National Research University Higher School of Economics

as a manuscript

Daniil Skorinkin

SEMANTIC MARKUP OF LITERARY TEXTS FOR QUANTITATIVE SCHOLARSHIP IN PHILOLOGY (ON THE BASIS OF LEO TOLSTOY'S WAR AND PEACE)

PhD Thesis Summary
for the purpose of obtaining academic degree
Doctor of Philosophy in Philology and Linguistics HSE

Academic Supervisor:
Candidate of Sciences
Bonch-Osmolovskaya A.A.

Moscow 2018

Overview

Digital literary studies are a major part of contemporary literary research. Growing availability of text in digital form and novel methods of electronic text analysis create new frontiers in the studies of literary heritage. Aside from extracting grammatical information, contemporary natural language processing tools allow for semantic as well as pragmatic analysis of texts.

Unlike many other branches of today's digital humanities, computational literary studies have a strong tradition from the non-digital era: since at least late XIX century scholars were applying quantitative methods to authorship attribution, establishing creation dates of texts and other forms of what would later be known as 'computational criticism'. The beginning of the XX century saw the emergence of Russian formalism, with its positivistic tendencies, desire to use 'scientific' methods and study literary works as formal object. In the 1950-es a wave of structuralism (together with semiotic literary criticism) came into the humanities, following the lead of structuralist tradition in linguistics, but also revisiting Russian formalism. Both periods saw some remarkable formalized, sometimes even computational researches (e.g. by B. Yarkho, Y. Lotman), despite the lack or outright absence of actual computers.

Today our capabilities for working with date increased manifold. With the development of computational tools, many research operations (esp. involving different sorts of corpora analyses and linguistic statistics) take seconds instead of months. However, the analysis of current research in digital literary studies shows that there are still considerable hindrances that severely restrict the development of this promising field.

For instance, it is still a challenge to extract clean structured data directly from the text. Various kinds of textual elements relevant for literary scholars tend to differ in terms of availability for computational analysis. It is fairly easy, for instance, to count word or n-gram frequencies in a text, and this might be enough for some applications in computational stylistics. At the other extreme, it is not yet possible to produce universal automatic extraction tools (or even consistent formal model) for the elements of the plot.

Fictional characters are positioned somewhere in between. On the one hand, a character in literature can almost always be tracked down to the very concrete sequence of words (names and name phrases, pronouns etc.). On the other hand, measuring and modeling literary characters is much harder than counting the frequencies of words. Even counting the number of occurrences of a single character in a big text might be a considerable challenge — one needs to account for different names and aliases, anaphoric mentions and so on. Things get even more complicated if one is interested in capturing actions of a character: speech acts, interactions with other characters and so on. For this reason, a lot of research that attempts computational modeling and analysis of character systems tends to focus on dramatic texts, which are a much easier target for

computational processing due to their specific structure.

One could hope for the development of natural language processing algorithms sophisticated enough to extract the necessary information. A more sustainable solution and realistic solution, however, would be to use **standardized semantic markup**. Textual markup adds a machine-readable semantic annotation layer to the text, e.g. all mentions of a character in the text, identified with a unique ID, or all instances of direct speech. This layer can be automatically converted into structured data, e.g. a table containing all instances of direct speech, each row associated with the speaker character. This allows for easy and reproducible quantitative research of character systems and character spaces.

This thesis is dedicated to the creation of a semantic markup layer for Leo Tolstoy's War and Peace. We then use the produced markup to test several methods of character space modeling and analysis. Thus, the **goal** of the research is to develop and test a markup-based method of character space analysis in a large work of prose that has a well-developed character system. To reach this goal, we had to fulfil the following **objectives**:

1. Analyze related work of both non-computational and computational literary scholarship related to modeling and formalization of character.
2. Produce markup for character mentions in Leo Tolstoy's War and peace. Connect mentions of a single character to a unique ID.
3. Produce markup for speech instances in Leo Tolstoy's War and Peace. Connect each speech instance with the speaker and the addressee(s).
4. On the basis of this markup
    a. Perform quantitative analysis of character idiolects.
    b. Perform analysis of character interactions through network analysis. Compare existing methods of network analysis to demonstrate the difference on well-known material (War and Peace).

The **scientific novelty** of the work is, firstly, in testing different methods of character space modeling on a single work (a standard practice in computational linguistics, it has not been used in this particular field of digital literary studies), secondly, in applying state-of-the-art natural language processing tools to automate a large share of the markup procedure, and thirdly, in introducing new parametric features for fictional characters which became the objects of the study.

