

ЎЗБЕК ТИЛИ МАТНЛАРИНИ РАЗМЕТКАЛАШ

Турсунов Муҳаммадсолиҳ Саъдин ўғли

muhammadsolih927@gmail.com

СамДУ мустақил тадқиқотчиси

Аннотация. Ушбу мақолада ўзбек тили миллий корпуси (uzbekcorpora.uz)га матн юклаш ва сўзларни разметкалаш келтирилган.

Калит сўзлар: *MS Word, разметка, МББТ, лугат, автоматик разметка, грамматик ҳарактеристика.*

Abstract. This article describes how to upload text to the National Corpus of the Uzbek language (uzbekcorpora.uz) and organize words.

Key words: *MS Word, markup, MBBT, dictionary, automatic markup, grammatical characteristics.*

Абстрактный. В этой статье описывается, как загрузить текст в Национальный корпус узбекского языка (uzbekcorpora.uz) и упорядочить слова.

Ключевые слова: *MS Word, разметка, МББТ, словарь, автоматическая разметка, грамматические характеристики.*

uzbekcorpora.uz платформасида грамматик разметкалаш дастури *Nuxt JS, Python* ва *PostGreSql* маълумотлар базасини бошқариш тизимида (МББТ) тузилган. Матнларни корпусга киритиш жараёнини соддалаштириш учун грамматик разметкалаш икки босқичда бажарилади: матнни дастлабки форматлаш босқичи ва разметкалаш босқичи. Дастлабки форматлаш босқичида матннинг тузилмавий компонентлари аниқланади, яъни матнда сўзлар, гаплар, абзацлар ажратилади. Шунингдек, сўзларнинг матндаги ўрни аниқланади. Иккинчи босқичда грамматик разметкалаш амалга оширилади. Разметкалаш натижасида матндаги ҳар бир сўзшаклга ўзининг морфологик параметрлари ва сатрли атрибутлари тўплами бириктириб чиқилади. Қуйида бу иккала босқичнинг грамматик разметкалаш дастури орқали бажарилиши батафсил баён қилинади.

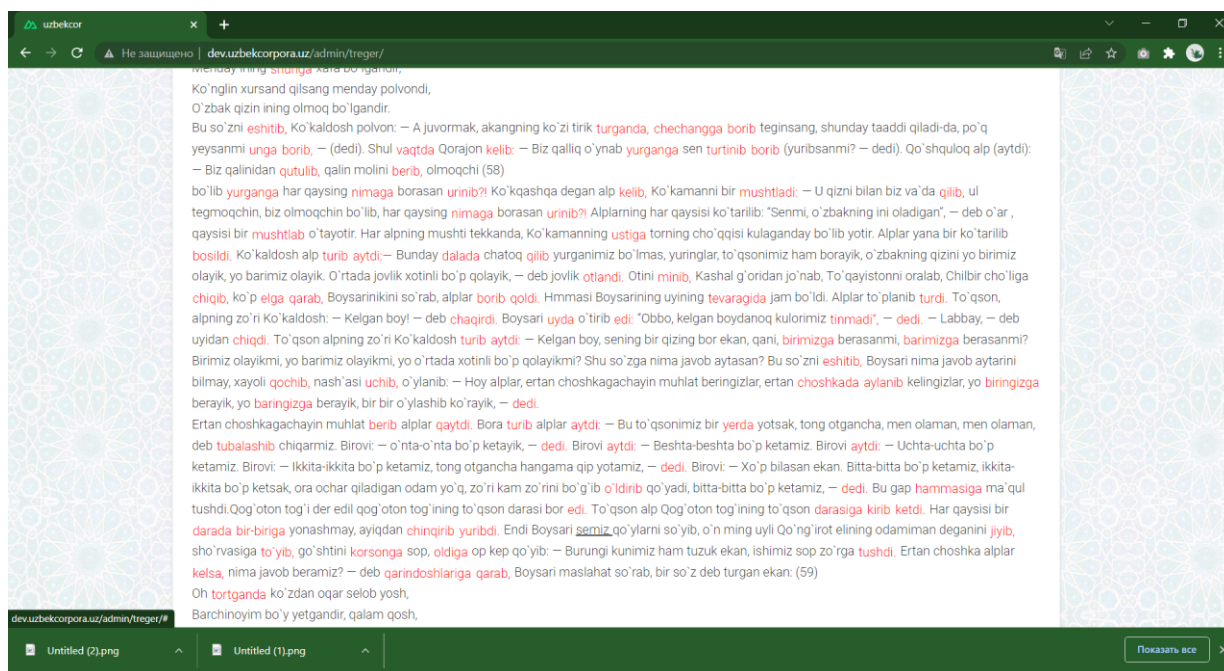
Дастлабки форматлаш босқичи. Бу босқичда корпусга киритиладиган файл дастлаб қандай шаклда бўлишидан қатъий назар, у *MS Word* дастури форматида ўтказилиши керак. *MS Word*нинг 2007 ва ундан юқори версияларини ишлатиш зарур, чунки уларда матн **.docx* форматда бўлиб, стандарт тизим асосида автоматик равишда тегланган ҳолда бўлади. Матнни тузилмавий компонентларга (сўзлар, гаплар, абзацлар) бўлиш *MS Word* дастури томонидан автоматик тарзда бажарилади. Матнда сўзларнинг ўрнини билдирувчи маълумотлар алоҳида файлга ёзилади.

