

O‘ZBEK TILIDA FE’LNING VAZIFA SHAKLLARI KORPUSINI YARATISH VA ANNOTATSIYALASH TAMOYILLARI

To‘xtayev Shuhrat Furqatovich
ORCID:0009-0005-3807-0498
shuhrattoxtayev22@gmail.com
Buxoro viloyat Kogon shahar IM

Annotatsiya: Mazkur maqolada o‘zbek tilida fe’lning vazifa shakllari - sifatdosh, ravishdosh va harakat nomi - asosida maxsus lingvistik korpus yaratish hamda ushbu birliklarni morfologik, grammatik va sintaktik jihatdan annotatsiyalash tamoyillari tadqiq etiladi. Tadqiqotning nazariy-metodologik asosini o‘zbek kompyuter lingvistikasi sohasida amalga oshirilgan korpusshunoslik, morfologik analiz, avtomatik tahlil va lingvistik annotatsiyaga oid ilmiy izlanishlar tashkil etadi. Maqolada o‘zbek tilining agglutinatив tabiati sababli fe’lning vazifa shakllarini faqat tashqi affiksial ko‘rsatkich asosida aniqlash yetarli emasligi, bunda morfologik, sintaktik va kontekstual belgilarni kompleks hisobga olish zarurligi asoslanadi. Shuningdek, sifatdosh, ravishdosh va harakat nomini korpusda belgilash uchun ko‘p qatlamli annotatsiya modeli taklif qilinadi.

Kalit so‘zlar: o‘zbek tili, kompyuter lingvistikasi, korpus lingvistikasi, lingvistik annotatsiya, PoS-tegging, morfologik analizator, fe’l, vazifa shakllari, sifatdosh, ravishdosh, harakat nomi, NLP, korpus

PRINCIPLES OF CREATING AND ANNOTATING A CORPUS OF VERBAL FUNCTIONAL FORMS IN THE UZBEK LANGUAGE

To‘xtayev Shuhrat Furqatovich
ORCID: 0009-0005-3807-0498
shuhrattoxtayev22@gmail.com
Kogon City Specialized School

Abstract: This article examines the development of a specialized linguistic corpus based on functional forms of the verb in the Uzbek language - participles, converbs, and action nouns - as well as the principles of morphological, grammatical, and syntactic annotation of these units. The theoretical and methodological framework of the study is based on research in Uzbek computational linguistics concerning corpus linguistics, morphological analysis, automatic processing, and linguistic annotation. The article substantiates that, due to the agglutinative nature of the Uzbek language, identifying functional forms of the verb solely on the basis of external affixal markers

is insufficient. Therefore, morphological, syntactic, and contextual features need to be taken into account in an integrated manner. In addition, a multilayer annotation model is proposed for the identification and annotation of participles, converbs, and action nouns in the corpus.

Keywords: Uzbek language, computational linguistics, corpus linguistics, linguistic annotation, PoS tagging, morphological analyzer, verb, functional forms of the verb, participle, converb, action noun, NLP, corpus

Kirish. Bugungi kunda kompyuter lingvistikasi til hodisalarini matematik va algoritmik modellashtirish, katta hajmdagi elektron matnlarni qayta ishlash hamda tabiiy tilni avtomatik tahlil qilishga yo'naltirilgan fanlararo soha sifatida jadal rivojlanmoqda. Tabiiy tilni qayta ishlash texnologiyalarining takomillashuvi esa, avvalo, muayyan tilning grammatik qurilishini raqamli shaklda ifodalovchi sifatli lingvistik resurslarni yaratishni taqozo etadi. Shu ma'noda korpus, morfologik analizator, so'z turkumlarini avtomatik belgilash tizimi va sintaktik analizator bir-biri bilan uzviy bog'langan resurslar sifatida namoyon bo'ladi.

