

O'zbek tilida pos tegging masalasi: muammo va takliflar

Botir Elov¹,
Shahlo Hamroyeva²,
Oqila Abdullayeva³,
Mohiyaxon Uzoqova⁴

Abstrakt:

Til korpusi qurilishida lingvistik ta'minot masalasi muhim va murakkab hisoblanadi. Korpuslarda matnlardagi nutq bo'laklariga mos identifikatori belgilash jarayoni muammolidir, sababi tilni modellashtirish teglash qoidasi va tilda mavjud qonuniyat bilan bog'liq. Teglash, xususan, grammatik teglash yoki PoS tegging o'zbek korpus lingvistikasi uchun ham dolzarb masaladir. Chunki maxsus "kodlangan" belgilar tizimi o'zbek tili bilan bog'liq NLP masalalarini yechishda birlamchi kalit bo'lib xizmat qiladi. Maqolada teg va PoS tegging masalasining jahon tilshunosligidagi o'rganilishi tahlil qilindi, o'zbek tilshunosligidagi hozirgi teglash jarayoniga nazar tashlandi. Jahonda keng qo'llanib kelinayotgan tegsetlardan oziqlangan holda o'zbek tili qonuniyatlari asosida muqobil tegset taklif qilindi.

¹*Elov Botir Boltayevich* – texnika fanlari bo'yicha falsafa doktori (PhD), dotsent. Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universiteti Kompyuter lingvistikasi va raqamli texnologiyalar kafedrası mudiri.

E-pochta: elov@navoiy-uni.uz

ORCID: 0000-0001-5032-6648

²*Hamroyeva Shahlo Mirdjanovna* – filologiya fanlari doktori (DSc), Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universiteti Kompyuter lingvistikasi va raqamli texnologiyalar kafedrası dotsenti v.b.

E-pochta: shaxlo.xamrayeva@navoiy-uni.uz

ORCID: 0000-0002-5429-4708

³*Abdullayeva Oqila Xolmo'minovna* – filologiya fanlari bo'yicha falsafa doktori (PhD). Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universiteti Kompyuter lingvistikasi va raqamli texnologiyalar kafedrası katta o'qituvchisi.

E-pochta: abdullayeva.oqila@navoiy-uni.uz

ORCID: 0000-0002-2524-4832

⁴*Uzoqova Mohiyaxon Tuyg'un qizi* – Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universiteti Kompyuter lingvistikasi va raqamli texnologiyalar kafedrası magistranti.

E-pochta: uzoqovamohiyaxon@navoiy-uni.uz

ORCID: 0000-0001-7102-0824

Iqtibos uchun: Elov B., Hamroyeva Sh., Abdullayeva O., Uzoqova M. 2022. "O'zbek tilida PoS tegging masalasi: muammo va takliflar". *O'zbekiston: til va madaniyat. Amaliy filologiya*. 2 (5): 51-68.

Kalit so'zlar: *Teg, razmetka, annotatsiya, tegset, NLP, korpus, Braun korpusi, CLAWS, PoS tagging.*

Kirish

Ma'lumki, istalgan turdagi korpusni tuzish mavjud datani teglashdan boshlanadi. Teglash jarayoni annotatsiya, razmetka kabi so'zlar bilan turli adabiyotlarda yonma-yon, ba'zan sinonim sifatida ishlatilganligiga tez-tez guvoh bo'lish mumkin [Zaxarov, 2005, 6]. Tadqiqotimizda til korpuslarida nutq bo'laklarini teglash masalasini muhokama qilishdan oldin, teg, razmetka va annotatsiya tushunchalari to'g'risidagi fikrlarni kuzatamiz. Mazkur terminlarni sinonim sifatida baholash to'g'rimi? Biri o'rnida ikkinchisi qo'llanilganda bir mazmunni bera oladimi? Annotatsiya, teg va razmetkaning bir-biridan farqlovchi aniq chegarasi mavjudmi?

Annotatsiya – Korpus lingvistikasi doirasida qaralganda berilgan matnga bevosita aloqasi bo'lmagan, ammo uning qaysidir qismi haqidalingvistikyoki ekstralingvistik axborot beruvchi umumiy ma'lumot. Annotatsiya o'z ichiga metama'lumot va teglarni qamrab olishi mumkin. Ayrim manbalarda [Leech, Vilson, 1994] korpus annotatsiyasi deganda, til korpusida matnning elektron shakliga kodlash qo'shilgan izohlovchi, lingvistik ma'lumotlarni qo'shish amaliyoti ham tushuniladi. McEnery annotatsiyalashni uch usulda amalga oshirish mumkinligini yozadi: to'la avtomatlashtirilgan, yarim avtomatlashtirilgan va qo'l mehnati yordamida. Ammo hech bir usul mukammal ish bermasligi, mutlaqo xatosiz natija olib bo'lmasligini aytgan.

Razmetka – McEnery razmetka va meta ma'lumotni annotatsiyalash jarayonining bir qismi sifatida baholagan [McEnery, Hardie, 2012, 47]. Boshqa manbalar bilan tanishganimizda esa razmetkaga annotatsiyalash jarayonining o'zi sifatida baho berilganligiga guvoh bo'ldik [Myfilology.ru]. Ingliz manbalarida razmetka markup termini bilan yuritiladi [McEnery, Hardie, 2012, 47]. Jahon amaliyotida razmetkalashning standart prinsiplari ishlab chiqilgan. U SGML yoki Standard Generalized Markup Language [<https://www.techtarget.com/>] deb ataladi. E'tibor qaratish lozim bo'lgan jihat shundaki, SGMLda standart annotatsiya emas, balki annotatsiyalash jarayonini tashkil etish metodologiyasi aks etadi. SGML asosida ishlab chiqilgan va keng ommalashgan tillarga HTML yoki XMLni misol qilish mumkin.

