

ЎЗБЕК ТИЛИ МАТНЛАРИНИНГ РАЗМЕТКАЛАНГАН КОРПУСИНИ МОДЕЛЛАШТИРИШ

Қаршиев Абдували Беркинович

Физика-математика фанлари номзоди,
ТАТУ Самарқанд филиали профессори
abduvalikarshiyev@gmail.com

Аннотация. Ушбу ишда, Ўзбек тили миллий корпусини қурилишида пайдо бўлган назарий ва амалий масалалар келтирилган. Корпусни қуриш вақтни ва меҳнатни талаб қилувчи жараёндир. Мавжуд корпусларни таҳлил қилиш натижасида қўриш мумкинлиги, корпусларни қуришда барча тиллар учун ягона бўлган методология мавжуд эмас. Ушбу мақолада корпусга қўйиладиган талаблар, корпус қуриш модели ва босқичлари таҳлил қилинган.

Калит сўзлар: *корпус, модел, маълумотлар базаси, матн, разметка, тег, JSON, сўз.*

Abstract. In this work, the theoretical and practical issues arising in the construction of the national corpus of the Uzbek language are presented. Building a hull is a time-consuming and labor-intensive process. As a result of the analysis of existing corpora, it can be seen that there is no single methodology for building corpora for all languages. This article analyzes the hull requirements, hull construction model and steps.

Keywords: *corpus, model, database, text, markup, tag, JSON, word.*

Абстракт. В данной работе представлены теоретические и практические вопросы, возникающие при построении национального корпуса узбекского языка. Строительство корпуса — трудоемкий и трудоемкий процесс. В результате анализа существующих корпусов видно, что единой методики построения корпусов для всех языков не существует. В этой статье анализируются требования к корпусу, модель конструкции корпуса и этапы.

Ключевые слова: *корпус, модель, база данных, текст, разметка, тег, JSON, слово.*

Кириш. Мавжуд корпусларни таҳлил қилиш натижасида қўриш мумкинлиги, корпусларни қуришда барча тиллар учун ягона бўлган методология мавжуд эмас. Бунинг сабаби шундаки, турли тилларда турли хил қоидалар, технологик жараёнлар амал қилади. Морфологик тавсифлар тўпламлари турлича бўлишига қарамасдан, юқорида қараб чиқилган корпусларнинг барчасида морфологик (ёки грамматик) разметка таъминланган. Синтаксис разметка эса баъзи корпусларда бор, лекин турлича ёндошувлар асосида амалга оширилган бўлса, баъзиларида умуман йўқ. Корпусларни тилни қамраб олиш даражаси бўйича солиштирганда, рус тили миллий корпусини ҳам хронология, ҳам жанрлараро энг мувозанатлашган ва матнлар хилма-хиллиги таъминланган корпус деб ҳисоблаш мумкин. Лекин рус тили миллий корпусининг бу ютуғи

унинг асосий камчилигини ҳам юзага келтирган. Ҳар қандай йирик корпусда тўлиқ ва аниқ разметкаланишнинг имконияти чекланган бўлади. Ноаниқликлар, асосан, автоматик разметкалашда омоним сўзлар туфайли пайдо бўлади.

Мавжуд корпуслар таҳлилидан қўриш мумкинки, миллий корпус тилнинг қонуният ва хусусиятларини иложи борида қўп тақдим эта олиши ва таркиби бўйича мувозанатлашган бўлиши керак. Матнларнинг бир тури тилдаги жами матнлар ичида қандай ҳажмий улушга эга бўлса, ана шундай миқдор миллий корпусда ҳам сақланиб қолиши керак [А.Б.Қаршиев, 2021].

