to large amounts of texts, Topic Modeling has proven immensely popular in Digital Humanities (for applications, see e.g. blevins2010, rhody2012, or jockers2013).

The Topic Modeling Workflow

Topic Modeling has been performed as part of a larger processing workflow described in this section (Figure 2 provides an overview of the process). The workflow is almost entirely automated using a custom-built set of Python scripts called tmw, for Topic Modeling Workflow. *[Note: The module tmw uses lxml to read the XML-TEI encoded files, calls TreeTagger and Mallet via the subprocess module, adapts code by Allen Riddell for aggregating per-segment topic scores, and uses the word_cloud and seaborn modules for visualisation.]* Starting from the original XML/TEI-encoded texts from Théâtre classique, speaker text and stage directions have been extracted from the plays in order to exclude interference from prefaces, editoral notes, or speaker names. In the same process, each play has been segmented into its individual acts and scenes, again using the structural markup contained in the TEI files. Very short scenes have been automatically merged with the preceding or the following scene and a small number of very long scenes have been manually divided into smaller parts. This results in an average of 15.6 text segments of comparable length per play, or 5872 segments in total. While these text segments are not of exactly identical length, they largely respect the original act and scene boundaries which can be assumed to be, in the majority of cases, locations of thematic shift. *[Note: The vast majority of the text segments have a length of around 800 to 1600 word tokens, with the maximal range lying between 500 to 2000 word tokens. In an earlier iteration of this research, a very similar number of arbitrarily delimited, equal-sized text segments was used without respect for scene boundaries. The results do not change strikingly, but working with strucurally motivated text segments, even when they are of unequal length, appears to be the more methodologically sound choice. It has been reported by jockers2013 (p. 134) that in his corpus of novels, text segments of 1000 words with arbitrary boundaries produced the best, i.e. the most interpretable topics.]*

Lemmatization (i.e. the transformation of each word form to its base form such as could be found in a dictionary entry's head word) has been performed because French is a highly inflected language, and differences in the word use in different inflectional forms may obscure the common semantic structure which is of interest here. Also, part-of-speech tagging using TreeTagger (schmid1994) was applied to the texts, something which not only allows to filter out (speaker) names mentioned by other speakers (which are not of interest here), but also allows for the specific selection, for topic modeling, of only a certain number of word categories. For the research presented here, only nouns, verbs, adjectives and adverbs were retained for analysis, under the assumption that those are the main content-bearing words.

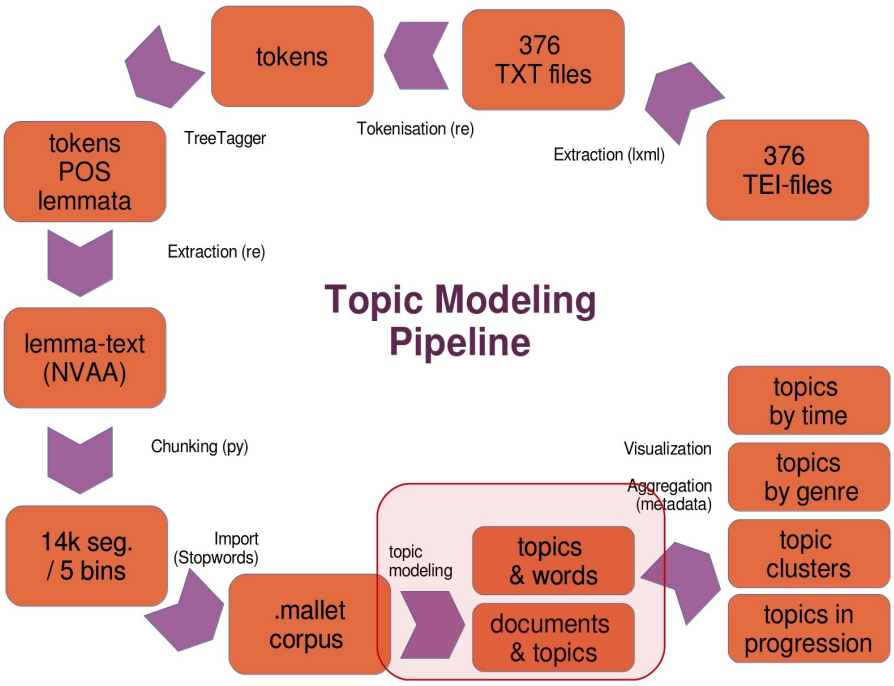

Figure 2 Topic Modeling Workflow

Ultimately, this means that instead of XML-encoded entire plays, short pseudo-text segments made up of sequences of noun, verb, adjective and adverb lemmata were submitted to the Topic Modeling procedure. Each of these segments can be identified as to the individual play it belongs to (which, in turn, is associated with decriptive metadata, such as the date of publication, the author, and the subgenre of the play) and as to the relative position in the play it comes from (with a granularity of five subsequent sections of the plays, corresponding roughly, if not structurally, to the five acts of the majority of the plays included). *[Note: The metadata table used can be found in the annex to this paper.]*

The Topic Modeling procedure itself has been performed using MALLET. Following considerable experimentation, the number of topics was set to 80. After (in this case) 10,000 iterations, the result of this process is, among other things, a table of all topics, ranked according to their overall probability in the entire text collection, with the words most strongly associated with each of them; a table containing the probability score of each topic in each of the 5872 text segments; and a table containing the probability score of each word in each topic. Post-processing involved visualizing topics as word clouds as well as using per-segment metadata to generate average topic scores for each complete play, for all plays in each subgenre, for all plays by each author, and for each of the five sections of all plays in a given genre. Finally, these averaged values were used to generate heatmap and lineplot visualisations.

It should be noted that all of these pre- and post-processing steps, however well argued for, involve a significant amount of parameter setting and, therefore, some degree of intuitive or arbitrary decisions. Such decisions may all have more or less direct influence on the results obtained and, while there is not always a clear rationale for choosing one specific value for a parameter, it is all the more important to document the choices made to increase the transparency and reproducibility of the method. *[Note: In order to increase the transparency of this research and allow the results to be reproduced, the dataset as it has been used, the Python code employed in the workflow, and a complete set of graphs produced for all topics, have been archived on Zenodo.org as a supplement to this article: [# DOI] and [# DOI]. Further development of the workflow happens on GitHub, at https://github.com/cligs.]*

Results and Discussion

In the following three sections, three types of results will be discussed and related to the hypotheses described above. First of all, results relating to the topics found in the collection of plays. Then, results pertaining to topics which are distinctive for the three subgenres contained in the collection, including results relating to genre-specific plot-related patterns. Finally, results from clustering based on topic scores as well as raw word frequencies will be presented.