Разметкалаш босқичи. Разметкалаш деб матнга ва унинг компонентларига махсус тегларни бириктириб чиқишни тушуниш керак. Махсус теглар икки хил бўлади: лингвистик теглар ва экстралингвистик (ташқи) теглар. Лингвистик теглар матн элементларининг лексик, грамматик ва шунга ўхшаш бошқа хусусиятларини тавсифлайдиган маълумотлардан иборат. Экстралингвистик теглар эса муаллиф ҳақидаги ва матн ҳақидаги маълумотларни (муаллиф, номланиши, нашр йили ва жойи, жанр, тематика ва ҳақозо) тавсифлайди [Николаев И.С., 2016:320].

Корпуснинг улкан ҳажми билан мутаносиб равишда разметкалаш кўп вақт ва меҳнат талаб қиладиган жараёндир. Шунинг учун у автоматик тарзда бажарилиши зарур. Агарда разметкаланишнинг бирор бир тури ҳали автоматлаштирилмаган бўлса, у ҳолда бу ишни қўлда бажаришга тўғри келади [M.S.Tursunov, 2022].

Бу босқичда, дастлаб, автоматик разметкалаш дастури ишга туширилади.

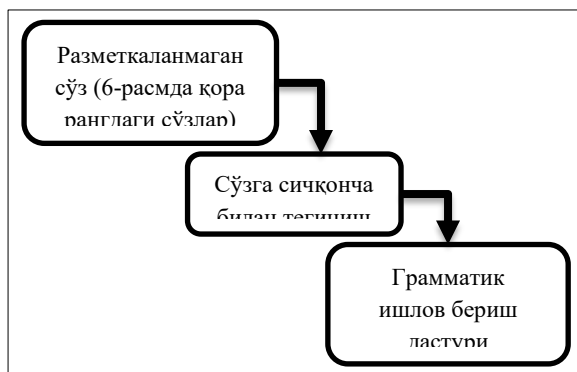
Натижасида матн *.docx форматдаги файл (1-расм) кўринишида тақдим этилиб, унда разметкаланган сўзлар қизил ва разметкаланмагани қора ранг ажратилган бўлади. (Грамматик луғат бўш бўлса, ҳамма сўзлар қора рангда чиқарилади.)



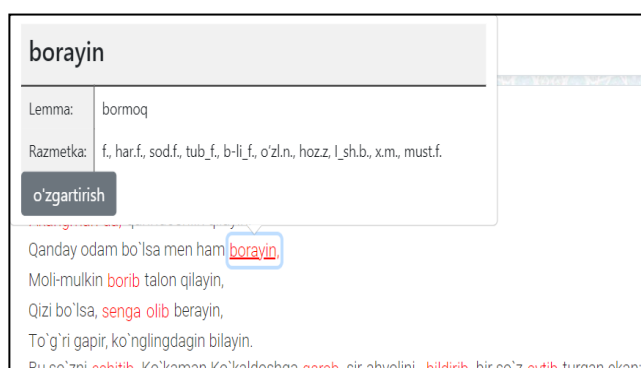
1-расм. Файлнинг дастлабки ҳолати

Бунинг учун, сўз устига курсор келтириб босилганда, сўзнинг морфологик характеристикалари ёзилган ойна экранда пайдо бўлади (2-расм). Бу ойнадаги маълумотлардан