O'zbek kompyuter lingvistikasining shakllanishida korpus lingvistikasi alohida o'rin egallaydi. B. Elov, M. Abjalova va R. Alayev tomonidan o'zbek tili korpusi va uning imkoniyatlariga bag'ishlangan tadqiqotda korpusning til birliklarini real kontekstda o'rganish, statistik ma'lumotlarni olish va turli lingvistik tadqiqotlarni avtomatlashtirishdagi ahamiyati ko'rsatiladi [7]. O'zbek tilining ta'limiy korpusini shakllantirish borasidagi tadqiqotlarda ham korpusning lingvistik va dasturiy ta'minotini yaratish, matnlarni teglar bilan belgilash, qidiruv va tahlil imkoniyatlarini kengaytirish masalalariga e'tibor qaratilgan [8]. Bu jarayonda korpusni oddiy elektron matnlar to'plami sifatida emas, balki lingvistik jihatdan qayta ishlangan ma'lumotlar bazasi sifatida tushunish muhimdir. M. Abjalova lingvistik annotatsiyani korpus birliklariga maxsus identifikatorlar - teglar berish orqali ularni lingvistik jihatdan tavsiflash jarayoni sifatida ko'radi [6]. Bunday yondashuv bizning tadqiqotimiz uchun muhim metodologik asosdir. Chunki fe'lining vazifa shakllari korpusida *kelgan*, *kelib*, *kelish* kabi birliklarning mavjudligi yetarli emas; ularning qanday grammatik shaklga mansubligi va gapda qanday vazifa bajarayotgani ham mashina o'qiy oladigan shaklda qayd etilishi lozim.

Tadqiqot metodologiyasi. O'zbek tilining agglutinatив xususiyati korpus annotatsiyasiga alohida talablar qo'yadi. Agglutinatив tillarda grammatik ma'nolar, asosan, affikslar ketma-ketligi orqali ifodalanadi. Shu sababli bitta so'z shakli tarkibida bir necha grammatik ma'no mujassamlashishi mumkin. B. Mengliyev va Sh. Hamroyeva turkiy tillardagi morfologik analizatorlarni tahlil qilib, o'zbek tilining agglutinatив tabiati grammatik shakllarning ko'pligi va ularni avtomatik qayta ishlashda yagona, izchil grammatik tavsif tizimining zarurligini ko'rsatadilar [9].

Ularning fikricha, turkiy tillar uchun yaratilgan morfologik analizatorlar tajribasidan foydalanish va o'zbek tili analizatorini yaratishda ikki bosqichli morfologik tahlil nazariyasiga tayanish maqsadga muvofiq [9].

Bu qarash fe'l vazifa shakllari korpusini yaratishda ham muhimdir. Masalan: *kelmaganlaringizdan* shaklini faqat bitta leksik birlik sifatida qayd etish uning ichki grammatik tuzilishini ochib bermaydi. Korpusda: *kel-ma-gan-lar-ingiz-dan* tarzidagi morfemik segmentatsiya ham saqlanishi mumkin. Bunda: *kel-* - *fe'l negizi*; *-ma-* - *bo'lishsizlik*; *-gan* - *sifatdosh ko'rsatkichi*; *-lar* - *ko'plik*; *-ingiz* - *egalik*; *-dan* - *chiqish kelishigi*. Demak, fe'l vazifa shakllari korpusida morfemik va morfologik qatlamni ajratish muhim metodologik tamoyil bo'ladi.

Mualliflar morfologik analiz jarayonida token, lemmatizatsiya, stemming, affikslarni aniqlash va morfologik generator kabi tushunchalarning o'zaro aloqasiga e'tibor qaratadilar [2]. Bu qarashdan kelib chiqib, fe'l vazifa shakllari korpusini yaratishda quyidagi ketma-ketlikni taklif etish mumkin: tokenizatsiya → morfemik segmentatsiya → lemma aniqlash → grammatik tahlil → vazifa shaklini aniqlash → sintaktik annotatsiya. Masalan: *o'qigan* → token: *o'qigan* → lemma: *o'qi-* → negiz: *o'qi-* → vazifa shakli: *sifatdosh* → grammatik belgi: *-gan* → sintaktik funksiya: *aniqllovchi*. Yuqoridagi modelda vazifa shaklini aniqlash morfologik analizning alohida bosqichi sifatida qaraladi.

PoS-tegging va fe'lining vazifa shakllari. O'zbek tilida PoS-tegging masalasiga bag'ishlangan B. Elov, Sh. Hamroyeva, O. Abdullayeva va M. Uzoqova tadqiqotida grammatik teglash o'zbek korpus lingvistikasi uchun muhim va murakkab masala sifatida baholanadi [1]. Mualliflar ta'kidlaganidek, korpusdagi lingvistik birliklarga mos identifikatorlarni belgilash tilning ichki qonuniyatlarini hisobga oluvchi teglash qoidalarini ishlab chiqishni talab qiladi [1]. Mazkur fikr fe'l vazifa shakllarini annotatsiyalash uchun bevosita ahamiyatlidir. Chunki *-gan*, *-ib*, *-ish* kabi shakl ko'rsatkichlarini avtomatik ravishda aniqlashning o'zi yetarli emas. Masalan: *U kelgan odamni ko'rdi*. Bu gapda *kelgan* - sifatdosh. *U kecha kelgan*. Gapida esa *kelgan* kesim tarkibida kelib, boshqa sintaktik vazifani bajargan. Shuning uchun annotatsiya: shakl + kontekst + sintaktik vazifa tamoyilida bo'ladi.

