

# TNA ACTIVITY REPORT

## New Metrics for Computational Drama Analysis

Author: Botond Szemes

Current position: Research Fellow

Affiliation: UN-REN Research Centre for the Humanities, Institute for Literary Studies

Host institution: Potsdam University

Mentor(s): Peer Trilcke

Period of stay: 01.04.2024 – 30.04.2024.

## Introduction

The aim of the Fellowship was to create a method that compares the utterances of characters in dramatic texts according to whether a character tends to talk innovatively compared to others, or vice versa: to what extent she or he tends to repeat others. My paper presenting a first version of the method was accepted with proposals for modifications for the *3rd Annual Conference of Computational Literary Studies* (<https://jcls.io/site/ccls2024/>); during the Fellowship I was able to make these modifications in consultation with the DraCor team at the University of Potsdam (<https://dracor.org>). As a result, I was able to make both methodological and conceptual improvements to the study.

Another important part of the Fellowship was the development of the Hungarian sub-corpus of the DraCor database (HunDraCor: <https://dracor.org/hun>). Previously, 41 dramas were available here, but the underlying corpus, which is being developed at Eötvös Loránd University (ELTE-DH), has since grown significantly. During the fellowship period, I have been working on the alignment of these two corpora, the cleaning of the TEI XML files, and the creation of a workflow that can be used automatically when changes are made to one of the corpora (ELTE-DH or DraCor).

## Methodology

For calculating the novelty/innovativeness of character's speeches, first we have to assign to each character's sentence a timestamp, the character's name, and an embedding score, which describes the position of the sentence in the "semantic space" of the drama according to a language model.

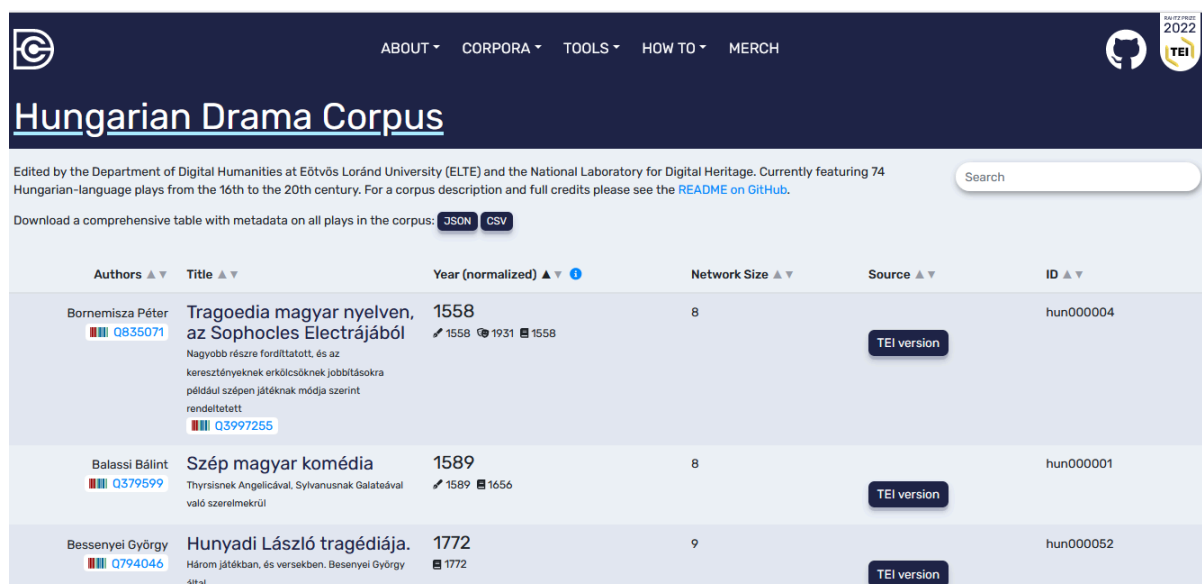


During the Fellowship I developed a method for choosing the best model for the task and to evaluate its performance. A pairwise analysis is then performed: the (cosine) similarity of a given sentence of character A is compared to all the preceding sentences of character B. From these, the most similar, i.e. semantically closest, is selected; then the maximum cosine similarity scores for all sentences of character A are averaged. In this way, we can get how semantically similar a character's sentences are to another character's preceding sentences, or the extent to which he or she talks about something new. We had to consider using other metrics (such as the average of all similarities) instead of maximum cosine similarity, but after testing, the original approach seemed to be the best choice. Finally, the results should be weighted/normalized in different ways: first of all, in terms of the time of utterance (because the later a utterance is made, the more likely it is to be similar to earlier ones) and the position of the characters in the information network (because what matters is not only how similar character A's sentences are to character B's previous sentences, but also how character B relates to other ones.)

For the development of the HunDraCor, an automatic transformer was needed to be created to adopt the files of the ELTE-DH corpus to the DraCor specification. Carsten Milling and Frank Fischer were of great help in this task. You can read more about the process on the corpus GitHub: <https://github.com/dracor-org/hundracor>. The cleanup of the TEI XML files primarily involved reconciling the ID values assigned to the characters, as in several cases there were errors in the original files.