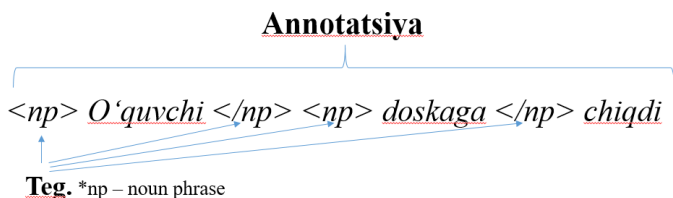
Teg – Kompyuter yordamida matn tahlilini amalga oshirish jarayonini tezlashtirish va osonlashtirishga xizmat qiluvchi shartli belgi yoki maxsus kod. Teglar bir necha turlarga bo'linadi [<https://>

ucrel.lancs.ac.uk/]: semantik teg, sintaktik teg va grammatik teg. Grammatik teg, shuningdek, PoS (Part of speech) tegging nomi bilan ham mashhur.

Annotatsiya (razmetka) va tegning farqini quyidagi jadvalda aniqroq ko'rish mumkin [<https://knowledge.autodesk.com/>]:

1-jadval. Annotatsiya va teg terminining ta'riflanishi

Annotatsiya	Teg
Muayyan komponent yoki segment haqida jadval yoki grafik ko'rinishidagi ma'lumot	Muayyan komponent yoki segment uchun unikal identifikator
Komponent yoki segment uchun bir nechta annotatsiya bo'lishi mumkin	Komponent yoki segment uchun yagona nom beriladi va u teg deb ataladi
Bir qancha ta'riflar to'plamini matn shaklida o'zida jamlaydi va unda teglar ham aks etadi.	Gap emas, uning bo'laklari haqida ma'lumot beruvchi muayyan formatdagi data hisoblanadi



1-rasm. SGMLda annotatsiyalash va teglash jarayoni.

1-rasmda SGMLda ishlab chiqilgan standart asosida gap annotatsiyalangan. Ko'rish mumkinki, gapdagi har bir bo'lak o'zi tegishli bo'lgan turkum asosida teglanib, boshlanish va tugash jarayoni qirrali qavs `</>` yordamida aks ettirilmoqda. Bu jarayonni umumiy qilib annotatsiyalashning bir ko'rinishi sifatida baholash mumkin.

Demak, annotatsiya va razmetka kengroq tushuncha bo'lib, teg uning bo'lagi sifatida xizmat qilishi mumkin. Teg – unikal kod, ammo havola qilayotgan komponentasining qisqartmasi sifatida baholanmasligi lozim.

Asosiy qism

Izlanishlarimiz davomida [Mitkov, 2003, 228] shu narsa ayonlashdiki, teglashga bo'lgan ehtiyoj, ya'ni nutq bo'laklarini unikal kodlashtirish kompyuterni avtomatik tahlil qilishga yordam beruvchi manbalar, jumladan, korpus tuzish ishlaridan keyin boshlangan. Shuning uchun ham teglash jarayoni birinchi marta jahon tajribasida

Braun korpusini tuzish bilan qayd etiladi. 1971-yilda 77 xil turdagi teglar TAGGIT dasturi uchun Braun korpusi mualliflari tomonidan ishlab chiqilgan [Mitkov, 2003, 228]. Bunday teglashning asosiy yutug'i nafaqat asosiy grammatik kategoriyalar, balki ularning ichki guruhlarigacha teglanganida ko'ringan. Keyinroq, Lankaster universiteti tomonidan CLAWS teggerini yaratish uchun tegsetlar (teglar to'plami) ham ishlab chiqila boshlangan [<https://ucrel.lancs.ac.uk>]. Tegger – bu berilgan matnni yoki segmentni avtomatik teglashga xizmat qiluvchi dastur. Shunday qilib, 1980-yillarda tuzish boshlangan CLAWS1 tegseti [<https://ucrel.lancs.ac.uk>] jami 137 ta tegni o'zida qamrab olgan. Teglarining aksariyati Braun korpusida foydalanilgan teglardan oziqlangan, ammo ulardan farqli bo'lgan. CLAWS tegsetlarini ishlab chiqishda bevosita ishtirok etgan Robert Garside bu haqida shunday yozgan [<https://ucrel.lancs.ac.uk>]: *“CLAWS1da qo'llanilgan teglar to'plami Braun korpusini teglashda foydalanilgan teglardan oziqlangan. Biz teglangan LOB korpusini Braun korpusi bilan taqqosladik, ammo Braun tegsetidagi atoqli otlar va olmoshlar teglariga o'zgartirish kiritgan holda sintaktik birliklar (so'zlar va tinish belgilari) uchun jami 137 ta ehtimoliy teglarni ishlab chiqdik”*. CLAWS tegsetlari yillar davomida takomillashtirilib, bugungi kunda 8 taga yetdi [<https://ucrel.lancs.ac.uk>]. Bunda har bir tegset keyingisi uchun asos sifatida xizmat qilavergan.

Jahon miqyosida keng tarqalgan va mashhur bo'lgan tegsetlardan Pen Tree Bank Tagsetni [<https://www.sketchengine.eu>] ham e'tirof etishni lozim topdik. Mazkur tegset 1993-1994-yillarda Shtutgart universiteti tomonidan ishlab chiqilgan [<https://www.ims.uni-stuttgart.de>]. Keyinroq bu tegsetlardan Sketch Engine korpusida ham foydalanilgan [<https://www.sketchengine.eu>]. Hozirda Sketch Engine tarkibidagi tegsetlar versiyasi 3 taga yetgan, ammo ular Pen Tree Bank tegsetlarining aynan o'zi emas, biron takomillashgan versiyasi hisoblanadi. CLAWS va Pen Tree Bank tegsetlarini solishtirganimizda, aksariyat teglar bir xil shaklga ega ekanligiga guvoh bo'ldik (Masalan, *CC* = *Coordinating conjunction*, *JJ* = *adjective*, *MD* = *modal*). Demak, yaratilish yilini hisobga olgan holda Pen Tree Bank tegsetlari CLAWS tegsetlaridan oziqlangan, deyish mumkin.

Yana bir ommalashgan tegger va tegsetlar Python kutubxonasiga tegishli NLTK dasturida aks etgan. NLTK – Natural Language Toolkit tokenizatsiya, stemming, lemmatizatsiya va avtomatik teglay oluvchi universal hamda NLP olamidagi eng ommabop dastur bo'lib, unda taqdim etilgan tegsetlarni o'rganib chiqqanimizda Pen Tree Bank va CLAWS tegsetlaridan deyarli farq

qilmasligiga guvoh bo'ldik [<https://www.learntek.org>]. Hatto BNC (British National Corpus) namunaviy tegsetlari [<https://ucrel.lancs.ac.uk>] ham CLAWS tegsetlari asosida tuzilgan.

