Detecting semantic or structural similarities for theater play comparison

Pierre Bourhis[1], Aaron Boussidan[2], Céline Fournial[3], Philippe Gambette[4]

[1]Université de Lille, CNRS, Inria, Centrale Lille – pierre.bourhis@inria.fr

[2]Université Gustave Eiffel – aaron.boussidan@univ-eiffel.fr

[3]Université Clermont Auvergne – celine.fournial@uca.fr

[4]Université Gustave Eiffel – philippe.gambette@univ-eiffel.fr

Abstract (in English)

From the very beginnings of modern theater in France, dramatic creation has been using several practices of imitation and adaptation of earlier texts, from biblical, antique or historical sources. Here, we focus on plays inspired by other plays, and on ways of automatically detecting similarities between the texts of such plays. We present approaches based on structural similarities and others based on semantic similarities between play texts, using word embedding methods. Finally, we look at the advantages of developing hybrid approaches that combine these two criteria, using on the one hand the corpus of sources on dramatic creation provided on the *Hyperpièces* website for comedies, tragi-comedies and tragedies published in France from 1550 to 1650, and on the other hand the *Dracor* corpus.

Keywords: theater plays, text comparison, textual alignment, word embeddings, parameterized words

Résumé

Dès les débuts du théâtre moderne en France, la création dramatique s'est nourrie de plusieurs pratiques d'imitation ou d'adaptation de textes antérieurs, de sources bibliques, antiques ou historiques. Nous nous intéressons ici aux pièces inspirées par d'autres pièces de théâtre, et aux moyens de détecter automatiquement les similarités entre les textes des pièces dans ce cas. Nous présenterons des approches fondées sur des similarités de structure et d'autres fondées sur des similarités sémantiques entre les textes des pièces, en utilisant des méthodes de plongements de mots. Enfin, nous étudierons l'intérêt de développer des approches hybrides qui mêlent ces deux critères, en mobilisant d'une part le corpus des sources de la création dramatique fournies sur le site *Hyperpièces* à propos des comédies, tragi-comédies et tragédies publiées en France de 1550 à 1650, et d'autre part le corpus *Dracor*.

Mots clés : théâtre, comparaison de textes, alignement de textes, plongements de mots, mots paramétrés

1. Introduction

The goal of this work is to define methods to detect similarities between theater plays. We aim to detect similarities on a global scale, rather than local similarities which reflect intertextuality between plays, which was already studied for example for plays by Molière or Corneille (Bourqui, 2014; Douguet, 2018; Douguet, 2020). We also provide tools which go beyond

computing a simple distance between two plays, which could be done using various formulas of intertextual distance (Brunet, 1988; Luong, 2003; Tempestt et al, 2017). Instead, we want to follow the usual principles of exploratory textual analysis (Lebart et al., 2020) by allowing interactions between distant reading, highlighting high-level similarities between the plays, and going back to the text using the aligned speeches among the compared plays.

Our main tool to compare two plays is therefore creating alignments, that is to say create association between elements of the first and the second play, where those elements can be characters, characters lines, scenes, etc. The benefit of technique is twofold : first, it often yields a similarity score or distance, a simple numerical value, which can quantify synthetically the similarity between both plays. Second, the alignment in itself is more fine-grained, and highlights the parts that are common to both plays, and the parts that significantly differ, allowing for a closer analysis of the results.

Working with drama specifically has a few practical advantages : those texts are classically very structured, as they are often split into acts, scenes and character lines. Notably, for those texts, the NLP task of speaker identification is easily resolved, since the name of speakers is always written clearly before their lines. What is more, many seventeenth-century European plays tried to obey some composition rules (notably by reducing time and unifying place and action). Though the actual following of those rules vary strongly with time and authors, they induce general structure properties, or regularities on plays, which may be checked automatically. Those rules also play an important role in the translation or rewriting of plays from a foreign language. As an example, relevant to the current article, when French authors adapted plays from the "Spanish Golden Age", they often did so with very little modification to the story, but with changes to the structure, fitting the three long *jornadas* to the five acts. Moreover, character names were generally translated to French, and even sometimes replaced by completely different French names.

This motivates to study similarities among theater plays which go beyond lexical proximity. How can we take advantage of the structure of plays to compare them? Can we use it to extend recent statistical approaches to compute the semantic similarity between two texts? Does it help to compare plays written in different languages?