Yuqoridagi misollarda o'zbek olimlarining PoS-teggingga oid qarashlaridan kelib chiqadigan muhim metodologik xulosa shuki, tegset tilning real grammatik xususiyatlarini aks ettirishi kerak [1].

Fe'ning vazifa shakllarini annotatsiyalash. O'zbek tilining grammatik tizimida fe'ning vazifa shakllari sifatdosh, ravishdosh va harakat nomini qamrab oladi. Ularning umumiy xususiyati fe'llik ma'nosini saqlagan holda boshqa sintaktik vazifalarni bajarishidir.

Shuning uchun korpusda uch asosiy guruhni ajratish taklif qilamiz:

Vazifa shakli	Tavsiya etiladigan belgi	Misol
---------------	--------------------------	-------

Sifatdosh	PART	<i>kelgan</i>
Ravishdosh	CONV	<i>kelib</i>
Harakat nomi	VNOUN	<i>kelish</i>

Bunda xalqaro korpus formatlari bilan muvofiqlikni ta'minlash maqsadida Universal Dependencies tizimidagi Part, Conv va Vnoun kategoriyalaridan foydalanish mumkin. Ammo o'zbek tilining milliy grammatik tasnifi bilan xalqaro standartni aralashtirib yubormaslik kerak. UD - universal formal model, milliy grammatik tasnif esa o'zbek tilining ichki qonuniyatlarini ifodalovchi lingvistik modeldir. Shu sababli eng maqbul yechim - ikki qatlamli belgilash mumkin.

Milliy grammatik belgi + xalqaro korpus tegi.

Masalan: *kelgan* VERB | SIFATDOSH | Part

Kelib VERB | RAVISHDOSHI | Conv

Kelish VERB | HAKARAT_NOMI | Vnoun

Sifatdoshni annotatsiyalash tamoyillari. Sifatdoshni aniqlashda uchta asosiy omil hisobga olinishi kerak: morfologik ko'rsatkich, fe'llik semantikasi, sintaktik vazifa. Masalan: *O'qigan bola savolga javob berdi.*

Annotatsiya: o'qigan | o'qi- | VERB | SIFATDOSH | Part | Aniqlovchi

Bunda *o'qigan* shakli fe'l asosida hosil bo'lgan, sifatdosh shakliga ega va *bola* otini aniqlab kelgan.

Sifatdoshni faqat ADJ deb belgilash noto'g'ri bo'lishi mumkin, chunki bunda uning fe'lga xos grammatik xususiyatlari yo'qoladi. Aksincha, faqat VERB deb belgilash ham uning sifatdosh sifatidagi funksional xususiyatini yetarli aks ettirmaydi. Shuning uchun asosiy so'z turkumi + maxsus vazifa shakli modeli maqbul hisoblanadi.

Ravishdoshni annotatsiyalash tamoyillari. Ravishdosh boshqa harakat bilan munosabatga kirishadigan fe'l shaklidir:

Talaba topshiriqni bajarib, o'qituvchiga yubordi. Bu gapda *bajarib* asosiy harakat *yubordi* bilan bog'langan.

Annotatsiya: bajarib | bajar- | VERB | RAVISHDOSHI | Conv | Hol

Ravishdoshni aniqlashda affiks ko'rsatkich muhim bo'lsa-da, yakuniy qaror kontekst asosida berilishi lozim. Shu sababli avtomatik annotator uchun:

affiks → *morfologik kandidat* → *sintaktik tekshiruv* → *yakuniy teg algoritmini taklif qilish* mumkin.

Harakat nomini annotatsiyalash tamoyillari. Harakat nomi fe'lning otlashgan shakli sifatida gapda otga xos sintaktik vazifalarni bajaradi:

Kitob o'qish foydali. Bu yerda: o'qish | o'qi- | VERB | HAKARAT_NOMI | Vnoun | Ega tarzida annotatsiya qilish mumkin.

Harakat nomining asosiy muammosi shundaki, ayrim hollarda shaklning otlashganligi va fe'llik xususiyatlari chegaradosh bo'lishi mumkin. Shuning uchun uni

aniqlashda:shakliy ko'rsatkich, lemma, fe'l semantikasi, gapdagi sintaktik pozitsiya birgalikda hisobga olinishi lozim.