Biroz farqli tegsetlarni 25 xil tegsetni taqqoslash asosida universal tegset ishlab chiqqan bir guruh olimlar tadqiqotida ko'rinadi [Petrov, Das, McDonald, <https://static.googleusercontent.com>]. Tadqiqotda bayon etilishicha, 22 xil tilga xos umumiy grammatik kategoriyalar o'rganib chiqilgan va universal teglar taklif qilingan. Universal tegsetdagi [<https://universaldependencies.org>] teglar yuqorida tahlilga olingan tegsetlardan mutlaqo yangi teglarni taklif qilishi bilan farq qiladi. Ammo faqat eng yirik grammatik guruhlar uchun teglar taklif etilgani bois muayyan til uchun to'liq bu tegsetlarni joriy qilish imkonsiz, degan fikrdamiz.

Shunday qilib, jahondagi eng ommalashgan tegsetlarni tadqiq va tahlil qilish natijasida shunday xulosaga keldikki, **har qanday til uchun joriy qilinishi mumkin bo'lgan standart teg mavjud emas**. Bu xulosani quyidagi sabablar bilan dalillasak:

1) Tegsetlarni ishlab chiqishda har bir maktab o'zining yondashuvlariga asoslangan. Davr o'tishi bilan biri ikkinchisidan oziqlangan va ilgarigi tegsetlar yangi teglar bilan takomillashtirib borilgan.

2) Dunyodagi hech bir til tuzilishi va tabiatiga ko'ra to'laqonli o'xshash emas, garchi o'xshashlik darajasiga qarab umumiy til oilalari va guruhlariga tasniflansa ham.

3) Annotatsiyalashning turlari kengaygani sari yangi tegsetlar ishlab chiqishga ehtiyoj tug'ilaveradi. Natijada tegsetlar statik holatdan dinamik holatga o'tib qoladi.

Shular sababli ham, bizning fikrimizcha, butun dunyoga tatbiq qilish mumkin bo'lgan mukammal va standart tegsetlar hanuz mavjud emas va buni ishlab chiqish ham mantiqsizlik bo'ladi.

O'zbek tilshunosligida razmetka, annotatsiyalash masalasi turli tadqiqotlarda monografik planda o'rganilgan. Ba'zi tadqiqotlarda teglar to'plami taklif etilgan. Teg va teglash masalasini ba'zi ishlarda razmetkalash shaklida kuzatish mumkin [Hamroyeva, 2020, 83]. Sh.Hamroyeva tadqiqotida mukammal teg – keng imkoniyatli, universal korpus garovi ekanligini ta'kidlaydi. Razmetkalash jarayoni uchun bir nechta teg lozimligi, teglarni lingvistik modellash tirish maqsadga muvofiqligi va aynan mana shu modelda morfologik teg shartli qisqartma shaklini olishi ko'rsatilgan. Har bir nutq bo'lagi uchun morfologik va semantik teglarning o'zbek tilida lingvistik modellari taklif qilingan [Hamroyeva, 2020, 205].

O'zbek tilida avtomatik tahlil va tahrir uchun so'z shakllarini formallashtirish va modellashtirish yuzasidan e'lon qilingan ishlarda lingvistik modullar ishlab chiqilgan [Abjalova, 2019, 178]. A.Abjalova tadqiqotida har bir shaklga qo'yilgan belgini razmlar sifatida baholagan. Masalan, ko'plik affiksiga k_a, egalik affiksiga e_a, kelishik affiksiga ke_a va hokazo ramzlarni kiritgan [Abjalova, 2019, 96]. M.Abjalova tadqiqotida grammatik kategoriyalarni formallashtirish uchun modullar taklif qilgan. Formal grammatika nuqtayi nazaridan kategorial shakllarni quyidagicha formallashtirgan: zamon kategoriyasini z deb belgilab, o'tgan zamon – z1, hozirgi zamon – z2, kelasi zamon – z3 shaklida [Abjalova, 2019, 106]. Shuningdek, har bir kategorial shakl o'z ichida grammatik affikslariga ko'ra belgilash davom ettirilgan. Oxirgi amalga oshirilgan ishlarda o'zbek tilining morfologik bazasini shakllantirish yuzasidan ham nazariyalar mavjud. O'Xoliyorov fe'l shakllarining morfologik lug'atini tuzib, unda fe'l shakllarining nisbat, zamon, mayl, shaxs-son, bo'lishli-bo'lishsizlik, vazifa shakllari aranjirovkasi uchun lingvistik modellar taklif qilgan [Xoliyorov, 2021, 22]. Masalan, Fe'l so'z turkumi uchun *F* belgisi, fe'l so'z turkumi bo'lishsizlik shakli uchun *bb2* yoki ravishdosh shakli uchun *v2* belgisi, ot so'z turkumi kichraytirish shakli uchun *kich*, birlik shakli uchun *s1* belgisi tavsiya etilgan. O.Abdullayeva tadqiqotida matnda nutq bo'laklarini teglash uchun 65 ta morfologik va 40 ta semantik teg taklif qilgan, foydalanuvchilar uchun qulay bo'lishi, korpus foydalanuvchilarining kutilayotgan keng xalqaro auditoriyasini hisobga olgan holda xalqaro standardlashtirilgan teglarni ham teg jadvalida berilgani ta'kidlangan [Abdullayeva, 2022, 93]. O'zbek tilshunosligida e'lon qilingan ko'plab ilmiy tadqiqot natijalarida teg va tegsetlarning bir-biriga yaqin va o'xshash jihatlari ko'zga tashlanadi. Albatta, oldingi ishlardan namuna olish, oziqlanish holatini kuzatish mumkin. Tadqiqot ishlarida grammatik shakl uchun qaysi tegning belgilanishi emas, balki korpus platformasida PoS teglash masalasining hal qilinganligi, lingvistik tahlillar amalga oshirilganda muhim natijalarning olinishi bilan belgilanishi ham aytib o'tilgan [Abdullayeva, 2022, 71].