Лингвистик корпусларга талаблар. Корпус қўлланилишига доир қуйидаги талаблар қўйилиши мумкин:

1. Корпус ўқув ва маълумот қўлланмалар ёзишда, луғатлар тузишда ҳамда лингвистик экспертизалар ва таҳлилларни ишлаб чиқишда, тил маълумотлари сифатида ишончли ва ваколатли манба ҳисобланади;

2. Корпус турли хил лингвистик фанларни ўқитишда худди тилнинг кўргазмали (иллюстрация) манбаси, ҳамда таҳлилларни тақдимотлилик қўриниши сифатида ишлатиш мумкин;

3. Корпус илмий тадқиқот ишларида, шу ўринда илмий фаразларни текширувчи сифатида қўлланилиши мумкин;

4. Корпус тил фаолиятининг универсал маълумотлар қўлланмаси сифатида бўлиши мумкин.

Корпус яратишда қуйидаги саволларга асосланган ҳолда талабларни белгилаб олиш муҳим:

1. Корпус фойдаланувчиси ким? (индивидуал, гуруҳ, лингвистик жамият).

2. Корпуснинг мантиқий ғояси нимада?

3. Корпусни қуришда қанча ҳажмдаги маълумотлар билан ишламоқчисиз?

4. Нима ишлатилади: матн парчалари, тўлиқ матнлар ёки иккаласи ҳамми?

5. Корпусдаги матнларни танлаш тартиби қандай?

Корпус яратишда турли хил разметканинг бўлиш ёки бўлмаслигини ва агар бўлса, унинг аниқлик даражасини ҳисоб олиш мақсадга мувофиқ. Корпус дастурий таъминоти таркибида автоматик разметкалаш дастурлари албатта бўлиши керак. Разметканинг юқори аниқликда бўлишини фақат филолог мутахассис таъминлаши мумкин [Kennedy G., 1998]. Филолог мутахассис ҳар бир матннинг разметкаланишини сўзма-сўз синчиклаб текшириши, автоматик разметка хатоларини тузатиши ва омоним ҳолатларни бартараф қилиши керак. Ишнинг асосий қисми қўл меҳнати орқали бажарилиши мумкин. Шунинг учун филолог мутахассиснинг компьютердаги ишига мўлжалланган махсус разметкалаш дастурлари яратилиши зарур. Бу дастурлар қуйидаги сифатларга эга бўлиши керак:

- фойдаланувчи сўзлар ёки гапларнинг грамматик ёки синтаксис характеристикаларини етарли даражада тўлиқ аниқлаш ва ўзгартириш имкониятига эга бўлсин;

- таҳлил максимал даражада қулай ва содда бўлиб, тез бажарилсин ва бу таҳлилнинг тўлиқлигини камайтирмасин [M.S.Tursunov, 2022].

Корпус моделининг схематик шакли 1-расмда тасвирланган [Седов А.В., 2013].

1-расм. Корпус моделининг схематик тасвири



Корпус модели. Корпусни T_i матнларнинг жамланмаси деб қараймиз. Матн эса p_{ik} тиниш белгилари, пробел ёки сатр охири белгилари билан ўзора ажратилган w_{ij} сўзлар кетма-кетлигидан иборат [Седов А.В., 2013]:

$$Corpus = \{T_1, T_2, \dots, T_n\},$$

$$T_i = \{w_{ij}\} U \{p_{ik}\},$$

ерда j - w_{ij} сўзнинг i -матндаги ўрнини, k - p_{ik} тиниш белгининг ўрнини билдирувчи қиймат. Шунингдек, ҳар бир сўзга унинг матндаги жойлашув ўрни мос қўйилади:

$$Disposition: w \rightarrow positions \in Positions, Positions = Pos1, Pos2, \dots, Posq,$$

бу ерда w сўз учун унинг матндаги координатаси q аниқланади. Матндаги боблар, параграфлар, абзацлар, гаплар, гап бўлаклари - унинг қўшимча

тузилмавий бирликларидир. Модел матннинг энг кичик тузилмавий бирликларидан (сўзлардан) сўз бирикмалари ҳосил бўлишини инобатга олади. Гапни сўзлар тўплами деб қаралса, ҳар бир сўз бирикмани қисмий тўплам деб ҳисоблаш мумкин:

$$SS_j = w_{i1}, w_{i2}, \dots, w_{ik} \subset S, \text{ бунда } U_j SS_j = S.$$

Ҳар бир сўз фақат битта сўз бирикмаси таркибига кириши мумкин, шунинг учун иккита бирикманинг кесишмаси бўш тўпламдан иборат бўлади: $SS_k \cap SS_j = \emptyset$, бунда $k \neq j$.

