Character-distinguishing features in fictional dialogue: the case of *War and Peace*

Daniil Skorinkin
skorinkin.danil@gmail.com
National Research University
Higher School of Economics, Russia

Introduction

The study of character speech is a topic of fairly consistent interest among digital literary scholars. It is usually acknowledged that voices of characters are essentially different from narrator's own voice and should be treated separately. Some researchers have fictional dialogue removed from the texts they studied before any tools of computational investigation are applied (Hoover, 2004). Quite a lot of effort has been made recently to address the problem of identifying character speech in prose and attributing it to the correct speaker (ссылки!). One of the outcomes of such research is the possibility to study voices of different characters on relatively large scale and apply computational tools that measure their recurring stylistic parameters.

Method

The study of character speech has traditionally had strong ties to the fields of stylometry and authorship attribution, as their methods proved quite useful for studying idiolect of a fictional speaker. Suffice it to say that one of the seminal works in stylometry, Computation into criticism# by Burrows (Burrows, 1987), was focused on the study of character speech in Jane Austen's novels. The method developed by Burrows grew into what is currently known as Delta, a widely-adopted standard for authorship attribution. Delta has been consistently and successfully applied to identifying the author of an unattributed text of different languages and genres, but at the same time it saw considerable usage as a purely stylometric tool for the study of text where authorship is undisputed. Among other things, this included research into the specific idiolects of fictional characters (see, for example, Rybicki, 2006).

In our research Delta was used as one of the two possible approaches to studying character voices in Leo Tolstoy's War and peace. Much like in case of Senkewic (Rybicki, 2006), there's certain critical opinion (Eikhenbaum, 2009) that Tolstoy's characters are quite distinct from each other in their speech. Our own experience of carefully reading speech instances extracted from War and peace (for details on extraction procedure see (Skorinkin, Bonch-Osmolovskaya, 2015) supports the opinion. So it seemed natural to try and test computational methods that already showed their applicability to precisely such task. We used R package stylo by (Eder *et al,* 2013)

Testing the method on Russian material

Surprisingly enough, we were unable to find any work that applied Delta to any Russian material. Therefore we felt obliged to conduct a couple of experiments that would test its general applicability to Russian before we proceed with character speech. At the first stage we tried Delta's ability to distinguish between Tolstoy and Dostoevsky. The training set contained one of the six parts of Dostoyevsky's Crime and Punishment and three of the fifteen books of Tolstoy's War and Peace. The remaining 18 pieces of text (5 by Dostoevsky and 13 by Tolstoy) constituted the test set. The results with different settings can be seen in Table 1 and Figures 1,2:

N most frequent	Words	Character 3-grams	Character 3-grams	Character 3-grams
25	80% (4/5)	60% (3/5)	60% (3/5)	100% (5/5)
30	80% (4/5)	80% (4/5)	60% (3/5)	80% (4/5)
35	80% (4/5)	60% (3/5)	60% (3/5)	80% (4/5)
40	80% (4/5)	60% (3/5)	60% (3/5)	80% (4/5)
45	80% (4/5)	60% (3/5)	80% (4/5)	100% (5/5)
50	80% (4/5)	60% (3/5)	80% (4/5)	100% (5/5)
55	80% (4/5)	60% (3/5)	80% (4/5)	80% (4/5)
60	100% (5/5)	60% (3/5)	80% (4/5)	80% (4/5)
65	100% (5/5)	60% (3/5)	80% (4/5)	80% (4/5)
70	80% (4/5)	60% (3/5)	80% (4/5)	100% (5/5)
75	80% (4/5)	80% (4/5)	80% (4/5)	80% (4/5)
80	80% (4/5)	60% (3/5)	80% (4/5)	100% (5/5)
85	80% (4/5)	80% (4/5)	80% (4/5)	100% (5/5)
90	100% (5/5)	80% (4/5)	80% (4/5)	100% (5/5)
95	80% (4/5)	80% (4/5)	80% (4/5)	100% (5/5)
100	80% (4/5)	80% (4/5)	80% (4/5)	100% (5/5)

