jcls. Journal of Computational Literary Studies

conference version

Repetition and Innovation in Dramatic Texts

An attempt to measure the degree of novelty in character's speech

Botond Szemes[1]
Mihály Nagy[2]

1. Institute for Literary Studies, HUN-REN Research Centre for the Humanities, Budapest, Hungary.
2. Doctoral School of History, Eötvös Loránd University, Budapest, Hungary.

OPEN ACCESS

Citation
Botond Szemes and Mihály Nagy (2024). "Repetiton and Innovation in Dramas. An attempt to measure the degree of novelty in character's speech". In: *CCLS2024 Conference Preprints* 3 (1). 10.26083/tuprints-00027395

Date published 2024-05-28
Date accepted 2024-04-03
Date received 2024-01-25

Keywords
computational drama analysis, information theory, innovation, sentence embedding, Shakespeare

Note
This paper has been submitted to the conference track of JCLS. It has been peer reviewed and accepted for presentation and discussion at the 3rd Annual Conference of Computational Literary Studies at Vienna, Austria, in June 2024.

Abstract. In the following, we develop a method to study dramas as information networks. We examine how innovative characters are in relation to each other, i.e. whether they tend to repeat the utterances of others or introduce new information to the discourse of the play. Our method captures the role of characters in this discourse, and through pairwise comparisons, we can also construct networks that represent character relationships in a new way compared to existing approaches. By examining some of Shakespeare's plays, we also identify general patterns regarding the structural differences of the networks and gender roles in comedies and tragedies/non-comedies.

1. Introduction

In dramatic works, the flow of information maintained by the speech acts of the characters is particularly important. In terms of the *internal communication system*, the flow (or the withholding) of information between characters is the driving force of the plot (Andresen et al. 2022, 2024); in terms of the *external communication system*, the audience/readers gain access to the storyworld also mostly through the dialogues (for theoretical description of the two types of systems, see Pfister 1988). Accordingly, co-presence or co-occurrence networks (Trilcke 2013; Trilcke et al. 2015), which have become increasingly popular in recent years, are also often interpreted from the perspective of the internal information flow, although usually implicitly, as in the case of using betweenness centrality as a metric to infer the mediating, even "conspiratorial" role of characters (e.g. Algee-Hewitt 2017; Szemes and Vida 2024). Benjamin Krautter, however, points out that knowledge networks, which represent the transfer of knowledge between characters, and which may well show a different arrangement than co-presence networks, are more helpful and theoretically better grounded in such an investigation of information flow (Krautter 2023, see also Andresen et al. 2022).

In contrast to these approaches, the present study analyses the information value of characters' speeches in Shakespeare's works from the perspective of the *external communication system*, i.e. from the perspective of the recipient. Andresen et al. 2022 also took this aspect into account in their research, albeit in less detail and focusing on just a specific type of knowledge transmission. Furthermore, we do not follow Manfred Pfister's theory (Pfister 1988) strictly in our analysis as they did. That is, we do not only consider

utterances when a character conveys specific knowledge to the audience;[1] rather, we consider all utterances according to the extent to which they add new meanings to the storyworld. When in *Hamlet*, for example, Claudius raises the idea of Hamlet's exile, the information value of the speech is increased by the mentioning of England (and its relationship to Denmark) for the first time in the play – the horizon of the storyworld is literally expanded. However, Denmark's foreign policy relations (with Norway) have been discussed before, so the difference from the earlier discourse is not that great. Equally, it can be informative if a character speaks in a new register, different from previous ones, since this shows that such ways of speaking are in fact possible in the represented world, and that these as contexts influence the interpretability of other utterances as well. Consider, for example, the differences between the royal speech at the beginning of *Hamlet*'s second scene and the sentences exchanged between Horatio and his companions in the first scene, or the dialogue of the Gravediggers in Act 5. The tensions between the royal propaganda and the friendly or humorous remarks create the framework in which the tragedy unfolds. The Gravediggers' sentences about Hamlet's exile are less novel, however, as this is already mentioned earlier in the play (see the comparison of sentences from these characters in Appendix 2.) Together, we refer to these types of differences from the previous discourse as *semantic difference*, which according to our experiments can be captured well with the use of BERT-based language models. The term indicates a focus on the content of the dialogues, but also a consideration of the semantic components of style (for example, a highly metaphorical utterance is usually more distinct from sentences that elaborate the meaning less metaphorically.)

In light of this, we are interested in the role that a character plays in shaping the storyworld. Two general functions can be distinguished according to the extent to which they contribute to the creation of new meanings by often deviating from what has been said before, or to the extent that they repeat and thus reinforce an already established discourse. *Innovative characters* are responsible for the elaboration of new (semantically distinct) meanings, while *repeaters* or *maintainers* contribute to the development of the central themes and the general ways of speaking in the drama. There is, of course, also a duality of innovation and repetition within each individual character. This can also be detected with our method, since we calculate the semantic difference between each sentence and its preceding discourse for each character, which makes it possible to examine the distribution of both functions in the cast separately. This sentence-level approach can also help us to answer the question of what the innovative function of a character means in a specific case beyond the broad definition. In this paper, we argue that Shakespeare's innovative characters can be divided into two groups: those who are in fact responsible for transmitting knowledge, and those who speak in a different way from the dominant discourse in the drama, usually expressing uncertainty and/or emotion, or using metaphorical language. Our results, furthermore, provide a novel way of describing the difference between comedies and tragedies (or more preciesly "non-comedies"[2]). Namely that female characters in Shakespeare's comedies are more likely to have innovative functions and be repeated by others compared to tragedies.

1. Pfister's example is Prospero's speech to Ariel in the beggining of *The Tempest* (I/ii, 250-293), which is more informative for the audience, since Ariel already knew everything that was in the speech.
2. Dramas labelled as „comedy" are those that are listed as such in the First Folio (1623). All others are labelled as „non-comedy" or sometimes in the paper as "tragedy" for the sake of simplicity. For the structural similarities of the "non-comedies" (and their resemblance to tragedies) see Szemes and Vida 2024

Finally, the paper also addresses the question of the network representation of character relations. Benjamin Krautter has pointed out that the interpretability of networks is significantly affected by the type of relations they represent – different methods lead to different conclusions (Krautter 2023). In the following, we present a new method intended to complement already existing ones. It is based on defining the innovativeness of a character's speech along pairwise comparisons, i.e. comparing characters with each other separately. On the one hand, this makes it possible to measure the similarities between two characters at sentence level. On the other hand, it allows us to represent the relationships on a directed graph, showing which character in the pairwise comparison is more likely to repeat the other. Similarly to Andresen et al. 2022, we attempt to use "a more content-based form of character networks [...] to chart a path to better integrate quantitative analysis and interpretative reading." In the resulting networks, the role played in the whole discourse of the drama and the relationship between two characters can be examined simultaneously.