Topics: Structure and semantic coherence

An initial inspection of the 80 topics obtained with their top-40 words, visualized as word clouds, shows that most of the topics display a relatively high level of (subjective) coherence. *[Note: As noted, varying numbers of topics have been experimented with before settling on 80: If the number of topics is set to a smaller value, several distinct and well-defined topics fail to be included among the results and several topics become aggregates of more than one semantic field. If it is set to a larger value, an overly large number of very similar topics emerge. The number of topics chosen may also interact with the length of the text segments used. It should be noted that this was a purely subjective assessment that did not involve any measure of topic coherence, topic dissimilarity, or model quality, as a function of both segment size and number of topics.]* A first selection of topic word clouds is shown in Figure 3. As is typical for such topics, the topics with the highest probability scores (i.e. the ones best represented in the collection) are less interpretable than most others, in the sense that they are rather generic and vague. This is the case, for example, for tp32 as well as for tp55 which has a score of 0.405 and whose top four topic words are "esprit, dessein, amour, âme" ("spirit, intention, love, soul"). These topics are present in many of the plays and are rather hard to interpret. Topics with very low probability scores (i.e. those appearing only in a few plays) are typically highly specific, but tell us less about the collection as a whole. This is the case, for example, for tp37 as well as for tp54, which has a score of merely 0.024 and whose top four topic words are "auteur, jouer, comédie, théâtre" ("author, to play, comedy, theater). Both topics are precisely focused and interesting, but occur only in very few plays or a single author. For instance, the metafictional topic just mentioned (tp54) is strongly present only in the comedy Les Chinois by Regnard and Dufresny, and a bit less strongly in a small handful of other plays. If one happens to be interested in one of these very specific topics, Topic Modeling provides a great way of identifying plays which should be included in a more detailed analysis. The most relevant topics for the research presented here, however, are those with less extreme probability scores, because subgenre distinctions are located by definition somewhere between individual plays and an entire collection of plays.

Figure 3 Word cloud visualization for selected topics (1)

The selection of topics in Figure 4 shows another phenomenon, related to the internal structure of topics. Most topics show a small number of quite important words (i.e. with high probability in the topic, displayed in a very large font size), with a relatively smooth drop-off and a long tail of less important words (displayed in very small font size). However, some topics show a different internal structure: for instance, in tp13, only "haine" and "venger" ("hate" and "to avenge") have a very large score, with subsequent words much less important in that topic. The effect is even more marked in tp42, in which only a single word, "loi" ("law"), has a very high score, with an extremely clear drop in scores for all other words in the topic. Inversely, tp35, has a large number of words with relatively high scores, the first five being "monsieur, oncle, enfant, ami, madame" ("mister, uncle, child, friend, madam"). The same phenomenon can be observed in tp51. The word cloud visualisations bring out this internal structure of the topics quite nicely.

Figure 4 Word cloud visualization for selected topics (2)

What, then, are topics characteristic for this collection of plays, what are the themes most commonly found in them? Figure 5 shows two types of topics. Many of the topics found are related to clear, abstract themes, such as love, death, crime, marriage, which are also themes we can expect to appear in plays of the seventeenth and eighteenth centuries. Two such topics are shown here for illustration, namely tp21, clearly related to life and death ("mourir, mort, vivre, vie" / "to die, death/dead, to live, life"), and tp78, clearly related to marriage ("hymen, époux, foi, heureux" / "marriage, husband, belief, happy"). Such topics typically come from the upper region of the probability scores. However, quite a number of topics are rather more concrete, such as those related to a quite specific setting, as in tp75 ("mer, vaisseau, vent, eau" / "sea, vessel, wind, water") or focusing on a very specific activity, as in tp45 ("lettre, écrire, billet, lire" / "letter, to write, note, to read"). These topics typically come from a somewhat lower range of probability scores. The latter type of topics actually shows that taking a method such as topic modeling, developed initially for collections of non-fictional prose such as scholarly journal articles or newspapers, and adapting it to the domain of literary texts, actually changes the meaning of the word "topic": to put it another way, the "topics" found in literary texts are not only abstract themes such as justice, human relations, or crime, but also more concrete activities typically performed by fictional characters, like writing, eating, drinking, hunting, speaking and thinking, or literary settings or motives such as the sea, the forest, or interiors. While the presence of the former is related to the choice of including verbs into the analysis, the same is not true for the latter. Obviously, the semantic coherence of both activity- and setting-related topics is more one of narrative structure than one of conceptual coherence.

Figure 5 Word cloud visualization for selected topics (3)

In a small number of cases, several topics are related to very similar semantic fields. This indicates that there may be a hierarchical structure to the topics, and that it may be useful to think of topics in a hierarchical relationship or as clusters. *[Note: This has not been assessed here systematically and could be investigated in the future using hierarchical topic modeling.]* This is the case for topics concerning family members and relations (not shown), and especially for no less than 10 topics related in some way to love, and containing among their top-ranked words items such as "amour" and "coeur" ("love, heart"). Four such topics are shown in Figure 6. Interestingly, the words following these top-ranked words differ significantly in each of these topics: for instance, in tp15, words like "gloire, espoir, foi, ardeur, flamme" ("glory, hope, belief, ardor, flame") link this kind of love to ideas of intensity and absoluteness, something which in turn is related to tragedy. Conversely, tp56 prominently contains words such as "parler, demander, refuser, croire, penser, sentir" ("to talk, to demand/ask, to refuse, to believe, to think, to feel"), linking this kind of love to activities of communication, negotiation, and cognition. What seem to be very similar topics, at first, turn out to contextualize the top words in very different ways. These differences are related, in addition, to subgenres, because each of the four topics described is associated more strongly with one subgenre than with others, an issue that will be addressed in the next section.

Figure 6 Word cloud visualization for selected topics (4)

Topics and genre: distinctiveness and plot-related patterns

While it is possible (and interesting) to extract, from the data, information showing which topics are over-represented or under-represented in certain authors or in certain decades, this paper focuses on the relation between topic and dramatic subgenres. One way to discover such distinctive topics is to proceed as follows. The topic scores obtained for each text segment are averaged, taking into account the genre of the play that each text segment belongs to. This yields, for each topic, a score for its average importance (technically: its probability) in each of the three subgenres. The topics with the highest probability in a given subgenre are not necessarily also the most distinctive ones, i.e. those which are over-represented in one subgenre relative to other subgenres, because some topics are just highly present in all subgenres. Therefore, these genre-related topic scores should be sorted according to the amount of variability they display across subgenres (i.e., according to their the standard deviation). Then, the topics with the highest variability across topics can be displayed, which are also the topics which are the most distinctive of different genres. *[Note: An alternative to this strategy is to perform clustering of topics with the scores of each topic in each genre as the features (not shown). This yields similar results but also shows that each genre seems to have two distinct groups of characteristic topics: those that are highly distinctive of them, and those which, although on a significantly lower level, also tend to be more associated with one genre than with the two others.]* Figure 7 shows a heatmap of such subgenre-related average topic scores for the thirty topics with the highest cross-genre variability.

