UDC 004.9

O‘ZBEK TILI MILLIY KORPUSI UCHUN MATNLARNI FORMATLASH

Qarshiyev A.B.[1], Karimov S.A.[2], Tursunov M.S.[2]

[1] Muhammad al Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti Samarqand filiali, Samarqand, O‘zbekiston

[2] Sharof Rashidov nomidagi Samarqand davlat universiteti, Samarqand, O’zbekiston

muhammadsolih927@gmail.com

Annotatsiya. *Ushbu maqolada o’zbek tili milliy korpusiga matnlarni kiritishda foydalanilgan usullarni tavsiflash va kodlashga umumiy yondashuv muhokama qilinadi. Umumiy format mavjud matn formatlarining xilma-xilligi va nomuvofiqligi bilan asoslanishi mumkin. Korpusda matnlarni saqlash uchun JSON formatdan foydalanish orqali korpus qidiruv tezligini oshirish va kengayuvchanlikdagi nazariy va texnik muammolarni bartaraf etish mumkin. Korpusga Alpomish dostoning matnlari kiritilishi tavsiflangan.*

Kalit so‘zlar: *korpus, formatlash, fayl, matn, Alpomish dostoni, token, razmetka, teg, tegger, JSON format, DOCX format.*

I. KIRISH

Zamonaviy korpuslar – bu ma’lum bir tilda, elektron shakldagi matnlar to‘plamiga asoslangan axborot-ma’lumot tizimidir. Har bir korpus filologik yondashuv bilan matnlar ustida ishlash uchun, albatta, lingvistik apparat va dasturiy ta’minot bilan ta’minlanishi lozim.

Bugungi kunda korpuslar lug‘atlar va grammatika kabi tilshunoslikning ajralmas qismiga aylandi. Korpus paydo bo‘lganidan so‘ng tilshunoslik fanlari o‘zgarib ketdi, aytish mumkinki, butun tilshunoslik korpus tilshunosliliga aylandi. Eng taniqli va tan olingan lingvistik korpuslarga namuna sifatida quyidagilarni keltirish mumkin: Rus milliy korpusi (https://ruscorpora.ru/new/), Britaniya milliy korpusi (http://www.natcorp.ox.ac.uk/, https://www.english-corpora.org/bnc/), Turk milliy korpusi (https://www.tnc.org.tr/), Amerika milliy korpusi (http://www.anc.org) va boshqalar[1].

O‘zbek tili uchun milliy korpuslarni yaratish va u orqali tilni tadqiq qilish dolzarb masala bo‘lib turibdi. Ushbu maqolada biz ishlab chiqayotgan (https://uzbekcorpora.uz) platformasiga korpus tarkibiga matnlarni kiritish qoidalari va sabablari tushuntirilgan.

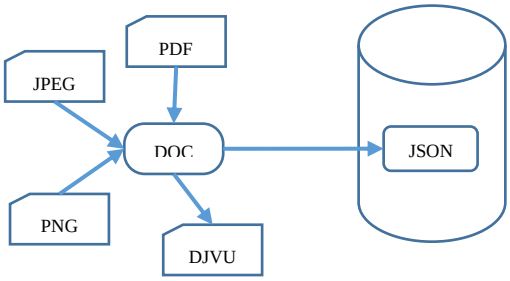
II. ASOSIY QISM

Raqamli va raqamli bo‘lmagan matnlar korpus uchun matn manbalari sifatida ishlatilishi mumkin. Tabiiyki, ikkinchi holatda, qandaydir tarzda matnni kompyuterga kiritish kerak bo‘ladi: uni qayta yozish yoki skanerlalash kerak. Misol uchun, Alpomish dostonining Fozil Yo’ldosh variantining *DOCX* formati bizda mavjud emas. Shuning uchun, bu kitobni raqamlashtirishimiz kerak bo‘ladi. Bunda qo‘lyozma manbani yoki kitobni skanerlash jarayonida *PDF* formatga