Before addressing those questions, we start by presenting our corpus, which focuses on modern European plays, published between 1550 and 1650. We then introduce two methods to compare plays: a structural approach, using the formalism of parameterized matching, and a semantic approach, using word embedding tools.

2. Corpus and pairs of plays

In the following, we present our corpus: we detail both the sources of the plays themselves, and the sources of the similarities. All of our data and scripts are available at the following url: https://github.com/AaronFive/jadt2024

2.1. Play data

To make full use of both the structure and the text of the plays in our corpus, they must be available in a digital text format where acts, scenes and speakers are annotated, ideally in XML-TEI (Burnard, 2014). Here is an example of the typical encoding of the beginning of a play in this format:

```
<div type="act" n="1">
```

```
<head>ACTE I</head>
  <div type="scene" n="1">
    <head>SCÈNE PREMIÈRE. Philinte, Alceste.</head>
      <sp who="#philinte">
        <speaker>PHILINTE.</speaker>
        <l n="1" part="I">Qu'est-ce donc ? Qu'avez-vous ?</l>
      </sp>
      <sp who="#alceste">
        <speaker>ALCESTE.</speaker>
        <l n="1" part="F">Laissez-moi, je vous prie.</l>
      </sp>
```

Although some tools have recently been developed, for example using BERT models, to automatically obtain such files in the interoperable XML-TEI format from scanned versions of plays in German (Pagel et al., 2021), such tools are not yet available for the French language. But we were able to gather plays from several sources, both in the XML-TEI and in the HTML formats.

28 of the TEI files from our corpus are taken from *DraCor* (Fischer, 2019), a European project collecting plays from several European countries, in the public domain, encoded in TEI. Dracor has corpora in several languages, the largest is in French with 1560 plays, adapted from the files provided on Paul Fièvre's website *theatre-classique.fr*.

8 plays come from the website *Bibliothèque dramatique*, a database of French plays maintained by the CELLF laboratory of Université Paris-Sorbonne. Plays originating from this database are generally accompanied with a preface written by a student, detailing among other things the inspirations and sources of the play. Finally, some Spanish plays were downloaded in HTML from the websites *Artelope*, *cervantesvirtual.com*, and *EMOTHE*. In those files, the structure of the play can usually be extracted from specific HTML tags and classes, which vary depending on the website.

2.2. Similarity data

Our corpus is composed of 32 pairs of plays, for 64 plays in total. Among those, 44 are in French, 18 in Spanish, and 2 in Italian. Their publication dates span from 1524 to 1697.

All pairs consist in two plays that are similar, in a way or another. We detail the type of similarities and give examples below. Among those 32 pairs, 30 are extracted from the *Hyperpièces* database, which gathers 593 classical plays in French and their sources in various types of texts, including other French plays, but also Spanish and Italian plays. This database was constructed from the appendix of Celine Fournial's doctoral thesis (Fournial, 2019). All these similarities have been established manually, using the scientific literature or the influences acknowledged by the authors of the plays. One of the goals of our work is therefore to extend this corpus of similar plays by designing methods, based on this training data, to help automate and speed up this kind of search for possible connections between some plays and other which inspired them.

Additionally, 2 pairs have been aligned more precisely: *La Dama Duende* and *L'Esprit Folet,* and Hardy's and Sallebray's version of *La Belle Égyptienne.* For both of these pairs, we created a correspondence between individual lines, highlighting expressions or words that clearly correspond between the two works. This allows for a more close reading of the similarity, and is useful in the NLP approach.

2.3. Similarity types

Plays can be similar in various ways. The most obvious similarity is between two versions of the same text. In our corpus, this is the case for the 1677 and 1697 editions of *Phèdre,* by Racine. Plays can also be similar in different languages if one is a translation, adaptation or imitation of the other. These are often very similar for the main plot, but vary with respect to the secondary characters. Some parts of the original plays are also often cut by more recent authors, resulting in shorter plays. In our corpus, this is typically the case for Spanish plays, by Calderón or Lope de la Vega rewritten in French. The similarity may also result from a specific narrative technique, like the metatheater or *mise en abyme* originally used in *La Comédie des Comédiens* that inspired *L'illusion Comique*. Some plays have also been so successful that they spanned a sequel, as is the case for *La Mort des Enfants d'Hérode,* written to be the sequel of *La Mariane.* Finally, some plays are inspired by a common source, like is the case for *Don Juan Alvaredo* and *Jodelet ou le Maitre Valet*, which are both inspired by Tirso de Molina's *Don Juan*.