The **theoretical significance** of the thesis consists in comparing various methods of quantitative analysis and modeling of the character space. This comparison takes place on the well-known material, and the semantic markup, which allows easy reproduction, is freely available to other researchers. The results of the comparison enable us to demonstrate for each method the particular

feature of the character system that it highlights or otherwise ignores. We show some limitations that were not previously taken into account or reported by researchers.

The **practical significance** of the thesis is, first and foremost, in the creation of a freely available semantic markup. As markup is based on an international standard (TEI/XML), it allows researchers from all over the world to reproduce the work, adjust it to its' own needs and build further research upon it. In addition to that, network data and visualization created in the course of this research proved successful as teaching material. They were used by the author of the thesis and other teachers at the Higher School of Economics lyceum (2017/2018 academic year), during the April crash-course in Digital Humanities at Helsinki University (2018), and in the lectures organized by the Higher School of Economics Centre for Digital Humanities.

Public demonstrations of the results.

The main findings of the research were presented at:

- International conference for young philology scholars in Tartu (twice, in 2015 and 2017),
- Dialogue — International conference on computational linguistics and intellectual technologies (twice, in 2015 and 2017)
- Digital Humanities 2015 — Annual Conference of the Alliance of Digital Humanities Organizations (Sydney, July 2015)
- Digital Humanities 2016 — Annual Conference of the Alliance of Digital Humanities Organizations (Krakow, July 2016),
- TEI Conference and Members' Meeting 2016 (Vienna, September 2016),
- 6th AIUCD Conference 2017 (Rome, January 2017)
- DH Russia 2017 conference (Krasnoyarsk, September 2017)
- Natural Science Methods in the Digital Humanitarian Environment conference (Perm, May 2018).

The following propositions are submitted for the defense:

1. Modern natural language processing tools are suitable for extracting meaningful information about character system and storing it in the form of semantic markup.
2. The produced markup allows the analysis of the character system using quantitative methods (frequency analysis, multivariate statistical analysis, correlation analysis, network analysis).
3. The choice of a specific method for analyzing the data obtained from the markup defines the exact properties of the character system that will be reflected in the resulting model.

This thesis consists of an introduction, main part with three chapters, conclusion section, bibliography and supplementary materials.

Summary of the main body of the thesis

The **first chapter** is dedicated to theoretical aspects of building a formal model of character and character system. This chapter includes analysis of related works and the description of markup procedure for War and Peace. In the first section of the chapter we describe approaches to the formalization of characters that have been developed in the pre-digital era. We demonstrate that the view of fictional character as model dates back to the ancient world. In the XX century a major split occurred between formal and psychological approaches to the study of fictional character. The formal approaches developed by formalism and later structuralism viewed fictional character as a set of 'text spans' — a reductionist idea, which, nevertheless, turned out to be very relevant for digital analysis. Psychological approaches, on the other hand, considered character to be a more or less accurate model of personality. One of their particular foci was the direct speech of fictional characters, which was sometimes praised as the most 'straightforward' way for a writer to describe a character. In the second half of the XX century, hybrid approaches started to emerge.

The second section of the first chapter describes modern computational approaches to the modeling of fictional characters and character systems. In this section we show how ideas of non-digital scholarship (described in the first section) re-emerge in digital environment. For instance, structuralist approaches influenced contemporary field of literary network analysis, while some ideas of the psychologically-oriented approaches affected current practices of researching character speech. However, all modern computational approaches to the study of character systems face the challenge of reliable data extraction from the text. Only the latest papers tend to use semantic XML-based markup to address the issue, but such markup is mostly produced for dramatic texts.

The third section of the chapter describes the markup we created for the text of War and Peace. The markup is consistent with TEI/XML — an international standard for encoding texts in the humanities and digital preservation sphere. TEI/XML has been used to markup two layers of semantic data. The first layer consists of character mentions, merged into coreference links through unique character identifiers. To produce this layer automatically, we used ABBYY Compreno named entity recognition and information extraction tools, and then added our own list of names to help coreference resolution. To test the quality of the resulting markup we used stander measures from natural language processing and information retrieval, such as precision, recall, and F-measure. The evaluation of the resulting markup demonstrated 78.2% overall F-measure on the

task of character extraction and identification, with precision reaching 94% and recall being around 67%. The second layer of the markup consists of direct speech annotation. The speaker was extracted partly automatically (in cases where s/he was mentioned explicitly, as in 'said Natasha'), but most speakers and all addressees of direct speech were later marked up by hand.