o‘tkaziladi va undagi yozuvlarni tanib oluvchi konvertor (masalan, Fine Reader) dasturlar yordamida Microsoft Office Word dasturida ochadigan *DOCX* formatga o‘tkaziladi. Bunda albatta konvertor skaner qilingan elektron kitobni *DOCX* formatga o‘tkazishda kitobning asl holati bo‘yicha o‘tkazishi juda qiyin. Shuning uchun qo‘l mehnati asosida *DOCX* formatdagi matn, asl holatdagi matn bilan birxillikka keltiriladi. Bunda imloviy buzilishlar, matndagi belgilarni noto‘g‘ri tanib olish kabi xatoliklar tuzatiladi.

Matnlar turli xil PDF, rasmli, dokumentli va boshqa formatlarda bo‘ladi. Korpusga matnlarni kiritishdan avval, mavjud matnli fayllarni Microsoft Office ning 2010 yil va undan yuqori bo’lgan versiyasidagi *.docx formatiga o‘tkazish kerak bo‘ladi. Boshqa formatdagi matnlarni *.docx formatiga maxsus dasturlar yordamida o‘tkaziladi va *.docx formatiga o‘tkazish jarayonida matnning asl holati buzilishi mumkin. Bunda matndagi imloviy xatolar qo‘l mehnati yordamida matnning asl holati bilan bir xillikka keltiriladi. Undan so’ng matnni korpusga yuklash mumkin bo’ladi. Ushbu tadqiqotda korpusda matnlarni saqlash uchun JSON formatdan foydalanilgan (1-rasm).

Lekin deyarli barcha korpuslarda XML formatdagi matnlar saqlanadi. Xuchen Yao o’zning maqolasida XML format uchun quyidagi fikrlarni bildirgan. “Uning tilshunoslik sohasidagi vazifalarga qo‘shgan hissasi asosan standart va moslashuvchan tilning afzalliklariga asoslanadi. XML aniq va izchil ma’lumotlar strukturasini taklif etadi: tahlil qilingan ma’lumotlar bilan ishlash tajribasiga ega bo‘lmagan yangi boshlanuvchilar uchun uni o‘rganish juda oddiy. Moslashuvchan ma’lumotlarni tashkil etish foydalanuvchiga o‘z ehtiyojlariga qarab hujjat tuzilishini qo‘shish va aniqlash imkonini beradi. Nihoyat, uning izchil kodlash formati turli korpuslarning kombinatsiyasini va bundan tashqari, yagona so‘rovlar tilini osonlashtiradi” [2].



1-rasm. Matnlarni korpusga saqlash formati

Biz tadqiqot natijalarimizga ko‘ra korpusda matnlarni saqlash uchun XML formatdan ko’ra JSON formatda saqlashni tavsiya beramiz. JSON va XML formatlarni umumiy jihatlari va farqli jihatlari bor.

XML formatning JSON formatga o‘xshash tomonlari:

- JSON ham, XML ham "o‘zini tavsiflovchi" (odam tomonidan o‘qilishi mumkin);
- JSON ham, XML ham ierarxik (qiymatlar ichidagi qiymatlar)
- JSON ham, XML ham ko‘plab dasturlash tillari tomonidan tahlil qilinishi va ishlatilishi mumkin
- JSON ham, XML ham XMLHttpRequest bilan olinishi mumkin [2].

JSON formatning XML formatdan afzallik tomonlari:

- JSON yakuniy tegdan foydalanmaydi;
- JSON qisqaroq;

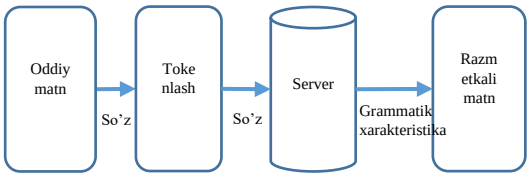
- JSON tezroq o‘qish va yozish;
- JSON massivlardan foydalanishi mumkin.