Ҳар бир сўз учун матндаги ўрнидан ташқари унинг морфологик параметрлари тўплами ҳам аниқланиши керак. От сўз туркумига тегишли сўз учун бу параметрлар – келишик, сон (бирлик/кўплик), жонли/жонсиз.

Морфологик параметрларнинг ўзи яна бир векторни ташкил қилади:

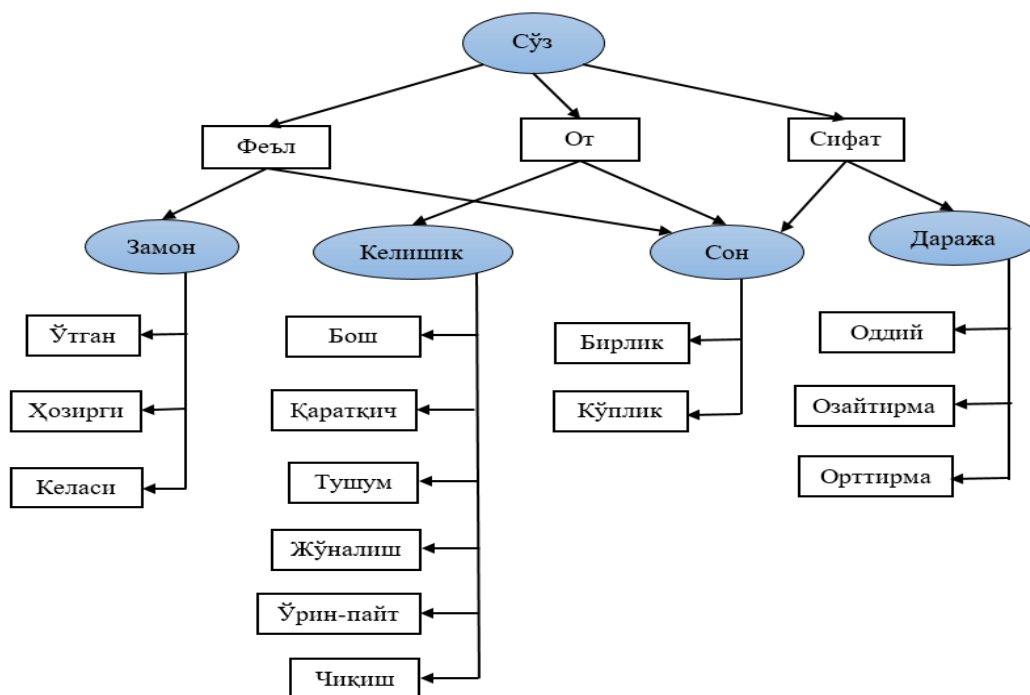
$Morfology: w \rightarrow intParams \in Nm, intParams = Param_1, Param_2, \dots, Param_m$, бу ерда m – сўзнинг грамматик белгиларининг сони. Морфологик белгилар натурал сонлар билан кодланган. Ушбу белгиларнинг декодланиши алоҳида конфигурация файлда сақланади ва дастур администратори ёки фойдаланувчининг ўзи ҳам белгиларни ўзгартириши мумкин.

Грамматик белгиларнинг ўзаро боғлиқлигини график кўринишда тасвирлаш мумкин:

$$MorfAttr = (W, E), W = MorfParam \cup MorfItems, MorfParam, MorfItems \neq \emptyset, MorfParam \cap MorfItems = \emptyset, E = E_1 \cup E_2 \quad E_1 = u, v: u \in MorfParam, v \in MorfItems, \quad E_2 = v, u: u \in MorfParam, v \in MorfItems,$$

бу ерда $MorfParam$ – морфологик параметрлар тўплами, $MorfItems$ – морфологик параметрлар қийматлари тўплами. Морфологик параметр, масалан, от сўз туркумига тегишли сўз учун – бу келишик бўлса, морфологик қийматлар – бош, қаратқич, тушум, жўналиш, ўрин-пайт, чиқиш.