O'zbek tilida olimlar tomonidan PoS uchun tegning belgilanishini quyidagi jadvalda ham ko'rish mumkin.

2-jadval. Fe'llarda nisbat kategoriyasi linivistik modelining teganishi

	Aniq nisbat	O'zlik	Majhul	Orttirma	Birgalik
Sh.Hamroyeva	an.n.	o'zl.n.	maj.n.	ort.n.	bir.n.

O'Xoliyorov	n1	n2	n3	n4	n5
M.Abjalova	x1	x2	x3	x4	x5
O.Abdullayeva	Aniqn	O'zn	Majn	Ortn	Birn
Turklang	-	REFL	PASS	CAUS	RECP

Tegsetlarni shakliy tomondan ishlab chiqishda, guvoh bo'lamizki, shu vaqtga qadar uch usuldan foydalanilgan: 1) teglar faqat harflardan iborat; 2) teglar faqat sondan iborat; 3) teglar harf va sonning kombinatsiyasidan iborat. Ammo bunda teglarni abreviaturadan farqlash kerak [Abdullayeva, 2022, 70]. Teglar muayyan kategorianing qisqartmasi bo'lishi shart emas, eng muhimi inobatga olinishi lozim bo'lgan jihat ularning sodda, qisqa va eslab qolishga qulay bo'lishi lozim.

Biz yuqoridagi barcha tahlillar asosida chiqargan xulosalarimizga asoslangan holda esda saqlash uchun tushunarli bo'lgan, sodda va qisqa grammatik tegsetni taklif etishga qaror qildik. Ushbu tegsetni ishlab chiqishda, dastlab, o'zbek tilida mavjud bo'lgan barcha grammatik kategoriyalar o'rganib chiqildi va to'plandi. Dunyo miqyosida eng keng tarqalgan tegset CLAWS tegsetlari bo'lgani bois biz ham ulardan oziqlanishga qaror qildik. Teglarni nomlashda esa quyidagi prinsipga asoslanildi:

1. Xalqaro teg o'zbek tilidagi mavjud grammatik kategoriyaga to'liq mos tushsa, o'zgarishsiz qabul qilindi.

2. Xalqaro teg o'zbek tilidagi mavjud grammatik kategoriyaga qisman mos tushsa, uni to'ldirish uchun eslab qolishga qulay qilib inglizcha so'zlar qisqartmasiga tayangan holda yangi teg yasaldi.

I. Xalqaro teg o'zbek tilidagi mavjud grammatik kategoriyani o'zida umuman aks ettirmasa, tegsetda bir tillik prinsipini saqlash maqsad qilingani uchun inglizcha tushunchalarga tayangan holda mutlaqo yangi teg yasaldi.

	Ot so'z turkumi	N
1.1.	Bevosita kuzatish mumkin bo'lgan/ bo'lmaganligiga ko'ra	
1.1.1.	Mavhum ot	NAbs
1.1.2.	Aniq ot	NCnc
1.2.	Atalish/atalmasligiga ko'ra	
1.2.1	Atoqli ot	NP
1.2.2	Turdosh ot	NC (Noun Cognate)
1.3	Atoqli otlarning leksik-semantik ma'no turlari	
1.3.1	Shaxs nomi	NPTitl

1.3.2	Geografik nom	NPGeo
1.3.3	Tashkilot, muassasa, korxona nomi	NPOrg
1.3.4	Samoviy yoritgichlar nomi	NPCos
1.3.5	Tarixiy sana yoki bayram nomi	NPDt
1.3.6	Hayvon nomi	NPAnm
1.3.7	Mahsulot nomi	NPPrd
1.3.8	Ilohiy tushuncha nomi	NPDvn
1.3.9	Mansab nomi	NNA
1.3.10	Unvon nomi	NNB
1.3.11	Joy nomlari	NPTop
1.4.	Turdosh otlarning leksik-semantik ma'no turlari	
1.4.1.	Narsa/buyum otlari	NCPr
1.4.1.1	Hayvon va uning a'zolari otlari	NCAAn
1.4.1.2	Oziq-ovqat nomlari	NCPrd
1.4.1.3	O'simlik va uning qismlari otlari	NCPlt
1.4.1.4	Diniy-ilohiy otlar	NCDvn
1.4.1.5	Inson tana a'zolari nomlari	NBTitl
1.4.1.6	Fan nomlari	NCSub
1.4.2.	Shaxs oti	NCPrs
1.4.2.1	Qarindoshlikka ko'ra	NCRI
1.4.2.2	Yoshga ko'ra	NCAg
1.4.2.3	Kasbga ko'ra	NCJb
1.4.2.4	Jinsga ko'ra	NCSx
1.4.2.5	Joyga ko'ra	NCPlc
1.4.2.6	Ijtimoiy holatiga ko'ra	NCSoc
1.5.	Makon/maskan otlari	NCPlc
1.6.	Zamon otlari	NCTm
1.7.	Faoliyat/jarayon otlari	NCEv
1.2.11.	Miqdor otlari (Hisob so'zlar)	NNU
1.3.	Predmetning o'zini yoki to'dasini bildirishiga ko'ra	
1.3.1.	Yakka ot	NN1
1.3.2.	Jamlovchi ot	NN
	Tuzilishiga ko'ra	
1.3.4	Sodda ot	ND
1.3.5.	Qo'shma ot	NCmp
1.3.6.	Juft ot	NCpl
1.3.7.	Takroriy ot	NRep
1.4.	Yasalishiga ko'ra	
1.4.1.	Tub ot	ND0
1.4.2.	Yasama ot	ND1
1.5.	Yasama bo'lsa	