JSON bilan XML ning eng katta farqi shundaki XMLni XML parser yordamida tahlil qilish kerak. JSON standart JavaScript funksiyasi tomonidan tahlil qilinishi mumkin [4].

Kompyuter lingvistikasi, korpus lingvistikasi, tabiiy tilni qayta ishlash tadqiqtolarni takomillashtirish uchun JSON formatdan foydalanilayotgan bir qancha istiqbolli loyihalarni ko’rish mumkin:

- Hieber, Daniellar tomonidan raqamli lingvistikada lingvistik resurslar va matnlar uchun JSON asosini yaratish bo‘yicha keng qamrovli loyiha amalga oshirilmoqda [5].
- JSON uchun sintaksis va semantika taqdimotlari uchun Kaliforniya universiteti tadqiqotchilari Angus G., Forbeslar tomonidan ham ilmiy tadqiqotlar olib bormoqda [6]. Ushbu JSON yondashuvi haqida ma’lumotni GitHubdagi Angus Forbes Creative Coding Labdan olish mumkin [7].
- Spacydan Ines Montani lingvistik ma’lumotlar/NLP uchun JSON formatlashning yangi yondashuvi bo‘yicha qiziqarli taklifni taqdim etadi [8].
- Stenford NLP JSON formatlagichini tavsiflaydi [9].
- Rezonator loyihasida (https://github.com/johnwdubois/rezonat) Santa Barbara Corpus of Spoken American English (SBCSAE) 2-nashri uchun JSON formatini yaratish ustida ilmiy ishlar bajarilmoqda.

Matn korpusga kiritishga tayyor holatga keltirilganda, uni korpusga yuklash mumkin. Korpusga matn yuklanish jarayonida DOCX formatdagi matn JSON formatga almashtiriladi va bir vaqtning o‘zida matn ham tokenlanadi, ham har bir so‘zga grammatik ma’lumotlar biriktiriladi. Grammatik ma’lumotlarni har bir so‘zga biriktirishda dastur serverga murojaat qiladi. Serverda so‘zlar lug‘ati mavjud bo‘lib, har so‘zning grammatik harakteristakalari oldindan berilgan. Bunda matnni razmetkali matnga o‘tkazishda ikki bosqichdan o‘tadi. Birinchi bosqichda matn tokenlanadi, ikkinchi bosqichda token-langan matn razmetkalanadi. Razmet-kalash jarayonida dastur serverga so‘zni yuboradi va server shu so‘zni Grammatik harakteristikasini dasturga javob tariqasida qaytaradi va oddiy matnni razmetkalangan matn holatiga o‘tkazadi (2-rasm).



2-rasm. Oddiy matnni razmetkali matnga o’tkazish

Bunda matndagi so‘z razmetkalangan hisoblanadi va so‘zning razmetkasi kesh xotirada saqlanadi. Qachonki so‘zning ustidan bosilganda, razmetka kesh xotiradan foydalanuvchiga taqdim etiladi. Matn so‘zlarga ajratilganda, *n* iborat bo‘lsa dastur har bir tokenga Grammatik harakteristikasini berish uchun serverga 10 000 marta murojaat qiladi. Bu albatta dasturning ishlash tezligiga ta’sir ko‘rsatadi. Bunda internet tezligini ham hisobga olish zarur. Shu sababli korpusga matnni kiritishda agar uning hajmi katta bo‘lsa matnni bo‘laklarga ajratish tavsiya etiladi. Alpomish dostonining Fozil Yo‘ldosh varianti 2010-yilda “sharq”

