

**YETAKCHI JAHONKORPUSLARI VA RAAMLII ADABIYOT LOYIHALARIDA
BELGILASH METODOLOGIYASI: QIYOSIY TAHLIL (BNC, COCA, RUS MILLIY
KORPUSI, DTA, OPENITI)**

Aliqulova Mohinur O`ktam qizi

Toshkent Amaliy fanlar universiteti, Xorijiy til va adabiyoti o`qituvchisi

E-mail: mohinuraliqulova96@gmail.com, <https://orcid.org/0009-0009-8659-6346>

Turg`inova Lobar Shavkat qizi

Toshkent Amaliy fanlar universiteti, filologiya va tillarni o`qitish ingliz tili ta`lim yo`nalishi 1-kurs
bakalavr talabasi

E-mail: lobarturginova206@gmail.com

Annotatsiya

Tadqiqotda jahon va turkiy tillar korpuslari belgilash metodologiyasi qiyosiy tahlil etilib, o'zbek filologiyasini raqamlashtirishning *integrativ qatlamli modeli* asoslangan. O'zbek tili agglutinatib tabiati va uch alifboli muhitini hisobga olgan holda, Universal Dependencies va TEI P5 standartlarini moslashtirish, klassik matnlarni ko'p qatlamli belgilash hamda avtomatlashtirilgan morfologik tahlilni joriy etish bo'yicha aniq texnologik yechimlar taklif qilingan.

Аннотация

В исследовании проведен сравнительный анализ методологии разметки мировых и тюркских корпусов, на основе которого разработана *интегративная многоуровневая модель* цифровизации узбекского языка. С учетом агглютинативной природы и многоалфавитной среды предложены технологические решения по адаптации стандартов Universal Dependencies и TEI P5, внедрению многоуровневой поэтической разметки классических текстов и автоматического морфологического анализа.

Abstract

This study presents a comparative analysis of annotation methodologies in global and Turkic corpora to establish an *integrative layered model* for digital Uzbek philology. Considering the language's agglutinative nature and multi-script environment, concrete technological solutions are proposed for adapting Universal Dependencies and TEI P5 standards, implementing multi-layered poetic annotation for classical texts, and integrating automated morphological analysis.

Kirish

Xalqaro belgilash amaliyotini o'zlashtirish faqat nazariy standartlar tahlili bilan cheklanmay, real loyihalardagi texnik yechimlarda namoyon bo'ladi. Shu sababli, jahon tilshunosligidagi yirik milliy va ko'p tilli korpuslarning belgilash metodologiyasini qiyosiy tadqiq etish o'zbek filologiyasining amaliy modelini yaratishda metodologik zaruratdir [2, b-250]. Hozirgi kunda o'zbek tili uchun ilmiy standartlarga javob beradigan to'liq milliy korpusning mavjud emasligi ushbu qiyosiy tahlilning amaliy va dolzarb ahamiyatini yanada oshiradi [7, b-160].

BNC - ingliz tilining 100 million so'zlik ushbu matn korpusi Oxford University Press, Longman, Lancaster University va British Library hamkorligida yaratilgan. Korpus tarkibining

90%ini yozma, 10%ini ogʻzaki nutq tashkil etadi; unda 627 ta janr va 4124 ta matn jamlangan. Belgilash jarayoni CLAWS4 vositasi orqali avtomatik POS usulida (137 ta teg) bajarilib, keyinchalik qoʻlda tekshirilgan [5,b-500].

COCA (Corpus of Contemporary American English): Loyihaning asosiy innovatsiyasi - real vaqtda til oʻzgarishlarini kuzatish imkonini beruvchi doimiy yangilanib turadigan “jonli” korpus modelidir. Ushbu yondashuv soʻnggi oʻttiz yilda kirill-lotin alifbosi almashinuvi va yangi leksika hisobiga jiddiy oʻzgargan oʻzbek adabiy tili uchun juda dolzarbdir [3,b-348].

