The Russian language of the XVIII century: presentation of texts by M.V. Lomonosov

The Russian language of the XVIII century is heterogeneous one both in lexical aspects and in grammatical ones. There was no academic grammar until 1757, when M.V. Lomonosov, the great Russian scientist and literary man, published his famous "Russian grammar", which became the first reference book. It was the first attempt to fix standards of usage, to abolish obsolete word forms and syntactic constructions.

The aim of the paper is to demonstrate linguistic phenomena typical for the XVIIIth century language and particularly for the individual language of M.V. Lomonosov. Lomonosov's tragedy "Tamira and Selim" was chosen as the starting point for the work. It is common knowledge that poetic texts do not conform to grammatical rules in order to keep rhyme.

On the first stage of the work the text was tagged by the Dialing tagger. Its morphological dictionary is created on the basis of A. Zaliznyak's "Russian Grammar Dictionary" that represents the Russian grammatical system of the second half of the XXth century. When working with texts of previous centuries a scholar has to face a number of phenomena which are not inherent in the present morphological system. The Dialing tagger has a prediction algorithm for unknown words which searches in the dictionary for the most conterminous word form to an entrance word. When tagging, a "predictor" assigns to an unknown token such a set of possible tags which is determined by the similar dictionary entry. Nevertheless on the basis of the 18-19th century vocabulary this algorithm appears to be not so efficient (about 70%). In the output the processed text is stored in XML-file.

On the second stage the procedure of manual morphological disambiguation takes place. At the same time several lists are compiled. They contain information about:

- morphological ambiguity;
- "predicted" lexemes.

On this basis we form list of cases in which the program failed to attribute proper lexemes and grammatical characteristics to a token such as gender, case, and number for nouns; aspect, voice, and tense for verbs and so on. This register will be added to the dictionary of the tagger for its improving.

The following example illustrates one of such "predicted" lexemes:

Селим противу стен стоит напереди (Selim is standing opposite the walls).

Incorrect:

```
<w> противу
           <ana lemma="против?" pos="С" gram="|мр,дт,ед," />
</w>
```

Correct:

```
<w> противу
           <ana lemma="противу" pos="ПРЕДЛ" gram="| " />
</w>
```

The word "противу" is preposition and not noun as it was attributed. There is no such a word in present Russian, but it existed in Lomonosov's idiolect.

The language of M.V. Lomonosov bears traces of German syntax and sometimes the text is difficult to interpret because of this. We mean to show this interrelations in the text using technical labels. It is also intended to introduce special tags to label phenomena lacking in "Russian Grammar Dictionary", for instance obsolete words. Also we plan to work out system of tags according to Lomonosov's "Russian Grammar". This new approach will contribute to understanding the individual writer's language which cannot be explained within the framework of the XXth century grammar.

On the next stage this information from XML-file is converted into MySQL database. And a user can search through it by two types of parameters. The former is one that was attributed by the tagger. The latter is the system of "Lomonosov's" tags which is attributed manually. So the text will be interpreted from two different points of view. And comparing them a scholar will be able to observe language phenomena which are typical for the XVIIIth century texts, and also correlations between items.

We hope that this quite a new approach will be of benefit for investigations in the field of historical linguistics, and will help to trace the development of Russian grammatical system.