Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2018”

Moscow, May 30—June 2, 2018

BUILDING A CORPUS FOR THE QUANTITATIVE RESEARCH OF RUSSIAN DRAMA: COMPOSITION, STRUCTURE, CASE STUDIES

Skorinkin D. (dskorinkin@hse.ru),
Fischer F. (frafis@gmail.com),
Palchikov G. (rebel368@gmail.com)
National Research University Higher School of Economics

In this paper we introduce RusDraCor—an open corpus of Russian drama for digital literary & linguistic research. The corpus (rus.dracor.org) contains plays from the middle of XVIII to the first third of XX century provided with structural (plus some semantic) markup and metadata. Texts are encoded in the XML-based standard TEI, widely used in building corpora for the humanities. We describe the contents and annotation layers of our corpus, provide some details on its development and enrichment, and finally describe three research cases. Each case demonstrates the use of RusDraCor to answer specific questions about composition, structural features and historical evolution of Russian drama.

Keywords: corpora, TEI, XML, markup, drama, Russian drama, digital humanities, digital literary studies, stylometry, network analysis

РАЗРАБОТКА КОРПУСА ДЛЯ АНАЛИЗА РУССКИХ ДРАМАТИЧЕСКИХ ТЕКСТОВ: СОСТАВ, СТРУКТУРА, ИССЛЕДОВАТЕЛЬСКИЕ СЦЕНАРИИ

1. Introduction

The development of richly-encoded dramatic corpora for digital literary research has been on the rise recently. Some examples include “Shakespeare His Contemporaries” (510 English plays from the Shakespeare era, [Mueller 2014]) “Théâtre Classique” (1080 French dramas from the XVII and XVIII century, collected and encoded by Paul Fièvre at http://www.theatre-classique.fr/), the DLINA corpus (466 German-language plays from 1730 up to 1930, [Fischer et al. 2016a]) , the Dramawebben (66 Swedish plays at http://dramawebben.se). These corpora are encoded in TEI and have all proved their usefulness for the digital literary studies [Glorieux 2016], [Xanthos et al. 2016], [Fischer et al. 2016a,b]. Adding a Russian-language collection to the family of drama corpora will enable similar research on Russian material and boost cross-cultural studies on the structure and evolution of dramatic texts.

The ultimate goal of the RusDraCor project is to provide a corpus of at least 500 encoded Russian plays spanning for two centuries, roughly in between 1740 and 1940 (later plays are still under copyright). Currently the corpus (rus.dracor.org) features 89 plays, provided with semantic and structural annotation described below. The earliest play in RusDraCor is Horev (Хорев) by A. P. Sumarokov (1747); the newest—Ivan Vasilievich (Иван Васильевич) by M. A. Bulgakov (1936). The main sources for growing the corpus are Wikisource (wikisource.org), the Russian Virtual Library (rvb.ru), Online library of Alexei Komarov (ilibrary.ru) and Maxim Moshkov’s library (lib.ru).

2. Annotation and metadata available at RusDraCor

RusDraCor provides both structural and semantic markup for the plays included. It also contains certain meta information about the encoded texts. The corpus is encoded in accordance with the TEI guidelines (http://www.tei-c.org/Guidelines/), a widely used 30-year old XML standard comprising around 550 elements, specifically defined for digital editions and the demands of Digital Humanities research. Source TEI/XML files of the corpus are available at https://github.com/dracor-org/rusdracor. In this section we describe the layers of the corpus annotation implemented at this moment.

2.1. Structural markup

The structural markup asumes the hierarchical representation of all subdivisions in the play (acts, scenes, enters etc.). This is done with help of div tag:

```
<div type=”act”>
<head>Действие первое</head>
<div type=”scene”>
<head>Сцена первая</head>
<div type=”enter”>
<head>Выход первый</head>
<stage>Игроки, князь Звездич, Казарин и Шприх. За столом мечут банк
        и понтируют...
Кругом стоят.</stage>
```

```
<!--..text of the first enter..-->
</div>
</div>
</div>
```

The main components of a dramatic texts are character speeches and stage directions. Our markup uses TEI tags <p> (paragraph) or <lg> + <l>(verse line group and a single verse line) for speeches and TEI tag <stage> for stage directions. Speaker is encoded with <speaker> tag; each character gets a unique identifier by which s/he is referenced in the markup throughout the entire play:

```
<stage>Городничий, попечитель богоугодных заведений, смотритель
        училищ, судья, частный пристав, лекарь, два квартальных.
        </stage>
<sp who="#Gorodnichij">
<speaker>Городничий.</speaker>
<p>Я пригласил вас, господа, с тем чтобы сообщить вам пренеприятное
    известие: к нам едет ревизор.</p>
</sp>
<sp who="#AmmosFedorovichLjapkinTjapkin">
<speaker>Аммос Федорович.</speaker>
<p>Как ревизор?</p>
</sp>

***

<sp who="#Famusov">
<speaker>Фамусов</speaker>
<lg>
<l>Сказал бы я, во-первых: не блажи,</l>
<l>Именьем, брат, не упрекай оплошно,</l>
<l>А, главное, поди-тка послужи.</l>
</lg>
</sp>
<sp who="#Chatskij">
<speaker>Чацкий</speaker>
<lg>
<l>Служить бы рад, прислуживаться тошно.</l>
</lg>
</sp>
```

Cases of multiple speech authorship ('Все', 'Оба', 'Вместе') are resolved manually if possible at all:

```
<sp who="#Zagoretskij #PervajajaKnjazhna #VtorajajaKnjazhna #TretjajaKnjazhna #ChetviortajajaKnjazhna #PjatajajaKnjazhna #ShestajajaKnjazhna">
<speaker>Все вместе</speaker>
<lg>
```

```
<l>Мсьё Репетилов! Вы! Мсье Репетилов, что вы!</l>
<l>Да как вы! Можно ль против всех!</l>
<l>Да почему вы? стыд и смех.</l>
</lg>
</sp>
```

2.2. Semantic markup

As of now, semantic markup is mostly limited to the specification of each character's gender. Gender is first assigned automatically to each character during the initial conversion (relying on typical name endings). Then it goes through manual correction. We use standard way of specifying gender via the @sex attribute of each <person> element in <listPerson> of the <teiHeader> that is recommended by the TEI Consortium (see the 'TEI Header' section of the TEI P5: Guidelines at http://www.tei-c.org):

```
<listPerson>
<person xml:id="MarijaVasilevna" sex="FEMALE">
<persName>Мария Васильевна</persName>
<persName xml:lang="en">Mariâ Vasil'evna</persName>
<persName xml:lang="de">Mariâ Vasil'evna</persName>
</person>
<person xml:id="Telegin" sex="MALE">
<persName>Телегин</persName>
<persName xml:lang="en">Telegin</persName>
<persName xml:lang="de">Telegin</persName>
</person>
<!-- … rest of the list .. -->
<listPerson>
```

We are also working on adding the 'social status' information for each character—whether s/he belongs to the nobility, or is a servant, a serf, a soldier, a merchant and so on. This could give more material for formal analyses of character systems that we implement (see the research cases below).