2. Related Works

The paper draws from previous research within information theory that has likewise attempted to measure innovation and repetition in different communicative situations. However, these studies differ not only in their methods, but also in their theoretical assumptions. As well as in their understanding of the terms 'information', 'novelty', or 'innovation'. Therefore the paper must be situated within previous research and define its subject of measurement – i.e. how it considers the concept of 'innovation' to be operationalised in the study of dramas.

South et al. 2022 analyzed repeated linguistic elements to detect the flow of information between Twitter accounts of news organizations. They assume that when more words exist in the same order across two texts, the degree of novelty between them is lower, and vice versa that previously unused phrases and novel word order make a text innovative. Accordingly, their method is based on the identification of the longest repeated sequences of words. This approach functions well in the case of Twitter posts, however, when applied to less homogenous and considerably more poetic dramatic texts, it is less useful. This is because in such texts, repeating sequences almost in all cases are conventionalised expressions (e.g.: 'there are', 'good morning'). Therefore, the results would not primarily indicate semantic similarity.

Sims and Bamman 2020 also set out to explore recurring linguistic elements when determining the role of characters in a novel's social and information networks. Beyond considering the mere frequency of words, they also examined POS tags and grammatical relations. Using a selection of verbs that describe the most important events of a plot, they identified 'Subject – Verb – Object' triples (e.g.: 'Thomas – left – Vienna') – if a triple is mentioned by two characters, we can say that they refer to the same event so that the former has an *informational impact* on the later. The challenges of the method include inaccuracies in co-reference resolution (which assigns each utterance to the corresponding character, although this is much simpler in dramatic works) and in dependency analysis, as well as the somewhat arbitrary selection of the group of verbs to be considered. Whereas Sims and Bamman 2020 sought to explore the direct effect between characters

(internal communication system), we interpret innovation and repetition in relation to the entire discourse preceding an utterance (external communicational system): even though we make pairwise comparisons, we do not assume that the similarity of two characters' utterances indicates a direct causal relation; we just examine the extent to which the content of an utterance is similar to what was said before.

The same question was asked by Barron et al. 2018, who measured whether speeches by members of the Parliament during the French Revolution had raised new themes or contributed to maintaining previous ones. Their approach applies Kullback–Leibler Divergence (KLD), a measure often used in similar contexts due to its strong foundation in information theory. In short, with KDL the difference between the vector representation of texts is not calculated through the spatial metaphor of distance (how far one text is from another in a vector space), but through a model of *experience* (how surprising a text is when conditioned on prior knowledge - see Chang and DeDeo 2020). Barron et al. 2018 first determined the distribution of different topics across parliamentary speeches, then compared these distributions with the help of KLD. A similar attempt was made by Piper et al. 2023 who, on the other hand, used a simple distribution of word frequencies of equal-length chunks to calculate their divergence, through which they could measure the process of narrative revelation.

Since the comparison of texts in this study is based on their semantic relations, neither the consideration of the longest recurring sequences nor word frequency distributions proved to be useful approaches. Similarly, doing topic modelling like Barron et al. 2018 also proved impractical, because in the case of a drama, the utterances are usually too short to effectively identify themes in them. Nor does one drama provide enough data to distinguish the characters efficiently according to the distribution of themes. Therefore, we use Large Language Models (LLMs) to determine the position of each sentence of a drama within a vector space representing the semantic field of the given language. The embedding process is driven by the SBERT (Sentence-BERT) algorithm, which can quantitatively capture the meaning of larger units, such as sentences, compared to the word-level embeddings of previous BERT models (Reimers and Gurevych 2019). The vector representation of separate sentences makes their semantic comparison possible, which can be utilized in our research to examine the character speeches based on their content. *Semantic similarity* refers mainly to thematic similarities, but also includes the style of the sentences (e.g. terms belonging to the same style/register are semantically more similar). In light of this, we can say that semantically the less similar a sentence is to its predecessors, the greater the degree of information it conveys (innovativeness). Conversely, the more similar a sentence is to its predecessors, the more it contributes to the repetition of an already existing discourse.

This was the approach also used by Dubourg et al. 2023 in their study measuring the innovation of movie plots. Converting the plot summaries of over 19,000 films into vectors with the help of the SBERT algorithm, they calculated the cosine similarity between a summary and all preceding film summaries and averaged them to determine a film's Innovation Score, i.e. the average distance of the current embedding from previous ones. Our method compares the sentences spoken by characters in a similar way. It is important to note because Dubourg et al. 2023 also evaluated the method and found their results to be positively correlated with results from text mining of viewer

reviews (see Luan and Kim 2022). In our case such a comparison is not possible due to the lack of other results and because, as we have seen, the procedures mentioned so far cannot be adapted without problems to answer our research question.

Indeed, so far in the field of quantitative drama analysis, there have not yet been any attempts to answer such a question relating to repetition and innovation in a character's speech. Most of the previous research investigated primarily the structural characteristics of plays (for an overview: Szemes and Vida 2024); while other, more language-oriented investigations have mostly experimented with topic-modelling of larger corpora (and explore genre differences - see Schöch 2017), and regarding Shakespeare's works most attention has been paid to authorial style and keyword analysis (Craig and Kinney 2009), or uncovering changes in word use in the oeuvre (Hope and Witmore 2014). The closest to the research is that of Andresen et al. 2022 and Krautter 2023, with the differences already mentioned in the *Introduction*. It is also important to refer to the research of Šeļa et al. 2024, in which they used stylometric methods developed for authorship attribution to calculate the difference between characters' speeches. However, their focus was not on the semantic content of the texts and their degree of innovation, but exclusively on their stylistic differences. We hope, therefore, that our study will provide new perspectives to the field, and at the same time enrich the interpretability of certain plays.