Мисол сифатида ўзбек тили корпусига киритилган матнни грамматик разметкалашда (А иловага қаранг) сўз таҳлилининг бир қисмини схема шаклида келтирамиз (2-расм).



2-расм. Сўзнинг морфологик таҳлили иерархик дарахтини қуришга доир
мисол

Ҳар бир сўз учун морфологик параметрлардан ташқари сатрли параметрлар деб аталадиган қўшимча параметрлар ҳам киритилган. Бу параметрлар, хусусан, сўз ўзаги ва леммасидан иборат. Сўзларни разметкалашнинг универсал дастурини яратишда бундай параметрлар сонини ва таркибини олдиндан аниқлаб олиш имконияти назарда тутилиши зарур. Формал ҳолда $WStrAttr:w \rightarrow sParams=sParam_i, i=1,l$ деб ёзсак, $sParams$ тузилма фойдаланувчи томонидан алоҳида конфигурация файлида берилиши керак.

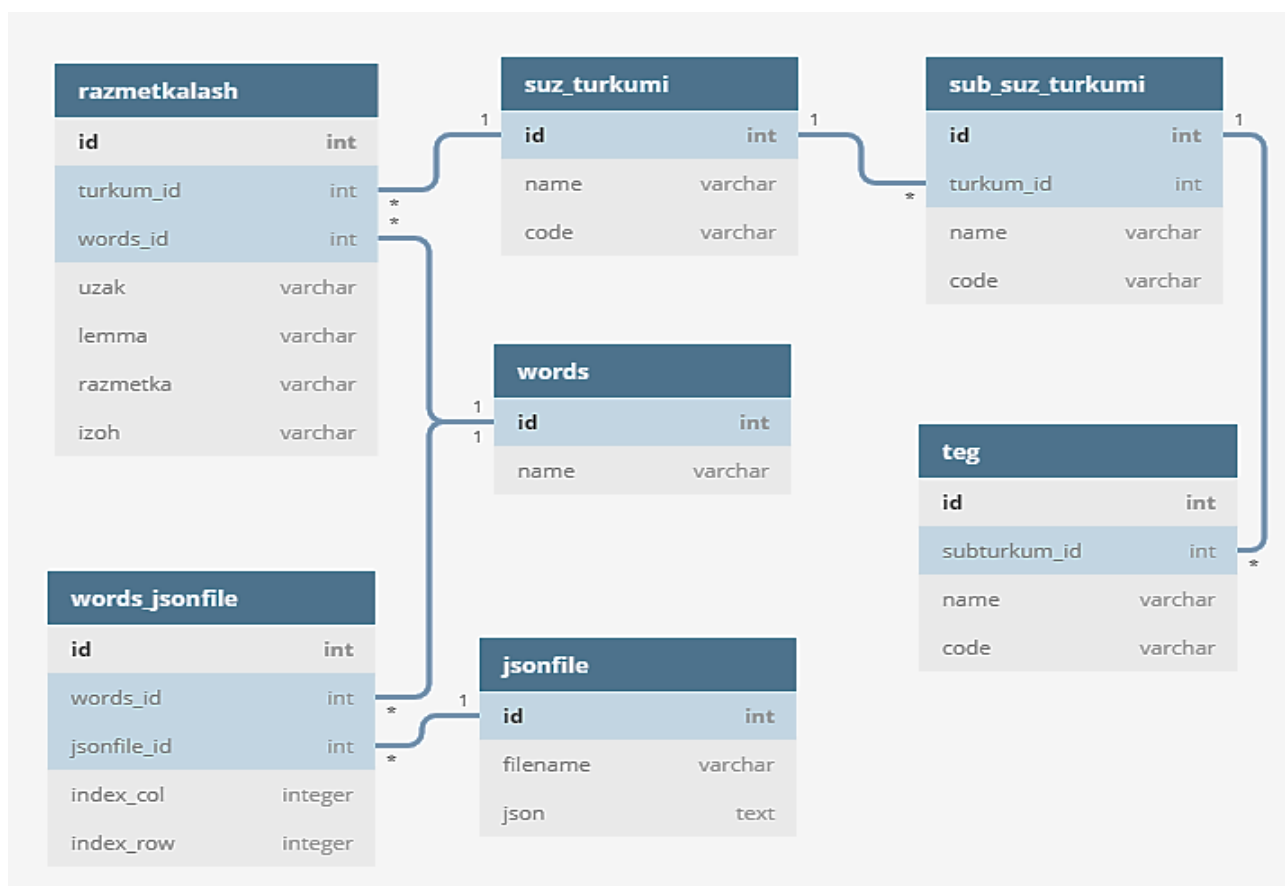
Шундай қилиб, ҳар бир сўзга қуйидаги тузилма мос қўйилади:

$$w \rightarrow \langle Spelling, Disposition, intParams, sParams \rangle$$

(сўз ўрни, морфологик параметрлар, сатрли параметрлар). Бу тузилмани сўзнинг грамматик разметкаси деб ҳисоблаш мумкин.

Маълумотларни компютерлаштиришда турли хил файл форматлари билан танланган матнлар, УТФ-8 кодланган матн форматларига ўтказилади [Гвишиани Н.Б., 2018, Грудева Е.В. 2012]. Мавжуд ўзбекча матнларни қайта ишлаш ва мониторинг қилиш учун вебга асосланган қидирув ва бошқарув корпуסי ишлаб чиқилган. Ўзбек тили миллий корпусикорпус бошқарув тизимида компютерлаштирилган файллар биринчи метаразметка ҳолатида базага (PostgreSQL) киритилади. Сўнгра матнли маълумотлар ва транскрипцион сўзлар тадқиқот гуруҳининг эксперт тилшунослари томонидан кузатилади ва таҳрир қилинади. Ўзбек тили миллий корпусига матн маълумотлари киритилаётганда батафсил текширилади.

Разметкаланган корпус сақланадиган маълумотлар базасининг тузилмаси 3-расмда тасвирланган.



3-расм. Маълумотлар базасининг тузилиши

Words жадвал корпусдаги сўзларни сақлайди. *Razmetkalash* жадвали корпусдаги сўзларнинг грамматик маълумотларни (ўзак, лемма ва разметкасини) ўз ичига олади. *Suz_turkumi* жадвали ўзбек тили сўз туркумларини ўз ичига олади ва сўзни разметкалаш жараёнида, унинг туркуми шу жадвалдан чақирилади. *Sub_suz_turkumi* жадвалида тилнинг грамматик маълумотларини ўз ичига олади. *Teg* жадвалида *Sub_suz_turkumi* грамматик маълумотларини ички грамматикасини ўз ичига олади. Сўзни разметкалаш жараёнида *razmetkalash* жадвалининг *razmetka* майдонига грамматик маълумотларни ёзиш жараёнида, даставвал *Suz_turkumi* жадвалига боғланиб, сўзнинг туркуми белгиланади. Ундан сўнг *Sub_suz_turkumi* жадвали билан боғланиб, бу жадвал *Teg* жадвалини чақиради ва *Sub_suz_turkumi* ҳамда *Teg* жадваллари фойдаланувчи учун танланган сўз туркумига мос равишда грамматик маълумотларни тақдим этади. Фойдаланувчи томонидан танланган грамматик маълумотлар *Razmetkalash* жадвалининг *razmetka* майдонига ёзилади. *Jsonfile* жалвали корпусдаги матнларни сақлайди. *Words_jsonfile* жадвали корпусдаги сўзларни, *Jsonfile* жалвалида жойлашган матнлардаги адреслар (ўрганилаётган сўз жойлашган матн ва ушбу матндаги учраш нуқтаси) харитаси ҳисобланади.