2.3. Metadata

The metadata is stored in the <teiHeader> element of each document. It contains information about the author; dates of origin, publication and premiere of the play (if available), character names and IDs, link to the source of the text. The following example demonstrates a part of a play metadata containing the information about the play and its author:

```
<titleStmt>
<title type="main" xml:lang="ru">Гроза</title>
<title type="main" xml:lang="en">The Storm</title>
<title type="sub" xml:lang="ru">Драма в пяти действиях</title>
<title type="sub" xml:lang="en">A Drama in Five Acts</title>
```

<**author** key="Wikidata:Q171976">Островский, Александр Николаевич
</**author**>
</**titleStmt**>

Second example shows the encoding of metadata related to dates of creation, premiere and first publication:

<**bibl** type="originalSource">
<**title**>А. Грибоедов. Горе от ума. А. Сухово-Кобылин. Пьесы. А. Островский. Пьесы. "Библиотека Всемирной литературы", М.: Художественная литература, 1974.</**title**>
<**date** type="print" when="1860">1860 год (wikipedia)</**date**>
<**date** type="premiere" when="1859">1859 год (wikipedia)</**date**>
<**date** type="written" when="1859">1859 год (wikipedia)</**date**>
</**bibl**>

3. Case study 1. Measuring the evolution of drama through stage directions

3.1. Rationale for the research

Our first case study is a research of the evolution of dramatic texts. We demonstrate the use of RusDraCor, which currently contains plays written in between the middle of XVIII and the first third of the XX centuries, for diachronic studies. Specifically, we analyze the changes in length and linguistic composition of *stage directions* (<stage> tag, see the 'Structural markup' section above). These changes, in our view, reflect the general 'epification' of drama—a process that later reaches its peak with the emergence of Brecht's 'epic theatre' theory [Брехт 1965].

"A stage direction can be detailed and evocative <…> More typically, however, is direction that lacks specific details but instead invokes a formula where the implementation of the onstage effect is left to the players or to the imagination of a reader" [Dessen 2011]. When one reads a play from the XVIII of early XIX century, s/he may not even notice stage directions at all. They are typically short and purely technical:

1) <**stage**>Те же и невольник.</**stage**> (А. П. Сумароков. Хорев. 1747)
2) <**stage**>Тамира и Клеона.</**stage**> (М. В. Ломоносов. Тамира и Селим. 1750)
3) <**stage**>Оскольд, Семира, Избрана, Возвед и воины.</**stage**> (А. П. Сумароков. Семира. 1751)
4) <**stage**>Слуги уходят.</**stage**> (А. А. Шаховской. Пустодомы. 1819)
5) <**stage**>Народ расходится.</**stage**> (А. С. Пушкин. Борис Годунов. 1826)

However, as new types of drama evolve, stage directions become more elaborate and content-rich, turning into a significant part of the dramatic narrative. Consider these examples from plays in our corpus:

6) <**stage**>Слышно, как к дому подъезжают два экипажа. Лопахин и Дуняша быстро уходят. Сцена пуста. В соседних комнатах начинается шум. Через сцену, опираясь на палочку, торопливо проходит Фирс, ездивший встречать Любовь Андреевну; он в старинной ливрее и в высокой шляпе; что-то говорит сам с собой, но нельзя разобрать ни одного слова. Шум за сценой все усиливается. Голос: «Вот пройдемте здесь...» Любовь Андреевна, Аня и Шарлотта Ивановна с собачкой на цепочке, одетые по-дорожному. Варя в пальто и платке, Гаев, Симеонов-Пищик, Лопахин, Дуняша с узлом и зонтиком, прислуга с вещами—все идут через комнату.</**stage**> (А. П. Чехов. Вишневый сад. 1905)
7) <**stage**>Прыгает в окно. Даль, видимая в окне, оказывается нарисованной на бумаге. Бумага лопнула. Арлекин полетел вверх ногами в пустоту. В бумажном разрыве видно одно светлеющее небо. Ночь истекает, копошится утро. На фоне занимающейся зари стоит, чуть колеблемая дорассветным ветром,—Смерть, в длинных белых пеленах, с матовым женственным лицом и с косой на плече. Лезвее серебрится, как опрокинутый месяц, умирающий утром. Все бросились в ужасе в разные стороны. Рыцарь споткнулся на деревянный меч. Дамы разроняли цветы по всей сцене. Маски, неподвижно прижавшиеся, как бы распятые у стен, кажутся куклами из этнографического музея. Любовницы спрятали лица в плащи любовников. Профиль голубой маски тонко вырезывается на утреннем небе. У ног ее испуганная, коленопреклоненная розовая маска прижалась к его руке губами. Как из земли выросший Пьеро медленно идет через всю сцену, простирая руки к Смерти. По мере его приближения черты Ее начинают оживать. Румянец заиграл на матовости щек. Серебряная коса теряется в стелющемся утреннем тумане. На фоне зари, в нише окна, стоит с тихой улыбкой на спокойном лице красивая девушка—Коломбина. В ту минуту, как Пьеро подходит и хочет коснуться ее руки своей рукой, между ним и Коломбиной просовывается торжествующая голова автора.**</stage>** (А. А. Блок. Балаганчик. 1906)
8) **<stage>**Грохот, взрыв, выстрел. Победоносиков распахивает дверь и бросается в квартиру. На нижней площадке фейерверочный огонь. На месте поставленного аппарата светящаяся женщина со свитком в светящихся буквах. Горит слово «Мандат». Общее остолбенение. Выскакивает Оптимистенко, на ходу подтягивает брюки, в ночных туфлях на босы ноги, вооружен. **</stage>** (В. В. Маяковский. Баня. 1929)

In a manner similar to the study of the evolution of novelistic titles by Moretti [Moretti 2009], we made an attempt to quantify and measure these changes in dramatic texts.

3.2. Analysis

To measure the evolution of stage directions in plays over time, one could use relatively simple & obvious features. Given the examples above, one obvious choice

would be to use a set of features measuring absolute and relative lengths of the stage directions. We implemented the following measures (calculated for *each play*):

1. total length of stage directions in a play (figure 1)

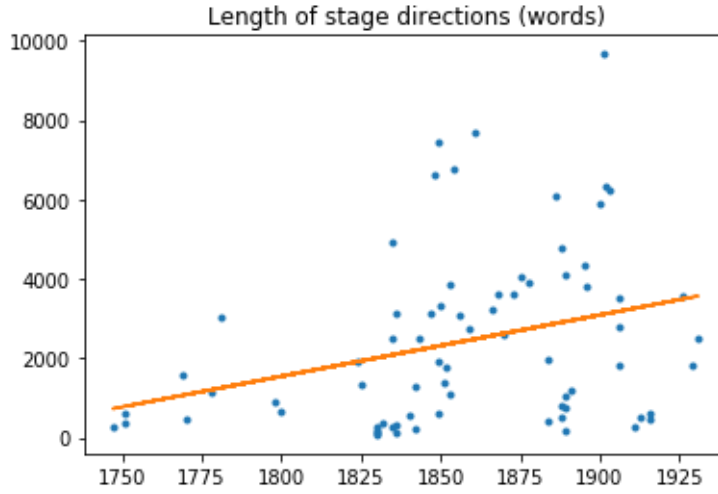

Figure 1

2. average length of a stage direction (figure 2)