1.5.1.	Affiksatsiya	ND1Af
1.5.2.	Kompozitsiya	ND1Cmp
1.6.	Kategoriyalari	
1.6.1	Kelishik kategoriyasi	Cs (Case)
1.6.1.1	Bosh kelishik	Nom (Nominative)
1.6.1.2	Tushum kelishigi	Acc (Accusative)
1.6.1.3	Qaratqich kelishigi	Gen (Genetive)
1.6.1.4	Jo'nalish kelishigi	Dat (Dative)
1.6.1.5	O'rin-payt kelishigi	Loc (Locative)
1.6.1.6	Chiqish kelishigi	Abl (Ablative)
1.6.2	Egalik kategoriyasi	PossP
1.6.2.1	Birinci shaxs egalik birlik	PossP1s
1.6.2.2	Birinci shaxs egalik ko'plik	PossP1p
1.6.2.3	Ikkinchi shaxs egalik birlik	PossP2s
1.6.2.4	Ikkinchi shaxs egalik ko'plik	PossP2p
1.6.2.5	Uchinchi shaxs egalik birlik	PossP3s
1.6.2.6	Uchinchi shaxs egalik ko'plik	PossP3p
1.7.1	Ko'plik	p
1.7.2	Birlik	s
	Sifat so'z turkumi	JJ
2.1.1.	Oddiy daraja	JJOrd
2.1.2.	Qiyosiy daraja	JJR
2.1.3.	Orttirma daraja	JJT
2.1.4	Ozaytirma daraja	JJMin
2.2	Tuzilishiga ko'ra	
2.2.1.	Sodda sifat	JJD
2.2.2.	Qo'shma sifat	JJCmp
2.2.3.	Juft sifat	JJCpl
2.2.4.	Takror sifat	JJRep
2.3	Yasalishiga ko'ra	
2.3.1.	Tub sifat	JJD0
2.3.2.	Yasama sifat	JJD1
2.4	Yasalgan bo'lsa	
2.4.1.	Affiksatsiya	JJD1Af
2.4.2.	Kompozitsiya	JJD1Cmp
2.5.	Semantik guruhleri	
2.5.1	Xususiyat-holat bildiruvchi sifat	JJFt
2.5.3	Hid bildiruvchi sifat	JJSc
2.5.4	Rang-tus bildiruvchi sifat	JJCol
2.5.5	Maza-ta'm bildiruvchi sifat	JJTs
2.5.6	Shakl bildiruvchi sifat	JJM
2.5.7	O'lchov bildiruvchi sifat	JJS
2.5.8	O'rin bildiruvchi sifat	JJL

2.5.9	Payt bildiruvchi sifat	JJT
III.	Fe'l so'z turkumi	VB
3.1.	Lug'aviy ma'noga ega/ega emasligiga ko'ra	
3.1.1	Mustaqil fe'l	VBI
3.2.1	Lug'aviy ma'noviy ifodasiga ko'ra	
3.2.1.1	Harakat	VBAc (Action)
3.2.1.2	Nutq	VBSp (Speech)
3.2.1.3	Holat	VBCon (Condition)
3.2.1.4	Natijali faoliyat	VBRes (Result)
3.2.1.5	Tafakkur	VBTh (Thinking)
3.2.1.6	Munosabat	VBAtt (Attitude)
3.2.2	Grammatik xususiyatiga ko'ra	
3.2.2.1	O'timli	VBTran (Transitive)
3.2.2.2	O'timsiz	VBIttran (Intransitive)
3.2.3	Tuzilishiga ko'ra	
3.2.3.1	Sodda	VBD
3.2.3.2	Qo'shma	VBComp
3.2.3.3	Juft	VBCpl
3.2.3.4	Takror	VBRep
3.2.4	Yasalish xususiyatiga ko'ra	
3.2.4.1	Tub	VBD0
3.2.4.2	Yasama	VBD1
3.2.5	Yasama bo'lsa	
3.2.5.1	affiksatsiya	VB1Aff
3.2.5.2	kompozitsiya	VB1Cmp
3.3	Nomustaqil fe'l	VBD
3.3.1	Vazifasiga ko'ra	
3.3.1.1	Ko'makchi fe'l	VBH
3.3.1.2	Yetakchi fe'l	VB�
3.3.1.3	To'liqsiz fe'l	VBNotF
3.3.1.4	Ko'makchi fe'lli so'z qo'shilmasi	KFSQ
3.4	Zamon qo'shimchalari	T (Tense)
3.4.1	O'tgan zamon	TPast
3.4.2	Hozirgi zamon	TPres
3.4.3	Kelasi zamon	TFut
3.5	Shaxs-son	Prs
3.5.1	I shaxs birlik	Prs1s
3.5.2	I shaxs ko'plik	Prs1p
3.5.3	II shaxs birlik	Prs2s
3.5.4	II shaxs ko'plik	Prs2p
3.5.5	III shaxs birlik	Prs3s

3.5.6	III shaxs ko'plik	Prs3p
3.6.1	Fe'ning bo'lishli shakli	XY
3.6.2	Fe'ning bo'lishsiz shakli	XNOT
IV.	Son so'z turkumi	NUM
4.1.	Tuzilishiga ko'ra	
4.1.1	Sodda	NUMD
4.1.2	Murakkab	NUMCmp
4.2.	Ma'noviy guruhlariga ko'ra	
4.2.1	Dona son	NUMC (Count)
4.2.2	Jamlovchi	NUMCol (Collection)
4.2.3	Chama son	NUMG
4.2.4	Taqsim son	NUMDiv (Dividing)
4.2.5	Kasr son	MF
4.2.6	Sanoq son	NUM
4.2.7.	Tartib son	MD
V.	Olmosh so'z turkumi	P
5.1	Ma'no turlari	
5.1.1	Kishilik olmosh	PP
5.1.2	Ko'rsatish olmosh	PDem (Demosntrative)
5.1.3	O'zlik olmosh	PRef (Reflexive)
5.1.4	Gumon olmosh	PPred (Prediction)
5.1.5	So'roq olmosh	PQues (Question)
5.1.6	Belgilash olmoshi	PInd (Indefinitive)
5.1.7	Bo'lishsizlik olmoshi	PDis (Distributive)
5.2	Tuzilishiga ko'ra turlari	
5.2.1	Sodda olmosh	PD
5.2.2	Qo'shma olmosh	PCmp
5.2.3	Juft olmosh	PCpl
5.2.4	Takror olmosh	PRep
5.3	Yasalishiga ko'ra	
5.3.1	Tub olmosh	PD0
5.3.2	Yasama olmosh	PD1
5.4	Yasalishiga ko'ra	
5.4.1	Affiksatsiya	PD1Aff
5.4.2	Kompozitsiya	PD1Comp
VI	Ravish so'z turkumi	RR
6.1	Lug'aviy ma'noviy guruhi	
6.1.1	Payt ravishi	RT
6.1.2	O'rin ravishi	RL