Mazkur tadqiqot asosida quyidagi annotatsiya modelini taklif qilish mumkin:

Token	Lemma	UPOS	Milliy kategoriya	UD belgisi	Zamon	Bo'lishlilik	Sintaktik funksiya
kelgan	kel-	VERB	Sifatdosh	Part	Past	Pos	Aniqlovchi
kelayotgan	kel-	VERB	Sifatdosh	Part	Pres	Pos	Aniqlovchi
keladigan	kel-	VERB	Sifatdosh	Part	Fut	Pos	Aniqlovchi
kelib	kel-	VERB	Ravishdosh	Conv	-	Pos	Hol
kelgach	kel-	VERB	Ravishdosh	Conv	-	Pos	Hol
kelguncha	kel-	VERB	Ravishdosh	Conv	-	Pos	Hol
kelish	kel-	VERB	Harakat nomi	Vnoun	-	Pos	Ega
kelmoq	kel-	VERB	Harakat nomi	Vnoun	-	Pos	Ega
kelmaslik	kel-	VERB	Harakat nomi	Vnoun	-	Neg	Ega

Bu modelning yangiligi shundaki, u o'zbek tilining milliy grammatik terminologiyasini xalqaro korpus annotatsiyasi bilan parallel qo'llashni ko'zda tutadi.

Korpus yaratish bosqichlari. Fe'l vazifa shakllari korpusini yaratishni quyidagi bosqichlarda amalga oshirish mumkin:

1-bosqich. Matnlarni yig'ish. Badiiy, ilmiy, publitsistik, rasmiy va ta'limiy matnlar tanlanadi.

2-bosqich. Matnlarni tozalash. HTML kodlar, ortiqcha belgilar, texnik elementlar va takroriy matnlar chiqarib tashlanadi.

3-bosqich. Tokenizatsiya. Matn so'z va boshqa tokenlarga ajratiladi.

4-bosqich. Lemmatizatsiya. Fe'l shakllari umumiy lemma ostida birlashtiriladi.

5-bosqich. Morfemik segmentatsiya. So'z tarkibidagi o'zak va affikslar aniqlanadi.

6-bosqich. Morfologik annotatsiya. Fe'lning zamon, mayl, nisbat, bo'lishlilik, shaxs-son kabi kategoriyalari belgilanadi.

7-bosqich. Vazifa shakli annotatsiyasi. Sifatdosh, ravishdosh va harakat nomi belgilanadi.

8-bosqich. Sintaktik annotatsiya. Vazifa shaklining gapdagi sintaktik vazifasi aniqlanadi.

9-bosqich. Ekspert tekshiruvi. Noaniq va bahsli holatlar lingvist tomonidan tekshiriladi.

10-bosqich. Sifat nazorati. Annotatorlararo kelishuv va avtomatik annotatsiya aniqligi baholanadi.

Korpusni yangilashning sanab o'tilgan bosqichlaridan xulosa qilish mumkinki, mazkur tadqiqotning asosiy ilmiy yangiligi o'zbek tilida fe'lning vazifa shakllarini maxsus korpus obyekti sifatida ajratib, ularni milliy grammatik tasnif va xalqaro korpus standartlari asosida ko'p qatlamli annotatsiyalash modelini taklif etishdan iborat. Bunda o'zbek kompyuter lingvistikasida shakllangan ilmiy tajriba - xususan, G.Toirova, B.Elov va Sh.Hamroyevaning PoS-tegging va morfologik analizatorga oid

ishlari, M. Abjalovanning avtomatik lingvistik modullar va annotatsiyaga oid tadqiqotlari, N.Abdurahmonovanning elektron korpusning kompyuter modellari haqidagi qarashlari, B.Mengliyev va Sh.Hamroyevanning turkiy tillar morfologik analizatorlari bo'yicha tadqiqotlari - yagona metodologik tizimga birlashtiriladi [1-6; 9]. Shunday qilib, taklif qilinayotgan model tayyor xalqaro teglarni mexanik tarzda o'zbek tiliga ko'chirish emas, balki o'zbek tilining lingvistik xususiyatlarini hisobga olgan holda xalqaro standart bilan uyg'unlashtirish tamoyiliga asoslanadi.