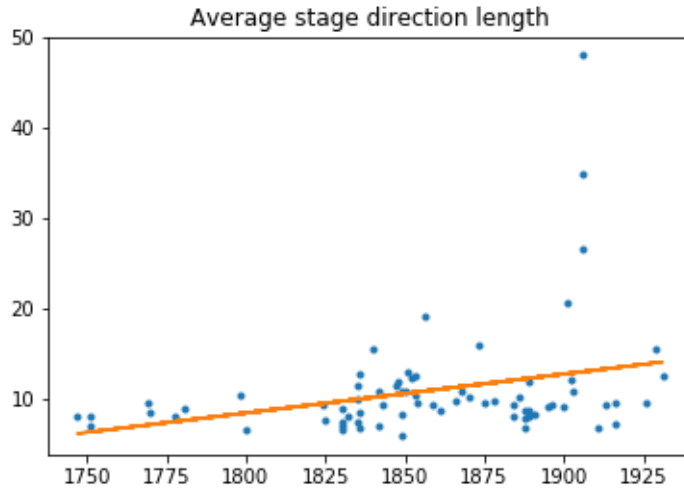

Figure 2

3. ratio of stage directions text to the direct speeches text (measured in word tokens). (figure 3)

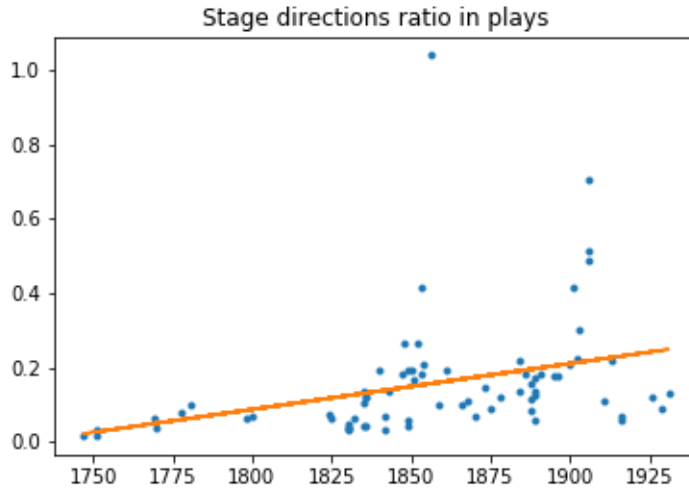

Figure 3

Another set of features verb usage in stage directions (we used MyStem (https://tech.yandex.ru/mystem/) to obtain PoS tags). As one may notice, earlier stage directions (examples 1–5) seem to contain only few verbs. These verbs usually describe the technical dynamics of the play: characters *entering* or *leaving*, also in some cases *laughing*, *crying*, *dying* and so on. In later stage directions (examples 5–8) there is a greater abundance diversity of verbs, which can be considered a marker of a *narrative* stage direction. Therefore, we also implemented the following verb-related measures:

1. the total share of verbs per all words in stage directions texts in a play (figure 4)

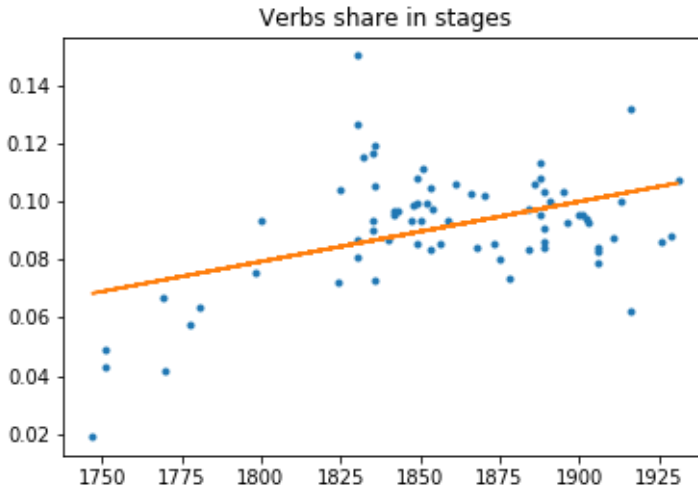

Figure 4

2. the total number of unique verbs in all stage directions (figure 5)

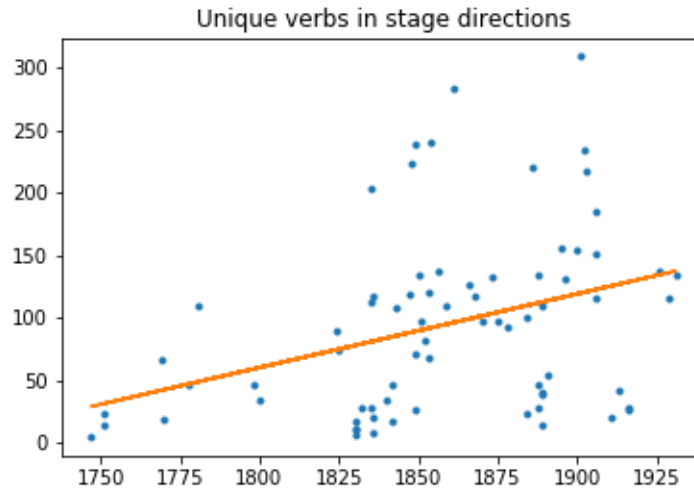

Figure 5

As one can see from figures 1–5, all measures show steady increase over time. And though the dependency is not linear in the strict sense, we can at least claim that no play in XVIII or early XIX century has long diverse stage directions of the narrative kind. In the late XIX and early XX centuries, on the other hand, we have a lot of plays with traits of 'epification' in them. Of course, this is only preliminary research, but the result could be a visible trace of a cultural evolutionary process that lead to the emergence of Brecht's epic theatre theory.

4. Case study 2. Gender specifics in character speech

4.1. Rationale for the research

In our second case study we switch to the analysis of direct speech in drama. Using the structural markup for speeches (<sp> tag and @who attribute) we can easily extract every speech instance for each character. And since we have gender information in the metadata, one obvious research goal could be to perform the statistical comparison of male and female speeches in the corpus. Similar research on movie subtitles [Schofield, Mehr 2016] produced fruitful results.

4.2. Tools and preprocessing

The earliest quantitative work on character speech is probably [Burrows 1987]—the now-famous book that laid the foundations of contemporary stylometry

(computational stylistics). In this research we also chose to use stylometric tools, namely the widely used Stylo package for R [Eder et al. 2016]. Stylo has a number of built-in stylometric functions for statistical exploratory analysis of differences in text/speech styles. To eliminate the effect of morphology we performed analysis on lemmatized text, using MyStem (https://tech.yandex.ru/mystem/) for lemmatization.