6.1.3	Holat ravishi	RCon (Condition)
6.1.4	Miqdor-daraja ravishi	RMes (Measurment)
6.1.5	Maqsad ravishi	RPur (Purpose)
6.1.6	Sabab ravishi	RRs (Reason)
6.2	Tuzilishiga ko'ra	
6.2.1	Sodda ravish	RRD
6.2.2	Qo'shma ravish	RRComp
6.2.3	Juft ravish	RRCpl
6.2.4	Takror ravish	RRRep
6.3	Yasalishiga ko'ra	
6.3.1	Tub ravish	RRD0
6.3.2	Yasama ravish	RRD1
VII	Bog'lovchi	C
7.1	Sof bog'lovchi	C0
7.2	Vazifadosh bog'lovchi	C1
7.3	Teng bog'lovchi	CEq (Equeal)
7.3.1	Biriktiruvchi bog'lovchi	CC
7.3.2	Ayiruv bog'lovchi	CDis (Distributive)
7.3.3	Zidlov bog'lovchi	CCB
7.3.4	Inkor bog'lovchi	CNeg (Negative)
7.4.	Ergashtiruvchi bog'lovchi	CS
7.4.1	Aniqlovchi bog'lovchi	CSDet (Determine)
7.4.2	Sabab bog'lovchi	CSRs (Reason)
7.4.3	Shart bog'lovchi	CSIf (If)
7.4.4	Chog'ishtiruv bog'lovchi	CSCmp (Comparing)
7.4.5	Inkor bog'lovchi	CSNeg (Negative)
7.4.6	Maqsad bog'lovchi	CSPur (Purpose)
VIII	Ko'makchi	II
8.1	Sof ko'makchi	II0
8.2	Vazifadosh ko'makchi	II1
IX	Yuklama	Prt (Particle)
9.1	Tuzilishiga ko'ra	
9.1.1	So'z yuklama	PrtD
9.1.2	Qo'shimchasimon yuklama	PrtComp
9.2	Ma'no turi	
9.2.1	So'roq va taajjub yuklama	PrtQ
9.2.2	Kuchaytiruv-ta'kid yuklama	PrtEm (Emphasis)
9.2.4	Ayirish-chegalarash yuklama	PrtB
9.2.5	O'xshatish-qiyoslash yuklama	PrtCmp
9.2.6	Gumon yuklama	PrtPred

9.2.7	Inkor yuklama	PrtNeg
9.3	Vazifasiga ko'ra	
9.3.1	Sof yuklama	Prt0
9.3.2	Vazifadosh yuklama	Prt1
X	Modal so'z	MD
10.1	Ma'nosiga ko'ra	
10.1.1	Ishonch	MDTr (Trust)
10.1.2	Gumon	MDPred (Predicting)
10.1.3	Tasdiq ma'nosi	
10.1.4	Fikrni xulosalash	MDSum (Summary)
10.1.5	Fikrning tartibi	MDOr (Order)
10.1.6	Fikrning dalillanishi	MDEvd (Evidence)
10.1.7	Achinish	MDPt (Pity)
10.1.8	Quvonch	MDJ (Joy)
10.1.9	Tasdiq so'zlar	MDCnf (Confirmation)
10.1.10	Inkor so'zlar	MDDnl (Denial)
10.1.11	Mavjudlik	MDEx (Existence)
10.1.12	Zaruriyat	MDNec (Necessety)
10.1.13	Ruxsat	MDPer (Permission)
10.1.14	Esga olish	MDRm (Rememebering)
10.2	Tabiatiga ko'ra	
10.2.1	Sof	MD0
10.2.2	Vazifadosh	MD1
XI	Undov so'z	UH
11.1	His-hayajon undovlari	UHEm (Emotion)
11.1.1	Quvonch	UHJ
11.1.2	Erkalash-erkalanish	UHSp (Spoiling)
11.1.3	Rohatlanish	UHEj (Enjoyment)
11.1.4	Hayron qolish	UHSrp (Surprize)
11.1.5	Norozilik	UHDiss (Dissatisfaction)
11.1.6	G'azab-nafrat	UHH (Hate)
11.1.7	Og'riq	UHPn (Pain)
11.1.8	Do'q-po'pisa	UHTh (Threat)
11.1.9	Afsuslanish	UHReg (Regret)
11.1.10	Cho'chitish, qo'rqitish	UHF (Fear)
11.1.11	Ta'kid-kuchaytirish	UHEm (Emphasis)
11.2	Buyruq-xitob undovlari	UHImp (Imperative)
11.2.1	Shaxsga qaratilgan	UHPrs (Person)

11.2.2	Hayvon/parrandalarga qaratilgan	UHA _n (Animal)
11.2.3	Nutqiy-odat	UHD _{ly} (Daily)
XII	Taqlid soʻzlar	IM (Imitative)
12.1	Tovushga taqlid	IMV (Voice)
12.2	Holatga taqlid	IMCon

Xulosa

Xulosa qilib aytganda, oʻzbek tilshunosligida teglovchi dasturlar uchun teglarning kelishib olingan va meʼyorga solingan varianti eʼlon qilingunga qadar tadqiqotlarda oʻxshash va farqli teg-modellarni kuzatish mumkin. Taklif qilinayotgan teglar leksik yozuvlardan farq qilishi tabiiy. Hozirda nafaqat grammatik teglar, semantik teglar, sintaktik teglar, shu bilan birga uslubiy va prosodik teglar ham ishlab chiqilishi kerak. Tildagi umumiy qonuniyatlarni oʻrgangan holda, til modelini elektronlashtirish ustida ishlanmoqda. Tillarda yaratilayotgan korpuslar koʻp mehnat va vaqt talab qiladi va aynan mana shu korpuslarning ahamiyati va qiymatini tilda ishlab chiqilgan tegsetlar oshiradi. Biz tadqiqotimizda taklif qilgan tegset koʻplab ilmiy izlanishlar natijasi sifatida vujudga keldi va keyingi tadqiqotlar uchun, korpuslar uchun yechim boʻlib xizmat qiladi.