The **second chapter** describes the experiment in which a character space was built through quantitative analysis of direct speech. Having extracted direct speech from the markup, we then applied two different methods of quantitative analysis to demonstrate the difference between them. The first method used was Delta [Burrows, 2002][1], a baseline stylometric tool for authorship attribution and other tasks in the field of computational stylistics. The method relies on the frequency distribution of the most frequent words (lemmatized in our case) in the characters' speech. The size of the list of words was established during an experiment fundamentally similar to experiments on authorship attribution. Each character's speech was treated as a corpus of works by one author. The 'corpora' of each character were randomly divided into two comparable collections — our training and test sets. We then classified the documents in the test collection using a Delta classifier trained on the train set with different number of lemmas. The most successful classification (13 out of 14 samples identified correctly) happened when we used 130 most frequent lemmas. In further experiments we consistently used 130 most frequent lemmas to calculate Delta distances.

As stylometric tools are sensitive to the size of texts, we only use those characters who speak at least 1000 words over the course of the novel. This initially gave us a list of 16 characters. We then removed two characters speaking predominantly in French, and performed all subsequent experiments with the remaining 14 characters. These were Andrey Bolkonsky, Natasha Rostova, Nikolai Rostov, Pierre Bezukhov, Marya Bolkonskaya, Vasily Kuragin, old prince Nikolai Bolkonsky, countess Natalya Rostova (mother), count Ilya Rostov, Dolokhov, Denisov, Kutuzov, Anna Mikhaylovna Drubetskaya, Anna Pavlovna Scherer.

We performed stylometric analysis of character speech using 'stylo' package for R, the most widely used Delta implementation. To visualize Delta distances and character groupings we used multidimensional scaling, principal components analysis and hierarchical clustering.

[1] Burrows J. 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship. Literary and Linguistic Computing. 2002. v. 17. no 3. pp. 267–287.

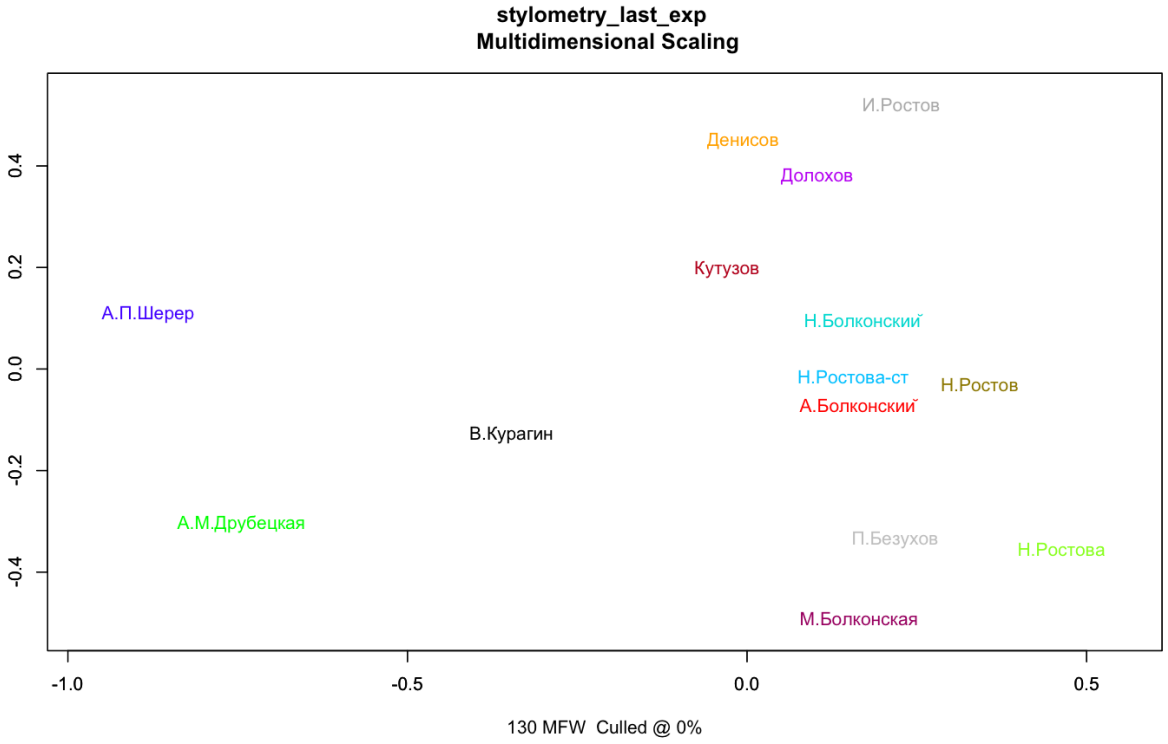

Fig. 1. 14 top speakers of war and peace in the stylometric space reduced to two dimensions with help of MDS

The most obvious division we identified is between the high-society group of Vasili Kuragin, Anna Shcherer and A. M. Drubetskaya, and the rest of the characters. This group is visible through different kinds of multivariate analysis that we applied.