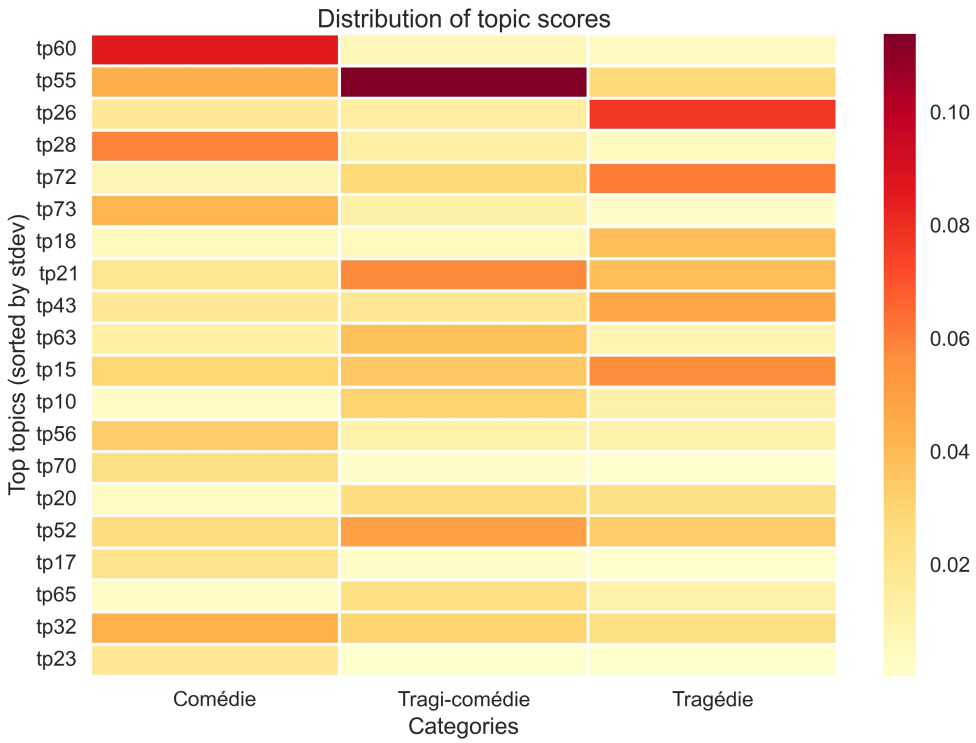

Figure 7 Heatmap of top-30 topics by subgenre

As can be seen, each subgenre has several distinctive topics, i.e. topics which in one genre have a score significantly higher both in relation to the other genres for the same topic (across rows), and to other topics for the same subgenre (across columns). The two most distinctive topics for comedies, with their four top-ranked words, are tp60 ("homme, monsieur, parler, donner" / "man, mister, to speak, to give") and tp32 ("beau, croire, rire, esprit" / "beautiful, to believe, to laugh, spirit"). The first is an expected topic for drama in general, since it relates to drama's structural foundations of personal relations and dialogue, but it seems that comedy plays out this fact more clearly than tragedy, which tends to use more extended monologues. The second could be related to activities typical of social gatherings, but also to (possibly misleading) outer appearance, a known abstract theme in comedy. Overall, however, these topics do not represent very sharply defined, abstract themes, but rather modalities of interaction. The two most distinctive topics for tragedy are tp26 ("cœur, ciel, cruel, malheureux" / "heart, heaven, cruel, unhappy") and tp72 ("sang, mort, main, crime" / "blood, death, hand, crime"). Both of these topics are much more immediately thematic than the two top comedy topics. The first combines the dimension of personal experience ("heart") with the idea of cruel authority figures causing unhappiness. The second relates univocally to violent, physical, emotional and deadly crimes, something which seems to indicate that the tragedies analysed here fulfil, to a significant extent, rather stereotypical genre expectations. Tragicomedy, interestingly, only has one clearly distinctive topic, tp55 already mentioned, which however is also quite widespread in comedy, followed at a large distance by tp21 which, however, is also of some importance in tragedy. This seems to indicate that at this level of analysis, tragicomedy is a mix of both tragedy and comedy rather than a genre of its own, something which confirms established knowledge about the genre, but does not yet give us more detailed information about which of the two genres is related more closely, topically and hence, thematically, to tragi-comedy. Does tragi-comedy have more topic-based overlap with comedies or with tragedies? Finally, comedy and tragedy each have one highly distinctive topic related to “love”, i.e. topics in which either "amour", "aimer" or "coeur" are among the top two words. As has already been seen, each of the "love" topics actually represents quite a different perspective on the theme of love, when looking at some of the words in the top-40 range (see Figure 6, above). Comedy seems to be favoring a communicative notion of love (tp56) and tragedy one with rather negative associations

(tp26).

To summarize, not only are very different topics associated with different subgenres, but they also seem to be different types of topics: rather vague interactional topics for comedy, and quite focused, abstract topics for tragedy. This seems to indicate that while tragedy remains defined by the topics it addresses, possibly quite explicitly, for example in monologues, comedy does not have abstract defining topics (not even marriage seems to be distinctive, while expected themes like deception, misunderstanding, or humor do not appear prominently among the topics), but is defined by its highly dialogic, interactive modality. Also, some topics which seem quite similar at first glance but are distinctive, one for comedy, one for tragedy, actually reveal deep-running differences which explain their association with different subgenres.

While topic-related patterns across subgenres, then, are very clearly present, the question remains whether this is also the case for plot-related patterns. This question can be investigated in the following manner: for each topic, average topic scores are calculated not by grouping the values of each topic in each of the 5,872 text segments only by their genre association, but also by the section of the play they belong to. This was done here with a relatively rough granularity of five sections per play corresponding, numerically if not structurally, to the five acts of most of the plays, and for comedy and tragedy only. *[Note: Further experiments varied this setup: Notably, instead of all plays, only those with exactly five acts have been used; and instead of sections, the actual acts have been used. The results change, but the trends in the data do not become more interpretable.]* In this way, the rising or falling importance of a given topic over the course of the average comedy or tragedy can be assessed and compared with existing knowledge about characteristic features of plot in these subgenres.