## Results

As a result of the work with the DraCor team, the HunDraCor collection now contains 74 dramas instead of the 41 previously available, and the files are in much better condition. This was an important step forward for future quantitative analysis of Hungarian dramas, as their study will now lead to more reliable results.



Authors	Title	Year (normalized)	Network Size	Source	ID
Bornemisza Péter Q835071	Tragoedia magyar nyelven, az Sophocles Electrájából Nagyobb része fordított, és az keresztényeknek erkölcsöknek jobbitásokra például szépen játéknak módja szerint rendeltetett Q3997255	1558 1558 1931 1558	8	TEI version	hun000004
Balassi Bálint Q379599	Szép magyar komédia Thyrsisnek Angelicával, Sylvanusnak Galatédával való szerelmekről	1589 1589 1656	8	TEI version	hun000001
Bessenyei György Q794046	Hunyadi László tragédiája. Három játékban, és versekben. Bessenyei György által.	1772 1772	9	TEI version	hun000052

Figure 1. Screenshot of the front page of the HunDraCor collection

In addition to methodological improvements and a procedure for evaluating language models, important conceptual considerations helped to further develop the study on the innovation of characters. These include, following Manfred Pfister's description, the distinction between the internal information network of a drama and the external communication of which the audience is also a part. It is important to clarify that in the paper we are concerned with the latter: we are looking at the relation of an utterance to the whole of the previously established discourse (which overall perspective is equivalent to the audience's knowledge), rather than the specific transfer of information between characters (that builds up the internal communication system). We have found that previous research has tended to focus on the internal system, so our method could be a useful addition to the field of quantitative drama analysis.

Another significant outcome of the fellowship period was to place the network visualizations that can be generated by the method in relation to existing network models: what these new arrangements show, what differences and similarities can be seen with the previous ones.

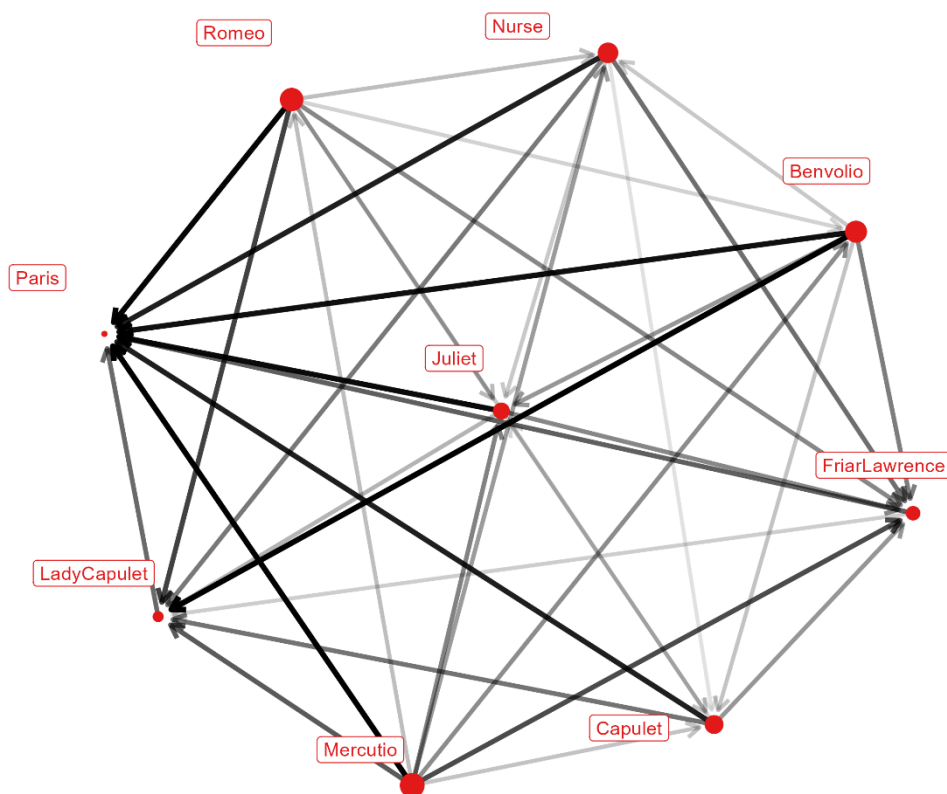


Fig. 2. The network of Romeo and Juliet based on the innovativeness of characters. The arrows indicate which character is more likely to repeat the other, their thickness is determined by the degree of similarity/repetition, and the size of the nodes as an innovation score indicates how often a character is considered innovative in pairwise comparisons.

## Future work

In the future I plan to finalize the paper and prepare it for publication in the Journal for Computational Literary Studies. The paper focuses on Shakespeare's works as a case study, but it seems worthwhile to apply the method to other corpora – possibly to Hungarian dramas. Regarding HunDraCor, with the workflow we have developed during the period, I plan to increase the collection, initially up to 100 dramas.