The IX International Conference on Computer Processing of Turkic Languages “TurkLang 2021”

SOFTWARE DEVELOPMENT FOR UZBEK TEXT CORPUS

Lecturer: Abduvali KARSHIEV, Mukhammadsolikh TURSUNOV, Bakhtiyor KHOLMUKHAMEDOV (Uzbekistan)

Abstract. This article is dedicated to theoretical and practical issues that arose during the creation of corpus for Uzbek language. Foreign experience and software have been investigated while creating the corpus of the Uzbek language. The modern corpus will be designed and developed in the Uzbek language as a balanced, large-scale and universal corpus. The theoretical and practical methods have been studied before the creation of the corpus are. Throughout the process, different softwares have been used to solve specific problems and The created corpus will be an open source for non-commercial use. The article describes the initial stages of the structure of the corpus and the requirements for the creation of modern corpus.

MARKING

Marking up the text is the most basic stage; as a result of its work, the corpus will be formed. Marking is the process of attaching special tags to text and its components. There are two types of special tags: linguistic tags and extralinguistic ‘external’ tags. Linguistic tags contain information that describes the lexical, grammatical, and other characteristics of text elements. The markup information will be presented as a structure. The morphological markup of a fragment of the Uzbek text *Boysari farzandli bo‘ldi. Shohimardon bilan tunab qoldi* is presented in Table 2.

Table 1. Marking words: XML-form. Маркировка слов: XML-форма.

```
<?xml version="1.0" encoding="windows-1251"?><text><p>
<s>
<w> boysari <ana lemma="BOYSARI" pos="[ot],[at.ot.],[shaxs_n.]"/> </w>
<w> farzandli <ana lemma="FARZAND" pos="[ot],[tur.ot],[qar.LMG],[shaxs.ot], [b.k.], [birl.son],[3-sh.]"/> </w>
<w> bo`ldi <ana lemma="BO'L" pos="[f],[must.f.],[har.f],[14.2]"/> </w></s>
<s>
<w> shohimardon <ana lemma="SHOHIMARDON" pos="[ot],[at.ot.],[shaxs_n.]"/> </w>
<w> bilan <ana lemma="BILAN" pos="[ko'm.], [sof ko'm.], [vos.m.]"/> </w>
<w> tunab <ana lemma="TUNAB" pos="[f], [must. f.], [har. f.], [o'-siz f.], [b-li f.], [an.n.], [sod. f.], [t.f.], [rav-sh]"/> </w>
<w> qoldi <ana lemma="QOL" pos="[f], [must. f.], [holat. f.], [o'-siz f.], [b-li f.], [an.n.], [sod.f.], [t.f.], [x.m.], [o'.z.], [III sh.b.]"/> </w>
</s></p></text>
```

Corpus control system

If special software - corpus manager exists, the text corpus becomes a powerful tool in the hands of a linguist.

A modern corpus manager should be able to following:

- concordance formation;
- search for contexts not only by words, but also by phrases;
- sort lists by several criteria selected by the user;
- providing the ability to describe the found word forms in an extended context;
- provide statistics on individual corpus elements;
- save and print results;
- the ability to work not only with individual files, but also with an unlimited number of corpuses;
- quickly respond to inquiries and get results quickly;
- be convenient and understandable for both beginners and experienced users.

Modern corpus systems should be able to solve more complex problems, such as compiling frequency dictionaries of words, compiling lists of colloquial expressions ‘ideoms’, and forming lexical-semantic groups.

The corpus of the Uzbek language has been developed *uzbekcorpora.uz* and preliminary results were obtained. according to the preliminary result, the epic “Alpomish” has been included in the database.