Table 1. Delta authorship attribution, Tolstoy vs Dostoevsky

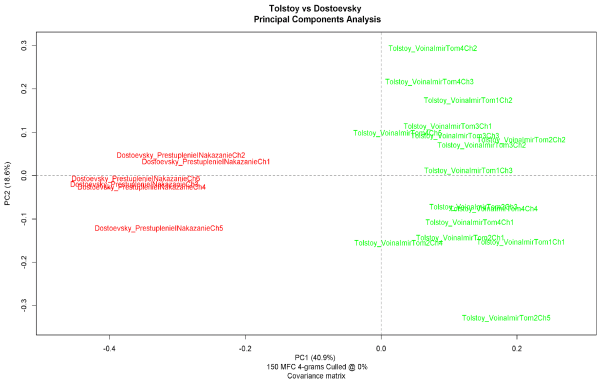


Fig. 1. Delta PCA on 150 most frequent character 4-grams, Tolstoy vs Dostoevsky

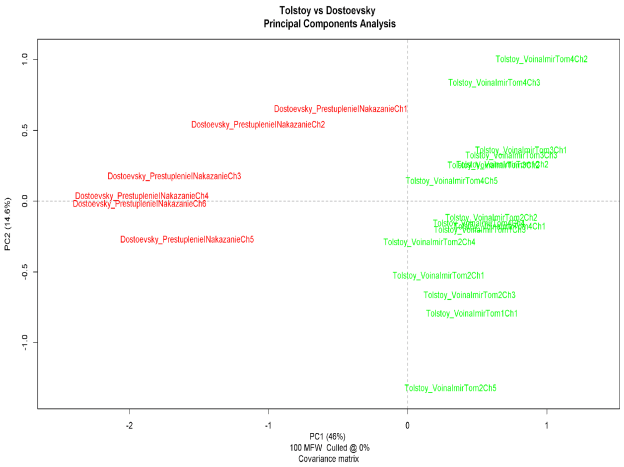


Fig. 2. Delta PCA on 100 most frequent words, Tolstoy vs Dostoevsky

The second experiment involved four Russian authors Tolstoy, Dostoevsky, Goncharov and Turgenev. All four represent (roughly) the same epoch of Russian literature and all four are recognized as masters of realistic prose. We used three novels by each author for our experiment. At the first stage two out of each three were placed in the training corpus, and Delta was supposed to attribute the remaining one. All four novels from the test corpus were attributed correctly. At the second stage we reverted the experiment and left only one novel by each author in the training set. In this case Delta consistently showed 7 out of 8 correct attributions (the only mistake being Tolstoy's Family Happiness incorrectly attributed to Dostoevsky.A possible explanation could be that Family Hap-piness is written in first person from the point of view of a young woman, something uncommon for Tolstoy; and the only Dostoevsky's work the training corpus contained was The Insulted and Humiliated , also a first-person narrative). Fig. 3 shows Delta scores for all the novels visualized with help of principal component analysis.

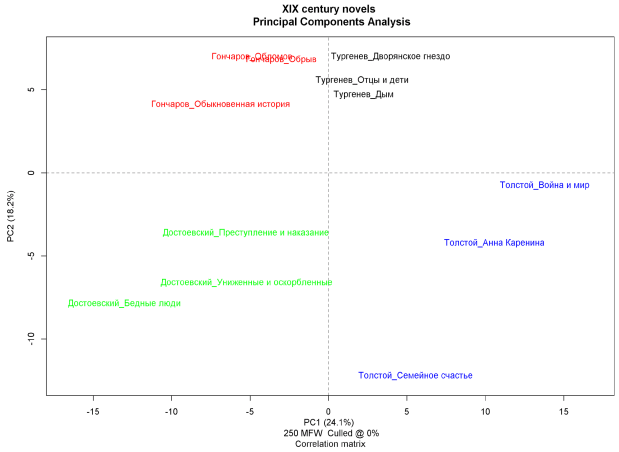


Fig. 3. Delta PCA for 12 Russian novels of 1850-1870-ies, 250 most frequent words

Applying Delta to Tolstoy's characters

Having thus shown that Delta is applicable to Russian, we proceeded with our experiment. In the first place we applied the method to top 5 characters by the total number of speech instances. We split the total sets of speeches by each character and then tried authorship attribution The results are shown in Table 2.