Designing a unique algorithm to capture all these different kinds of similarities is, of course, too ambitious. This is why, among those types of similarities, we focus on those based on story or text similarity, where designing alignment approaches is relevant. To begin with, we start with similarities in structure.

3. Similarities in structure

3.1. Models: character networks and parameterized words

A first possible approach to compare two plays is to focus only on their respective structure, i.e only keeping the information of which character is speaking when, and the structure of acts and scenes. This, of course, gets rid of most of the substance of the play, but has the advantage of simplicity, and of avoiding the problem of the language of the plays. Furthermore, focusing only on the structure of plays is sometimes sufficient to find similarities between them, as we will see later. We present two models which may use only their structure, and provide more details for a new approach based on parameterized words.

3.1.1. Character networks

A first possible methodology is to study character networks of the plays and try to align them, based on character names, textual content, relationships between characters, or number of words spoken in the presence of other characters, in order to find similarities between the plays. A character network is a mathematical directed or undirected graph, constructed by considering each character of the play, and drawing links between each of them based on their interaction during the play. Different techniques can be used to quantify this interaction: the number of scenes where both characters appear, the number of lines said by those characters, or in total, where both are present on stage, etc. Such character networks have appeared on several theater electronic libraries online, such as *DraCor* or *Dramagraph* (Glorieux, 2016) and used in studes about theater (Moretti, 2013; Santa Maria et al., 2021; van der Deijl, 2023).

Note that such networks are static visualizations of the play, which remove the notion of time. Another model keeping information about the chronologic aspects of the relationships between character in a play was also introduced in the literature. Matrices encoding the presence or absence of characters on stage during the play, representing the characters by lines and the scenes by columns, were studied for example by Marcus (1970) and Douguet (2016).

3.1.2. Parameterized words

The new model we introduce here is a bit more refined than the character matrix, keeping

information about the successions of speeches of the characters of the play. The idea is the following: given a play and its characters, forgetting at first the separation between scenes and acts, we consider the chronological succession of speakers. We can encode this with a simple string of letters, where each letter is, for example, the initial of each character (see Figure 1).

Figure 1. In this excerpt from Tartuffe (Molière, Act IV, Scene 3) the succession of speakers (Dorine, Orgon, Elmire, Orgon, Elmire, Orgon) can be represented as a parameterized word DOEOEO.

DORINE *:Mais quoi...*
ORGON *:Taisez-vous, vous. Parlez à votre écot,*
Je vous défends, tout net, d'oser dire un seul mot.
ELMIRE*:Si par quelque conseil, vous souffrez qu'on réponde…*
ORGON*: Mon frère, vos conseils sont les meilleurs du monde,*
Ils sont bien raisonnés, et j'en fais un grand cas;
Mais vous trouverez bon que je n'en use pas.
ELMIRE*:(à son mari.): À voir ce que je vois, je ne sais plus que dire,*
Et votre aveuglement fait que je vous admire.
C'est être bien coiffé, bien prévenu de lui,
Que de nous démentir sur le fait d'aujourd'hui.
ORGON:*Je suis votre valet, et crois les apparences [...]*

Doing this for two plays we want to compare yields two strings that are now to be matched. To compare them, we proceed in two steps : first, we try to match each character of the first play, to a character of the second play. The underlying hypothesis, here, is that for two close plays, some characters will play a similar role on both sides, even if they change names.

Figure 2a. The parameterized words associated with the theater and opera versions of Médée*.The letter-character correspondence is the following : O : Cléone , A : Créon , U : Créuse, J : Jason, M : Médée, N : Nérine, T : Théodar*

Act V, *Médée*, Pierre Corneille :
TMTMTMTMTMMAUAUAUAUOUOJUJUJMJMJMJMJ

Act V, *Médée*, Thomas Corneille :
NMNMUMUMUMUMUMKUKUOUOKUUMUMUJUJUJJUJMJMJMJM

Figure 2b. An alignment after replacing ***T****heodar by* ***N****érine, giving a distance of 19. Plus signs represent the insertion of a character, equality signs represent a match between both strings, and pipes represent the substitution of a character by another one.*

```
TMTMTMTMTM_MAU_AUAUAUO_______U_OJU_JUJMJMJMJMJ
====|=|=|=+=-=+|=|=|==+++++=+|==+==========-
NMNMUMUMUMUM_UMKUKUOUOKUUMUMUJUJUJJUJMJMJMJM_
```