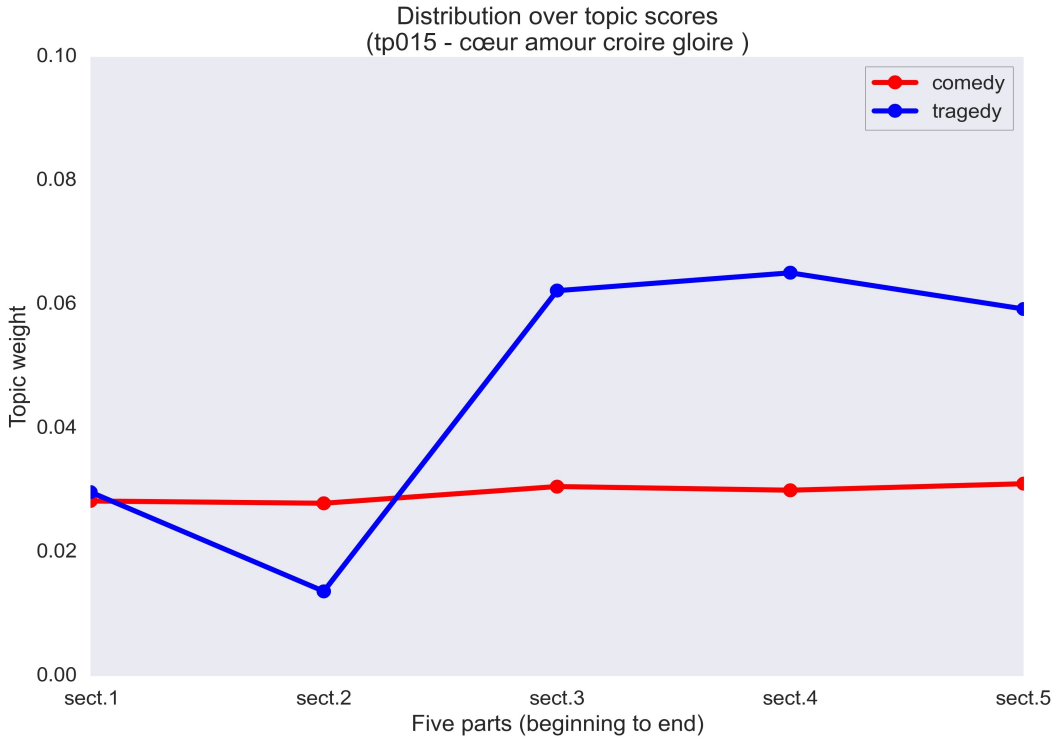

Figure 8 Distribution of plot-related topic scores (1)

Overall, while many topics do not show strong changes in importance over the course of plays in a given subgenre, approximately one quarter of the topics in the study presented here do. In several cases, a topic rises or falls in one genre while remaining stable in the other. This is the case, for example, for tp15 (one of the love-topics already discussed) and obviously related to tragedy (see Figure 8). Not only is this topic indeed more prevalent in tragedy than in comedy, but it also gains importance over the course of many tragedies: while it has a score of less than 0.03 in the first two fifths, this score more than doubles to over 0.06 in the three last fifth of an average tragedy. The

same pattern can be found for tp39, related to fear, suffering and death, as well as for tp26, related to the cruelty of authority figures (not shown). Quite obviously, the fact that tragedies frequently end with threats of cruelty and actual death manifests itself in the topic scores here.

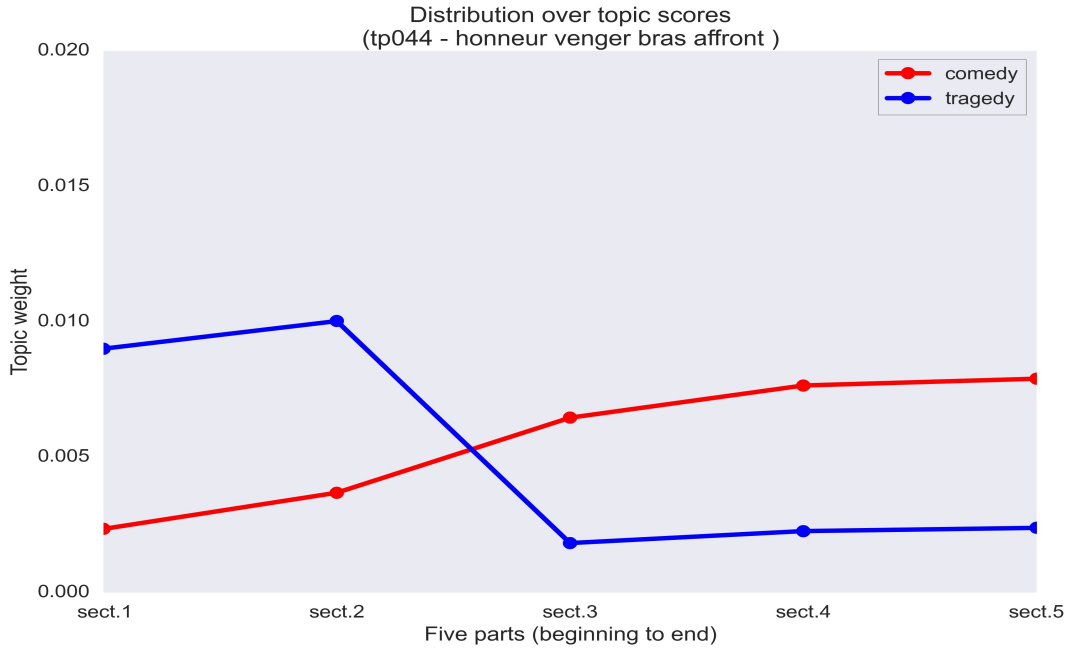

Figure 9 Distribution of plot-related topic scores (2)

In other, rare cases, an inverse relation between the topic scores in two genres can be observed. This is the case for tp44 ("honneur, venger, bras, affront" / "honor, to avenge, arm, insult"), i.e. a topic related to conflictual personal relations in the context of a societal code of honor and vengeance (see Figure 9). Interestingly, this topic is not clearly distinctive of tragedy, something which could have been expected. Rather, the overall scores of this topic for tragedy and comedy are quite similar, but the distributions of the topics scores over the course of an average tragedy and an average comedy are very different and do point to strong plot-related genre differences (even though the change observed here is of a much smaller amplitude than that for tp15, above). In tragedy, tp44 is of a much larger concern initially but then continually drops in importance, possibly because an initial affront is raised in the first two fifth of the plays. In comedy, on the contrary, tp44 is of constantly raising relevance over the course of an average play. It is not immediately apparent what the underlying reason for this pattern is. It is relevant, however, on a methodological level, because this is an example of a genre-related characteristic which may have been obscured by overall per-genre topic scores: it only becomes clearly visible when looking at the development of topic importance over the course of many plays. A number of plot-related trends exist in the data, some stronger than others, and some of which lend themselves more easily to interpretation than others.

Topic-based clustering

The differences observed so far rely on the *a priori* subgenre classification of plays, accepting the historical subgenre labels as given. This, it may be argued, is problematic not only because the labels and their use may not be uncontroversial, but also because some seemingly significant differences in topic score distribution across such a small number of categories may always be found, no matter what the categories are. Moreover, if the aim is to detect the dramatic genre's internal structure, such preconceived categories are not helpful. To explore whether the topics are indeed structuring the collection along the lines of subgenres, the perspective should also be inverted and a topic-based, unsupervised clustering method be performed on the data. This also allows one to move away from predefined, historical, potentially problematic genre labels, which

allow discovering distinctive topics but may obscure other structure in the data, whether it is related to genre or other factors, be it topic-based commonalities between several subgenres or additional divisions within one subgenre.

