Hi everyone,

my name is Daniel and I represent Higher school of economics, which is a National research university in Moscow, Russia. I'll be talking about a project on Leo Tolstoy that we've recently launched and show the results of some experiments we carried out as a preparatory part for the project.

Here's the plan of my presentation. The first part is about the source material were working with, i.e. Tolstoy's legacy, and different DH initiatives around it. In the second part i'll describe some experiments on automatic text analysis and information extraction on the basis of Tolstoy's texts. And I also have a couple of slides with conclusion and further work.

As you probably know, Tolstoy was a very prolific writer, and he also lived quite a long life. So it should surprise you that his entire body of works published form a pile of books 2,7 meters high. And it is a very diverse pile for a 'one writer corpus' too. These two things - size and diversity - pose great challenges, but they also provide plenty of very interesting material to work with.

Sidenote: we also love Tolstoy for his rejection of copyright. He did open source 'before it was cool', so to speak.

So the first project that gave a start to the whole thing was 'All of Tolstoy in one click' - a joint effort by ABBYY software company, Higher School of Economics and Tolstoy museum. The idea was to automatically recognize and then manually proofread the 90-volume edition of Tolstoy's collected works. The second part of the project was made possible with help of crowdsourcing. About 4000 volunteers from all over the world responded to the initiative and offered their help. The project was completed way ahead of schedule, received a lot of media attention (journalists called it a 'crowdsourcing wonder') and was an overall success.

Now we had a purified electronic version of the texts, and it opened shiny new possibilities. As an intention to employ these possibilities, a second project arrived - the 'live pages' mobile app, created jointly by Higher School of Economics and Samsung. The application enables the reader to explore the text interactively with help of story timelines, geotagging and character quotes.

But the app was overall 'a toy', which did not seem enough to us. We still were not satisfied with the way Tolstoy's texts were represented on the web. They were digitized and made available, but is this really the end? There are basically two questions we have to ask ourselves when creating a digital edition of a book. The first one is how to preserve (i.e. invert into markup) all the existing features of the book, and the second one is what can we add to make it a better web resource, truly digital resource that would demonstrate clear benefits from all the new shiny possibilities of the web and comp technologies. Our attempts to produce an answer to the second question gave birth to yet another project, which is now underway - the so-called 'Tolstoy Digital' initiative.

The main goal of the project is to produce a TEI-based edition of Tolstoy's entire body of works. Such an edition should preserve the existing features of the print primary source (extremely rich critical apparatus, notes of all sorts etc.) and at the same time employ the recent advances in DH & NLP to turn the book into a truly digital resource for both literary scholars and general readers. Among the main challenges we encounter are size and diversity of works, existence of numerous versions of the same text, lack of uniformity in volumes, pre-

reform orthography and the peculiarities XIX century language that requires adjustment of NLP tools.

The first experiment carried out as part of the project was about character's speech activity in Tolstoy's epic novel 'War and peace'. We used the Compreno parser that generates syntactic-semantic trees to extract all instances of speech by the characters. On these slides you may see our preliminary statistics, which can also be called the 'biggest chatterbox' rating. In the first volume we extracted a total of 2392 speech activity occurrences. In 1705 cases some speaker was found, 530 of them were identified as a known character from the database. The most frequent speaker of the first volume is (rather unsurprisingly) prince Andrey Bolkonsky – according to our results, along the plot development he engages in speech activity at least 149 times.

Here's also the overall speech frequency distribution in the second volume of the novel. You can clearly distinguish the 'war' chapters with little talk (and a lot of action and philosophic thought) from the 'peace' ones with a lot of speech activity going on.

The second experiment was an attempt to find out if the predicate-argument structure of Tolstoy's sentences tell us anything about the characters? When mentioned, each character occupies certain semantic role within the phrase - usually agent or patient (called object in our model), but also experiencer, possessor, addressee and more. Having a semantic parser at our disposal, we decided to explore the typical roles Tolstoy's characters occupy in predicate-argument structure of phrases. Our hypothesis was that it could help us reveal (by means of quantitative analysis) some traits of different characters hidden within the text, understand the composition of Tolstoy's novel, automatically separate each character's line within the plot and demonstrate interactions between them. We counted the frequency of different roles for each character and got us some interesting results.

For instance, if we compare two obviously contrasting characters like princess Mariya Bolkonskaya and prince Vasili Kuragin, we might note that the latter occurs more often in Agent and Possessor positions, while Bolkonskaya shows more inclination towards the roles of Experiencer, Object and Addressee. This might be a reflection of character traits – the cunning and intriguing of profit-seeking prince Vasili versus the sensitivity and timidity of the shy and awkward princess Mariya. It is easy to notice that princess Anna Mikhailovna Drubetskaya is a much more active character than her son Boris, although he is mentioned no less often than her. This is also clearly the reflection of the plot of the first volume.

We realize that the results we have obtained so far require further analysis and verification, and the extraction system itself longs for more thorough adjustments to meet the specific needs of literary research. However, it appears that we were able to demonstrate the potential in such 'digital humanities' studies. We believe that syntactic and semantic representation of the text can and will bring us valuable knowledge about the great works of literature, reveal less obvious facts about well-known characters and plots, and, essentially, provide insights into the very process of creation.