4.3. Analysis

To perform contrastive analysis of male and female speech in our corpus we used the oppose () function of Stylo package. This function performs a contrastive analysis between two given sets of texts, using Burrows's Zeta [Burrows 2007] and its extensions by [Craig & Kinney, 2009]. The function takes two sets of texts as input and outputs words significantly preferred and avoided by texts in one set (as compared to the other). Figure 6 shows such words for female speech (most preferred words are top left, most avoided words, as compared to male speech,—bottom right).

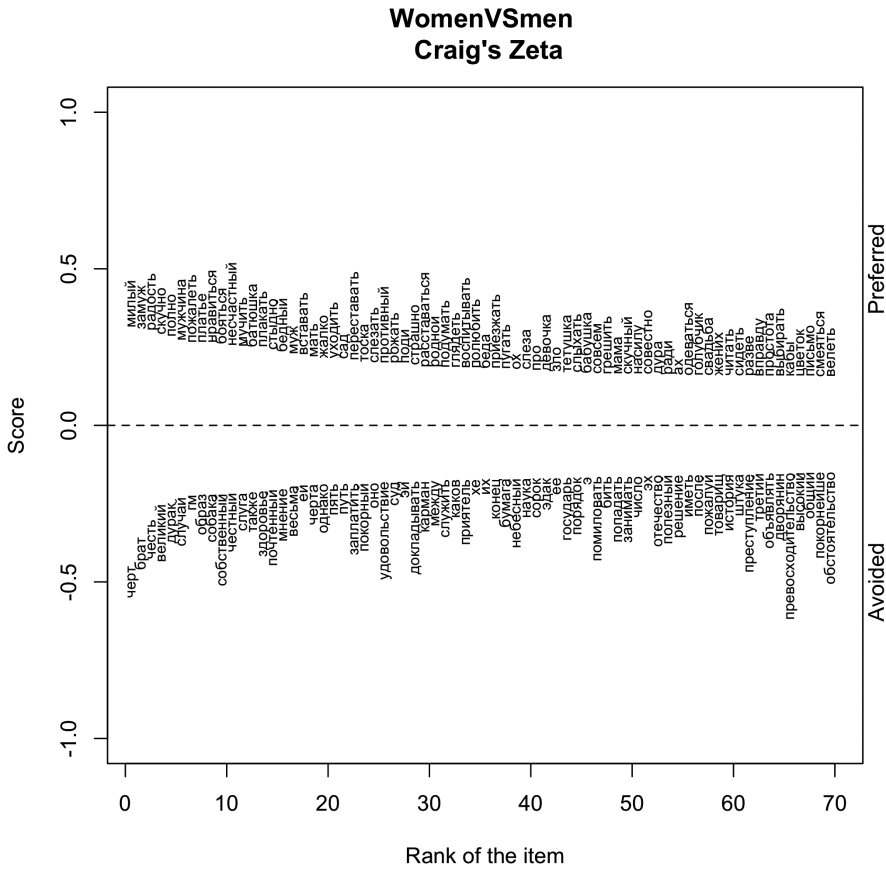

Figure 6. Results of the oppose() function applied to male and female character speech in RusDraCor

As one can see in figure 6, statistical analysis demonstrates results that can be called (awfully) stereotypical. Women tend to talk about marriages, matrimonial activity and procreation ('мужчина' = man, 'муж'=husband, жених = 'groom', 'замуж'= (to get) married, 'рожать'=to give birth, 'дитя'=child, 'воспитывать'= to bring up (a child)), feelings & emotions ('радость' = joy, 'весело'=cheerfully, 'стыдно'=ashamed, 'жаль'= it's a pity, 'счастие' =happiness), clothes ('платье'=dress, 'туалет'=clothes, 'одеваться'= get dressed) family ('бабушка'=grandmother, 'мать'=mother, папа=daddy, 'маменька'=mummy, 'тетушка'=auntie). Men, meanwhile, use swear words and offensive language ('черт'=devil/demon, usually an expressive interjection, 'дурак'=fool), talk about honor, honesty, affairs of the state and government service ('бумага'= paper/official document, 'превосходительство'= (your) excellency, standard address to an official of a certain rank, 'государь'=emperor/polite address to a person, 'докладывать'=to make an official report to a superior, 'служить'=to serve, 'служба'=service).

5. Case study 3. Network analysis

5.1. Rationale for the research

Literary network analysis is a sub-branch of digital literary studies that applies methods of network science to the study of fiction. The rise of literary network analysis is typically associated with the works of Moretti, who provided the philological rationale for this sort of digital formalism in [Moretti 2011] using Shakespeare's *Hamlet* as a showcase. However, there is also substantial amount of earlier research dedicated to network analysis of literary work. In [Schweize & Schnegg 1998] anthropologists analyze the network of characters in *Simple stories*, a contemporary novel by Ingo Shulz describing life in the former GDR after the unification of Germany. [Alberich et al 2002] explore the vast network of Marvel comics characters, extracted automatically from a total of 12,942 comics issues. The authors apply theoretical apparatus from graph theory, namely network density, clustering coefficients, average node degree, average path length and other formal metrics of the resulting network. This study demonstrates that fictional networks are structurally similar to the social networks of the real world and can be investigated with help of standard approaches from social network analysis.

In a follow-up study of the same Marvel universe [Gleiser 2007] all characters are additionally classified into heroes and villains, which enables authors to speculate on Marvel's marketing techniques. [Gleiser 2007] demonstrate that most heroes are connected to each other within one huge connected component of the network, whereas villains do not form a unified group. This, the authors suggest, could result from Marvel's attempts to popularize new and yet unknown characters by pairing them with older well-known superheroes, such as Captain America or Superman.

Other early network-related research includes several analyses of Shakespeare's plays [Stiller et al. 2003] [Stiller & Hudson 2005], analysis of community

structures in *Les Miserables* [Newman & Girvan 2004], comparison of rural and urban networks in XIX century British novels [Elson et. al 2010]. After [Moretti 2011] a lot more research on literary network analysis came around, see [Agarwal et al. 2012], [Lee & Yeung 2012], [Agarwal et al. 2013], [Ardunay & Sporleder 2014], [Lee & Wong 2016], [Grayson et al. 2016]. Literary network studies on Russian material include [Bodrova & Bocharov 2014] and [Skorinkin 2017].

Dramatic text with its inherent structure (acts, scenes, speeches) naturally becomes an easier target for automated network extraction and analysis. Studies like [Trilcke et al. 2015], [Trilcke et al. 2016], [Fischer et al. 2017] employ network analysis to large-scale digital exploration of drama (in a way following Moretti's lead with *Hamlet*).