Once this renaming is fixed, we compute an edit distance, with the following allowed operations: renaming a speaker for one individual speech, adding one speech for a speaker or removing it, between the two sequences of speeches. This corresponds to the Levenshtein distance computed between the two strings (Levenshtein, 1966), which gives a score for this renaming. The lower it is, the better the chosen renaming was. In particular, if the distance is 0, both plays have the exact same sequences of speakers, just with different names. Therefore, the idea is to find the best possible renaming of characters, which yields the lowest possible distance between the two strings. Once that optimal renaming is found, we call the associated distance the parameterized distance between both plays. Lastly, all the matchings we consider in this article are 1-to-1, *i.e.* characters are always associated to one unique counterpart in the other play. This is, of course, a simplification, as characters, notably secondary ones, are sometimes split or fused between

rewritings (we will not consider this generalization in this article).

A generalization addressing this remark, and more mathematical and computational aspects of this problem are addressed in Bourhis et al. (2023). The main results are that the problem is NP-complete (that is, the complexity to solve it explodes as the size of the data increases), but it can be solved in a reasonable amount of time when limiting the number of characters of each play to 8 or 9.

3.2. Results

Using this technique, we ran a parameterized matching comparison on our corpus. To be able to compare results between pairs of plays, we scaled the parameterized distance obtained between 0 and 1. We compared plays both act by act, and in their entirety. As expected, running the comparison on the complete plays yields better results, although it requires a longer computing time. After running our code, we ordered all the pairs of plays by distance.

Among the ten plays with the closest distances given in Figure 3, seven yield a correct alignment, confirmed by reading the plays, or their presentations in the prefaces of the *Bibliothèque dramatique* website. Do note that, even if some characters have obviously similar names, like between *Belissaire* and *El ejemplo mayor de la desdicha*, the actual name of the character is not taken into consideration by the algorithm, only the structure of the plays is.

Our algorithm seems to perform well on couples of rewritings from Spanish to French, with a limited number of false positives. An advantage of the 1-to-1 correspondence between characters is that when the main characters are matched correctly, the rest of the cast seems to also be correctly identified, down to the least active characters. For the computation of this matchings, each pair took less than 30 seconds of processing time, the cutting of characters to 8 ensuring an efficient computing time.

Below the point of the 10 closest pairs, we obtained distances ranging from 0.27 to 0.6, and no correct match found. In some cases, like between *La Comédie des Comédiens* and *l'Illusion Comique*, no real correspondence between characters of both plays exists, but our algorithm still outputs a pretty low distance, resulting in potential false positives. Optimizing this distance and defining a cutoff point is still an ongoing research at the time being, and most likely requires both theoretical and experimental data to pinpoint precisely. What is more, some "obvious" matchings are missed for some plays, as is the case for *Sophonisbe* shown in Figure 4.

These misses show that the structure similarity hypothesis is not always correct, or at least should be enriched with actual semantic data from the plays.

Figure 3. Ten closest pairs of plays, with their distance, and the character renaming inferred. All pair names written in green correspond to a correct guess of the algorithm, confirmed by human reading.

Pair name	**Dist.**	**Renaming**
Racine-phedre-77 racine-phedre-97	0.0	[('hippolyte', 'hippolyte'), ('theramene', 'theramene'), ('oenone', 'oenone'), ('phedre', 'phedre'), ('panope', 'panope'), ('aricie', 'aricie'), ('ismene', 'ismene'), ('thesee', 'thesee')]
rojas_zorrilla-dondescarron-jodelet	0.09	[('SANCHO', 'jodelet'), ('DON_JUAN', 'don-juan'), ('BERNARDO', 'etienne'), ('DON_LOPE', 'don-louis'), ('DOÑA_INÉS', 'isabelle'), ('BEATRIZ', 'beatrix'), ('DON_FERNANDO', 'don-fernand'), ('DOÑA_ANA', 'lucrece')]
hardy-belle_egyptiennes allebray-	0.13	[('PRÉCIEUSE', 'precieuse'), ('VIEILLE ÉGYPTIENNE', 'la-vieille'), ('CLÉMENT', 'le-poete'), ('DOM JEAN', 'don-jean'), ('Others', 'Others'), ('ANDRES', 'andres'), ('CARDUCHE', 'hipolite'), ('GUIOMAR', 'isabelle'), ('SÉNÉCHAL', 'ferdinand')]