а) сўз разметкаланмаган ҳолат

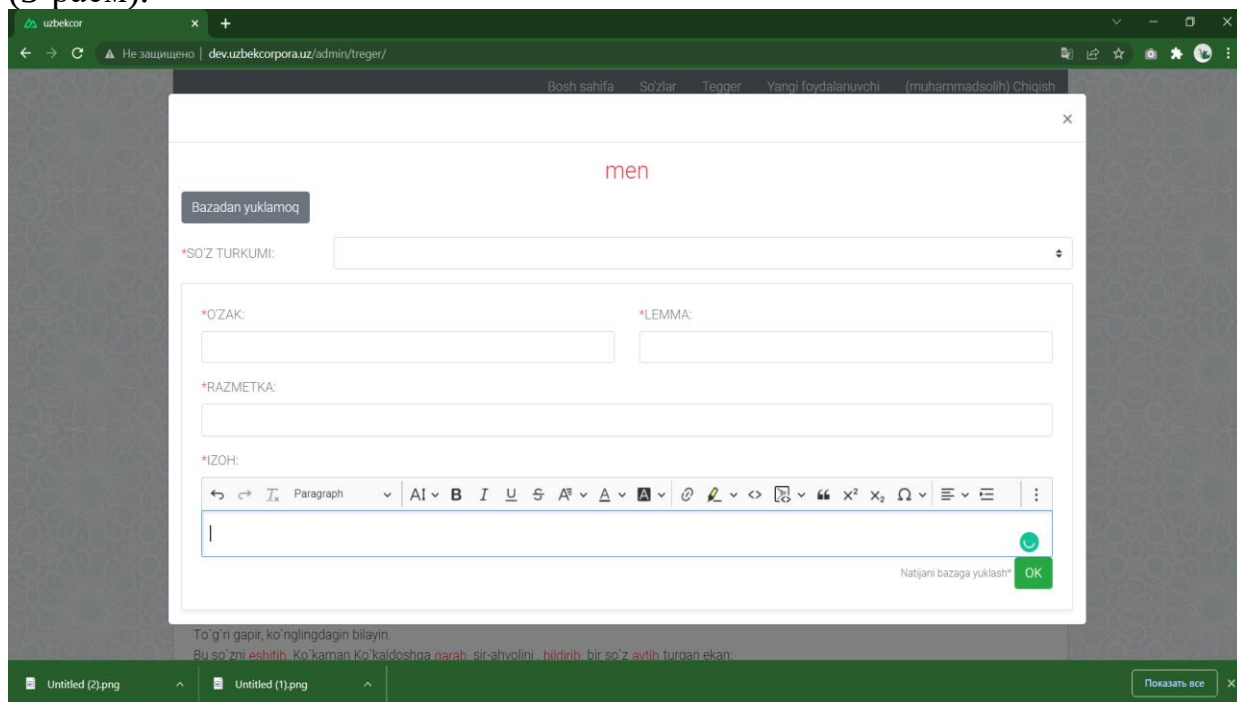


б) сўз разметкаланган ҳолат



2-расм. Сўзнинг грамматик теглари

Матндаги сўзларнинг ранглиридан фойдаланиб, сўз тегланган ёки тегланмагани аниқланади. Сўз ҳали разметкаланмаган бўлса, тегишли тугма босилади ва сўзнинг морфологик параметрларини белгилаш учун ойна очилади (3-расм).



3-расм. Сўзга морфологик характеристикалар бериш ойнаси

3-расмда сўзни морфологик разметкалаш мумкин бўлади. Бунда сўзга унинг туркуми, туркумга мос равишда грамматик белгилари, ўзаги, леммаси ва

изоҳ киритилади. Ойнадан туркумни танлаганимизда туркумга мос равишда грамматик белгилар рўйхати очилади. Рўйхатдан керакли белгилар танлаб олинади.

Сўзларни грамматик луғатга киритиш. Бу вазифани бажарувчи дастурий модул *Lug_Kirit* деб номланган. Луғатга янги бирлик киритиш фойдаланувчи томонидан интерактив режимда амалга оширилади. Бу жараёни *Lug_Kirit* дастури таъминлайди. Компьютер экранида *Suz_Rhati* дастури томонидан чиқарилган рўйхат мавжуд ва унда битта сўз бошқа ранг билан ажратиб кўрсатилган. *Lug_Kirit* дастури ёрдамида фойдаланувчи рўйхат бўйича ҳаракатланиб, жорий сўзни танлайди ва “Enter” тугмасини босади. Натижада экранда киритиш ойнаси пайдо бўлади. Фойдаланувчи бу ойнада сўралган маълумотларни клавиатурадан териб киритади ва “Базага киритиш” тугмасини босади. Шунда танланган сўз ва унга мос киритилган грамматик характеристикалар жамланиб, электрон луғат файли *LUGAT*га ва бошқа ёрдамчи файлларга ёзиб қўйилади. Шундан сўнг фойдаланувчи рўйхатдан навбатдаги сўзни танлайди ва базага киритади. Шу тариқа, берилган рўйхатдаги барча сўзлар *LUGAT* файлга киритилиб, грамматик луғат қадамма-қадам бойитиб борилади.