3. Method

For our study, we used dramas from Shakespeare in TEI-XML format provided by the Drama Corpus Project (Fischer et al. 2019).[3] As a first step we created a tabular representation of all the individual sentences from a play. We assigned to each sentence 1) the name of the character, 2) a timestamp representing the position of the spoken text within the whole drama (from 1 to the last sentence), 3) the number of the act in which the sentence is spoken, and 4) the embedding score provided by a language model. Regarding the last point, the selection of the right model is a primary concern. Using example sentences taken from the corpus, we experimented with several state-of-art best-performing SBERT models.[4] We selected sentences with similar and dissimilar meanings (at this stage we judged similarity intuitively and the selection was made manually), and calculated their cosine similarity in a pairwise manner. Subsequently, we calculated the standard deviation of the similarities. Although there was a minimal variation between the models, we chose to use the popular 'all-MiniLM-L6-v2', as its results showed the highest standard deviation, which means that the distribution among similar and dissimilar meanings are the largest in this case. See the experiment details and the performance of the chosen model in the project's GitHub repository (*Software availability*) where the performance can also be evaluated manually by looking at the most/least similar sentence pairs of the plays (see also the *Appendix* and the *Results* sections for further manual evaulation.) Regarding the most similar sentences, for example, character names seem to have a strong influence on sentence similarity. The names could have been therefore filtered out during the pre-processing stage, but it was considered worth keeping them because of their role in the creation of meaning. At the same time, sentences with fewer than four words (e.g., "Yes, sir") were excluded, as they

3. https://dracor.org/shake
4. See the list of best-performing models: https://www.sbert.net/docs/pretrained_models.html

are less likely to convey relevant meaning, but are rather conventionalised expressions.

We then created pairs from the most frequent speakers (i.e. the main characters[5]) in a specific order: the first member of the pair became the *Source*, and the second the *Target* character. During their comparison, we calculated the cosine similarity between a Target-sentence and all the preceding Source-sentences. In contrast to the method of Dubourg et al. 2023, we did not take the average of these similarities but only selected the largest of them to characterize semantic proximity. Thus, for each sentence of the Target character, we assigned a number indicating *how semantically similar it is to the most similar* of the previous sentences of Source (Maximum Cosine Similarity - MCS). It can be assumed that the higher the number, the less innovative the meaning of the sentence since it repeats previous content.

There are several arguments for using the Maximum Cosine Similarity instead of the average. Firstly, if a Source character speaks on many different topics in many different registers before the current Target-sentence, then on average this Target-sentence will be less similar, even if the Source character has spoken the same sentence before. MCS avoids this by focusing on the maximum value, however, this also means that the result does not report on *how often* the Source character has elaborated similar meanings. Secondly, MCS values can be used to find the most similar sentence pairs between Source and Target, contributing to the overall interpretability of the results. Thirdly, the average cosine similarity (as Dubourg et al. 2023 also point out) is strongly influenced by temporality: the later the utterance, the more similar it is on average to the earlier discourse (see Fig 1a). Therefore, by using the average cosine similarity, we would measure more the time in the plot at which a character speaks, than the novelty of his or her sentences. The MCS is also exposed to temporality, but to a much lesser extent (Fig 1b), and the effect can be compensated for by weighting/adjusting the values (Fig 1c). To do this, we first calculated the average MCS value for each act and for the drama as a whole, and then used the difference between the values for the acts and for the drama to weigh the scores according to the act in which the sentence was uttered. For example, the sentences in the first act were weighted by the difference between the average MCS for the first act and the drama as a whole. At the same time, a high degree of variation can be seen in the dataset: sentences with high MCS values can be found in the first act just as much as low ones at the end of a drama.

In the next step, we assigned the average of the weighted MCS scores to each Source-Target pair and performed network normalization on the dataset following the methodology developed by South et al. 2022. The key consideration here is that if character "B" frequently repeats character "A", but character "A" also repeats other characters, then character "B" is indirectly connected to such other characters as well. To conduct our network normalization, we determined the average score of a given character as Target, and then divided all similarity scores by this number where this character was the Source.

Finally, we calculated the differences for character pairs depending on which character

5. Main characters are considered those with more than 30 long sentences for shorter plays (less than 1000 long sentences), more than 40 for plays with mediium length (number of long sentences between 1000 and 1700), and more than 50 for longer plays. Occasionally, individual considerations may also come into play, for example if a character speaks a lot but only in one scene (e.g. the Gravediggers in *Hamlet*).

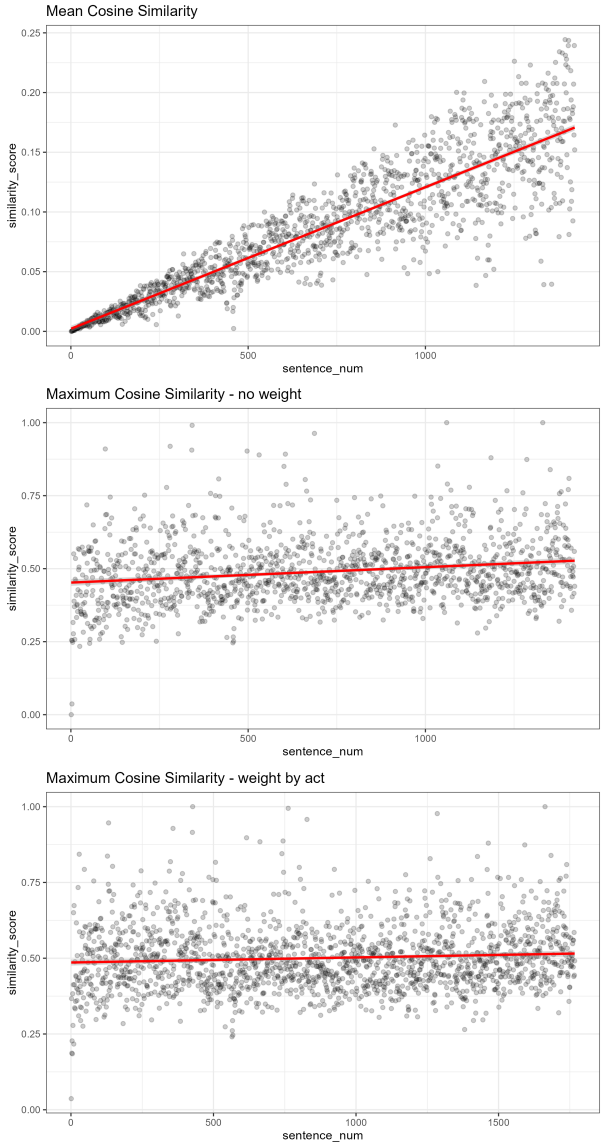
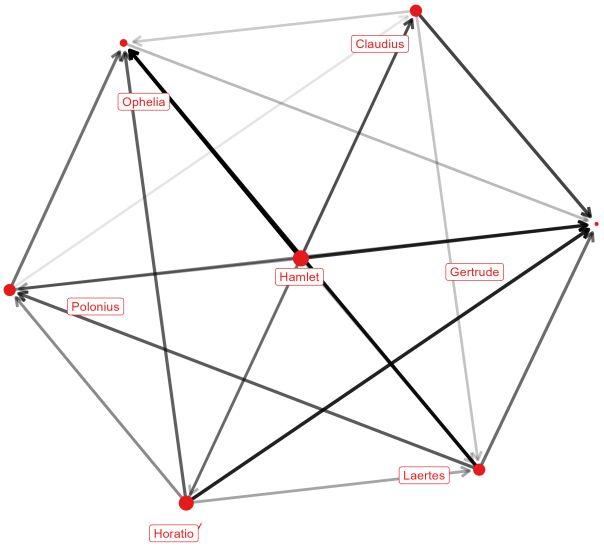


Figure 1: The relationship between time of utterance and similarity score in *Hamlet*. Up: Mean Cosine Similarity, Middle: Maximum Cosine Similarity - without weight, Down: Maximum Cosine Similarity - weight by act.

