Daniil Skorinkin, Evgeny Mozhaev

TEI markup for the 90-volume edition of Leo Tolstoy's complete works

In this paper, we describe our work on an ongoing project titled "Tolstoy Digital". Our chief objective is to convert the 90-volume collected works of Leo Tolstoy into a proper "digital edition" with help of TEI.

With a legacy of more than 46,000 pages of text that collectively contain 14,5 million words, Tolstoy is famed as one of the most productive writers ever. The preparation of the 90-volume print edition started in 1928 (Tolstoy's 100th anniversary) and took three decades, with the last volume published in 1958. Apart from finished works (prose, poetry, drama, essays; schoolbooks), the edition contains numerous drafts, about 8,500 letters, entire volumes of personal diaries, which Tolstoy diligently kept throughout his life, a certain number of facsimile manuscripts and all sorts of editorial comments. A separate volume (No. 91) is entirely dedicated to alphabetic and chronological indexes.

The volumes have been digitized a few years ago, but so far contain little electronic markup. The size and diversity of the edition, along with inevitable inconsistencies in editorial practices, present all sorts of challenges to markup attempts.

One of such challenges are footnotes, of which there are more than 30,000. Among them Tolstoy's own comments, explanations and translations, comments and translations by editors, plus all sorts of technical notes. The latter usually represent various 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".

Obviously, all these footnotes should somehow be distinguished from each other in the TEI markup. As the sheer size of the material suggests some automation is inevitable, currently our efforts are focused on automatic (or at least machine-aided) classification of notes and their subsequent conversion into TEI tags.