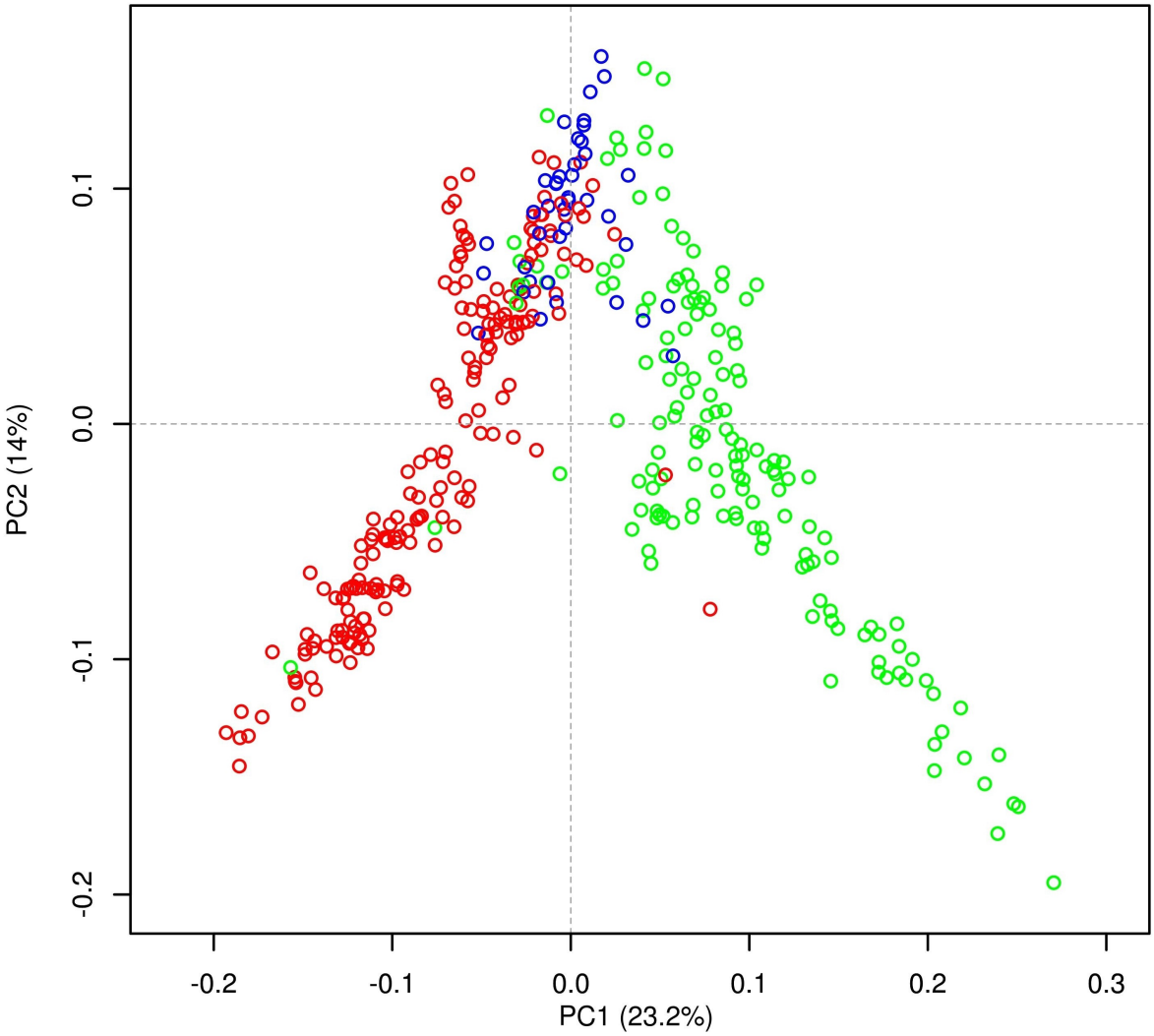

Figure 10 PCA plot of plays based on 80 topic scores

Figure 10 shows the result of a Principal Component Analysis (PCA) based on the topic scores. PCA is an unsupervised statistical method allowing the researcher to detect patterns in high-dimensional data. Ian Joliffe (joliffe2001, p. 1) defines it as follows: "The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables." That is, the purpose of PCA is to reduce the dimensionality of a given, high-dimensional dataset while retaining most of the information contained in the original dataset, in order to discover latent patterns in the data which may be related to some independent variable describing the elements in the dataset. Here, the unit of analysis is the play, and the features characterizing each play, are the probability scores of each of the 80 topics. *[Note: PCA as implemented in the "stylo" package for R (eder2013) has been used here.]* Figure 10 displays the two components in the data which summarize the greatest amount of variation in the data (i.e., the first two principal componentes which, together, account for roughly 37% of the variation in the data). Each circle in the plot represents one play, and their relative proximity or distance indicates topic-based, thematic similarity or difference in the two dimensions shown. The colors of the circles correspond to the conventional genre labels of each play, which however do not influence the positions the circles. The coloring only allows us to see to what degree the topic-based similarity of

the plays corresponds with their conventional genre label.

This correspondence is very high, and the first component (horizontal axis) clearly divides the plays into comedies (red circles in the two quadrants to the left) and the tragedies (green circles in the two quadrants to the right). It is true that to some extent, this may be an effect of correlations between authorship and genre (some authors, like Voltaire, having predominantly written only in one subgenre), but many authors in the collection have written plays in several subgenres. *[Note: This observation can be confirmed mathematically by computing the correlation between genre categories and positions on the first principal component (t-test correlation is very strong at -0.81, and highly significant at p < 0,0001).]* The second component seems to be related to the difference between the two main genres and tragi-comedy, which can only be found in the upper two quadrants of the plot. However, there is substantial overlap between tragicomedy and, especially, tragedy. *[Note: This overlap may not be present in one of the further dimensions; however, this has not been further assessed.]* When clustering plays based on their topic scores, then, tragicomedy appears to be more closely related to tragedy than to comedy. This fits one conventional description of tragicomedy as a tragedy that ends well, i.e. as a type of play with similar personnel, plot and themes, except that there is no disastrous ending. Another interesting feature of the PCA plot is that there seem to be two distinct groups of tragedies as well as two groups of comedies. These groups cannot be simply explained by chronology, and while the prose/verse distinction may possibly explain the two groups of comedies, it is not pertinent for tragedy, for which virtually all examples are written in verse (an exception is discussed below). A thorough investigation into what unites the plays of each subgroup and what distinguishes them from the other subgroup, thematically and/or otherwise, will need to be performed in further research.