belle_egyptienne		
calderon-casaouville-fausses_verites	0.14	[('MARCELA', 'florimonde'), ('SILVIA', 'nerine'), ('LISARDO', 'lidamant'), ('FÉLIX', 'leandre'), ('CELIA', 'julie'), ('LAURA', 'orasie'), ('FABIO', 'tomire'), ('CALABAZAS', 'fabrice'), ('Others', 'Others')]
lope-honrado_hermano corneille-horace	0.15	[('Julia', 'sabine'), ('Others', 'julie'), ('Flavia', 'camille'), ('Quirino', 'curiace'), ('Curiacio', 'horace'), ('Cayo', 'flavian'), ('Horacio', 'le-vieil-horace'), ('Eufrosina', 'valere'), ('Tulio', 'Others')]
lope-laura_perseguidar otrou-laure_persecutee	0.17	[('le-comte', 'RUFINO'), ('orantee', 'ORANTEO'), ('l-infante', 'Others'),, ('clidamas', 'ESTACIO'), ('le-roi', 'REY'), ('lydie', 'LEONARDA'), ('laure', 'LAURA'), ('octave', 'OCTAVIO'), ('Others', 'BELARDO')]
scudery-comedie_comedie nscorneille-illusion_comique	0.2	[('BEAU-SOLEIL', 'Others'), ('BELLE-ÉPINE', 'PRIDAMANT'), ('IRIS', 'ALCANDRE'), ('MONSIEUR_DE_BLANDIMARE', 'CLINDOR'), ('TANCREDE', 'MATAMORE'), ('BELLE-OMBRE', 'ADRASTE'), ('Others', 'ISABELLE'), ('ALCIDON', 'LYSE'), ('MADAME-BEAU-SOLEIL', 'GÉRONTE')]
lope-villana_xetaferotr ou-diane	0.2	[('FULGENCIO', 'diane'), ('DOÑA_ANA', 'dorothée'), ('LOPE', 'célirée'), ('DON_FÉLIX', 'orante'), ('PASCUALA', 'filémon'), ('HERNANDO', 'sylvian'), ('Others', 'Others'), ('INÉS', 'lysimant'), ('DOÑA_ELENA', 'orimand')]
mira_amescua-desdicharotrou-belissaire	0.2	[('FLORO', 'Others'), ('BELISARIO', 'belissaire'), ('LEONCIO', 'leonse'), ('Others', 'camille'), ('TEODORA', 'theodore'), ('ANTONIA', 'anthonie'), ('EMPERADOR', 'l-empereur'), ('NARSES', 'narses'), ('FILIPO', 'philippe')]
calderon-dama_duendeouvi lle-esprit_follet	0.22	[('MANUEL', 'florestan'), ('COSME', 'carrille'), ('ÁNGELA', 'angelique'), ('LUIS', 'licidas'), ('RODRIGO', 'ariste'), ('JUAN', 'lizandre'), ('ISABEL', 'isabelle'), ('BEATRIZ', 'lucinde')]

Figure 4 : an alignment guessed incorrectly by the algorithm, between two version of Sofonisba.

trissino-sofonisbamairet-sophonisbe	0.41	[('#sofonisba', 'syphax'), ('lelio', 'sophonisbe'), ('famiglio', 'Others'), ('massinissa', 'phenice'), ('scipione', 'corisbe'), ('messo', 'caliodore'), ('coro', 'massinisse'), ('Others', 'scipion'), ('erminia', 'lelie')]

4. Semantic approach

Looking at plays in a purely structural manner obviously deletes information that can be useful for comparison purposes. In the following, we present the beginning of our work in this direction.

4.1 *Methods and tools used*

To access the semantic meaning of the plays, we use the language model CamemBERT (Martin et al., 2019). On a surface level, CamemBERT allows us to get a semantically-informed numerical representation of our texts, called embeddings, on which computing distances and similarity measures is easier. CamemBERT has however been trained on a contemporary French dataset. Most of our plays are either in 17th century French, or in foreign languages. To solve this problem, we took a simple approach to test our methods: Spanish plays have been translated using an automatic tool (Google Translate), and French plays have been automatically modernized using a tool from Bawden et al. (2022). These add extra steps and approximations to the process, but turn out to be sufficient for our method.