N most frequent	Words	Character 3-grams	Character 3-grams	Character 3-grams
25	80% (4/5)	60% (3/5)	60% (3/5)	100% (5/5)
30	80% (4/5)	80% (4/5)	60% (3/5)	80% (4/5)
35	80% (4/5)	60% (3/5)	60% (3/5)	80% (4/5)
40	80% (4/5)	60% (3/5)	60% (3/5)	80% (4/5)
45	80% (4/5)	60% (3/5)	80% (4/5)	100% (5/5)
50	80% (4/5)	60% (3/5)	80% (4/5)	100% (5/5)
55	80% (4/5)	60% (3/5)	80% (4/5)	80% (4/5)
60	100% (5/5)	60% (3/5)	80% (4/5)	80% (4/5)
65	100% (5/5)	60% (3/5)	80% (4/5)	80% (4/5)
70	80% (4/5)	60% (3/5)	80% (4/5)	100% (5/5)
75	80% (4/5)	80% (4/5)	80% (4/5)	80% (4/5)
80	80% (4/5)	60% (3/5)	80% (4/5)	100% (5/5)
85	80% (4/5)	80% (4/5)	80% (4/5)	100% (5/5)
90	100% (5/5)	80% (4/5)	80% (4/5)	100% (5/5)
95	80% (4/5)	80% (4/5)	80% (4/5)	100% (5/5)
100	80% (4/5)	80% (4/5)	80% (4/5)	100% (5/5)

Table 2.

The most common mistakes are between princess Marya Bolkonskaya and Natasha Rostova and between prince Andrew Bolkonsky and Pierre Bezukhov. Their closeness can be seen in Figure 4:

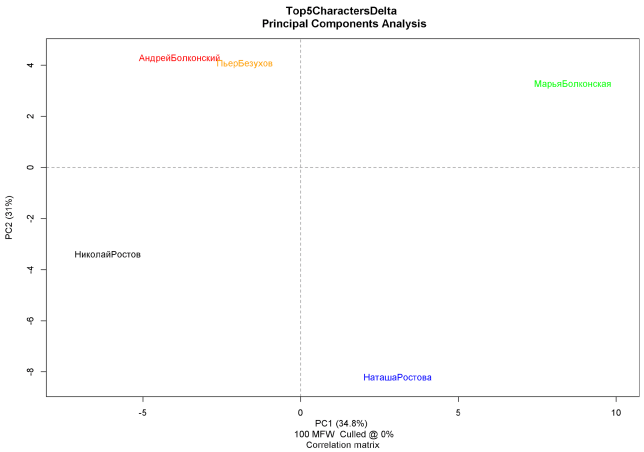


Fig. 4. Delta PCA for top 5 most talkative characters in War and Peace, 100 most frequent words

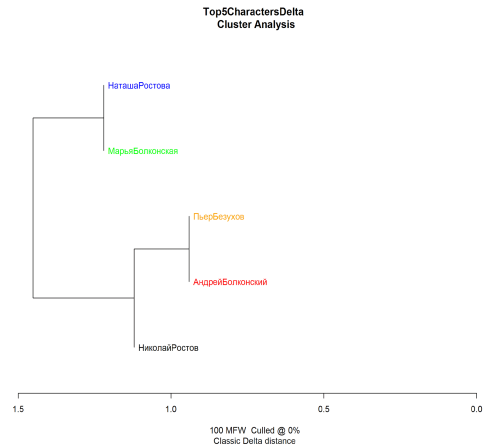


Fig. 5. Delta-based hierarchical clustering for top 5 most talkative characters in War and Peace, 100 most frequent words

The quality of speech authorship attribution inevitably got worse once we expanded our selection from 5 to 15 characters. The results were still quite tolerable reaching 10 out of 15 with certain settings. The analysis of mistakes showed that a) they're less likely to occur between characters of different gender and b) the mistaken characters have quite a lot of mutual conversations.

Further on, we decided to pay more attention to overall Delta scores of character voices and see if they give us any meaningful clustering of characters. Figure 6 shows PCA of characters based on Delta.

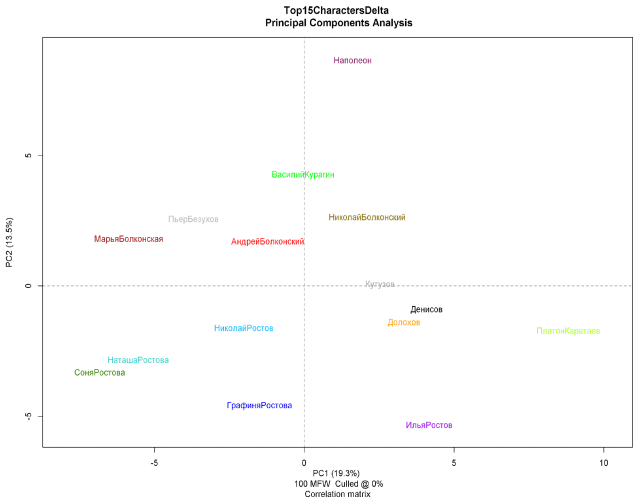


Fig. 6. Delta-based PCA for top 15 most talkative characters in War and Peace, 100 most frequent words

One can easily see the clustering of Rostov family, to a lesser extent this applies to Bolkonsky family as well. Dolokhov, Denisov and Kutuzov could constitute the 'war' cluster.