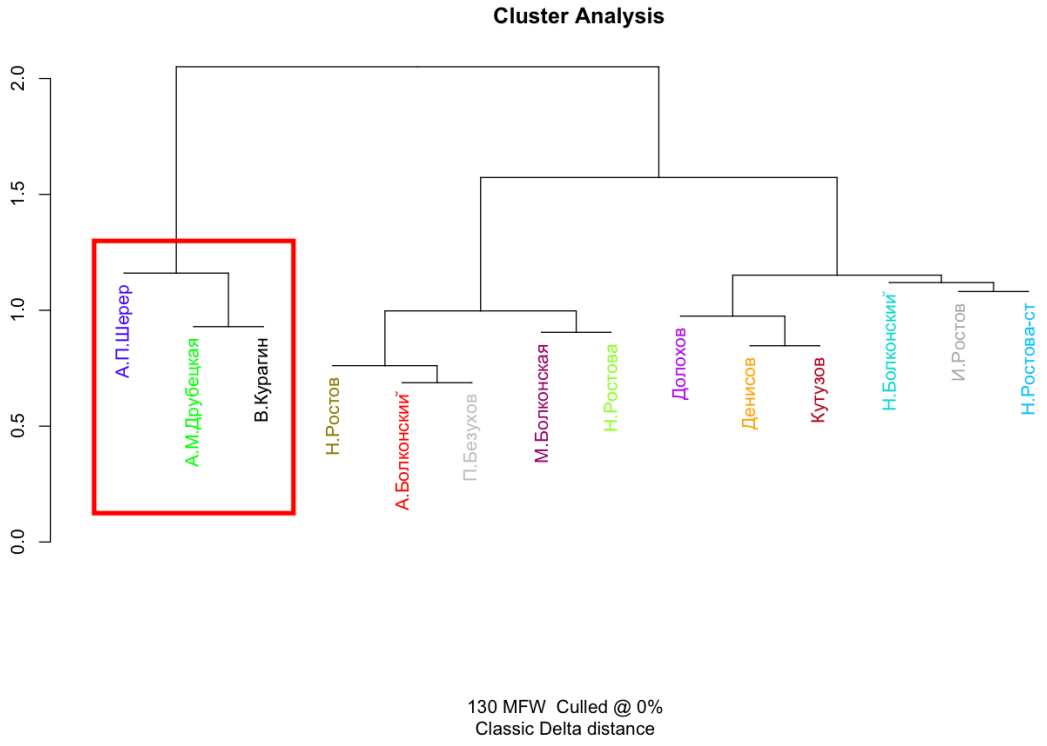

Fig. 2. Hierarchical clustering of 14 characters with Delta distance

Other characters seem to cluster into the group of main characters, and the group of non-evil secondary characters (with the possible exception of Dolokhov, whose overall image and role in the novel is quite complex).

The second method we used was our own homebrew approach specifically set to capture different features of character speech. The features here, unlike in stylometry, were mostly not connected to the lexical content of the speech. These were:

1. The share of exclamatory sentences
2. The share of question sentences
3. Punctuation marks to speech ratio
4. Discourse markers frequency (the only lexicon-related feature in this set)
5. Readability, as measured by http://ru.readability.io/

Thus, each character in the second experiment was represented as a 5-dimensional vector. We then used similar methods of multivariate statistics to visualize and compare the results.