4.1.1 Averaging embeddings

When using LLMs, the amount of data to handle quickly explodes. If every word in a play is

tokenized in two tokens, and every token corresponds to a 256-dimensional vector, it is easy to reach several million numbers to compare. The simplest way to reduce this information is to take the average over all words of the texts, to compute a unique vector. It is then easy to compare two plays, by simply computing the distance between both of their vectors.

We ran this comparison on all 1 216 020 pairs of plays available from the *French DraCor* corpus. A look at the closest pairs identified reveals that this method characterizes style, more than anything: the closest plays are, apart from duplicates, plays from the same authors and periods. Removing the stop words yields similar results.

4.1.2 Token matching method

A more precise approach than averaging over a whole play is to try and match token embeddings between two plays. For this part of our work, we started by manually aligning a pair of plays, at the level of character lines, highlighting lines, or parts of lines, that were very similar between both plays (see Figure 5). In this way, each line of a play was either matched with a corresponding line in the second play, or considered deleted.

Then, for each line of both plays, we computed a tokenization and a vectorization, filtering out punctuation, stopwords, and character names. Automatically comparing lines can then be done by checking for the closest tokens in each pair of lines. Again, this computation yields a distance for every pair of tokens. In our methodology, we kept the 5 closest tokens for every pair. Ordering those pairs of lines (either by taking the average of the 5 distances, or by taking the best one) allows one to check if this approach detects a similarity where the human alignment signaled one, and to find similarities it could have missed.

Figure 5 : Three consecutives lines aligned between La Dama Duende *and* L'Esprit folet

Pedro Calderón de la Barca, *La Dama Duende*, automatically translated to French	**Antoine Le Métel d'Ouville, *L'Esprit folet***
MANUEL| Arrêtez-le avec quelque chose| industrie; mais, si avec elle| Je ne peux pas, ce sera forcé **le recours à la force** sans qu'il en comprenne la cause.	Florestan| Elle est femme, il suffit, cherchons donc je te prie,|Pour en venir à bout quelque prompte industrie:| Et si par ce moyen je ne puis l'arrester,| **J'useray de la force.**
COSME |Si vous cherchez de l'industrie, attendez,| celui-là m'est offert.| **Cette lettre,** qui confie| Il appartient à un ami, je suis désolé.| (DON LUIS et RODRIGO, son domestique, sortent.)	Carrille| Il faut donc inventer|Quelque subtil* moyen, à propos cett**e lettre| No**us y pourra servir.| (Il tire une lettre de sa poche.)
LUIS |Je dois la **connaître**| pas plus que pour les soins| avec lequel je suis soupçonné.	Licidas|Nous y pourra servir. Ouy je la veux **cognoistre**| Avant qu'elle m'eschape, et veux sçavoir pourquoy| Elle fuit ma rencontre, et se cache de moy.

4.2 Results

Looking at the closest pairs of tokens found, our method correctly identifies most of the pre-aligned lines. Ordering by best distance favorites lines that have exact words matching between the two, but the method is still sensitive to semantic proximity between both lines. Moreover, some of the top-scoring lines pairs were missed by us during manual annotation, and are in fact very similar, which allowed us to refine the first alignment.

On a computational level, vectorizing two complete plays can be resource-intensive, and this

process, together with the modernization, took several hours on a laptop. The code is, however, not optimized yet, and could run much faster using GPUs and dedicated machines. The comparison process between each pair of tokens is however very fast, and can thus be tweaked very easily (removing or adding stop words, character names, word limit, etc).

5. Conclusion and future works

We showcased two techniques allowing us to compare plays: one using traditional algorithmic tools, and one using machine learning techniques. The parameterized matching approach seems to be fit for detecting structural similarities. Refining the distance computation and evaluation method is still needed to limit false positives, and is the subject of ongoing work. The NLP approach shows very promising results and could help get more precise alignments. The human input needed for this method is still high. Checking for distance between all pairs of lines instead of pre-aligned one is a possible generalization to this method, which needs to be accompanied by an efficient way to interpret results.

Our method makes it easier to detect sources on a large scale, and could be used to better understand and analyze the overall movements that lead authors to draw their inspiration from particular sources. In this particular application, it helps us refine the periodization of 16th- and 17th-century theater according to the preferred sources of inspiration. Moreover, it enables us to distinguish between different imitative practices, bringing to light different types of borrowing, both textual and structural, which is fundamental to a genetic approach to dramatic texts. The next step for this work is to go back to the texts, and complement our distant reading approach with a close reading one.