Корпус яратиш босқичлари. Корпусни яратиш жараёни бир неча босқичлардан иборат:

маълумотларни тўплаш: корпусга киритиладиган матнлар рўйхатини тузиш, уларнинг манбасини аниқлаш;

маълумотларни компьютерлаштириш: танланган маълумотларни (электрон матнлар, сканерланган матнлар, клавиатурадан терилган матн ва ҳ.к.) юклаш;

метамаълумотларни кодлаштириш: танланган матн ҳақидаги метамаълумотларни ёзиш;

маълумотларни разметкаш: матннинг ҳамма бирликларини теглаш;

корпусда қидирув интерфейсини ишлаб чиқиш: корпусдан маълумотлар олишни қулайлаштирадиган ва енгиллаштирадиган график веб-интерфейс яратиш;

корпусни ишга тушириш: дастлаб синов тариқасидаги версияни, сўнгра миллий ва халқора даражада фойдаланишга мўлжалланган биринчи версия ишини йўлга қўйиш.

тил корпуси нафақат қидирув тизими ёрдамида конкорданс тузиш, балки сўзларнинг морфологик маълумотларини тақдим этиши лозим. Разметкаланган матнлар JSON форматда корпус таркибига киритилади, бунда матн абзацлар, гаплар ва сўзлар каби таркибий бўлақларга бўлинади. Бунда ҳар бир сўзнинг ёнига, унинг морфологик маълумотлари ҳам бириктирилади.

Хулоса. Мавжуд корпуслар таҳлили асосида қуйидаги хулосалар олинди: корпусларни қуришда барча тиллар учун ягона бўлган методология мавжуд эмас;

морфологик характеристикалар тўплamlари турлича бўлишига қарамасдан, юқорида қараб чиқилган корпусларнинг барчасида морфологик (ёки грамматик) разметка таъминланган.

ҳар қандай йирик корпусда тўлиқ ва аниқ разметкалашнинг имконияти чекланган бўлиб, ноаниқликлар, асосан, автоматик разметкалашда омоним сўзлар туфайли пайдо бўлади;

миллий корпус тилнинг қонуният ва хусусиятларини иложи борида қўп тақдим эта олиши ва таркиби бўйича мувозанатлашган бўлиши керак;

матнларнинг бир тури тилдаги жами матнлар ичида қандай ҳажмий улушга эга бўлса, ана шундай миқдор миллий корпусда ҳам сақланиб қолиши керак.

Корпус яратишда турли хил разметканинг бўлиш ёки бўлмаслигини ва агар бўлса, унинг аниқлик даражасини ҳисоб олиш мақсадга мувофиқ. Корпус дастурий таъминоти таркибида разметкалаш дастурлари албатта бўлиши керак. Бу дастурлар қуйидаги сифатларга эга бўлиши керак:

- фойдаланувчи сўзлар ёки гапларнинг грамматик ёки синтаксис характеристикаларини етарли даражада тўлиқ аниқлаш ва ўзгартириш имкониятига эга бўлсин;

- таҳлил максимал даражада қулай ва содда бўлиб, тез бажарилсин ва бу таҳлилнинг тўлиқлигини камайтирмасин.



Фойдаланилган адабиётлар

1. Седов А.В., Математические модели, методы и алгоритмы построения размеченных корпусов текстов, дис. кан. наук, Петрозаводск, 2013.
2. А.Б.Қаршиев, С.А.Каримов, М.С.Турсунов, Development of a Modern Corpus of Computational Linguistics // Conference: 2020 International Conference on Information Science and Communications Technologies (ICISCT), DOI: 10.1109/ICISCT50599.2020.9351376, 2021.
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. Гвишиани Н.Б. Практикум по корпусной лингвистике: Учеб. пособие по английскому языку. М.: 2008.
5. Грудева Е.В. Корпусная лингвистика: Учебное пособие. 2-е изд., стереотип. М.: 2012.
6. Kennedy G., An introduction to corpus linguistics. London: Longman, 1998.