Rus milliy korpusi - Rossiya Fanlar akademiyasi tomonidan yaratilgan 600 milliondan ortiq soʻz hajmiga ega ushbu resurs koʻp qatlamli lingvistik belgilash tizimiga ega. Uning gibrid tizimida avtomatik morfologik tahlil ekspert nazorati bilan birlashtiriladi [7,b-160]. RMK sheʼriy korpusida (10 mln soʻz) sonet, triolet, oda kabi janrlar uchun maxsus metrik belgilash tizimi joriy etilgan [8,c-62–88]. Biroq, slavyan morfologiyasiga moslashgan ushbu teglar tizimi turkiy tillarning agglutinatив tabiatini aks ettira olmaydi [7, c- 160 .].

Deutsches Textarchiv - Berlin-Brandenburg Fanlar akademiyasi tomonidan rivojlantirilayotgan, 1650–1900-yillar nemis klassik adabiyotini qamrab olgan raqamli arxiv (160 mln soʻz). TEI-P5 standartiga asoslangan DTA tarixiy matnlardagi eski imloni normallashtirish, yoʻqolgan qismlarni tiklash va qoʻlyozma variantlarini tahlil qilish protokollariga ega [1,b-338]. Bu tajriba Alisher Navoiy asarlarining 200 dan ortiq nusxalari, imlo tafovutlari va matniy nuqsonlarini raqamlashtirishda oʻzbek klassik adabiyoti uchun muhim andoza boʻla oladi [1,b-338].

OpenITI - Arab va fors klassik adabiyoti raqamli arxivi boʻlib (1,8 mln sahifa), morfologik, intertekstual va janriy metamaʼlumot qatlamlarini qamrab oladi [4, 180]. OpenITI tarkibidagi tazmin, muxammas, taʼxir va naqsh kabi anʼanaviy belgilash kategoriyalari oʻzbek klassik sheʼriyatidagi konseptual hodisalarni korpuslashtirishda bevosita asos boʻladi. Gʻarb va Sharq korpus anʼanalari oʻzbek filologiyasi uchun metodologik zamin boʻlsa-da, eng yaqin qiyosiy material sifatida oʻzbek tili bilan tipologikdosh boʻlgan turkiy tillar korpuslari va treebanklari tajribasini oʻrganish zarur. Agglutinatив morfologiya, harakat feʼllari, ergash gap konstruksiyalari va zanjirli tokenizatsiya kabi umumiy muammolarni hal qilishda turkiy tilshunoslikning toʻrtta yetakchi loyihasi - Turk BOUN Treebank, Qozoq milliy korpusi, TüBa-D/Z va Apertium platformasi modullari qiyosiy tahlil qilinadi.

Turk BOUN Treebank: 2018-yildan keyingi davrda yaratilgan ushbu resurs Universal Dependencies tizimiga toʻliq integratsiyalashgan birinchi turk treebankidir [6,b-259–307]. Korpus 9.761 ta gap va 125 mingdan ortiq tokenni qamrab olib, ilmiy, biografik va jurnalistik matnlari ekspertlar tomonidan qayta tahrirlangan. Loyihaning muhim innovatsiyasi - agglutinatив morfologiyani UD sxemasida ifodalash protokolining yaratilishidir.

Qozoq milliy korpusi - A. Baytursinov nomidagi Til bilimi institutida yaratilgan va 40 milliondan ortiq tokenni qamrab olgan multifunksional resursdir. Unda zamonaviy adabiy, OAV, ilmiy va folklor matnlari jamlangan. Morfologik belgilash ikki bosqichda amalga oshiriladi: avval finite-state transducer vositasida avtomatik tahlil qilinadi, soʻngra ekspert tilshunoslar tomonidan tekshiriladi. Eʼtiborli jihati, KNC koʻp alifboli (kirill, lotin, arab) moslikka ega boʻlib, matnlarni parallel ravishda taqdim eta oladi.