References

Bourhis P., Boussidan A. and Gambette P. (2023). On Distances Between Words with Parameters. In Bulteau L. and Lipták Z. (Eds.), *Proceedings of the 34th Annual Symposium on Combinatorial Pattern Matching (CPM 2023)*. Leibniz International Proceedings in Informatics (LIPIcs) 259, Schloss Dagstuhl, 6:1-6:23.

Bawden R., Poinhos J., Kogkitsidou E., Gambette P., Sagot B. and Gabay S. (2022). Automatic Normalisation of Early Modern French. In *Proceedings of the 13th Language Resources and Evaluation Conference, European Language Resources Association (LREC 2022)*, Marseille, 3354-3366.

Bourqui C. (2014). Du corpus à l'outil herméneutique : la base de données intertextuelle *Molière 21*. In Pety D. (Ed.), *Patrimoine littéraire en ligne : la renaissance du lecteur ?*, Chambéry, 57-68.

Brunet É. (1988). Une mesure de la distance intertextuelle : la connexion lexicale. *Revue Informatique et Statistique dans les Sciences Humaines*, 24 (1-4), 82-116.

Burnard, L. (2014). *What is the Text Encoding Initiative? How to add intelligent markup to digital resources*. Marseille: OpenEdition Press, 2014.

Douguet, M. (2018). Les hémistiches répétés. In Celardo L., Fioredistella Iezzi, D. and Misuraca M. (Eds.), *Proceedings of the 16th International Conference on Statistical Analysis of Textual Data (JADT 2022)*, 215-222

Douguet, M. (2020). Les hémistiches répétés chez Corneille. In Dufour-Maître M. and Laurin C. (Eds.), *Pierre Corneille, la parole et les vers*, 9:1-9:12

Fischer F., Börner I., Göbel M., Hechtl A., Kittel, C., Milling, C. and Trilcke P. (2019). Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. In *Proceedings of DH2019: "Complexities"*, Utrecht University.

Fournial C. (2019). *Imitation et création dans le "théâtre moderne" (1550-1650) : la question des cycles d'inspiration*. Thèse de doctorat en littératures. Sorbonne université.

Glorieux F. (2016), *Dramagraph*, http://obvil.lip6.fr/Dramagraph/

Lebart L., Pincemin B., Poudat C. (2020), *Analyse des données textuelles*, Québec: Presses universitaires du Québec.

Levenshtein V.I. (1966), Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics. Doklady*, 10 (8), 707-710.

Luong X. (Dir.) (2003), *Corpus*, 2, La Distance intertextuelle.

Marcus, S. (1970), Chapter VIII, Mathematical Methods in the Study of Theatre, *Poetica matematică*. Bucharest : Editura Acad. Rep. Sot. Romania, 1970

Martin L., Muller B., Javier Ortiz Suárez P., Dupont Y., Romary L., De la Clergerie É., Seddah D., Sagot, B. (2020), CamemBERT: a Tasty French Language Model. In Jurafsky D., Chai J., Schluter N., and Tetreault J. (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, Online, 7203-7219.

Moretti F. (2013) *Pamphlet 6. "Operationalizing": or, the function of measurement in modern literary theory*, Stanford: Literary Lab.

Pagel J., Sihag N. and Reiter N. (2021). Predicting Structural Elements in German Drama. In Ehrmann M. et al. (Eds), *Proceedings of the Conference on Computational Humanities Research 2021*, Amsterdam, CEUR Workshop Proceedings, 2989, 217-227.

Santa María M.T., Calvo Tello J. and Jiménez C.M. (2021). ¿Existe correlación entre importancia y centralidad? Evaluación de personajes con redes sociales en obras teatrales de la Edad de Plata. *Digital Scholarship in the Humanities*, 36 (supp. 1), i81-i88.

Tempestt N., Kalaivani S., Aneez F., Yiming Y., Yingfei X. and Damon W. (2017), Surveying Stylometry Techniques and Applications. *ACM Computing Surveys*, 50(6), 86:1-86:36.

van der Deijl, L. (2023), Story patterns in early modern drama. Visualising character networks and plot speed in 22 Dutch plays by Nil Volentibus Arduum, *Proceedings of Digital Humanities Benelux 2023.*