**Digital Edition of the Complete Works of Leo Tolstoy**

Daniil Skorinkin, Higher School of Economics, dskorinkin@hse.ru

# Introduction

This paper presents a project aiming to create a complete digital edition of Leo Tolstoy’s works with rich structural, semantic, and metadata markup. The project is twofold: its first stage was a massive crowdsourcing effort to digitize Tolstoy’s 90-volume comprehensive print edition. That effort, known as ‘All of Tolstoy in One Click’, received considerable media attention (Bury 2013, McGrane 2013) and attracted more than three thousand volunteers from all over the world. Now that the first goal of ‘primary’ digitization had been achieved, an obvious next step was to provide the digitized texts with TEI-conformant markup. This work is in progress at the moment. Below we describe both stages of the project (the completed and the ongoing) with a special focus on their social and educational impact.

# Source description

More than 46 000 pages of text that collectively contain 14,5 mln words earned Tolstoy a place among the most productive writers of all times. The preparation of the 90-volume print edition started in 1928 (Tolstoy’s 100-th anniversary) and took three decades, with last volume published in 1958. The edition is rather diverse: apart from finished works of fiction (prose, poetry, drama), essays and schoolbooks, it contains numerous drafts, letters, volumes of personal diaries, which Tolstoy kept diligently throughout his life, certain number of facsimile manuscripts and drawings, and all sorts of editorial comments. A separate volume is dedicated entirely to alphabetic and chronological indexes. Each volume had 5000 copies, and none of them were ever reprinted, so by the second decade of the 21st century the whole edition was turning into a bibliographic rarity.

# OCR and primary digitization (aka ‘All of Tolstoy in one click’)

The ‘All of Tolstoy in one click’ project was a joint effort by the Leo Tolstoy State Museum and ABBYY, a Russian software company specializing in optical character recognition (OCR). The initial scanning of the print edition was performed by the Russian state library back in 2006. These images were recognized with help of ABBYY FineReader, proofread several times by volunteers, edited by professional editors and converted into e-books (now available at tolstoy.ru).

Proofreading was the most labour-intensive part of the whole project. Each volunteer was issued a special license for FineReader and a package of 20 unrecognized pages in PDF. Volunteers were supposed to recognize the PDF files using FineReader, correct the automatically identified areas on the pages if necessary (FineReader distinguishes between text, pictures, tables and so on) and then proofread the results of OCR. If the result was not uploaded back within 48 hours after the assignment, these 20 pages were returned in the initial assignments stack. The exchange was organized through a dedicated website readingtolstoy.ru, which now hosts a map with volunteers’ locations, press materials about the projects and other related information.

When the organizers announced the call for volunteers, they did not have very optimistic expectations and prepared to carry a fair share of the workload by themselves. The reality, however, proved their pessimism completely wrong. Within two hours after the launch of the crowdsourcing website (readingtolstoy.ru) more than two hundred people signed up and started working already, taking care of the first 5 volumes. In the end, the entire body of 46820 pages was recognized and proofread within 14 days (8,5 volumes per day) by 3249 volunteers from 49 countries. The most active volunteers processed up to two thousand pages. The leaders were awarded tours to Tolstoy’s family estate in Yasnaya Polyana, many other hardworking participants received free e-book readers and OCR software. When interviewed, many of the volunteers noted they could not stop working on the project because they were fascinated by Tolstoy’s text and experienced a surge of enthusiasm. Thanks to their hard work the organizers were able to prepare the entire electronic edition (91 original volumes plus 579 separate works extracted from these volumes) in all contemporary e-book formats in just 1,5 years.

# TEI markup (aka ‘Tolstoy.Digital’)

The diversity and scope of the 90-volume edition that we described above obviously call for various digital editorial practices (established or emerging), especially those associated with the TEI standards. To implement these, the second part of the project was launched under the codename Tolstoy.Digital. It is run jointly by the Leo Tolstoy State Museum and the National Research University ‘Higher School of Economics’. Though the main managers of the project are university professors and museum researchers, most of the actual research, planning, development and implementation is being done by students specializing in such fields as computational linguistics/NLP, digital humanities and (digital) literary studies. Some work is done in the form of student group projects for which credits are awarded, while other tasks are carried out by individual students as their personal course projects.

On one hand, a lot of effort is being put into re-encoding of the pre-existent metadata and editorial information in the digital environment. One particular example is the footnotes (more than 80 000 of them). Among them editorial and Tolstoy’s own comments, explanations and translations, plus all sorts of ‘critical edition’ style notes. The latter represent diverse editorial ‘secondary evidence’, e.g. ‘here Tolstoy wrote word A first, but then replaced it with an unclear word which is probably word B’ or ‘this phrase was crossed out with a dry pen, most likely by Tolstoy’s wife’ or ‘original page contained this addition on the margin’. As the size of the material suggests automation, currently our efforts are focused on automatic (or at least machine-aided) classification of notes and their subsequent conversion into TEI tags.

On the other hand, we are trying to augment the markup with new kinds of information that become available as text processing technologies advance. For instance, we have been experimenting a lot with reliable extraction of characters and identification of dialogue between them (with attribution of each speech utterance to its fictional speaker). This data later allows research on differences in the verbal behavior of different characters, which seems to have been a part o Tolstoy’s technique. Another area of active research is semantic role labeling within Tolstoy’s text (see Bonch-Osmolovskaya and Skorinkin, 2016).

The third major area of our work concerns letters (Bonch-Osmolovskaya and Kolbasov, 2015), which make up one third of the complete works. We have already extracted the metadata (addressee, date, place etc.) from the print edition in TEI format, and are building an extensive search environment/web interface upon it at the moment. Its current version is available at <http://digital.tolstoy.ru/tolstoy_search/>.

# Acknowledgements

This work was supported by grant 15-06-99523 from the Russian Foundation for Basic Research.

# References

Bonch-Osmolovskaya A., Kolbasov M. 2015. Tolstoy digital: Mining biographical data in literary heritage editions, in: 1st Conference on Biographical Data in a Digital World 2015, BD 2015; Amsterdam.

Bonch-Osmolovskaya, A.; Skorinkin, D. 2016. Text mining War and Peace: Automatic extraction of character traits from literary pieces. In: Digital Scholarship in the Humanities. Oxford: Oxford University Press.

Bury, L. 2013. Thousands volunteer for Leo Tolstoy digitization. In: The Guardian. <https://www.theguardian.com/books/2013/oct/16/all-leo-tolstoy-one-click-project-digitisation>

McGrane, S. 2013. Crowdsourcing Tolstoy. In: The New Yorker.
http://www.newyorker.com/books/page-turner/crowdsourcing-tolstoy