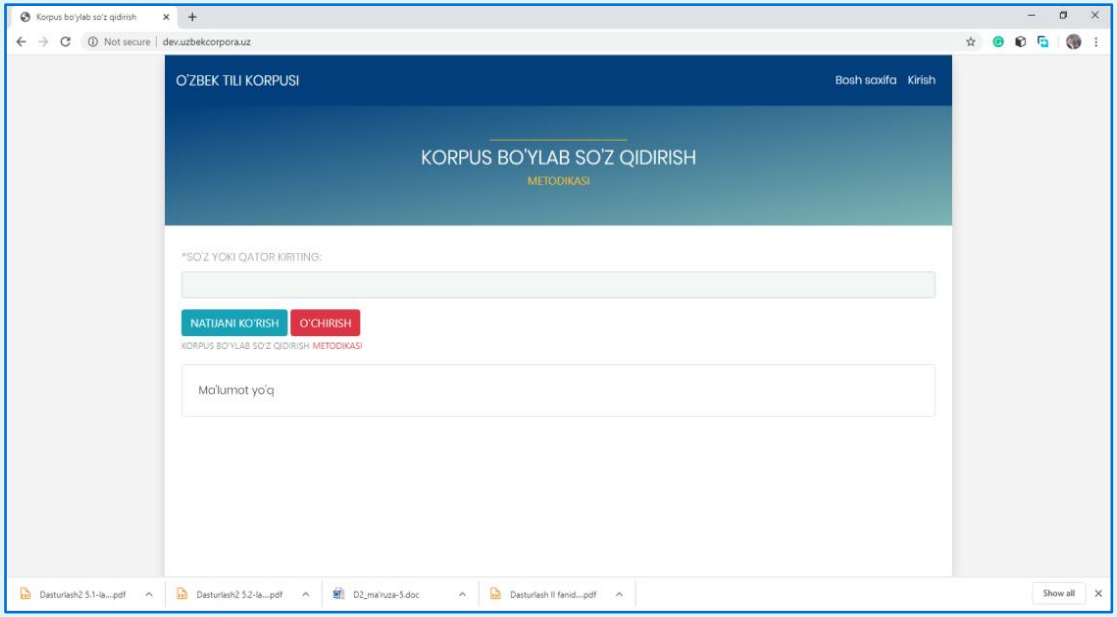

Fig. 1. Concordance

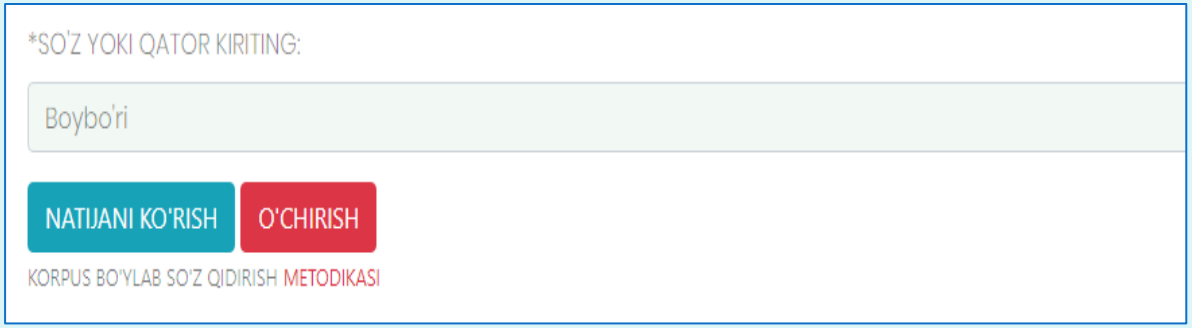

Fig. 2. Search word entry window

Fig. 3. Results of search

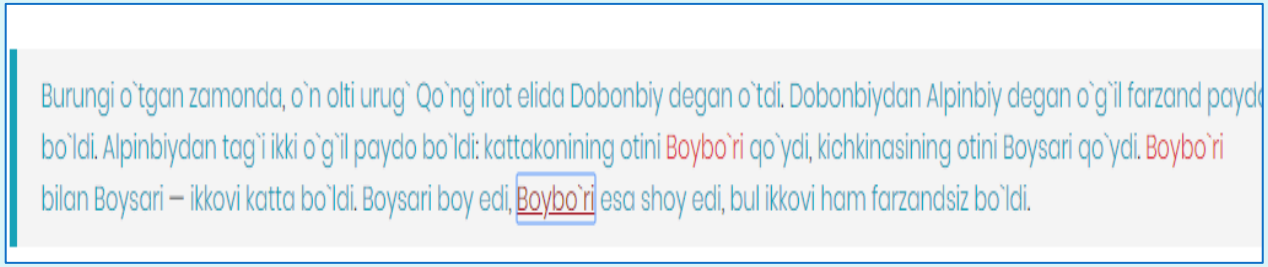

Fig. 4. Paragraph appearing in search results

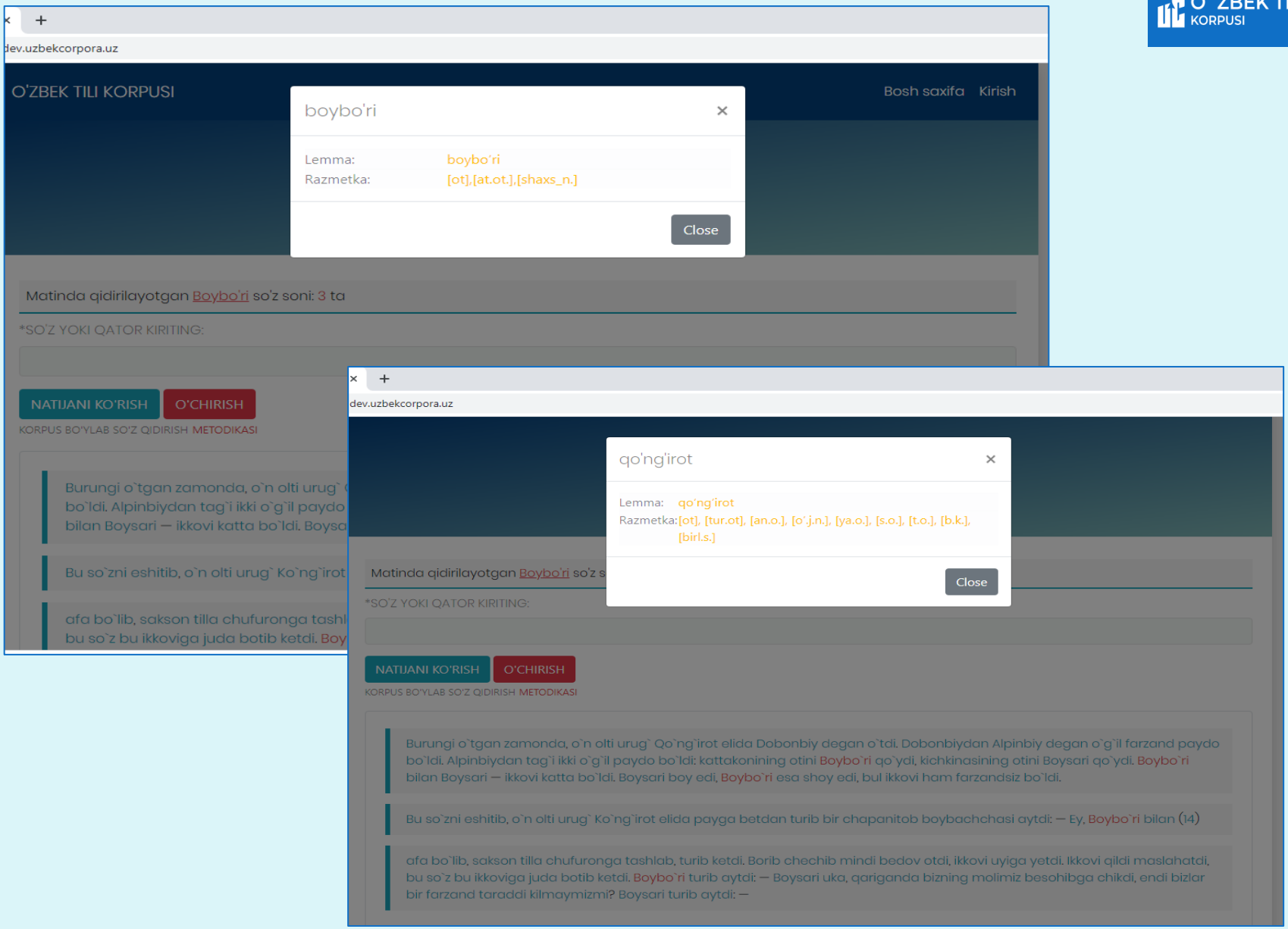

Fig. 5. *Boybo'ri* and *Qo'ng'irot* has been selected and lemma and markup of the word

The term ‘corpus quality’ varies depending on the intended use of the corpus. In empirically oriented theoretical linguistics, carefully chosen procedures and cleaning up indestructible redundant messages are important, however for many computational linguistics and language technology tasks, aggressive corpus cleansing is necessary to achieve good results. Technically, these differences are not only related to software architecture, but also to different software configurations. It is also important to focus on increasing the processing speed and data collection across all corpus-oriented projects