Another phenomenon that can be observed is that there is a small number of plays which appear in the vicinity of plays of a different subgenre. Notably, two tragedies by Voltaire, L'Envieux from 1738 (with identifier tc0701 in the metadata) and Socrate from 1759 (tc0723), appear one right among the upper comedy cluster, one in close vicinity to the lower comedy cluster. The latter of the two is actually a rare specimen of a tragedy written in prose, something which may explain its position separate from all other plays. In addition, two comedies appear among the tragedies, one being Le Glorieux by Philippe Néricault Destouches (tc0525), the other La Vie est un songeby Louis de Boissy (tc0055). The precise reasons why these plays appear in this context remain to be explored. However, it can be noted that Nericault's comedies are sometimes seen as precursors to the subgenre of *drame bourgeois*, so his plays may be atypical comedies. And Boissy's La Vie est un songe is not a plain comedy, but a so-called *comédie héroïque*, i.e. a play that has many of the features of comedy but combines them with noble personnel and illustrates values of the nobility, which may make it more similar to tragedy.

When looking not at the relative position of the plays in the first two dimensions, but at the underlying topics and their strength of association with the two dimensions (i.e., the so-called topic loadings, shown in Figure 11), some more interesting characteristics emerge. Some of the topics with extreme positions in the plot concur with the distinctive topics for comedies and tragedies (for comparison, see Figure 7 above). This is the case for topic 60 at the extreme south-east (related to comedy) or topic 26 in the south-west as well as topics 72, 43 and 18 in the west of the plot (related to tragedy). Finally, topic 55 is midway between east and west (i.e., not distinguishing between tragedy and comedy) but very much to the north of the graph (i.e., strongly distinctive of tragycomedy). This does not come as a surprise, of course, because it simply confirms the distinctiveness analysis performed earlier. It is striking, however, how sharp the drop-off is from a small number of topics with high variance in this plot, and the central cluster of all remaining topics which do not contribute much to principal components 1 and 2.

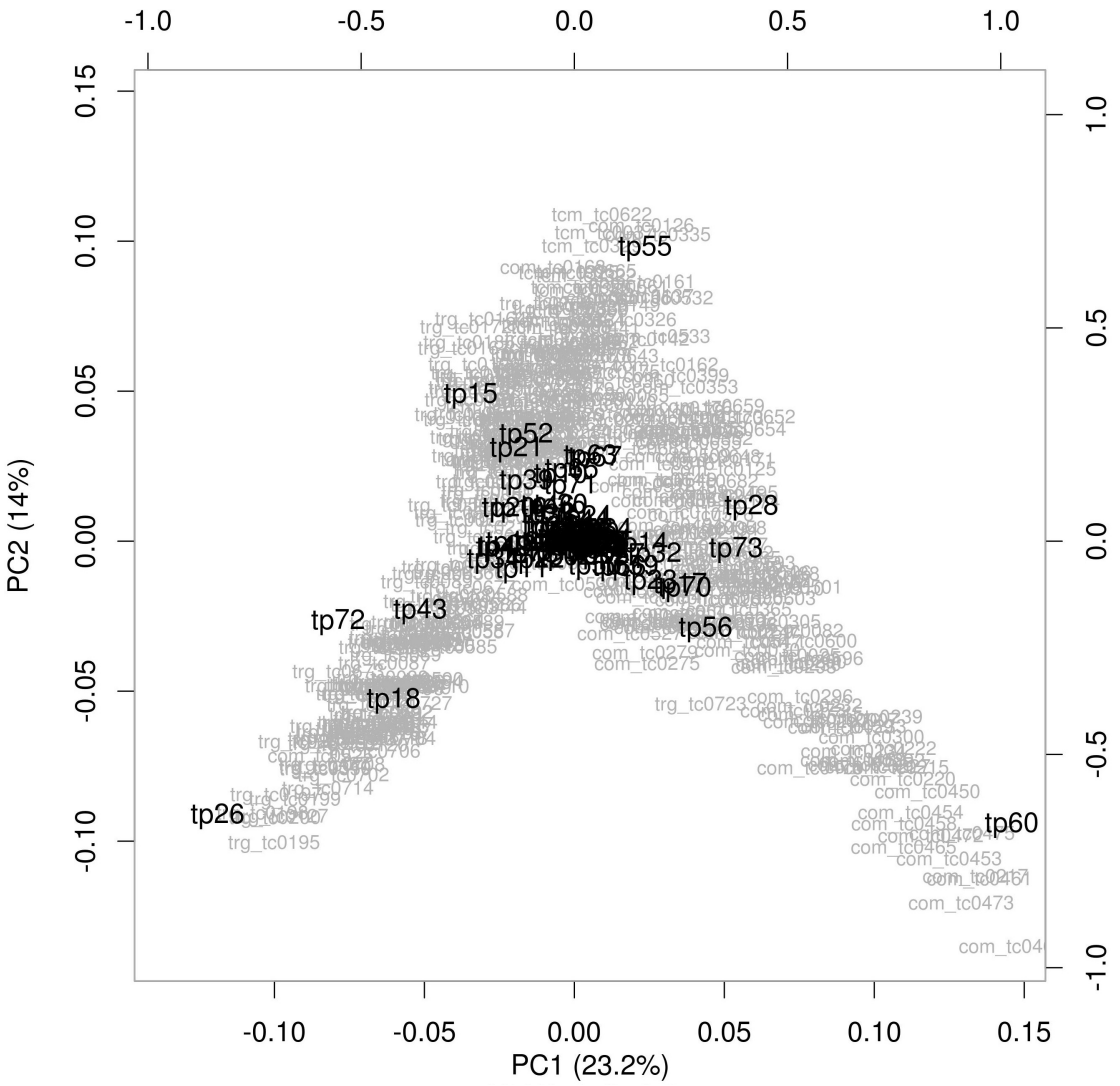

Figure 11 PCA based on topic-scores: loadings.

It remains to be seen whether topic-based clustering is more clearly genre-related than clustering based on word frequencies, as typically used in authorship attribution studies (for an overview, see stamatatos2009). It turns out that when comparing clustering based on topic scores on the one hand (Figure 10, above), and word frequencies on the other (Figure 12, below), there is a considerable amount of similarity. This is surprising, because both plots display very different views on the data: while the topic scores have been obtained based only on content-bearing words, are abstractions from individual words, and are directly related to themes, the word frequencies of the 1200 most frequent words used here also contain a large number of function words of particularly high frequency. Such function words are usually associated with authorship rather than genre or subgenre, and could have been expected to skew results or represent noise with regard to genre. However, the author signal is likely invisible here due to the large number of authors in the dataset compared to the very small number of subgenres. Interestingly enough, not only the general structure of the PCA plot is similar, but the overlap of tragicomedy and tragedy as well as the subclusters for comedy and tragedy are also visible in the word frequency-based plot. One advantage, however, of the topic-based clustering, is the fact that the features with high loadings are readily interpretable topics, while in the word-based clustering, the features with high loadings are individual words, many of them function words, so that their functional or thematic relation to literary subgenres is more difficult to assess.