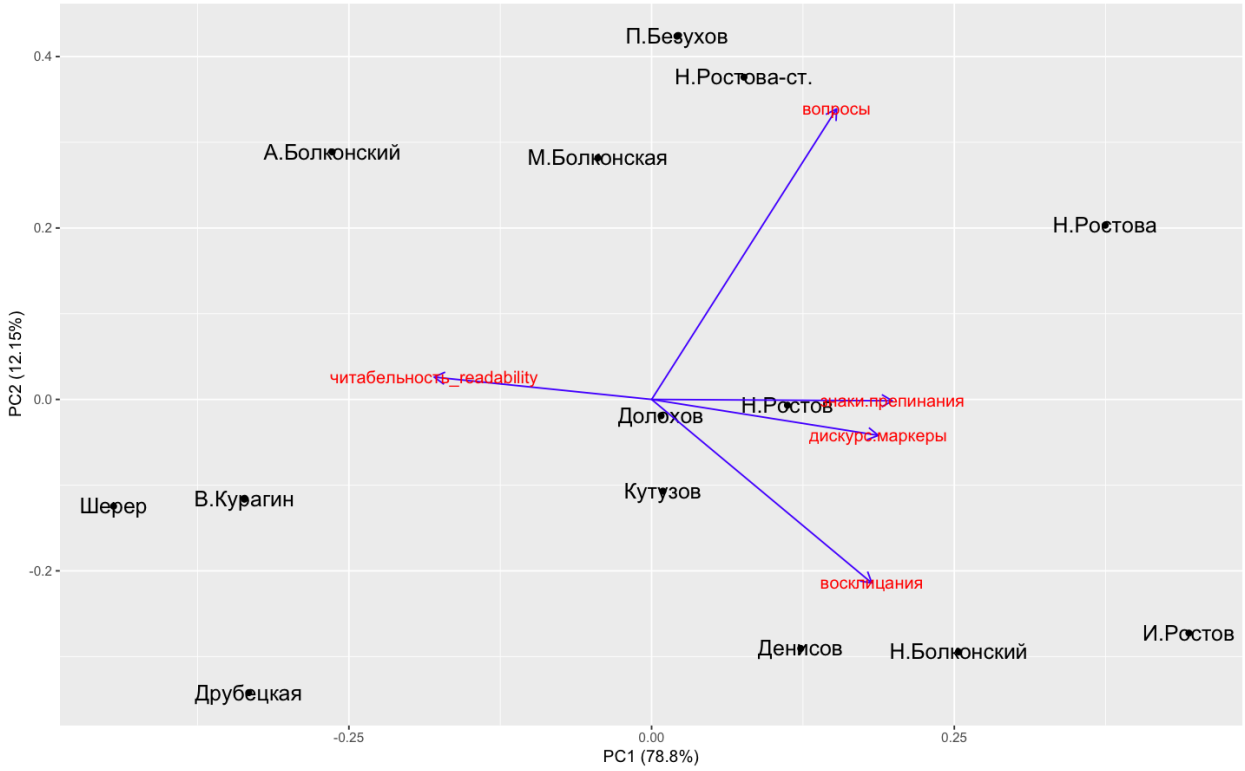

Fig 3. PCA of 14 characters with a set of alternative speech features.

Here we again observed the distinction in the speech of V. Kuragin, A. P. Scherer and A. M. Drubetskaya from the rest of the characters. However, in this case the features are interpretable. These characters tend to have speech with low readability, little share of exclamations and questions. Their complete opposite is Natasha Rostova, a character whose

speech is highly readable, abundant with discourse markers, punctuation marks, exclamations and questions.

We can also notice that the alternative method produces a rather different and more fine-grained division of characters into groups, as compared to stylometry. This is also visible in the results of hierarchical clustering (Fig. 4.)

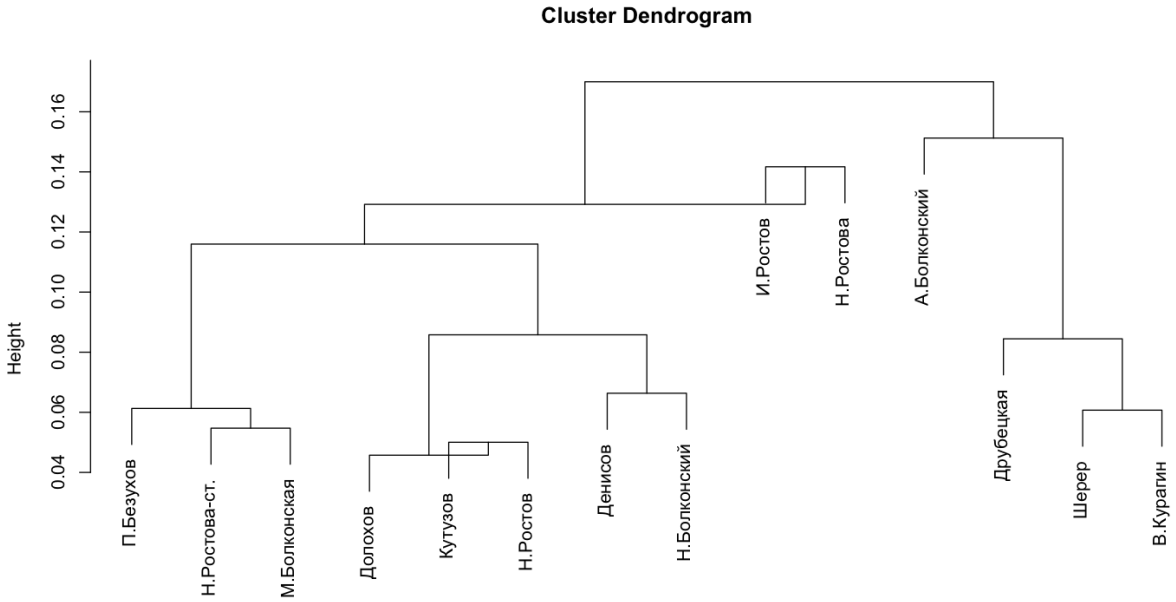

Fig 4. Hierarchical clustering of 14 characters with a set of alternative speech features.