Yetakchi xalqaro korpuslar va raqamli adabiyot loyihalarining qiyosiy tahlili

Standart / Loyiha	Ishlab chiqaruvchi	Qo'llanish sohasi	Belgilash darajasi	O'zbek uchun asosiy cheklov
TEI P5	TEI Consortium (xalqaro)	Adabiy matnlar, qo'lyozmalar	Morfologik, semantik, metama'lumot	Agglutinativ til va aruz kategoriyalari uchun standart modul yo'q
Universal Dependencies (UD)	UD Consortium (xalqaro)	Ko'p tilli sintaktik tahlil	Morfologik, sintaktik (dependency)	O'zbek agglutinativ morfologiyasi uchun maxsus qoidanoma yo'q
BNC	Oxford UP va boshq.	Ingliz tili (yozma+og'zaki)	POS (137 teg), janr metama'lumot	Bir tilli; agglutinativ tokenizatsiya yechimi yo'q
COCA	BYU (Mark Devis)	Amerika ingliz tili (dinamik)	POS, lemmatizatsiya, 8 janr qatlami	Bir tilli; ko'p alifboli o'zbek matni uchun protokol yo'q
RMK	IRL RAN (Rossiya)	Rus badiiy va ilmiy matni	Morfologik, sintaktik, semantik, she'riy	Slavyan morfologiyasiga yo'naltirilgan; turkiy til kategoriyalari yo'q
DTA	Berlin BBAW (Germaniya)	Nemis klassik adabiyoti XVII–XIX	TEI-P5: fonetik, morfologik, apparatus criticus	Yevropa adabiy an'anasi uchun; sharq poetikasi yo'q
OpenITI	KITAB (Harvard/Maryland)	Arab va fors klassik matni	Morfologik, intertekstual, janriy	Arabiy yozuv asosida; kirill/lotin versiyalar uchun adaptatsiya zarur

Turkiy tillar korpuslari va platformalari qiyosiy tahlili o'zbek filologiyasini raqamlashtirish loyihasi va taklif etilayotgan integrativ qatlamli model uchun quyidagi tayyor texnologik yechimlarni asoslab beradi:

1. Morfosintaktik standart: Agglutinativ morfologiyani *Universal Dependencies* sxemasida aks ettirishda Turkiyaning BOUN andozasini o'zbek tiliga moslashtirish.

2. Ko'p alifboli muvofiqlik: O'zbek tilining uch alifboli muhitini tartibga solishda Qozog'istonning KNC (ALT-FORM) protokolidan foydalanish.

3. Ko'p qatlamli arxitektura: O'zbek klassik adabiyoti matnlarida aruz, qofiya va sintaktik qatlamlarni TüBa-D/Z parallelizm tamoyili asosida alohida qayd etish.

3. Morfologik tahlil: O'zbek korpusining yarim avtomatik tahlil bosqichini Apertium platformasining turkiy *FST* (chekli holatli datchiklar) infratuzilmasi orqali texnologik ta'minlash. Turkiy tillar korpus tajribasi G'arb, Sharq va slavyan modellari bilan bir qatorda, o'zbek raqamli korpusi arxitekturasining amaliy va tipologik fundamentini tashkil etadi.

Foydalanilgan adabiyotlar.

1. Driscoll M. J., Pierazzo E. Digital Scholarly Editing: Theories and Practices. — Cambridge : Open Book Publishers, 2016. — 338 p.
2. McEnery T., Brezina V. Corpus Linguistics: Method, Theory and Practice. — Cambridge : Cambridge University Press, 2022. — 250 p.
3. Mengliyev B. O'zbek tilining amaliy grammatikasi. — Toshkent : Navro'z, 2018. — 248 b.

4. O'Sullivan J. (ed.) Digital Humanities in the Library: Challenges and Opportunities. — Chicago : ACRL, 2020. — 180 p.
5. Stefanowitsch A. Corpus Linguistics: A Guide to the Methodology. — Berlin : Language Science Press, 2020. — 500 p.
6. Türk U., Atmaca F., Özateş Ş. B., Köksal A., Öztürk Başaran B., Güngör T., Özgür A. Resources for Turkish dependency parsing: introducing the BOUN Treebank and the BoAT annotation tool // Language Resources and Evaluation. — 2022. — Vol. 56, № 1. — P. 259–307.
7. Захаров В. П., Bogdanova С. Ю. Корпусная лингвистика. — Иркутск : ИГЛУ, 2011. — 160 с.
8. Савчук С. О. Метатекстовая разметка в Национальном корпусе русского языка // Национальный корпус русского языка: 2003–2005. — Москва : Индрик, 2005. — С. 62–88.