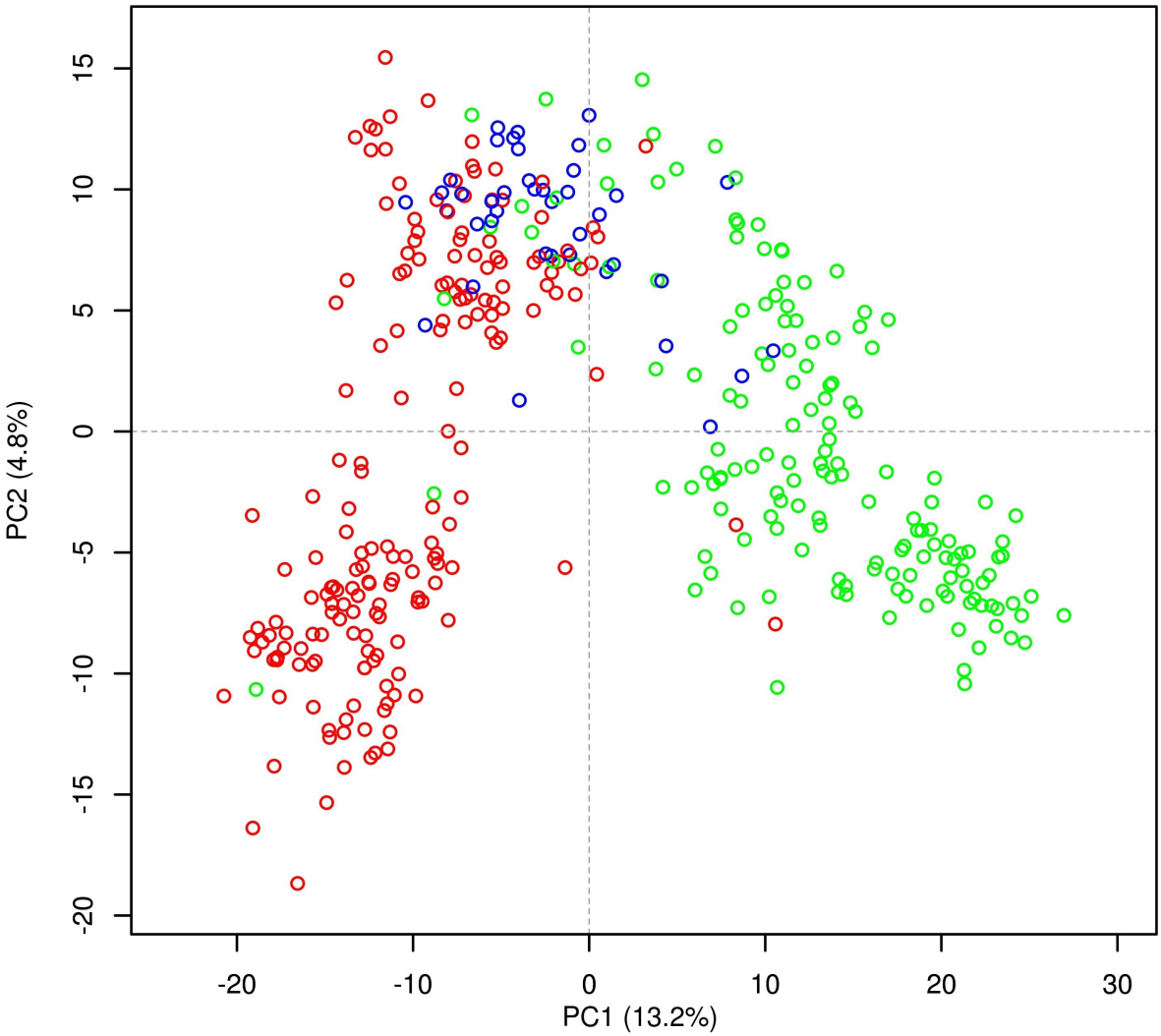

Figure 12 PCA based on word frequencies.

Overall, the results from clustering using PCA confirm that even when the algorithms do not know about subgenre categories, but cluster the plays only based on their topic scores, that is relating to their thematic similarity in a broad sense, subgenre turns out to be among the strongest factors that enter into the resulting clusters.

Conclusions

To summarize the major results of this study, one may state first that, as far as the topics obtained are concerned, there is a very high semantic coherence of topics. This coherence, however, depends in part on the number of topics defined for the analysis. Some of the topics are clearly thematic, others are related to setting or narrative motives. Also, based on metadata regarding the subgenre of each play in the collection, the strength and nature of topic-related patterns across subgenres can be observed, with each subgenre having a number of clearly distinctive topics. In addition, some topics show plot-related trends over the course of an average comedy or tragedy, trends which can in many cases be meaningfully related to existing knowledge about the subgenres' plot structure. Clustering based on the topic scores obtained also yields results which can be usefully related to subgenres and show that the distinctiveness of topics with regards to genre is not a projection but an actual pattern in the data. Overall, it appears that interesting patterns can be detected which, in many cases, confirm existing knowledge about subgenres themes and plot on the basis of a significantly larger amount of data than can usually be taken into account. In some cases, surprising results provide new insight into the history of French drama of the Classical Age and the Enlightenment or provoke new hypotheses which may need further inquiry in the future. This is the case, particularly, of the finding

that there may be two distinct groups or types of tragedy and comedy.

It has been established in previous (as yet unpublished) studies by the author of this paper that strong genre signals exist in the collection explored here on the levels of function words, content words and syntactic structure. It is interesting to see that they also exist on the level of theme or topic. Once each of these (and more) levels of description of genre have been analysed in more detail for the collection of French plays, new insights into the structure of the subgenres and into their development over time will become possible by analysing the ways in which these different levels of description correlate, interact, or possibly contradict each other. Ultimately, these results also show that far from merely being a projection or a social construct, literary genres do have a textual reality that can be assessed quantitatively. It is on the basis of such a reality and within the potentialities and constraints it offers, that different views of literary genre as a concept, and of literary history as the continual evolution of literary genres, become possible.