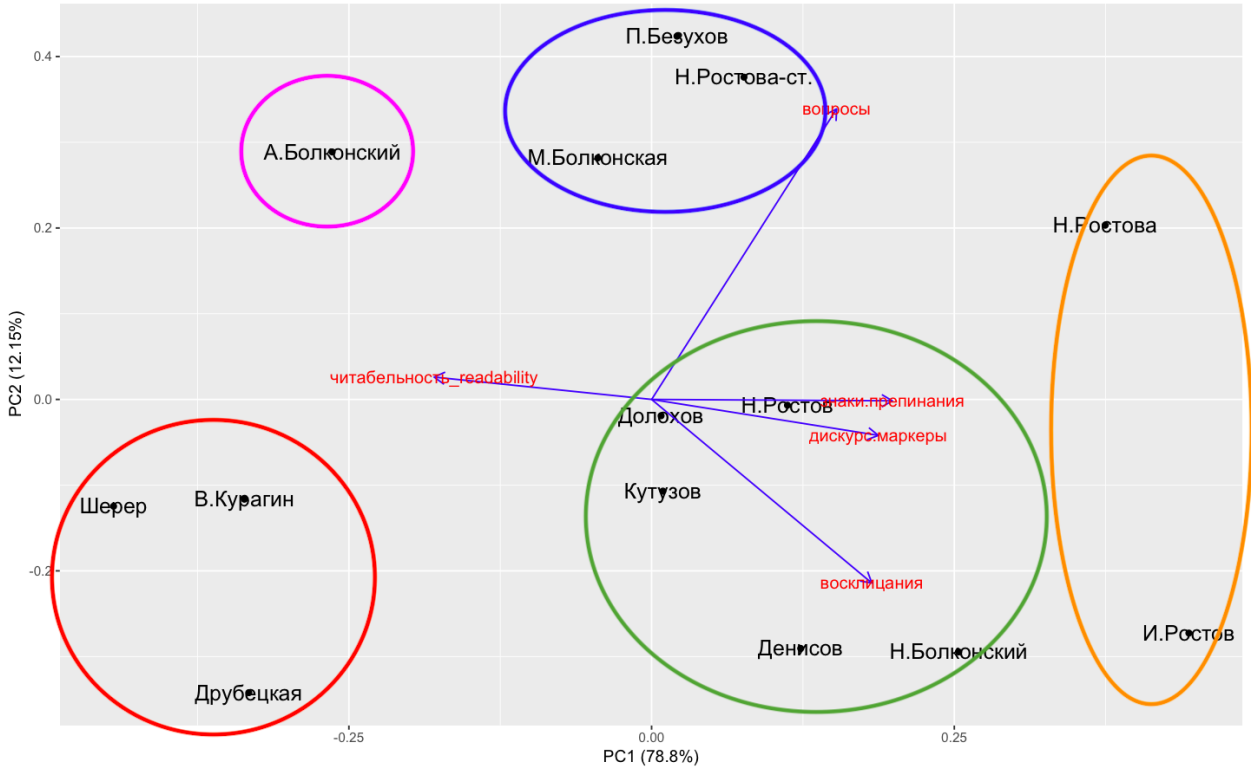

Fig 5. A combination of PCA and (higher-level) hierarchical clustering for 14 characters with a set of alternative speech features.

This kind of clustering seems to capture a different sort of similarity between characters. For instance, Ilya Rostov and Natasha Rostova represent the extreme of Rostov flamboyance and expressiveness. Denisov, towards the end of the book, becomes in many ways similar to the old prince Bolkonsky with whom he clustered — a retired general unhappy with the officials for his career misfortunes. But the most significant difference seems to be the separation of Andrey Bolkonsky from the rest of the character space. His speech is much less 'readable' (which essentially means longer and more formalized), less expressive and less abrupt (low punctuation ratio) than that of the other protagonists. This result seems very telling, as Tolstoy himself highlights several times that the young prince Bolkonsky speaks 'dryly' to people and is 'reserved'. His obvious dissimilarity with Natasha in our visualization might actually reflect the very difference between the two that caused countess Rostova fear that 'Natásha had too much of something, and that because of this she would not be happy' with Bolkonsky.

The **third chapter** describes the experiment of character space modeling through markup-based network analysis. As our research in chapter 1 showed, current approaches to literary network extraction fall into two major groups: co-occurrence-based approaches and conversational (dialogue-based) approaches. Our markup allows easy extraction of networks using both approaches, so we were able to compare them and demonstrate the difference.

At the first stage we built two networks — a co-occurrence one and a conversational one — for the entire novel. Both were too big to analyze visually, so we extracted the core of each network using network centrality measures.