5.2. Extracting networks

In our research we follow older formalist/structuralist approaches in literary studies [Ярхо 1997 (1930-ies)], [Сапогов 1974], [Лотман 1998]. We formalize interactions in drama as co-appearance of two characters in one scene of a play, in which both character speak at least once. This formalisation has its drawbacks, but its huge benefit is that it allows easy conversion of a play (provided with structural markup) into a network of characters and their interactions. And the availability of a multitude of plays in our corpus opens opportunities for large-scale research on the structure and evolution of different compositional types of plays. Figure 7 (also available attached as a separate file in scalable SVG format) contains a visualization of character networks extracted from all the plays currently included in RusDraCor.

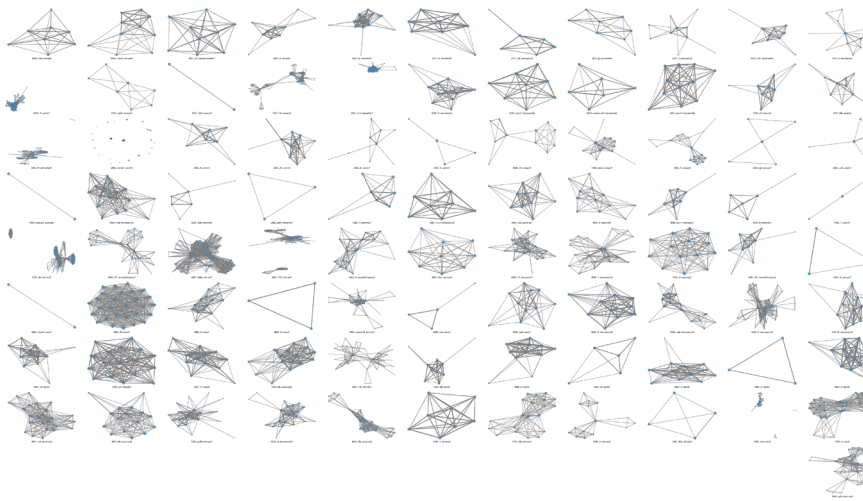

Figure 7. Visualization of character networks extracted from all plays currently included in RusDraCor (ordered chronologically)

5.3. Large-scale network analysis

Even simple visual analysis of figure 7 already tells something about certain changes in types of drama over time. For instance, one may notice from the first two rows that the networks of plays from XVIII and early XIX century all share certain traits. These traits include

1. relatively small number of nodes (characters)
2. single densely interconnected core, with few to none periphery characters (which are typically servants)

This structure apparently reflects the classicist tradition with its three unities of action, time and place. For better demonstration we provide figure 8, which features all plays in our corpus from 1747 (earliest) to 1825.

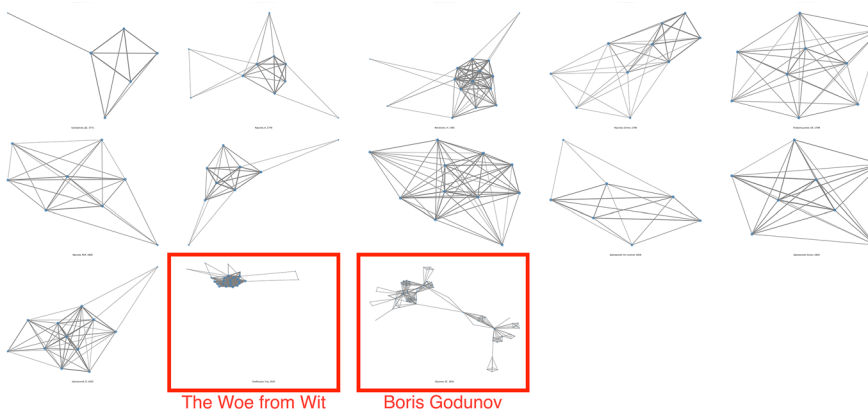

Figure 8. Visualization of character networks for all plays in our corpus from 1747 (earliest) to 1825. Ordered chronologically

The first two plays in our corpus to violate the standard structure are (marked with red) Griboedov's Woe from Wit ("Горе от ума") and especially Pushkin's Boris Godunov (Борис Годунов). Both plays are known to be a result of Shakespearean influence, and Shakespeare himself was an acknowledged breaker of the classical tradition [Dryden 1668]. Similar observations were made in [Fischer et al. 2016b] with regard to Shakespearean influence on Goethe and the structural evolution of German drama. A huge advantage of network formalization is the possibility to combine visual analysis with strict mathematical measures provided by graph theory. The visible difference of mentioned networks can be observed through networks metrics. Some of the metrics are

1. number of nodes
2. network density, which is the ratio of the number of edges in a graph to the maximum possible number of edges in that graph (i.e. if each node was connected to every other node).
3. network diameter, or the length of the longest path between one node and another in that network, measured in the number of edges

Figures 9, 10 and 11 present number of node, density and diameter measures for each play in our corpus. NB that networks with several components have no diameter measure in this implementation.