We then made another expansion and moved from 15 to 30 characters. Figure 7 demonstrates PCA of Delta scores for this selection.

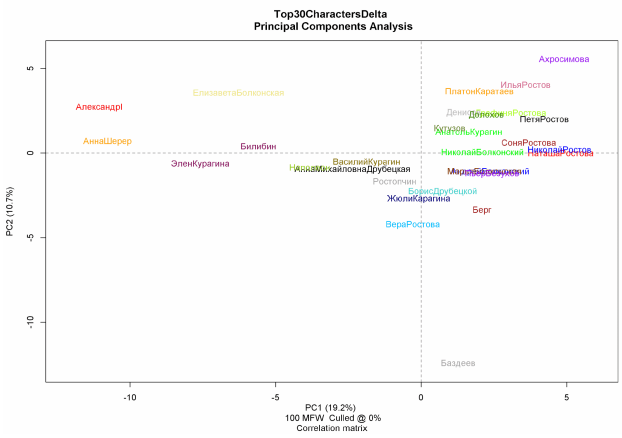


Fig. 7. Delta-based PCA for top 30 most talkative characters in War and Peace, 100 most frequent words

The most striking thing here is the obvious separation of Vera Rostova from the rest of Rostov family. The difference between cold, tempered and rational Vera and her emotional and very sentimental relatives is outspoken and obvious to the reader, but it seems valuable to have this potential quantiative support in the form of different Delta scores. What is even more striking is that Vera is quite close to Berg, a rationalizing careerist who becomes her husband. Note also the

closeness of Boris and Julie Karagine, another pragmatic couple happily united in a marriage of convenience.

Applying alternative features

Having tried Delta, we proceeded with our own set of alternative features for character voice analysis (a typical step, as dectibed in Eder, 2015). These features are not related to the lexical makeup of character speech and attempt to reduce the influence of gender-related morphological features of Russian language and the factor of mutual interactions between characters. At this stage we limited ourselves to four features only: the average number of words, the ratio of exclamatory sentences, the ratio oa question sentences and the ratio of punctuation marks (latter being a crude approximation of the 'disruptedness' of speech, which seems rather typical of certain more emotional and lively characters).

When the character set is limited to 5 characters these features even manage to distinguish character speech with some tolerable accuracy (though worse than Delta). However, the analysis of mistakes shows that they capture fundamentally different types of similarities than Delta does. For instance, joyful Natasha in this case is never mistaken for sentimental and melancholic Marya, but rather for her boisterous brother Nikolai. Pierre, on the other hand, is mistaken for Marya rather than for Andrey, who is distinct from them all. Figures 8 and 9 show the results of PCA and hierarchical clustering for these characters based on our own alternative features.

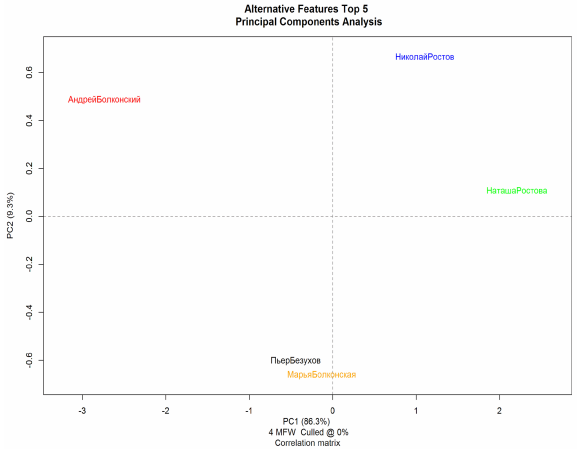


Fig. 8. PCA for top 5 most talkative characters in War and Peace, 4 alternative features