nashriyoti tomonidan chop etilgan. Kitob skanerlash dasturi yordamida elektron (PDF) variant yaratildi. Undan so‘ng *FineReader* dasturi yordamida *.docx formatga o‘tkazildi. Bunda matn imloviy xatoliklar tuzatilib, matnning asl holatiga mos holatga keltirildi. Matn 372 sahifadan iborat. Uni 10 sahifadan 37 bo‘lakga ajtarildi. Matnlar korpusga joylashtirildi. Korpusga matn kiritilganda dastavval matnni tokenlaydi undan so‘ng serverga murojaat qilib har bir so‘zga Grammatik xarakteristikani biriktiradi. Korpusda matnlar *PostGreSql* ma’lumotlar bazasida *.JSON formatida teglangan holatda bo‘ladi. Misol tariqasida Alpomish dostonining dastlabki bitta jumlasini olaylik va u korpusga kiritiladi. Dastlab matn tokenlanadi (3-rasmning a) qismi). Bunda har bir so‘z teglanadi va bunda asosiy teglarning (*uzak*, *lemma*, *izoh*, *turkum*, *razmetka*) qiymati bo‘sh bo‘ladi. Keyingi qadamda so‘zlar avtomatik tarzda razmetkalanadi (3-rasmning b) qismi).

Razmetkalangan matn o‘z o‘rnida teglardan iborat bo‘ladi. [] – tashqi to‘rtburchak qavslar matnning boshlanishi va oxirini anglatadi, ichki to‘rtburchak qavslar abzasning boshlanishi va oxirini bildiradi; {} – figurali qavslar bitta so‘zning tarkibini bildiradi; “” – qo‘shtirnoq belgisining ichida har bir tegning nomi va uning qiymati ko‘rsatiladi. : - ikki nuqta belgisi teg va uning qiymatini ajratib turadi.

Teglar quyidagilardan iborat:

- “word” – bu teg matndagi so‘z.
- “index_col” – bu teg matnda so‘zning satrda joylashish o‘rni;
- “index_row” – bu teg matnda so‘zning ustunda joylashish o‘rni;

“class” – bu teg so‘zga Grammatik xarakteristika berilgan yoki yo‘qligini anglatadi. Agar so‘zga Grammatik xarakteristika berilmagan bo‘lsa uning qiymati “link-black”, aks holda “link-red” bo‘ladi;

3-rasm. Oddiy matnni razmetkali matnga o‘tgan holati

- “uzak” – bu tegga so‘zning asosi yoziladi;

- “lemma” – bu tegga so‘zning lemmasi yoziladi;
- “izoh” – bu tegda so‘zning izohi keltiriladi;
- “turkum” – bu tegda so‘zning turkumi keltiriladi. Bunda turkumlar o‘n ikkitaga ajratilgan. Teg qiymatni turkum nomi bilan emas, balki uning tartib nomeri bilan oladi. Bunda 0-fe’l, 1-ot, 2-son, 3-sifat, 4-ravish, 5-olmosh, 6-yuklama, 7-ko‘makchi, 8-bog‘lovchi, 9-modal so‘z, 10-taqlid so‘z, 11-undov so‘z;
- “razmetka” – bu teg so‘zning Grammatik xarakteristikasini qiymat sifatida oladi.

III. XULOSA VA TAVSIYALAR

O‘zbek tili milliy korpusiga matnlarni kiritish uchun maxsus dasturiy ta’minot ishlab chiqildi. Dasturiy ta’minot yordamida matnlarni korpusda saqlash uchun JSON format tanlangan. Bu formatdagi matnlar korpus tarkibida oddiy fayllar ko‘rinishida emas, ma’lumotlar bazasida saqlanadi. Bu esa korpus matnlari ustida amallar bajarilganda yoki so‘rovlar yuborilganda dasturning ishlash sifatiga ta’sir ko‘rsatmaydi.

ADABIYOTLAR

[1] *Qarshiyev A.B., Tursunov M.S., Maxmidov Sh.B.,* “O‘zbek tili milliy korpusini loyihalash”, “Kompyuter lingvistikasi: muammolar, yechim, istiqbollar” mavzusidagi xalqaro ilmiy-amaliy konferensiya materiallari, Toshkent: ToshDO‘TAU, 22.04.2022, Vol. 1 № 01 (2022), 82-88 betlar.