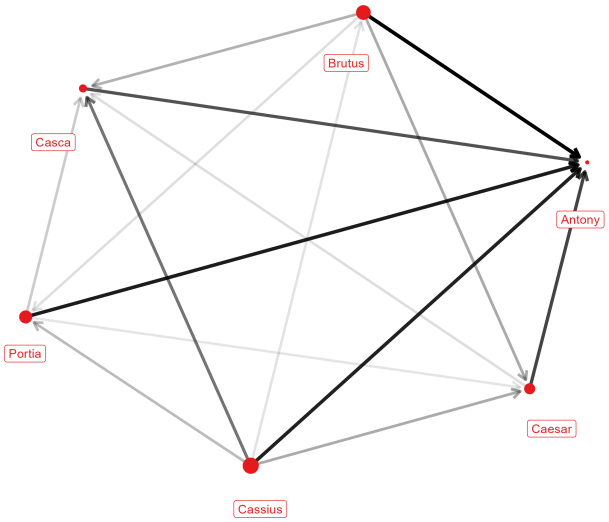
is listed as the Source or Target (e.g. Hamlet-Claudius vs. Claudius-Hamlet). If the difference is positive, then the Target character's sentences are more likely to develop a similar meaning to the Source character's earlier sentences than vice versa - i.e. the Source character is considered more innovative in their relationship. As a final result, only these positive values were retained and used for network visualization.

4. Results

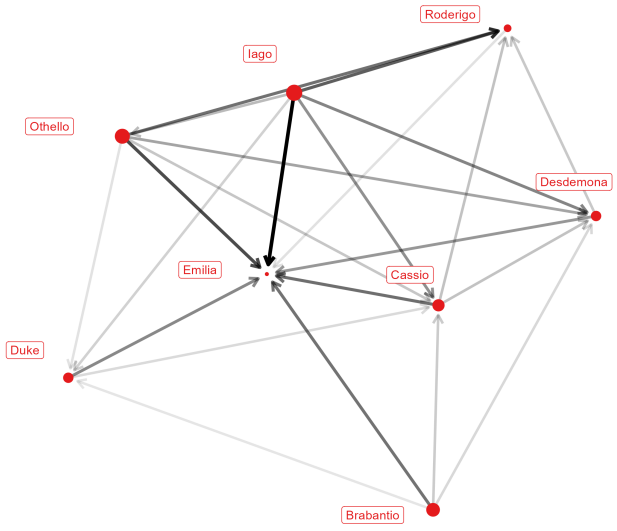
The results allow us to visualize the relationships between characters in terms of repetition and innovation as a network. In the example networks seen in Figure 2, the arrows go from Source to Target (indicating which character is more likely to repeat the other), their thickness is determined by the degree of similarity/repetition, and the size of the nodes as an innovation score indicates how often the character is listed as Source, i.e.



(a) *Hamlet*

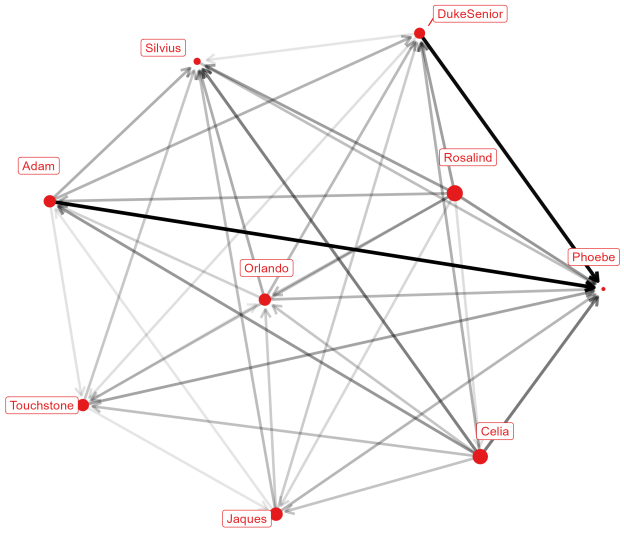


(b) *Julius Caesar*

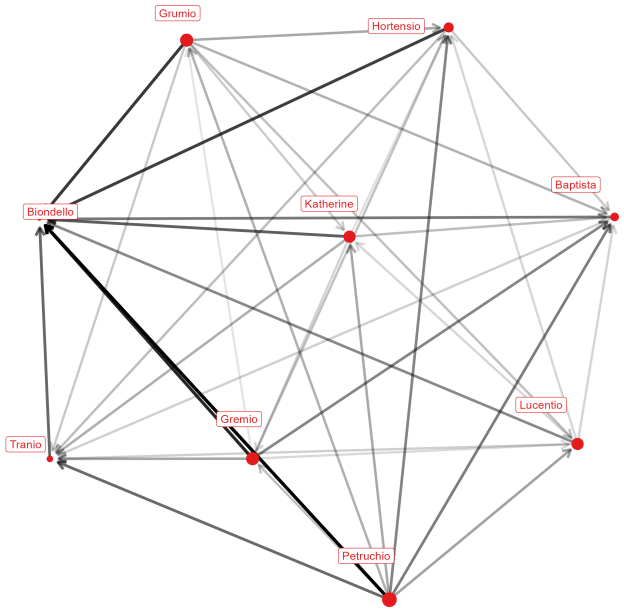


(c) *Othello*

Figure 2: Networks of Shakespeare's plays.