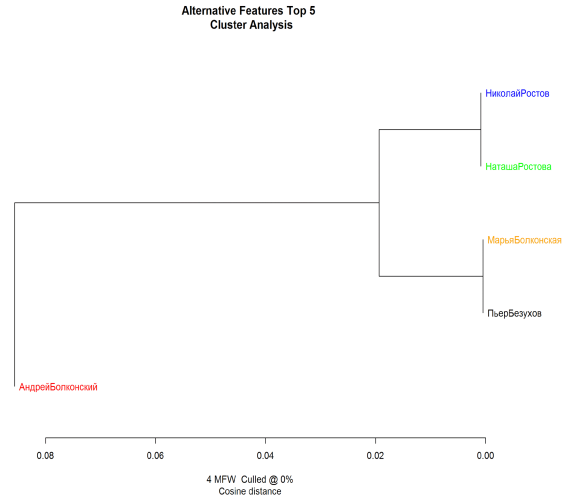


Fig. 9. Hierarchical clustering for top 5 most talkative characters in War and Peace, 4 alternative features

If we compare figures 8 and 9 to their counterparts from the Delta experiment (figures 4 and 5) we can see that the alternative features ignore gender or mutual interactions/ The hypothesis is that they enable us with a more indepth view of a character personality, his/her emotional type and so on.

Figures 10-12 show data on wider selections of characters using alternative features.

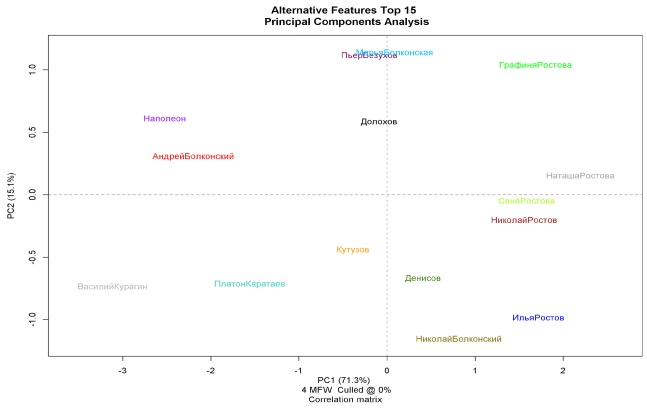


Fig. 10. PCA for top 15 most talkative characters in War and Peace, alternative features

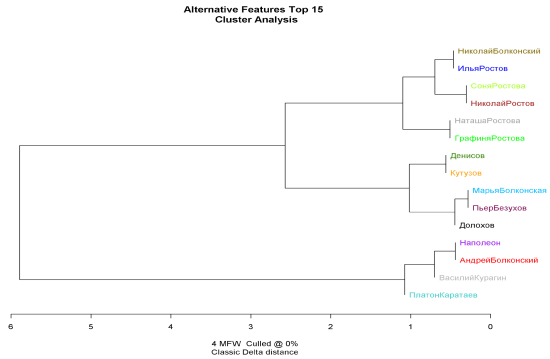
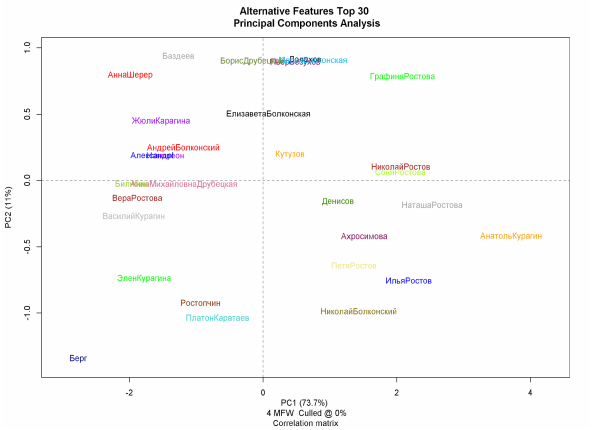


Fig. 11. Hierarchical clustering for top 15 most talkative characters in War and Peace, alternative features

Note that here we do not see any similarity between Andrey and Pierre. Moreover, Andrey is close to Napoleon, which is rather striking given Napoleon is his hero and role model for a considerable part of the novel.



The separation of Vera, on the other hand, is still rather visible - she is far from Moscow-centered Rostov world and close to Saint-Petersburg world of Kuragine family and berg.

Bibliography

Ipsum, L. (2017) "Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua*". Lorem Ipsum Quarterly*. 13.1: 27-45

Lorem, I. (2014) "Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua*". Lorem Ipsum Quarterly*. 7.1: 46-55

Amet, C. (1887) "Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua". Lorem Ipsum Quarterly. 3.1: 56-71