[2] *Xuchen Yao, Irina Borisova, Mehwish Alam,* PDTB XML: the XMLization of the Penn Discourse TreeBank 2.0.

[3] *Tobias Weisskopf,* In Digital Linguistics / Computational Linguistics, why is XML the preferred corpus format and not JSON?, ResearchGate, 2020.

[4] *https://www.w3schools.com/js/js_json_xml.asp*.

[5] *Hieber, Daniel W.* 2020. Data Format for Digital Linguistics. DOI:10.5281/zenodo.1438589.

[6] *Forbes, Angus G., Lee, Kristine, Hahn-Powell, Gus, Valenzuela-Escárcega, Marco A. & Surdeanu, Mihai.* Text Annotation Graphs: Annotating complex natural language phenomena, 2018. https://www.aclweb.org/anthology/L18-1169.

[7] *https://github.com/CreativeCodingLab/TextAnnotationGraphs*.

[8] *https://github.com/explosion/spaCy/issues/2928*.

[9] *https://nlp.stanford.edu/nlp/javadoc/javanlp/edu/stanford/nlp/pipeline/JSONOutputter.html*.

Поступила в редакцию 7.09.2022

Citation: *Qarshiyev A.B., Karimov S.A., Tursunov M.S.* O‘zbek tili milliy korpusi uchun matnlarni formatlash. // Raqamli texnologiyalarning nazariy va amaliy masalalari xalqaro jurnali. – 2022. – № 1(1). – B. 58-63.

FORMATTING TEXTS FOR THE NATIONAL CORPUS OF THE UZBEK LANGUAGE

Karshiev A.B.[1], *Karimov S.A.*[2], *Tursunov M.S.*[2]
[1] Samarkand branch of Tashkent University of information technologies named after Muhammad al-Khwarizmi, Samarkand, Uzbekistan
[2] Samarkand State University named after Sharof Rashidov, Samarkand, Uzbekistan
muhammadsolih927@gmail.com

Abstract. *This article discusses the general approach to the description and coding of the methods used in the inclusion of texts in the national corpus of the Uzbek language. A common format can be justified by the diversity and incompatibility of existing text formats. By using the JSON format to store texts in the corpus, it is possible to increase corpus search speed and overcome theoretical and technical problems of scalability. The inclusion of the texts of the Alpomish epic into the corpus is described.*

Keywords: *Corpus, formatting, file, text, Alpomish epic, token, markup, tag, tagger, JSON format, DOCX format.*

ОФОРМЛЕНИЕ ТЕКСТОВ ДЛЯ НАЦИОНАЛЬНОГО КОРПУСА УЗБЕКСКОГО ЯЗЫКА

Каршиев А.Б.[1], *Каримов С.А.*[2], *Турсунов М.С.*[2]
[1]Самаркандский филиал Ташкентского университета информационных технологий имени Мухаммада ал-Хорезми, Самарканд, Узбекистан
[2] Самаркандский государственный университет имени Шарофа Рашидова, Самарканд, Узбекистан
muhammadsolih927@gmail.com

Аннотация. *В данной статье рассматривается общий подход к описанию и кодированию методов, используемых при включении текстов в национальный корпус узбекского языка. Общий формат может быть оправдан разнообразием и несовместимостью существующих текстовых форматов. Используя формат JSON для хранения текстов в корпусе, можно увеличить скорость поиска в корпусе и преодолеть теоретические и технические проблемы масштабируемости. Описано включение в состав корпуса текстов эпоса «Алпомыш».*

Ключевые слова: *Корпус, форматирование, файл, текст, Алпомишский эпос, токен, разметка, тег, теггер, формат JSON, формат DOCX.*