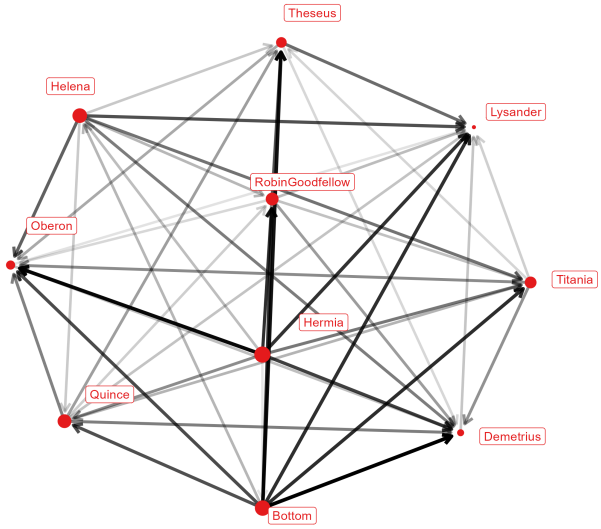


(d) *As You Like It*



(e) *The Taming of the Shrew*

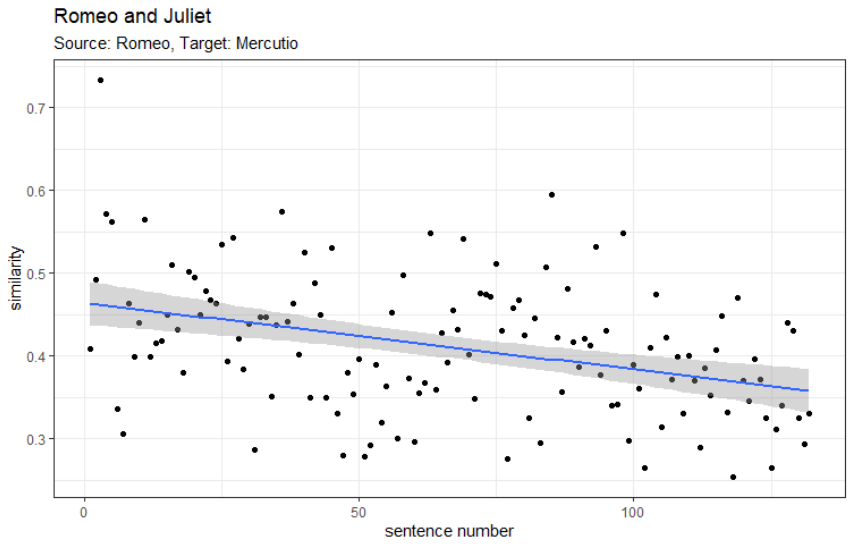
Figure 2: Networks of Shakespeare's plays.



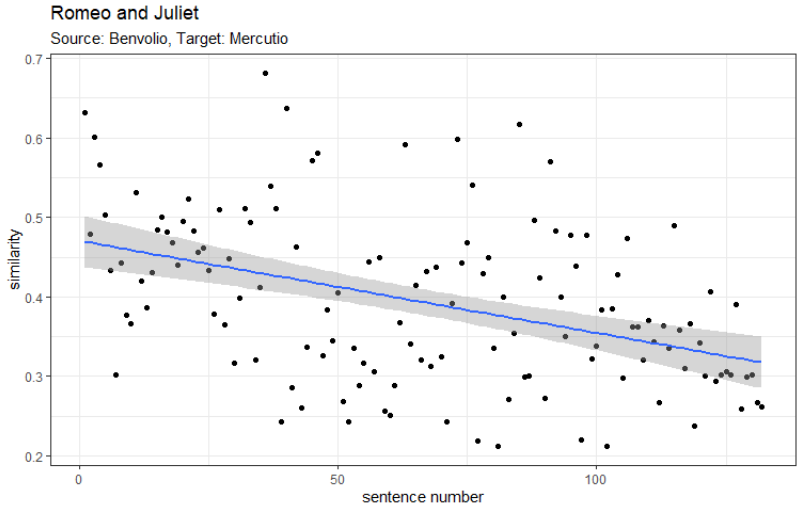
(f) *A Midsummer's Night Dream*

Figure 2: Networks of Shakespeare's plays. The arrows go from Source to Target (indicating which character is more likely to repeat the other), their thickness is determined by the degree of similarity/repetition, and the size of the nodes indicates how often the character is considered innovative in pairwise comparisons.

how often it is considered innovative in pairwise comparisons. The latter is influenced by both the number of observed sentences and partly the time of utterance: the chance of a character being novel is increased by speaking both earlier, and on more occasions. Even though we applied the above-mentioned weighting method, characters that speak mainly in the second half of the plot generally received lower innovation points (e.g. Antonius in *Julius Caesar* or Emilia in *Othello*). We do not see this as a measurement bias but as a characteristic of a character type. This is supported by the fact that there are also examples where as the plot progresses one character becomes increasingly different from another, such as Mercutio, the character with the highest innovation score in *Romeo and Juliet*, compared to both Romeo and Benvolio, the characters with the second and third highest scores, respectively (Figure 3).



(a) Target = Mercutio, Source = Romeo



(b) Target = Mercutio, Source = Benvolio

Figure 3: Changes in maximum cosine similarity over time between the most innovative characters in *Romeo and Juliet*. Mercutio's sentences become less similar to others.

The overall examination of Shakespeare's dramas shows that the relationship between characters is in most cases hierarchical (i.e. the characters can be ordered hierarchically according to their innovation scores). This is particularly true for tragedies/non-comedies, where the characters with the highest innovation scores can almost always be arranged in a hierarchical way, and only at lower levels can equal scores be found. Equal scores mean that there is a degree of circularity in the dramas: character "A" tends to repeat "B", "B" repeats "C", whereas "C" repeats "A" etc. At a higher level, this happens mainly in comedies (among non-comedies, in *Cymbeline*, *Macbeth* and *Pericles*, a play with much debated genre). For example, in *The Taming of the Shrew* Grumio and Gremio, and also Lucentio and Katharine; in *As You Like It* Orlando, Adam and Touchstone; in *Measure for Measure* Duke, Lucio and Angelo take on the same values. This difference between genres is in line with previous results based on co-occurrence networks, which show that comedies are characterized by a denser system of relationships, while tragedies by one or two characters with a connecting function who control the social relations (more hierarchical distribution of node degrees). This also means that in comedies there are many misunderstandings and parallelisms (two characters connected by different paths) during the interactions, however, for the same reason such networks are "protected" from falling apart when a certain piece of information is revealed to be untrue. In contrast, information flow is effective and fast in tragedies, but the networks themselves are fragile, as the failure of a connecting character can lead to the disintegration of the whole system (cf. Szemes and Vida 2024).