Fig. 6. The top 10 characters by eigenvector centrality in conversational network (left) and co-occurrence network (right)

It turned out that regardless of centrality measure conversational network seems to downgrade military and political commanders, which constitute the ‘historical’ dimension of War and Peace, in terms of centralities. Napoleon, Kutuzov, Alexander the I tend to be much more central in the co-occurrence network. This applies to all three centralities that we used.

The same effect was also observed when we analyzed the networks dynamically. In book 10 of War and Peace, where the Borodino battle is being prepared and then takes place, the distribution of centralities is very explicitly different — see figures 7 and 8.

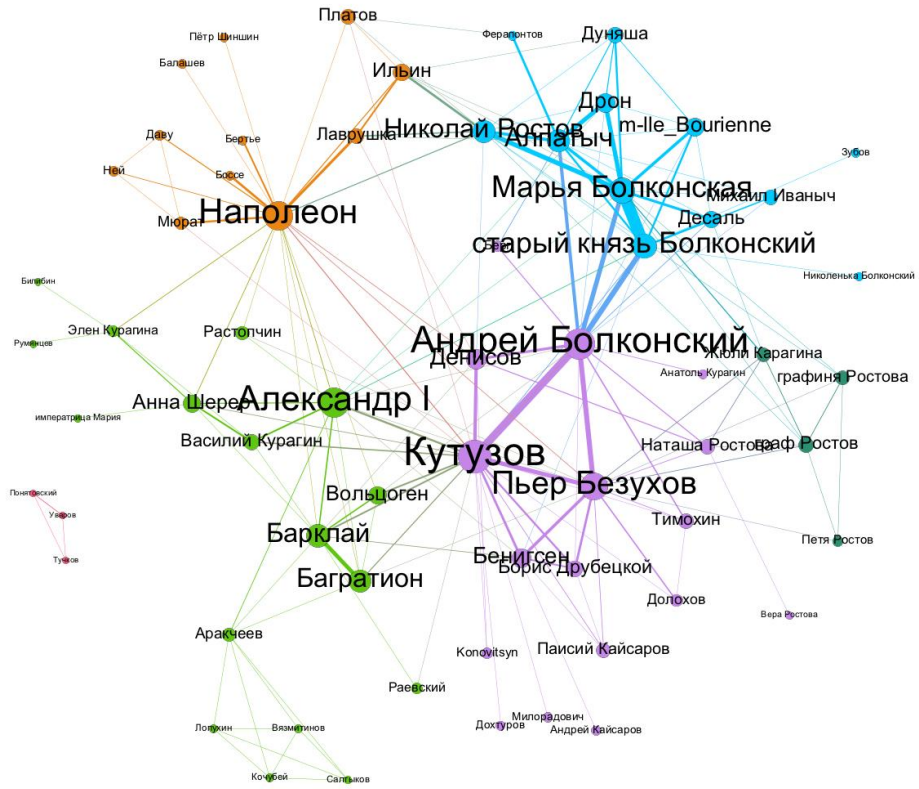

Fig 7. Co-occurrence network for book 10 of War and Peace. Nodes proportional to eigenvector centrality.