Adabiyotlar:

- Захаров В.П. Корпусная лингвистика. Санкт-Петербург, 2005. – С. 48.
- Leech G., Wilson A. Recommendations for the morphosyntactic annotation of corpora. // EAGLES Document EAG-TCWG-MAC/R, 1994. www.ilc.cnr.it/EAGLES/browse.html.
- McEnery T., Hardie A. Corpus linguistics: Method, theory and practice. Cambridge: Cambridge University Press, 2012.
- Корпусы: аннотированные и неаннотированные. Лингвистическая аннотация (разметка) и метаданные. — Текст: электронный // Myfilology.ru – информационный филологический ресурс: [сайт]. – URL: <https://myfilology.ru/177/korpusy-annotirovannye-i-neannotirovannye-lingvisticheskaya-annotacziya-razmetka-i-metadannye/> (дата обращения: 2.12.2022).
- [https://www.techtarget.com/whatis/definition/SGML-Standard-Generalized-Markup-Language#:~:text=SGML%20\(Standard%20Generalized%20Markup%20Language\)%20is%20a%20standard%20for%20how,It%20is%20metadata.](https://www.techtarget.com/whatis/definition/SGML-Standard-Generalized-Markup-Language#:~:text=SGML%20(Standard%20Generalized%20Markup%20Language)%20is%20a%20standard%20for%20how,It%20is%20metadata.)
- <https://ucrel.lancs.ac.uk/annotation.html#POS>
- <https://knowledge.autodesk.com/support/autocad-plant-3d/learn-explore/caas/CloudHelp/cloudhelp/2018/ENU/Plant3D-UserGuide/files/GUID-1E4D93B3-A6C0-4CF4-BA9A-23EC87522B54-htm.html>

- [https://e-library.namdu.uz/81%20%D0%A2%D0%B8%D0%BB%D1%88%D1%83%D0%BD%D0%BE%D1%81%D0%BB%D0%B8%D0%BA/%D0%9C%D0%B0%D1%88%D0%B8%D0%BD%D0%B0%20%D1%82%D0%B0%D1%80%D0%B6%D0%B8-D0%BC%D0%B0%D1%81%D0%B8/1mitkov_ruslan_editor_the_oxford_handbook_of_computational_li%20\(1\).pdf](https://e-library.namdu.uz/81%20%D0%A2%D0%B8%D0%BB%D1%88%D1%83%D0%BD%D0%BE%D1%81%D0%BB%D0%B8%D0%BA/%D0%9C%D0%B0%D1%88%D0%B8%D0%BD%D0%B0%20%D1%82%D0%B0%D1%80%D0%B6%D0%B8-D0%BC%D0%B0%D1%81%D0%B8/1mitkov_ruslan_editor_the_oxford_handbook_of_computational_li%20(1).pdf)
- <https://ucrel.lancs.ac.uk/annotation.html#POS>
- <https://ucrel.lancs.ac.uk/claws1tags.html>
- <https://ucrel.lancs.ac.uk/papers/ClawsWordTaggingSystemRG87.pdf>
- <https://ucrel.lancs.ac.uk/claws8tags.pdf>
- <https://www.sketchengine.eu/penn-treebank-tagset/>
- <https://www.ims.uni-stuttgart.de/en/research/projects/textkorpora-werkzeuge/>
- <https://www.sketchengine.eu/english-treetagger-pipeline-2/#toggle-id-1>
- <https://www.learntek.org/blog/categorizing-pos-tagging-nltk-python/>
- https://ucrel.lancs.ac.uk/bnc2sampler/guide_c7.htm#m2num
- <https://static.googleusercontent.com/media/research.google.com/ru//pubs/archive/37072.pdf>
- <https://universaldependencies.org/u/pos/>
- Ҳамроева Ш. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари: Монография. – Тошкент, 2020. – 229 б.
- Холиёров О. О'zbek tili ta'limiy korpusini tuzishning lingvistik asoslari. Filol. fan. bo'yicha falsafa doktori PhD avtoref. – Termiz, 2021.
- Abdullayeva O. O'zbek tilining internet axborot matnlari korpusini shakllantirishning nazariy va amaliy asoslari. filol. fan. bo'yicha falsafa doktori (PhD) diss. Andijon, 2022.
- Abdullayeva O., Bozorqulova O'. Korpus lingvistikasida lingvistik tahlillar: muammo va yechimlar // O'zbek tili tadqiqi va ta'limi muammolari. Toshkent, 2021.27.10. 296-302-b.
- Abdullayeva O.X. O'zbek tili korpusini yaratish bosqichlari va muammoli jihatlari // So'z san'ati. – Toshkent, 2021. ISSN- 2181-9297, 4-jild, 1-son. – B. 153-158.

Evaluation of part of speech tagging in uzbek language: problems and proposals

Botir Elov¹

Shahlo Hamroyeva²

Oqila Abdullayeva³

Mohiyaxon Uzoqova⁴

Abstract:

Speaking of a language corpus, the issue of building a linguistic database becomes the subject of concern because of its complexity and importance at the same time. The process of assigning appropriate identifiers to speech fragments in corpus texts is problematic since language modeling is associated with the rules and patterns of tagging existing in the language. Tagging, especially grammatical tagging or PoS tagging, is also a topical issue for Uzbek corpus linguistics. Because a special “encoded” symbol system serves as the primary key in solving NLP problems related to the Uzbek language. The article analyzes the studies of tagging and PoS tagging in world linguistics and considers the current tagging process in Uzbek linguistics. Based on the rules of the Uzbek language, an alternative

¹Elov Botir Boltayevich – doctor of philosophy in technical sciences (PhD), associate professor. Head of the Department of Computer Linguistics and Digital Technologies of Tashkent State University of Uzbek Language and Literature named after Alisher Navoi.

E-mail: elov@navoiy-uni.uz

ORCID: 0000-0001-5032-6648

²Hamroyeva Shahlo Mirdjanovna – doctor of philology(DSc), associate professor of the Department of computer linguistics and digital technologies of Tashkent State University of Uzbek language and Literature named after Alisher Navoi.