All of this is further nuanced by another distinction between genres based on our measures. It is striking that in the 23 non-comedies the characters most repeated by others are males (except Imogen in *Cymbeline* and Lady Macbeth who is as innovative as Macbeth and Banquo), while in comedies, female characters are more likely to be the most innovative (six times out of 14). In *As You Like It* Rosalinda (and Celia in the second place) has the highest score; in *All's Well That Ends Well* the Countess (and Helen in the second place), in *The Comedy of Errors* Adriana; in *A Midsummer Night's Dream* Hermia (and Helena in the third place, while their counterparts, Lysander and

Demetrius have the lowest innovation scores among the main characters); in *Much Ado About Nothing* Beatrice, and maybe most surprisingly in *The Tempest* Miranda ahead of Gonzalo and Prospero. We can say, that in the two kinds of communities, those who thematise the discourse (or at least who is repeated more than he or she repeats others) appears to differ, although not exclusively, in terms of gender. Women are more likely to play that role in the protected networks of the comedies, and men in the effective but vulnerable tragedies.

It is also worth looking at the results of pairwise comparisons in more detail and identifying the most and least similar sentences between characters. In addition to a qualitative evaluation of the method, this can also contribute to a close reading of the dramas and a deeper understanding of the characters. As an example, in *Hamlet*, the model grasps exactly the essential duality of the main character: he is striving to define himself and others but, at the same time, is constantly doubting such identifications. Hamlet's sentences which are most similar to the earlier utterances of the other characters, are often about defining his own and others' identity; while his most different and innovative sentences report doubt and uncertainty, often in a conditional or interrogative mood (Table 1; see our GitHub repository for all the sentences and their most/least similar pairs from other characters).[6]

High similarity, low innovation	**Low similarity, high innovation**
This is I, Hamlet the Dane.	I doubt some foul play.
The King is a thing -	I would I had been there.
O God, Horatio, what a wounded name, Things standing thus unknown, shall I leave behind me!	Do they hold the same estimation they did when I was in the city?
If Hamlet from himself be ta'en away, And when he's not himself does wrong Laertes, Then Hamlet does it not; Hamlet denies it.	The time is out of joint.
Here comes the King, The Queen, the courtiers.	These foils have all a length?

Table 1: Examples of the least and most innovative sentences spoken by Hamlet as Target (Hamlet)

Hamlet's speech is most similar to the discourse of the court when he names or identifies someone/something, and most divergent when he questions or is uncertain. Since he is considered the most innovative in the drama, we can say that his sentences about doubt are predominant, and they give the essence of his character – but it is also important to see his statements in the opposite direction. Conversely, the most innovative sentences by Horatio, the second most innovative character in the drama, do not express uncertainty. He is rather the one who brings news to others and often speaks as an *eyewitness* – in this sense, he really creates new information, not just develops semantically divergent meanings (Table 2). These sentences illustrate well his dramaturgical function of linking events and communities (cf. Moretti 2011).

6. The example sentences reported here have been hand-picked for interpretation from the 10 sentences with the highest and lowest cosine distance in the pairwise comparisons. The selection is therefore somewhat arbitrary: it is analogous to a researcher trying to make sense of the output of keyword analysis or topic modelling. The full list is given in the project's GitHub repository.

Low similarity, high innovation
Not when I saw 't.
My lord, I think I saw him yesternight.
Indeed, I heard it not.
It was as I have seen it in his life, A sable silvered.
It would have much amazed you.

Table 2: Examples from the most innovative sentences spoken by Horatio (*Hamlet*)

Utterances expressing doubt, reflecting on either mental states like emotions or the outside world appear as most divergent in other characters from other dramas as well. One example is Hermia in *A Midsummer Night's Dream* (Table 3), who is the most innovative character in the drama precisely because of questioning the nature of things around her (even compared to Bottom who appears in a subplot separate from the majority of the cast and, therefore often speaks about something else). Furthermore, the duality observed in Hamlet is also characteristic of Brutus in *Julius Caesar*. His most similar sentences to the previous discourse are predominantly about the murder; whereas the least similar ones are about doubts and emotions (Table 4). It is worth comparing this with the utterances of Caesar, who only briefly expresses doubt, specifically about going to the Senate (his most innovative utterances), and instead accepts his death to maintain the conventional image of the emperor. This is shown by the fact that he often speaks of himself in the singular third person: "Caesar shall forth."; "Danger knows full well/ That Caesar is more dangerous than he." etc.

Characters with connecting functions like Horatio can be found also in other plays, whose novelty lies in their reports about specific events. Such is Cassius in *Julius Caesar*, who can be seen as an innovator even compared to Brutus. His sentences with the highest/lowest MCS score show an opposite pattern to Brutus: he repeats the others when he uses terms referring to emotions and inner values, while his sentences about concrete events differ the most (Table 5). Cassius is in charge of moving the plot forward, bringing news and argument – he also recruits the wavering Brutus into the conspiracy. Part of it is that when Cassius speaks of emotions, he is not talking about himself, but about others. On the other hand, the sentences of Brutus that mark specific events, refer not to the conspiracy but to the murder itself; they are often retrospective and thus less novel. Until the murder takes place, or until he is determined to commit it, he speaks of more abstract topics, demonstrated by one of his most divergent sentences relative to Caesar: „Between the acting of a dreadful thing/ And the first motion, all the interim is/ Like a phantasma or a hideous dream."

Low similarity, high innovation
Who is 't that hinders you?
Then I well perceive you are not nigh.
I understand not what you mean by this.
Too high to be enthralled to low.
Nothing but "low" and "little"?

Table 3: Examples of the most innovative sentences spoken by Hermia (*A Midsummer Night's Dream*)

High similarity, low innovation	Low similarity, high innovation
Mark Antony, here, take you Caesar's body.	I would not, Cassius, yet I love him well.
And for Mark Antony, think not of him, For he can do no more than Caesar's arm When Caesar's head is off.	That you do love me, I am nothing jealous.
I killed not thee with half so good a will.	If I have veiled my look, I turn the trouble of my countenance Merely upon myself.
Hold, then, my sword, and turn away thy face While I do run upon it.	But if these – As I am sure they do - bear fire enough To kindle cowards and to steel with valor The melting spirits of women, then, countrymen, What need we any spur but our own cause To prick us to redress?
But, alas, Caesar must bleed for it.	Enjoy the honey-heavy dew of slumber.

Table 4: Examples of the most and least innovative sentences spoken by Brutus (*Julius Caesar*)

High similarity, low innovation	Low similarity, high innovation
Yet I fear him, For in the engrafted love he bears to Caesar -	The clock hath stricken three.
Well, Brutus, thou art noble.	The morning comes upon 's.
I blame you not for praising Caesar so.	And I do know by this they stay for me In Pompey's Porch.
Caesar doth bear me hard, but he loves Brutus.	When went there by an age,] since the great flood, But it was famed with more] than with one man?
I know that virtue to be in you, Brutus, As well as I do know your outward favor	No, it is Casca, one incorporate To our attempts.