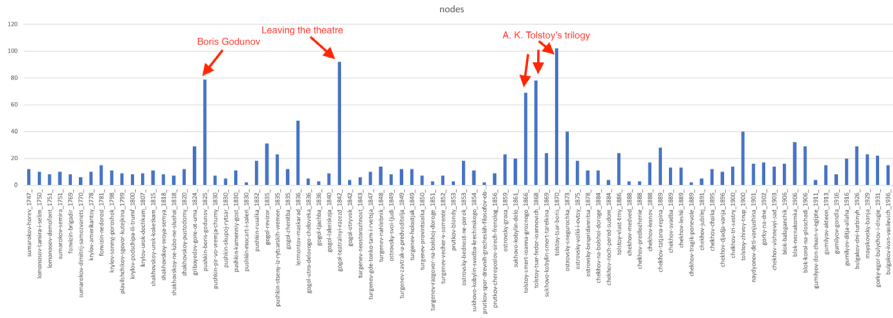

Figure 9. Number of nodes in play network

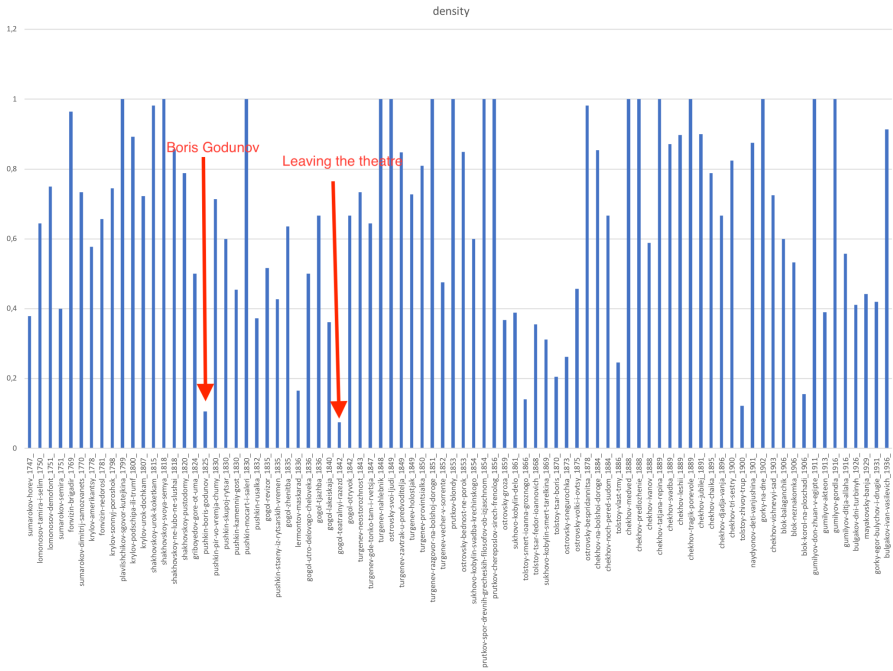

Figure 10. Network densities

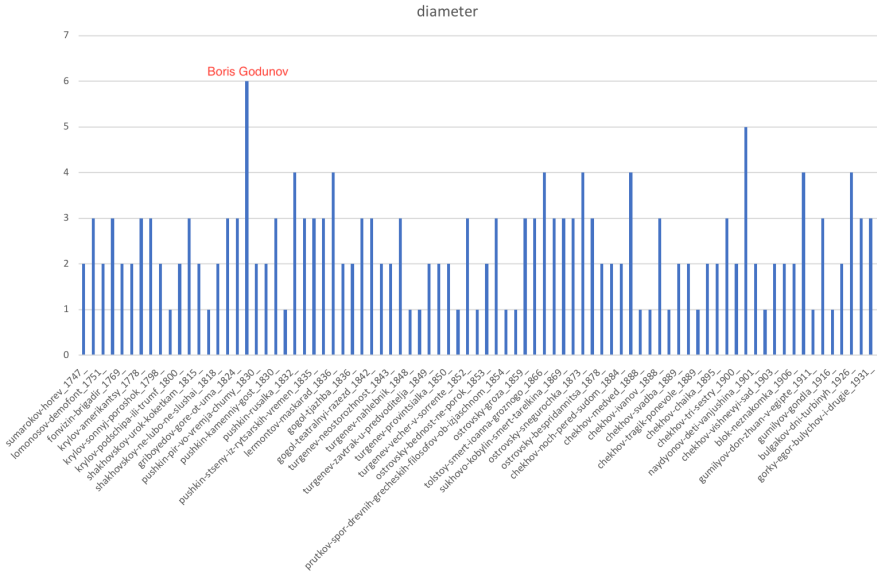

Figure 11. Network diameters

As one can see, the networks that are visually different also have extreme network measures. For instance, Boris Godunov, chronologically the first play with no single 'core' of main characters (see figure 7), is clearly an outlier in terms of the number of characters (much higher than the others), density (much lower) and diameter (much bigger). All these measures obviously reflect the specific structure of Pushkin's play, the fact that its plot takes place in different chronotopes and that the play itself was not meant to be staged (a specimen of the 'closet' plays). It is also no accident that similar measures are observed in the dramatic trilogy by A. K. Tolstoy, which also exhibits Shakespearean traits. Another example of a play with many characters and low density is Gogol's Leaving the theatre (Театральный разъезд)—a very specific meta-play of peculiar structure.

Other network measures can be used to track the evolution of plays in general. For instance, on figure 9 one can see that the average degree of a node in play (that is the number of connections each character has according to the chosen formalisation) gradually increases over time. This could also signify important changes in the types of drama produced in different time periods.

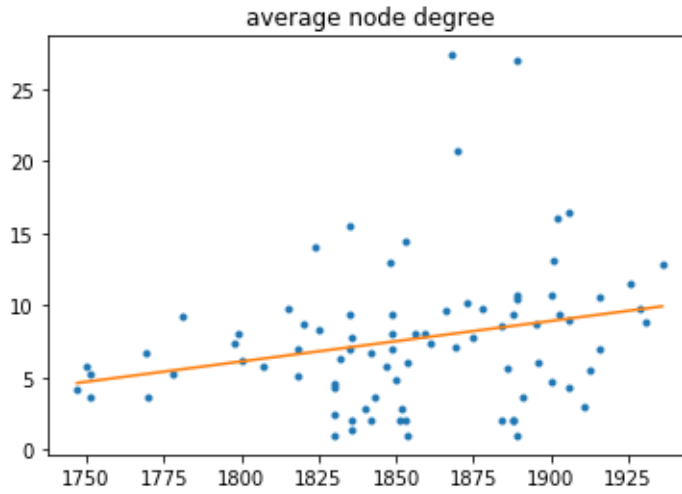

Figure 12. Average node degree in plays

5.4. Zooming in on Boris Godunov

All the research described above represents the Distant Reading approach [Moretti 2013] to literary studies, where distance (i.e. large-scale quantitative analysis) "is a condition of knowledge" [Moretti 2013: 129]. However, some people advocate a less radical, blended approach called Scalable Reading [Weitin 2017], where after large-scale analysis researcher might zoom in onto interesting samples. Here we take this step with Boris Godunov.

As we demonstrated above, Pushkin's play is one of the most expressive outliers in our corpus. The structure of the network, with several clearly distinguishable clusters, makes it an interesting target for more detailed network analysis.

Figure 13 shows the network for Boris Godunov visualized with Gephi [Bastian et al. 2009]. Colors represent automatic modularity clustering, which obviously captures Polish cluster with the False Dimitry in the middle, Moscow cluster with tsar Boris and his son Feodor, and the People cluster ('People/Народ' is an important 'group' character in the play). Node sizes are proportional to weighted degree, and the most central nodes generally correspond to the main characters (False Dimitry, Feodor, Boris etc.).

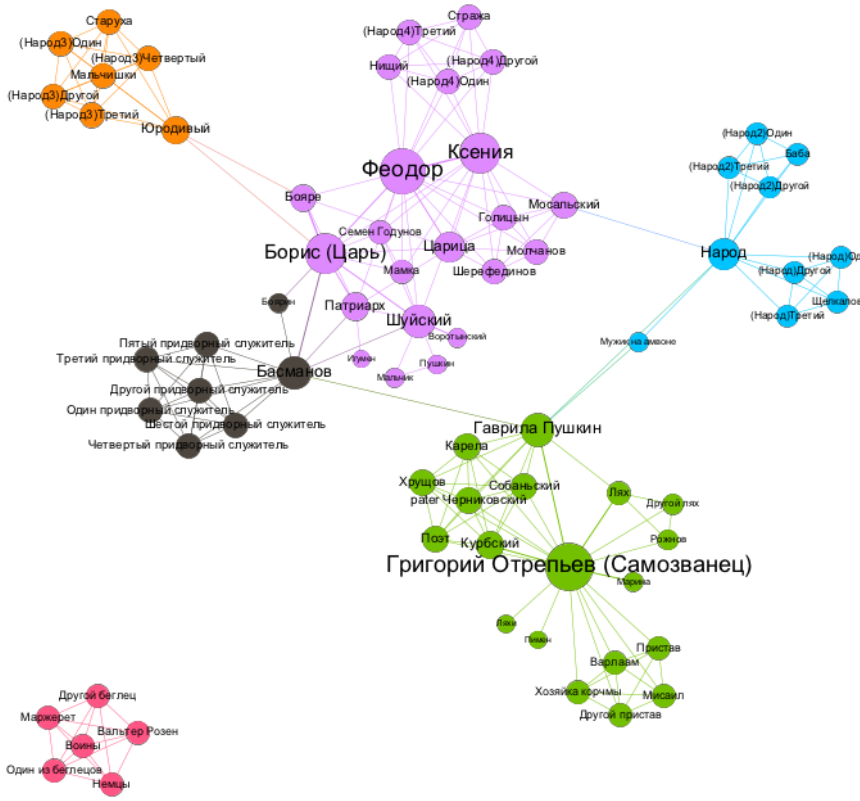

Figure 13. Boris Godunov network, nodes proportional to weighted degree

However, if we change the preferred centrality measure to betweenness centrality (нагрузка узла), i.e. the number of shortest paths going through this node, the picture changes significantly. The alternative visualization can be seen on figure 14.