Теггер дастури матнларни корпусга киритиш учун фойдаланилади. Матнларни корпусга киритиш жараёнида, матнга грамматик ишлов берилади. Матндаги сўзларга грамматик маълумотлар яримавтомат тарзида тўлдирилади ва грамматик маълумотга эга бўлган сўзларнинг ранги қизил рангга ўзгаради ва фойдаланувчига разметкаланмаган сўзлар ҳамда разметкаланган сўзлар якқол ажралиб кўринади. Сўзни разметкаланишнинг лингвистик таъминоти иловада (а илова) келтирилган. Бунини 1-расмда кўриш мумкин. Разметкаланмаган сўзларга грамматик маълумотлар бериш учун сўз устидан бир марта босилади ва сўзга грамматик ишлов бериш ойнаси ҳосил бўлади (8-расм). Сўз разметкалангандан сўнг сўзнинг ранги кўк рангга ўзгаради. Бу эса матн билан ишлаш вақтида разметкаси қўлда берилганлигини билдиради. Қизил рангдаги сўз устидан босилган вақтда унинг грамматик маълумотлари кўринади. Агар грамматик маълумотлар нотўғри берилган бўлса, 2-расмнинг б бандидаги ойнадан “O‘zgartirish” тугмаси босилади ва сўзга грамматик ишлов бериш учун 8-расмдаги ойна ҳосил бўлади. Матнга ишлов берилгандан сўнг матн қайта сақланади. Матнлар корпус таркибида *JSON* (б иловага қаранг) форматида сақланади.

Грамматик разметкаланиш автоматлаштириш. Сўзларни грамматик таҳлил дастури учун махсус автоматик разметкалаш модули яратилган. Бу модул қўлда теглаш дастуридан олдин ишга туширилади. Унда ҳар бир сўз мавжуд грамматик луғат (МБнинг *Words* жадвалидан) таркибидан изланади ва агар у топила, луғатда бор морфологик параметрлардан фойдаланиб, тег автоматик тарзда шакллантирилади ва сўзга бириктирилади. Грамматик луғатдан топилмаган сўзларнинг разметкаси бўшлиғича қолади. Матнда

такрорланувчи сўзлар учун луғатдан излаш фақат бир марта бажарилади ва натижа уларнинг ҳаммасига тадбиқ этилади.

Автоматик разметкашнинг ўзига хос камчиликлари ҳам мавжуд:

1) Матннинг грамматик луғатдан топилмаган сўзлари разметкасиз қолдирилади. Лекин, корпус яратиш жараёнида грамматик луғат кенгайиб бориши натижасида матннинг ундан топилмайдиган сўзлари улуши тобора камайиб боради. Шунинг учун, грамматик луғатни мумкин бўлган барча усулларни қўллаган ҳолда, мумкин қадар бойитиш зарур;

2) Грамматик луғатни кенгайтириб бориш натижасида омоним сўзлар муаммоси ҳам кўпроқ намоён бўла бошлайди. Бу муаммо шундан иборатки, луғатда бир сўз бир неча хил грамматик характеристикаларга эга бўлиб, улардан қайси бирини олиш керак деган савол пайдо бўлади. Муаммо сўзнинг контекстини таҳлил қилиш асосида ҳал қилинади.

Фойдаланилган адабиётлар:

1. Николаев И.С., Митренина О.В., Ландо Т.М., Прикладная и компьютерная лингвистика. Книга – Санкт Петербург: СПбГУ. 2016. – С. 320. - ISBN 978-5-9710-3472-8.
2. Kennedy G., An introduction to corpus linguistics. London: Longman, 1998.
3. M.S.Tursunov, Description of the management system programs of the national corpus of the Uzbek language // International journal of engineering mathematics, <https://iejemta.com/index.php/em/article/view/9/9>, volume 4 ISSUE 1, 2022, 32-42 p.
4. Elov B., Hamroyeva Sh., Abdullayeva O., Uzoqova M. O'zbek tilida pos tagging masalasi: muammo va takliflar. O'zbekiston: til va madaniyat (Amaliy filologiya), 2022, 5(4). 45-63.