Table 5: Examples of the most and least innovative sentences spoken by Cassius (*Julius Caesar*)

Finally, it is worth highlighting *Othello*, in which Iago is associated with the highest innovation score. This is not surprising as he increasingly controls the discourse as the plot develops, and in some cases even makes others, especially Othello, repeat his sentences (e.g. "Men should be what they seem" [Iago], "Certain, men should be what they seem." [Othello]; "Or to be naked with her friend in bed/ An hour or more, not meaning any harm?" [Iago], "Naked in bed, Iago, and not mean harm?" [Othello]). The sentences of Othello that differ most from Iago's previous utterances are at the end of the drama. In these, he describes his situation using more abstract language, which may indicate that by the end of the plot, he will be able to view events from an external and broader perspective (Iago's mastery of always focusing his attention on the concrete signs). However, this may also indicate that he is still incapable of introducing novel information about the concrete storyworld, and thus becomes innovative compared to Iago just when he refrains from naming things, as Iago does it instead of him. This is exemplified by one of Othello's less similar sentences said to Desdemona: "Let me not name it to you, you chaste stars."

5. Conclusion

Comparing sentence-level embeddings of character utterances can be useful both for interpreting specific dramas and for identifying general patterns in bigger corpora. According to the method proposed in the paper, characters whose sentences are the most semantically different from the previous sentences of other characters can be considered innovative. In this case, the degree of difference is measured by Maximum Cosine Similarity of embedding scores of a language model (how similar the most similar sentence is), rather than the average distance from all the previous sentences. The networks resulting from pairwise comparisons present the relationships between characters and provide at the same time a new way of describing the difference between Shakespeare's comedies and non-comedies. While in non-comedies that are more hierarchical in terms of the distribution of innovation scores, the male protagonists' speeches are repeated by others, whereas in more circular comedies, female characters are more likely to thematise the discourse of the play.

When analyzing the sentence pairs with the highest/lowest similarity scores, two types of characters seem to be distinguishable in Shakespeare's plays, both of which can be considered innovative. On the one hand, some characters often introduce new information into the discourse and report on events distant in time or space. For example, Horatio in *Hamlet* as an eyewitness to various events functions as a link between groups; Cassio in *Julius Caesar*, the main organizer of the conspiracy; and Bottom in *A Midsummer Night's Dream*, who also connects a subplot with the main characters. Others don't bring new information into the discourse in the traditional sense, i.e. they do not talk about something different, but in a *different way*. This may be the result of the doubt in the established relations and identities (for example, Hamlet on the question of identity, Hermia on the perception and interpretation of the outside world), the predominance of emotions (Brutus), or the use of puns and a language with erotic connotations (Mercutio). In this context, the difference between abstract and concrete sentences also seems to be a general pattern: the more poetic and abstract an utterance is, the more innovative it appears.

6. Appendix - Cosine Similarity Scores

6.1 Similar and Dissimilar Sentences from *Hamlet* Used to Model Comparison

Sentences:

1. How now, what noise is that?
2. Alack, what noise is this?
3. Exchange forgiveness with me, noble Hamlet.
4. O Hamlet, speak no more!
5. To die, to sleep—\No more—and by a sleep to say we end\The heartache and the thousand natural shocks\That flesh is heir to—'tis a consummation\Devoutly to be wished.
6. This gentle and unforced accord of Hamlet\Sits smiling to my heart, in grace whereof\No jocund health that Denmark drinks today\But the great cannon to the clouds shall tell,\And the King's rouse the heaven shall bruit again,\Respeaking earthly thunder.
7. To be or not to be, that is the question:\Whether 'tis nobler in the mind to suffer\The slings and arrows of outrageous fortune,\Or to take arms against a sea of troubles And, by opposing, end them.
8. Though yet of Hamlet our dear brother's death\The memory be green, and that it us befitted\To bear our hearts in grief, and our whole kingdom\To be contracted in one brow of woe,\Yet so far hath discretion fought with nature\That we with wisest sorrow think on him\Together with remembrance of ourselves.
9. Ay, truly, for the power of beauty will sooner transform honesty from what it is to a bawd thanthe force of honesty can translate beauty into his likeness.
10. Could beauty, my lord, have better commerce than with honesty?
11. Rest, rest, perturbed spirit!
12. Their residence,both in reputation and profit, was better both ways.

Similarity scores:

2	0.85										
3	0.04	0.04									
4	0.11	0.09	0.59								
5	0.05	0.09	0.36	0.34							
6	0.12	0.13	0.52	0.47	0.54						
7	-0.04	-0.01	0.39	0.33	0.40	0.32					
8	-0.03	-0.04	0.53	0.53	0.53	0.55	0.39				
9	-0.05	-0.07	0.26	0.19	0.30	0.31	0.22	0.25			
10	-0.06	-0.09	0.26	0.14	0.19	0.28	0.21	0.18	0.72		
11	0.10	0.09	0.23	0.18	0.42	0.36	0.19	0.27	0.20	0.14	
12	0.04	-0.03	0.16	0.01	-0.02	0.09	0.10	0.05	0.07	0.24	-0.03
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**

6.2 Similar and Dissimilar Sentences from *Hamlet* – Examples from the First Scene, the King's Speech and the Gravediggers's Dialogue

Sentences:

1. He shall with speed to England\For the demand of our neglected tribute.

2. It was that very day that young Hamlet was born — he that is mad, and sent into England.
3. Th' ambassadors from Norway, my good lord,\Are joyfully returned.
4. Therefore our sometime sister, now our queen,\Th' imperial jointress to this warlike state,\Have we (as 'twere with a defeated joy,\With an auspicious and a dropping eye,\With mirth in funeral and with dirge in marriage,\In equal scale weighing delight and dole)\Taken to wife.
5. I think it be no other but e'en so.
6. Is not this something more than fantasy?
7. It harrows me with fear and wonder.
8. I like thy wit well, in good faith.
9. Cudgel thy brains no more about it, for your dull ass will not mend his pace with beating.

Similarity scores:

2	0.34							
3	0.27	0.22						
4	0.35	0.28	0.31					
5	0.10	0.12	0.15	0.19				
6	0.05	0.12	0.03	0.19	0.16			
7	0.19	0.23	0.09	0.29	0.19	0.17		
8	0.06	0.17	0.23	0.21	0.14	0.09	0.18	
9	0.26	0.23	0.08	0.20	0.10	0.10	0.23	0.20
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**

7. Data Availability

Data can be found here: https://github.com/dracor-org/shakedracor

8. Software Availability

Software can be found here: https://anonymous.4open.science/r/innovation-drama/

9. Acknowledgements

Botond Szemes was supported by the ÚNKP-23-4 New National Excellence Program of the Ministry for Culture and nnovation (Hungary) from the source of the National Research, Development and Innovation Fund.