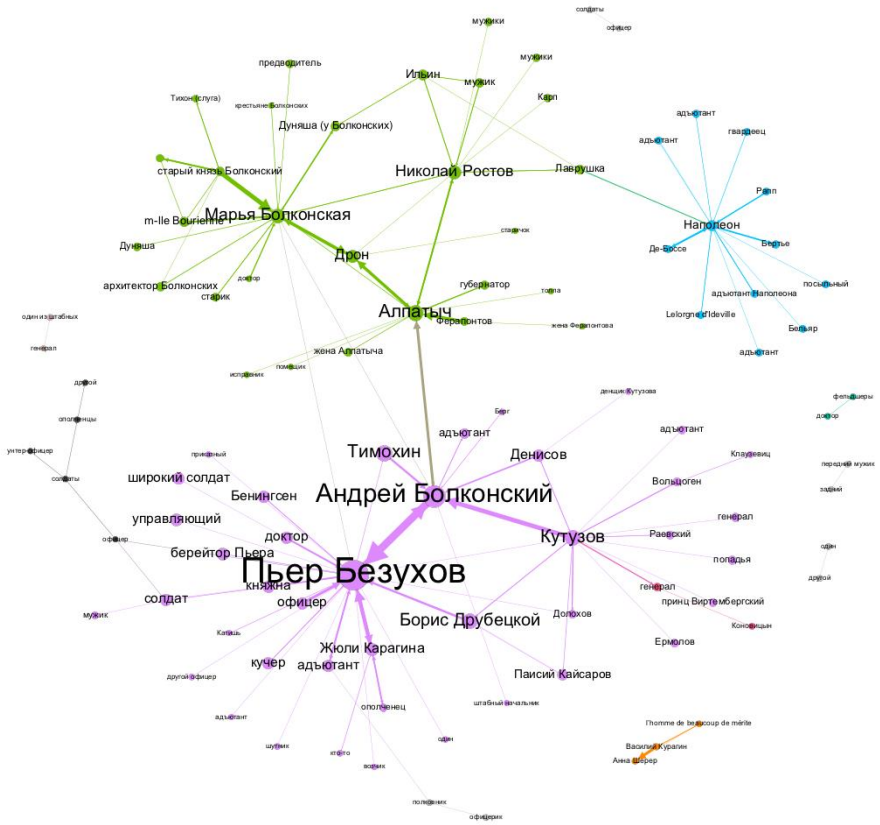

Fig 8. Conversational network for book 10 of War and Peace. Nodes proportional to eigenvector centrality.

Obviously, the co-occurrence network is better at depicting the central collision of this part of the novel between the Russian army headed by Kutuzov and the French invaders led by Napoleon. Russian emperor Alexander I is also among the central characters. On the other hand, the conversational network places the highest centrality on Pierre — probably the most important *spectator* of Borodino battle in Tolstoy's novel. From this we may speculate on how network centralities in different networks may reflect different functions of characters in a complex plot like the one of War and Peace.

We have also performed additional statistical comparative testing of the two networks using network density measure (the ratio of the number of edges in the network to the maximum possible number of edges, i.e. the measure of how densely the network is interconnected). It turns out that densities in the co-occurrence network tend to reflect the change of settings between war-related and peace-related ones. The conversational network is less sensitive to these changes, but its density peaks in the epilogue, reflecting the wrap-up of the events of the novel as two families are shown in (almost) happy co-existence.

We conclude the thesis by highlighting the **major findings and results**:

1. We have analyzed the existing body of research on computational and pre-computational modeling of characters and character spaces. The analysis of contemporary digital studies revealed a problem of data extraction from the text and a related problem of research reproducibility. This problem can be tackled with the creation of semantic markup.
2. We have created and published semantic markup for Leo Tolstoy's War and Peace, that is designed for computational modeling and analysis of characters in the novel. The markup contains 25,6 thousand identified character mentions and 6,3 thousand instances of direct speech with speaker and addressee identified.
3. The markup has been tested as the basis for character space modeling using the analysis of direct speech and network analysis.
4. The results of the analysis comply with certain findings of literary scholars who were researching War and Peace without the help of computer. In our view, this gives certain credibility to the applied method.
5. For each approach we have compared different methods and showed how they differ from each other while reflecting the same character system. These differences have not been reported previously by other researchers.

The main findings of the research were discussed in the following papers:

- Skorinkin D. Extracting Character Networks to Explore Literary Plot Dynamics // Komp'juternaja lingvistika i intellektual'nye tehnologii: po materialam ezhegodnoj mezhdunarodnoj konferencii «Dialog». Vyp. 16 (23): V 2 t. 2017. P. 257-270.
- Bonch-Osmolovskaya A., Skorinkin D. Text mining War and Peace: Automatic extraction of character traits from literary pieces// Digital Scholarship in the Humanities, 2017 Volume 32, Issue supplement 1, p. i17–i24
- Skorinkin D. A., Bonch-Osmolovskaja A. A. «Osobye primety» v rechi hudozhestvennyh personazhej: kolichestvennyj analiz dialogov v «Vojne i mire» L. N. Tolstogo // Jelektronnyj nauchno-obrazovatel'nyj zhurnal «Istorija» 2017. T. 7. # 7 (51)
- Skorinkin D. A. Jelektronnoe predstavlenie teksta s pomoshh'ju standarta razmetki TEI // Vestnik Moskovskogo universiteta. Serija 9: Filologija. 2016. # 5. S. 90-108
- Skorinkin D., Mozhaev E. TEI markup for the 90-volume edition of Leo Tolstoy's complete works, in: TEI Conference and Members' Meeting 2016 Book of Abstracts. Wien: Austrian Centre for Digital Humanities, 2016. P. 107-109
- Skorinkin D. Digital Edition of the Complete Works of Leo Tolstoy // 6th AIUCD Conference Book of Abstracts. Rome, 2017. P. 264-267

- Bonch-Osmolovskaya A., Skorinkin D., Sidorova E. Verbal Identity of a Fictional Character: a Quantitative Study with a Machine Learning Experiment // Digital Humanities 2016. Conference Abstracts, 11–16 July 2016. Kraków: Jagiellonian University, 2016. P. 747-749.