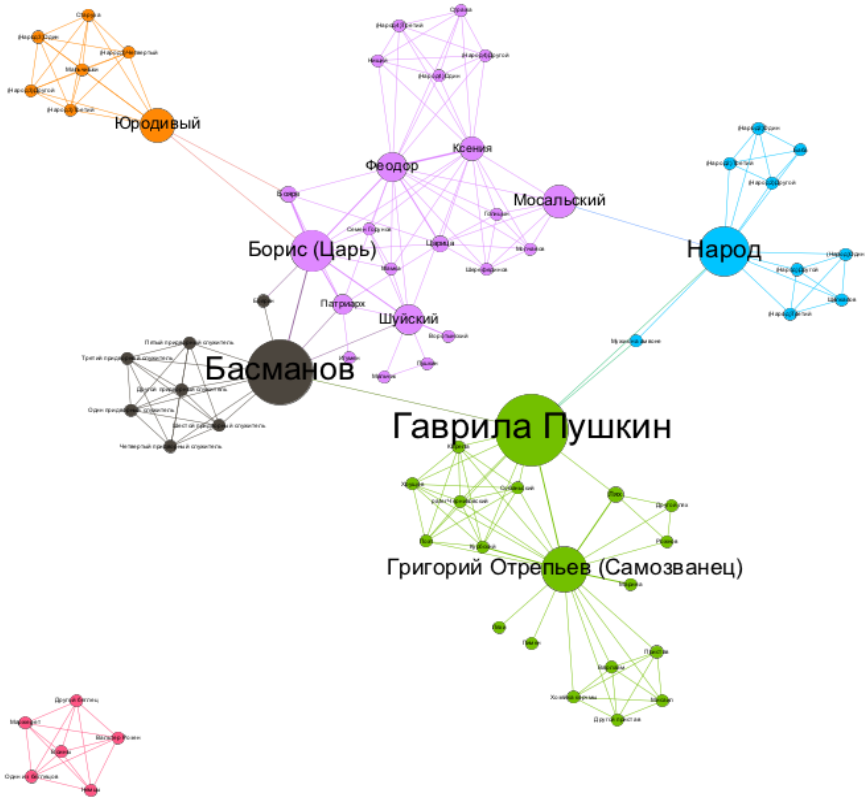

Figure 14. Boris Godunov network, nodes proportional to betweenness centrality

As one may notice, the most central character now is Gavrila Pushkin—clearly not one of the main heroes of the play. However, his high betweenness centrality is fairly obvious to those familiar with the plot. Gavrila Pushkin (one of the two Pushkin characters in the play) acts as a messenger ad mediator: he is being sent from Poland to Moscow to convey the False Dimitry's terms to Boris, and then he embarks on a mission to convince military chief Basmanov change sides—which eventually helps False Dimitry win the throne. After fulfilling this task, Gavrila Puskin as a follower of Dimitry, announces the decrees of the new tsar to the People ('Народ'), thus becoming the character that connects all cluster in the play.

This leads us to think that network metrics can sometimes reflect specific *functions* of characters in plays. And the function of Gavrila Pushkin may not be an *accidental* one, but rather deliberate: the idea that Pushkin's noble ancestors were actively *involved* in the Russian history, and especially the history of the Time of Troubles (Смутное время) can be traced throughout Pushkin's lyrics—see, for example, his famous poem 'My Pedigree' ("Мы к оной руку приложили"). All in all, such findings foreshadow new opportunities to network-oriented research of character systems in fiction.

Conclusions

In this paper we presented an open research-oriented corpus of Russian drama suitable for large-scale literary studies. Although the corpus is at the early stage of development, it can already serve as the basis for diverse research on structure and structural evolution of Russian drama, as we strove to highlight in our three cases studies above.

Later on, we hope to add more layers of annotation (e.g. named entities, classes of stage directions) and metadata (genres of the plays, social statuses of the characters etc.). This will open up new. The availability of compatible non-Russian corpora with similar markup obviously calls for cross-cultural research. RusDraCor is released under a free license, se we welcome derivation and enrichment efforts from third parties.

Acknowledgements

The article was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE) in 2017–2018 (grant № 17-05-0054) and by the Russian Academic Excellence Project “5-100”.

References

1. *Бертольт Брехт.* Теория эпического театра // Бертольт Брехт. Театр. Пьесы. Статьи. Высказывания. В пяти томах. Т. 5/2 М., Искусство, 1965
2. *Лотман Ю. М.* Структура художественного текста // Лотман Ю. М. Об искусстве. — СПб.: «Искусство — СПБ», 1998. — С. 14–285.
3. *Сапогов В. А.* Некоторые характеристики драматургического построения комедии А. Н. Островского «Лес» // А. Н. Островский и русская литература. Кострома, 1974
4. *Ярхо Б. И.* Распределение речи в пятиактной трагедии: (К вопросу о классицизме и романтизме): Подгот. текста, публ. и примеч. М. В. Акимовой; Предисл. М. И. Шапира // Philologica, 1997, т. 4, № 8/10, 201–284.
5. *Agarwal A., Kotalwar A., Rambow O.* (2013). Automatic Extraction of Social Networks from Literary Text: A Case Study on Alice in Wonderland, In Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013), Nagoya, Japan.
6. *Agarwal A., Corvalan A., Jensen J., Rambow O.* (2012). Social network analysis of Alice in Wonderland. In Proceedings of the NAACL HLT 2012 Workshop on Computational Linguistics for Literature, 88–96, Montreal, Canada.
7. *Alberich, R., Miro-Julia, J., Rossello, F.* (2002). Marvel universe looks almost like a real social network. Available at: https://arxiv.org/abs/cond-mat/0202174 (accessed December 29, 2017)
8. *Ardanuy M., Sporleder C.* (2014). Structure-based clustering of novels. In Proceedings of the EACL Workshop on Computational Linguistics for Literature, 31–39.
9. *Bastian M., Heymann S., Jacomy M.* (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.