Xulosa. O'zbek tilida fe'ning vazifa shakllari korpusini yaratish zamonaviy o'zbek kompyuter lingvistikasining korpusshunoslik, morfologik tahlil, lingvistik annotatsiya va tabiiy tilni qayta ishlash yo'nalishlari kesishmasida yuzaga keladigan muhim ilmiy masalalardan biridir. Fe'ning vazifa shakllari korpusida sifatdosh, ravishdosh va harakat nomini alohida annotatsiya kategoriyalari sifatida ajratish maqsadga muvofiq. Bunda morfologik belgi bilan bir qatorda sintaktik funktsiya ham qayd etilishi kerak. Tadqiqotda taklif etilgan modelga ko'ra, korpusning asosiy annotatsiya qatlamlari lemma, morfemik tarkib, umumiy so'z turkumi, vazifa shakli, grammatik kategoriyalar va sintaktik funktsiyadan tashkil topishi lozim. Xalqaro korpuslar bilan qiyosiylikni ta'minlash uchun Part, Conv, Vnoun kabi Universal Dependencies kategoriyalaridan foydalanish mumkin, biroq ularni o'zbek tilining milliy grammatik kategoriyalari bilan parallel qo'llash maqsadga muvofiq. Umuman olganda, o'zbek tilida fe'ning vazifa shakllariga ixtisoslashgan korpus yaratish nafaqat sifatdosh, ravishdosh va harakat nomining real matndagi qo'llanishini empirik o'rganish, balki o'zbek tilining morfologik analizatori, PoS-taggeri, sintaktik parseri, mashina tarjimasini tizimlari va boshqa NLP vositalari uchun muhim lingvistik baza yaratishga xizmat qiladi. Shu jihatdan mazkur korpusni yaratish o'zbek tilining milliy kompyuter lingvistikasi resurslarini xalqaro korpus standartlari bilan integratsiyalash yo'lidagi istiqbolli ilmiy-amaliy loyiha sifatida baholanishi mumkin.

Foydalanilgan adabiyotlar

1. Elov, B., Hamroyeva, Sh., Abdullayeva, O., Uzoqova, M. O'zbek tilida PoS tegging masalasi: muammo va takliflar. // *O'zbekiston: til va madaniyat. Amaliy filologiya*. 2022. Vol. 2, №5. B. 51-68.
2. Elov, B.B., Hamroyeva, Sh.M., Elova, D.Q. Morfologik analizatorni yaratish usullari. // *O'zbekiston: til va madaniyat. Amaliy filologiya*. 2022. 1(5). B. 67-86.
3. Abjalova, M.A. Tahrir va tahlil dasturlarining lingvistik modullari. Monografiya. Toshkent: Nodirabegim, 2020. 176 b.
4. Abdurahmonova, N. O'zbek tili elektron korpusining kompyuter modellari. Monografiya. Toshkent: Globe Edit, 2021.
5. Zaxarov, V., Mengliyev, B., Hamroyeva, Sh. Korpus lingvistikasi: korpus tuzish va undan foydalanish. Toshkent: Globe Edit, 2021. 185 b.

6. Abjalova, M.A. Tagging and Annotation of Corpus Units. *International Journal of Language Learning and Applied Linguistics*, 2(12), 2023. DOI: 10.51699/ijllal.v2i12.3228.
7. Elov, B., Abjalova, M., Alayev, R. O'zbek tili korpusi va uning imkoniyatlari. // *Informatika va energetika muammolari*. 2023. №2. B. 89-100.
8. Raupova, L., Elov, B., Abjalova, M., Alayev, R. O'zbek tilining ta'limiy korpusi va uning imkoniyatlari. // *O'zbekiston: til va madaniyat*. 2021. №4. B. 60-75.
9. Mengliyev, B.R., Hamroyeva, Sh.M. Turkiy tillardagi morfologik analizatorlar xususida. // *Kompyuter lingvistikasi: muammolar, yechim, istiqbollar* ilmiy-amaliy anjumani materiallari. 2020.
10. Akhundjanova, A., Talamo, L. Universal Dependencies Treebank for Uzbek. // *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*. 2025. P. 1-6.
11. Matlatipov, S.G., Aripov, M. UzUDT: Uzbek Universal Dependencies Treebank. // *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*. 2026. P. 11642-11649.
12. Elov, B., Hamroyeva, Sh., Alayev, R., Xusainova, Z., Yodgorov, U. O'zbek tili korpusi matnlarini qayta ishlash usullari. // *Digital Transformation and Artificial Intelligence*. 2023. Vol. 1, №3. B. 32-42.