It would be tempting to finish this paper on such an optimistic note. However, it seems necessary to also point to some of the challenges connected to research of the kind presented here. First of all, many questions remain open and while some results can readily be linked to established knowledge in traditional literary scholarhip, other phenomena remain to be explained. For example, the subgroups of tragedies and comedies found in the PCA plots will need to be investigated in future work. Also, the results obtained for plot-related patterns are promising, but so far relatively inconclusive. Possibly, the inclusion of three-act plays as well as plays with a prologue, in the collection used here, have obscured plot-related patterns in the majority of five-act plays. However, confirming established knowledge or widely accepted hypotheses on the basis of larger datasets is useful in itself, especially at a time when there is only a limited amount of experience with techniques newly applied to the domain of literary texts. This helps build confidence in results which may support entirely new hypotheses. A second, more technical, challenge is related to the method used here to obtain average values for topics depending on the subgenre of a play or the section in a play. It would be more appropriate, and may produce even better results, to include the genre or section information into the modeling process from the start, which would effectively mean practising something called "labeled LDA" or "(semi-)supervised LDA" (ramage2009) and "sequential LDA" (du2010), respectively. Also, as can be seen from the close relations between a number of topics, hierarchical LDA (blei2004) may be able to capture existing structure among the topics themselves. As has already been noted, however, no readily useable implementation of such techniques is currently available to the application-oriented research community in the humanities. A third, less technical but no less important challenge is the fact that there is still a possibly insufficient number of texts. 375 plays certainly is a substantial number, even compared to the total production of the time (see section "Data" above). Also, it is certainly more than a single researcher could read with benefit inside a limited time period. However, considering that the study covers a period of 150 years and three subgenres, 375 plays are actually not that many. For example, this effectively means there are only around 8 plays per decade and subgenre, on average, included in the collection (not considering the uneven distribution). Also, many important subgenres had to be discarded from the analysis, because only a handful of examples of them are currently available in the Théâtre classique collection. This is a challenge that only continued efforts in high-quality, full-text digitisation in standard formats such as TEI, and their open-access dissemination, will alleviate. However, trying to show that based on such data, interesting results can be obtained for literary history, may be the best way to motivate continued digitisation efforts. This is what this study, despite its modest thematic and methodological scope, has attempted to do.

Acknowledgements

Work on this paper was supported by funding received from the German Federal Ministry for Research and Education (BMBF) under grant identifiers FKZ 01UG1408 and 01UG1508. It was first presented at the workshop on Computer-based Analysis of Drama organised by Katrin Dennerlein in Munich, Germany, in March 2015. A special thank is due to Paul Fièvre who, by

relentlessly building his Théâtre classique collection over many years, has made this research possible.

Bibliography

Blei, David M. 2012. Probabilistic Topic Models. In: Communication of the ACM, 55.4.

Blei, David M., Jon D. McAuliffe. 2009. Supervised Topic Models. In: The Statistical Science.

Blei, David M., Tom Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. 2004. Hierarchical Topic Models and the Nested Chinese Restaurant Process. In: Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference, ed. Sebastian Thrun, Lawrence K. Saul, and Bernhard Schölkopf. Boston, MA: MIT Press.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. In: Journal of Machine Learning Research 3, 993-1022.

Blevins, Cameron. 2010. Topic Modeling Martha Ballard's Diary. In: Historying, http://historying.org/2010/04/01/topic-modeling-martha-ballards-diary/.

Burnard, Lou. 2014. What Is the Text Encoding Initiative? How to Add Intelligent Markup to Digital Resources. Encyclopédie Numérique. Marseille: OpenEdition Press. http://books.openedition.org/oep/426.

Chang, Jonathan, Jordan L. Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In: NIPS'09, 288–96.

Du, Lan, W.L. Buntine, and Huidong Jin. 2010. Sequential Latent Dirichlet Allocation: Discover Underlying Topic Structures within a Document. In: IEEE 10th International Conference on Data Mining (ICDM), 148–57, doi:10.1109/ICDM.2010.51.

Eder, Maciej, Mike Kestemont, and Jan Rybicki. 2013. Stylometry with R: A Suite of Tools. In: Digital Humanities 2013: Conference Abstracts,487–89. Lincoln: University of Nebraska.

Fièvre, Paul, ed. 2007-2015. Theatre classique. http://www.theatre-classique.fr.

Graham, Shawn, Scott Weingart, and Ian Milligan. 2012. Getting Started with Topic Modeling and MALLET. In: The Programming Historian. http://programminghistorian.org/lessons/topic-modeling-and-mallet.

Hempfer, Klaus W. 1973. Gattungstheorie. Information und Synthese. Munich: Fink.

Herrmann, J. Berenike, Karina van Dalen-Oskam, and Christof Schöch. 2015. Revisiting Style, a Key Concept in Literary Studies. In: Journal of Literary Theory, 9.1, 25-52.

Jockers, Matthew L. 2013. Macroanalysis: Digital Methods and Literary History. Champaign, IL: University of Illinois Press.

Joliffe, Ian T. 2002. Principal Component Analysis. 2nd edition. Berlin and New York: Springer.

Mazouer, Charles. 2010 / 2014. Le théâtre français de l'âge classique, vols. II and III. Paris: Champion.

McAuliffe, Jon D., and David M. Blei. 2008. Supervised Topic Models. In: Advances in Neural Information Processing Systems 20, ed. Neural Information Processing Systems Foundation. http://papers.nips.cc/paper/3328-supervised-topic-models.pdf.

McCallum, Andrew K. 2002. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

Ramage, Daniel, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-Labeled Corpora. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Volume 1, 248–56.

http://dl.acm.org/citation.cfm?id=1699510.1699543.

Rehuřek, Radim, and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 45–50. Valletta, Malta: ELRA.

Rhody, Lisa M. 2012. Topic Modeling and Figurative Language. In: Journal of Digital Humanities, 2(1). http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/.

Riddell, Allen B. 2014. Text Analysis with Topic Models for the Humanities and Social Sciences (TAToM). In: DARIAH-DE Portal. Göttingen: DARIAH-DE. https://de.dariah.eu/tatom/.

Riddell, Allen B. 2014. lda. In: The Python Package Index. https://pypi.python.org/pypi/lda.

Schaffer, Jean-Marie. 1989. Qu'est-ce qu'un genre littéraire. Paris: Seuil.

Scherer, Jacques. 2001. La dramaturgie classique en France. Nouvelle édition. Paris: Nizet.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of International Conference on New Methods in Language Processing. Manchester.

Schmidt, Benjamin. 2014. Typical TV episodes: visualizing topics in screen time . In: Sapping Attention, http://sappingattention.blogspot.be/2014/12/typical-tv-episodes-visualizing-topics.html.

Stamatatos, Efstathios. 2009. A Survey of Modern Authorship Attribution Methods. In: Journal of the American Society of Information Science and Technology, 60.3, 538-56.

Steyvers, Mark, and Tom Griffiths. 2006. Probabilistic Topic Models. In: Latent Semantic Analysis: A Road to Meaning, ed. by T. Landauer, D. McNamara, S. Dennis, and W. Kintsch. Laurence Erlbaum.