The authors are grateful for the help of Zsombor Komán in application of LLMs.

10. Author Contributions

Botond Szemes: Conceptualization, Methodology, Visualization, Writing - original draft

Mihály Nagy: Preprocessing, Methodology - LLM, Writing – editing

References

Algee-Hewitt, Mark (2017). "Distributed Character: Quantitative Models of the English Stage, 1550–1900". In: *New Literary History* 4.48, 751–782. https://doi.org/10.1353/nlh.2017.0038.

Andresen, Melanie, Benjamin Krautter, Janis Pagel, and Nils Reiter (2022). "Who Knows What in German Drama? A Composite Annotation Scheme for Knowledge Transfer. Annotation, Evaluation, and Analysis". In: *Journal of Computational Literary Studies* 1. https://doi.org/10.48694/jcls.107.

— (2024). "Knowledge Distribution in German Drama". In: *Journal of Open Humanities Data* 1.10, 1–7. doi:10.5334/johd.167.

Barron, Alexander T. J., Jenny Huang, Rebecca L. Spang, and Simon DeDeo (2018). "Individuals, institutions, and innovation in the debates of the French Revolution". In: *PNAS* 18.115, 4607–4612. https://doi.org/10.1073/pnas.171772911.

Chang, Kent K. and Simon DeDeo (2020). "Individuals, institutions, and innovation in the debates of the French Revolution". In: *Journal of Cultural Analytics* 2.5, 4607–4612. https://doi.org/10.22148/001c.17585..

Craig, Hugh and Arthur F. Kinney (2009). *Shakespeare, Computers and the Mystery of Authorship*. New York: Cambridge University Press.

Dubourg, Edgar, Andrej Mogoutov, and Nicolas Baumard (2023). "Is Cinema Becoming Less and Less Innovative With Time? Using neural network text embedding model to measure cultural innovation". In: *Proceedings of the Computational Humanities Research Conference 2023 Paris, France, December 6-8, 2023*. Ed. by Artjoms Šeļa, Fotis Jannidis, and Iza Romanowska. CEUR-WS. https://ceur-ws.org/Vol-3558/paper7806.pdf.

Fischer, Frank, Ingo Börner, Mathias Göbel, Angelika Hechtl, Christopher Kittel, Carsten Milling, and Peer Trilcke (2019). "Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama". In: *Proceedings of DH2019: "Complexities"*. Utrecht University. https://doi.org/10.5281/zenodo.4284002.

Hope, Jonathan and Michael Witmore (2014). "Quantification and the language of later Shakespeare". In: *Actes des congrès de la Société française Shakespeare* 31, 123–149. https://doi.org/10.4000/shakespeare.2830..

Krautter, Benjamin (2023). "Kopräsenz-, Koreferenz- und Wissens-Netzwerke. Kantenkriterien in dramatischen Figurennetzwerken am Beispiel von Kleists Die Familie Schroffenstein (1803)". In: *Journal of Literary Theory* 2.17, 261–289. 10.1515/jlt-2023-2012.

Luan, Yingyue and Yeun Joon Kim (2022). "An integrative model of new product evaluation: A systematic investigation of perceived novelty and product evaluation in the movie industry". In: *PloS One* 3.17. 10.1371/journal.pone.0265193.

Melanie, Andresen and Nils Reiter, eds. (2024). *Computational Drama Analysis*. Berlin: De Gruter.

Moretti, Franco (2011). "Network Theory, Plot Analysis". In: *Stanford Literary Lab Pamphlets* 2. https://litlab.stanford.edu/LiteraryLabPamphlet2.pdf.

Pfister, Manfred (1988). *The Theory and Analysis of Drama*. Trans. by John Halliday. Cambridge: Cambridge University Press.

Piper, Andrew, Hao Xu, and Eric D. Kolaczyk (2023). "Modeling Narrative Revelation". In: *Proceedings of the Computational Humanities Research Conference 2023 Paris, France,*

December 6-8, 2023. Ed. by Artjoms Šeļa, Fotis Jannidis, and Iza Romanowska. CEUR-WS. https://ceur-ws.org/Vol-3558/paper6166.pdf.

Reimers, Nils and Iryna Gurevych (Nov. 2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT- Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*. Ed. by Sebastian Padó and Ruihong Huang. Hong Kong, China: Association for Computational Linguistics. https://aclanthology.org/D19-1410.pdf.

Schöch, Christoph (2017). "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama". In: *Digital Humanities Quaterly* 2.11, 4607–4612. https://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html.

Šeļa, Artjoms, Fotis Jannidis, and Iza Romanowska, eds. (2023). *Proceedings of the Computational Humanities Research Conference 2023 Paris, France, December 6-8, 2023*. CEUR-WS.

Šeļa, Artjoms, Ben Nagy, Joanna Byszuk, Laura Hernández-Lorenzo, Botond Szemes, and Maciej Eder (2024). "From Stage to Page: Stilistic Variation in Fictional Speech". In: *Computational Drama Analysis*. Ed. by Andresen Melanie and Nils Reiter. Berlin: De Gruter.

Sims, Matthew and David Bamman (Nov. 2020). "Measuring Information Propagation in Literary Social Networks". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, 642–652. 10.18653/v1/2020.emnlp-main.47. https://aclanthology.org/2020.emnlp-main.0.

South, Tobin, Bridget Smart, Matthew Roughan, and Lewis Mitchell (2022). "Information flow estimation: A study of news on Twitter". In: *Online Social Networks and Media* 31, 100231. 10.1016/j.osnem.2022.100231.

Szemes, Botond and Bence Vida (2024). "Tragic and Comical Networks- Clustering Dramatic Genres According to Structural Properties". In: *Computational Drama Analysis*. Ed. by Andresen Melanie and Nils Reiter. Berlin: De Gruter.

Trilcke, Peer (2013). "Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft". In: Ajouri, Philip, Katja Mellmann, and Christoph Rauen. *Empirie in der Literaturwissenschaft*. Leiden, The Netherlands: Brill | mentis, 201–247. 10.30965/9783957439710_012. https://brill.com/view/book/edcoll/9783957439710/B9783957439710_s012.xml.

Trilcke, Peer, Frank Fischer, and Dario Kampkaspar (2015). "Digital Network Analysis of Dramatic Texts". In: *Digital Humanities 2015: Global Digital Humanities. Book of Abstracts*. Ed. by Anne Baillot, Toma Tasovac, Walter Scholger, and Georg Vogeler. University of Western Sydney.