10. *Blondel V. D., Guillaume J., Lambiotte R., Lefebvre E.* (2008). Fast unfolding of communities in large networks, In Journal of Statistical Mechanics: Theory and Experiment, 10, 1000.
11. *Burrows, J. F.* (2007). All the way through: testing for authorship in different frequency strata. Literary and Linguistic Computing, 22(1): 27–48.
12. *Burrows, J. F.* (1987) Computation into Criticism: A Study of Jane Austen's Novels. Oxford. Clarendod Press.
13. *Bodrova A., Bocharov V.* (2014). Relationship Extraction from Literary Fiction. In Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2014". Available at: http://www.dialog-21.ru/digests/dialog2014/materials/pdf/BodrovaAABocharovVV.pdf (accessed December 29, 2017)
14. *Craig, H. and Kinney, A. F.* (2009). Shakespeare, Computers, and the Mystery of Authorship. Cambridge: Cambridge University Press.
15. *Dessen A.* (2011). Stage Directions and the Theater Historian. In The Oxford Handbook of Early Modern Theatre. : Oxford University Press.
16. *Dryden, John* (1668). Jack Lynch (Ed.), ed. An Essay of Dramatick Poesie. Available at: http://andromeda.rutgers.edu/~jlynch/Texts/drampoet.html
17. *Eder, M., Rybicki, J. and Kestemont, M.* (2016). Stylometry with R: a package for computational text analysis. R Journal, 8(1): 107–121, url: https://journal.r-project.org/archive/2016/RJ-2016-007/index.html
18. *Elson, D. K., Dames, N. and McKeown, K.* (2010). Extracting Social Networks from Literary Fiction, In Proceedings of ACL 2010, Uppsala, Sweden.
19. *Fischer F., Göbel M., Kampkaspar D., Kittel C., Trilcke P.* (2017). Network Dynamics, Plot Analysis: Approaching the Progressive Structuration of Literary Texts, In Digital Humanities 2017 Book of Abstracts. Montréal: McGill University
20. *Fischer, F.; Göbel, M.; Kampkaspar, D.; Trilcke, P.:* (2016a)Theatre Plays as 'Small Worlds'? Network Data on the History and Typology of German Drama, 1730–1930. DH2016, Kraków. URL: http://dh2016.adho.org/abstracts/360.
21. *Fischer, F., Vogel, A., Göbel, M., Kampkaspar, D., Trilcke, P.* (2016b) Distant-Reading-Showcase: 200 Jahre deutscher Dramengeschichte auf einen Blick. In: Digital Humanities im deutschsprachigen Raum (DHd) 2016 Konferenzabstracts
22. *Gleiser. P. M.* (2007). How to become a superhero. In Journal of Statistical Mechanics: Theory and Experiment, 9
23. *Glorieux, F.* (2016) Dramagraphie 0.2. Online source, April 4th, 2016. URL: http://resultats.hypotheses.org/749.
24. *Grayson S., Wade K., Meaney G., Greene D.* (2016) The Sense and Sensibility of Different Sliding Windows in Constructing Co-occurrence Networks from Literature. In: Bozic B., Mendel-Gleason G., Debruyne C., O'Sullivan D. (eds) Computational History and Data-Driven Humanities. CHDDH 2016. IFIP Advances in Information and Communication Technology, vol 482. Springer, Cham. DOI: https://doi.org/10.1007/978-3-319-46224-0_7
25. *Lee J., Wong T.* (2016). Conversational Network in the Chinese Buddhist Canon. In Open Linguistics 2016, 2, 427–436. DOI 10.1515/opli-2016–0022

26. *Lee J., Yeung C. Y.* (2012). Extracting Networks of People and Places from Literary Texts. In Proceedings of 26th Pacific Asia Conference on Language, Information, and Computation (PACLIC). 209–218
27. *Moretti F.* (2013), Distant Reading. Verso, London
28. *Moretti F.* (2011). Network Theory, Plot Analysis. In Stanford Literary Lab Pamphlets, Stanford, CA. Available at: https://litlab.stanford.edu/LiteraryLabPamphlet2.pdf (accessed December 29, 2017)
29. *Moretti F.* (2009). Style, Inc. Reflections on Seven Thousand Titles (British Novels, 1740–1850). In: Critical Inquiry, Vol. 36, No. 1 (Autumn 2009), pp. 134–158
30. *Mueller M.* (2014). Shakespeare His Contemporaries: Collaborative Curation and Exploration of Early Modern Drama in a Digital Environment. Digital Humanities Quarterly. 8.3.
31. *Schofield A., Mehr L.* (2016). Gender-Distinguishing Features in Film Dialogue. In Proceedings of NAACL Workshop on Computational Linguistics for Literature 2016
32. *Schweizer T., Schnegg M.* (1998). The social structure of Simple Stories: network analysis [Die soziale Struktur der. „Simple Storys": Eine Netzwerkanalyse]. Available at: https://www.ethnologie.uni-hamburg.de/pdfs-de/michael-schnegg/simple-stories-publikation-michael-schnegg.pdf (accessed December 29, 2017)
33. *Schöch, C.; Henny, U.; Calvo Tello, J.*(2017) cligs/textbox: Spring is coming release. Data set, Zenodo, March 10th, 2017. URL: http://doi.org/10.5281/zenodo.376666.
34. *Skorinkin D.* (2017) Extracting Character Networks to Explore Literary Plot Dynamics, in: Computational Linguistics and Intellectual Technologies: papers from the Annual conference 'Dialogue '(Moscow, may 31—june 3 2017 r.). Issue. 16 (23): ed.: V. Selegey. V. 1. M.: RSUH, 2017. pp. 257–270.
35. *Stiller J., Hudson M.* (2005). Weak Links and Scene Cliques Within the Small World of Shakespeare. In Journal of Cultural and Evolutionary Psychology 3, no. 1
36. *Stiller J., Nettle D., Dunbar R.* (2003). The Small World of Shakespeare's Plays. In Human Nature, 14(4), 397–408
37. *Trilcke P., Fischer F., Göbel M., Kampkaspar D.* (2015a), Comedy vs. Tragedy: Network Values by Genre. Network Analysis of Dramatic Texts. Available at: https://dlina.github.io/Network-Values-by-Genre/ (accessed December 29, 2017)
38. *Trilcke P., Fischer F., Kampkaspar D.* (2015b). Digitale Netzwerkanalyse dramatischer Texte, In DHd2015. Von Daten zu Erkenntnissen 23. bis 27. Graz. Book of Abstracts. Austrian Centre for Digital Humanities.
39. *Trilcke P., Fischer F., Göbel M., Kampkaspar D.* (2016). Theatre Plays as 'Small Worlds'? Network Data on the History and Typology of German Drama, 1730–1930. In Digital Humanities 2016: Conference Abstracts. Jagiellonian University, Pedagogical University, Kraków, 385–387.
40. *Weitin T.* (2017), Scalable Reading. Zeitschrift für Literaturwissenschaft und Linguistik, Volume 47, Issue 1, pp 1–6
41. *Xanthos, A. et al.* (20016) Visualising the Dynamics of Character Networks. Proceedings, DH2016. Jagiellonian University & Pedagogical University, Kraków, 417–419.