E-mail: shaxlo.xamrayeva@navoiy-uni.uz

ORCID: 0000-0002-5429-4708

³Abdullayeva Oqila Xolmo‘minovna – doctor of philosophy in philology(PhD), Senior teacher. Senior lecturer of the Department of Computer Linguistics and Digital Technologies of Tashkent State University of Uzbek Language and Literature named after Alisher Navoi.

E-mail: abdullayeva.oqila@navoiy-uni.uz

ORCID: 0000-0002-2524-4832

⁴Uzoqova Mohiyaxon Tuyg‘un qizi – Graduate student of the Department of Computer Linguistics and Digital Technologies of Tashkent State University of Uzbek Language and Literature named after Alisher Navoi.

E-mail: uzoqovamohiyaxon@navoiy-uni.uz

ORCID: 0000-0001-7102-0824

For reference: Elov B., Hamroyeva Sh., Abdullayeva O., Uzoqova M. 2022. “Evaluation of Part of speech tagging in uzbek language: problems and proposals”. *Uzbekistan: language and culture. Applied philology.* 2 (5): 51-68.

set of tags was proposed taking sets of tags widely used in the world into consideration.

Keywords: *Tag, markup, annotation, tagset, NLP, corpus, Brown's corpus, CLAWS, PoS tagging.*

References:

- Zaxarov V.P. Korpusnaya lingvistika. Sankt-Peterburg, 2005. – 48 s.
- Leech G., Wilson A. Recommendations for the morphosyntactic annotation of corpora. // EAGLES Document EAG-TCWG-MAC/R, 1994. www.ilc.cnr.it/EAGLES/browse.html.
- McEnery T., Hardie A. Corpus linguistics: Method, theory and practice. Cambridge: Cambridge University Press, 2012.
- Korpusi: annotirovannie i neannotirovannie. Lingvisticheskaya annoatatsiya (razmetka) i metadannie. — Tekst: elektronniy // Myfilology.ru – informatsionniy filologicheskiiy resurs: – URL: <https://myfilology.ru//177/korpusy-annotirovannye-i-neannotirovannye-lingvisticheskaya-annotatsiya-razmetka-i-metadannye/> (data obrasheniya: 2.12.2022).
- [https://www.techtarget.com/whatis/definition/SGML-Standard-Generalized-Markup-Language#:~:text=SGML%20\(Standard%20Generalized%20Markup%20Language\)%20is%20a%20standard%20for%20how,It%20is%20metadata.](https://www.techtarget.com/whatis/definition/SGML-Standard-Generalized-Markup-Language#:~:text=SGML%20(Standard%20Generalized%20Markup%20Language)%20is%20a%20standard%20for%20how,It%20is%20metadata.)
- <https://ucrel.lancs.ac.uk/annotation.html#POS>
- <https://knowledge.autodesk.com/support/autocad-plant-3d/learn-explore/caas/CloudHelp/cloudhelp/2018/ENU/Plant3D-UserGuide/files/GUID-1E4D93B3-A6C0-4CF4-BA9A-23EC87522B54-htm.html>
- [https://e-library.namdu.uz/81%20%D0%A2%D0%B8%D0%BB%D1%88%D1%83%D0%BD%D0%BE%D1%81%D0%BB%D0%B8%D0%BA/%D0%9C%D0%B0%D1%88%D0%B8%D0%BD%D0%B0%20%D1%82%D0%B0%D1%80%D0%B6%D0%B8%D0%BC%D0%B0%D1%81%D0%B8/1mitkov_ruslan_editor_the_oxfordhandbook_of_computational_li%20\(1\).pdf](https://e-library.namdu.uz/81%20%D0%A2%D0%B8%D0%BB%D1%88%D1%83%D0%BD%D0%BE%D1%81%D0%BB%D0%B8%D0%BA/%D0%9C%D0%B0%D1%88%D0%B8%D0%BD%D0%B0%20%D1%82%D0%B0%D1%80%D0%B6%D0%B8%D0%BC%D0%B0%D1%81%D0%B8/1mitkov_ruslan_editor_the_oxfordhandbook_of_computational_li%20(1).pdf)
- <https://ucrel.lancs.ac.uk/annotation.html#POS>
- <https://ucrel.lancs.ac.uk/claws1tags.html>
- <https://ucrel.lancs.ac.uk/papers/ClawsWordTaggingSystemRG87.pdf>
- <https://ucrel.lancs.ac.uk/claws8tags.pdf>
- <https://www.sketchengine.eu/penn-treebank-tagset/>
- <https://www.ims.uni-stuttgart.de/en/research/projects/textkorpora-werkzeuge/>
- <https://www.sketchengine.eu/english-treetagger-pipeline-2/#toggle-id-1>
- <https://www.learn2tek.org/blog/categorizing-pos-tagging-nltk-python/>
- https://ucrel.lancs.ac.uk/bnc2sampler/guide_c7.htm#m2num
- <https://static.googleusercontent.com/media/research.google.com/ru//>

Hamroyeva Sh. O'zbek tili mualliflik korpusini tuzishning lingvistik asoslari: Monografiya. – Toshkent, 2020. – 229 b.

Xoliyorov O'. O'zbek tili ta'limiy korpusini tuzishning lingvistik asoslari. Filol. fan. bo'yicha falsafa doktori PhD avtoref. – Termiz, 2021.

Abdullayeva O. O'zbek tilining internet axborot matnlari korpusini shakllantirishning nazariy va amaliy asoslari. filol. fan. bo'yicha falsafa doktori (PhD) diss. Andijon, 2022.

Abdullayeva O., Bozorqulova O'. Korpus lingvistikasida lingvistik tahlillar: muammo va yechimlar // O'zbek tili tadqiqi va ta'limi muammolari. Toshkent, 2021.27.10. 296-302-b.

Abdullayeva O.X. O'zbek tili korpusini yaratish bosqichlari va muammoli jihatlari // So'z san'ati. – Toshkent, 2021. ISSN- 2181-9297, 4-jild, 1-son. – B. 153-158.