O‘ZBEK TILI KORPUSI MATNLARINI POS TEGLASH USULLARI

Botir Elov Boltayevich
Texnika fanlari bo‘yicha falsafa doktori PhD, dotsent
elov@navoiy-uni.uz
ToshDO‘TAU Kompyuter lingvistikasi va raqamli texnologiyalar kafedrasi mudiri
Nizomaddin Xudayberganov Uktambay o‘g‘li
nizomaddin@navoiy-uni.uz
ToshDO‘TAU Kompyuter lingvistikasi va raqamli texnologiyalar kafedrasi o‘qituvchisi

Annotatsiya. Til korpusi qurilishida lingvistik ta’minot masalasi muhim va murakkab hisoblanadi. Korpuslarda matnlardagi nutq bo‘laklariga mos identifikatorini 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. Ushbu maqolada jahonda keng qo‘llanib kelinayotgan tegsetlardan foydalanib o‘zbek tili korpusi matnlarini POS teglash usullari ko‘rib chiqiladi.

Abstract. In the construction of the language corpus, the issue of linguistic support is important and complex. The process of assigning appropriate identifiers to speech fragments in texts in corpora is problematic, because language modeling is related to tagging rules and regularities in the language. Tagging, especially grammatical tagging or PoS tagging, is also a relevant issue for Uzbek corpus linguistics. In this article, the methods of POS tagging of texts of the Uzbek language corpus using tagsets that are widely used in the world are considered.

Аннотация. При построении языкового корпуса важным и сложным является вопрос языкового обеспечения. Процесс присвоения соответствующих идентификаторов речевым фрагментам в текстах в корпусах проблематичен, поскольку моделирование языка связано с маркировкой правил и закономерностей в языке. Маркировка, особенно грамматическая маркировка или маркировка PoS, также является актуальной проблемой для узбекской корпусной лингвистики. В данной статье рассмотрены методы POS-разметки текстов корпуса узбекского языка с использованием наборов тегов, которые широко используются в мире.

Kalit so‘zlar: *Teg, razmetka, annotatsiya, tegset, NLP, korpus, PoS teglash.*

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 ishlatilgan.

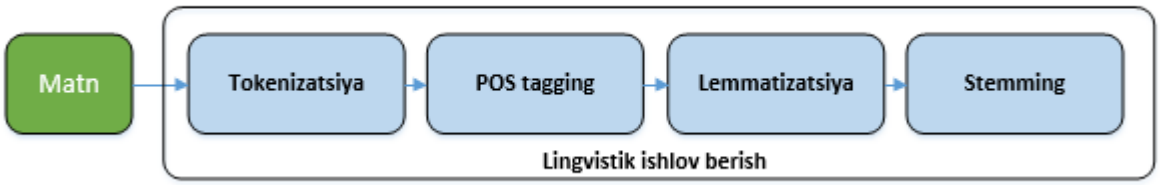
Annotatsiya (annotation) – Korpus lingvistikasi doirasida qaralganda berilgan matnga bevosita aloqasi bo‘lmagan, ammo uning qaysidir qismi haqida

lingvistik yoki ekstralingvistik axborot beruvchi umumiy ma’lumot. Annotatsiya o‘z ichiga metama’lumot va teglarni qamrab olishi mumkin.

Razmetka (markup) – razmetka va meta ma’lumotni annotatsiyalash jarayonining bir qismi sifatida baholanadi. Jahon amaliyotida razmetkalashning standart prinsiplari ishlab chiqilgan.

Teg – Kompyuter yordamida matn tahlilini amalga oshirish jarayonini tezlashtirish va osonlashtirishga xizmat qiluvchi shartli belgi yoki maxsus kod. Teglar bir necha turlarga bo‘linadi: *semantik teg*, *sintaktik teg* va *grammatik teg*. Grammatik teg, shuningdek, PoS (Part of speech) tegging nomi bilan ham mashhur.

POS teglash – bu berilgan gapdagi har bir so‘zshaklga uning turkum (*ot, fe’l, sifat, son, ravish* yoki *olmosh)*ga mansubligini belgilash (teglash) vazifasidir. POS teglash tabiiy tilni qayta ishlash (Natural Language Processing, NLP)ning asosiy vazifalaridan biri bo‘lib, pipeline konveyerining muhim bosqichi hisoblanadi (1-rasm).



1-rasm. Matnga boshlang‘ich ishlov berish bosqichlari

Tabiiy tilni qayta ishlash (NLP) vositalari axborot-kommunikatsiya texnologiyalarining jadal rivojlanishi tufayli katta qiziqish uyg‘otdi. Natijada bugungi kunda turli xil NLP usullari va vositalari yaratilmoqda. Biroq, tabiiy tillardagi matnlarni qayta ishlaydigan samarali NLP vositalarini ishlab chiqishda lingvistik teglash kabi vazifalarni hal qilish lozim.

Lingvistik teglash tavsiflovchi yoki analitik belgilarni til ma’lumotlari bilan bog‘lashni o‘z ichiga oladi. Strukturlanmagan (qayta ishlanmagan) ma’lumotlar matnli yoki har qanday manba yoki janrdan olingan yoki vaqt funksiyalari (audio, video va/yoki fiziologik yozuvlar) shaklida bo‘lishi mumkin. Teglar barcha turdagi transkripsiyalarni (fonetik xususiyatdan nutqiy qurilmagacha), POS teglash, ma’no belgilari, sintaktik tahlil, NER obyektlar, semantik rol belgilari, vaqt va hodisalarni aniqlash, so‘zlarning sintaktik zanjirlarini, nutq darajasini o‘z ichiga olishi mumkin.

Til korpuslari ishlab chiqilgandan so‘ng ularni teglash lozim. Jumladan, NLP tadqiqotchilari tomonidan **Brown korusi** matnlari POS teglangan [*Maverick, 1969: 35(1)]. Ushbu tadqiqiot K.V.Gruch [Church, 1989)], S.J.Derose[Derose, 1988: 14(1)], R.Garside[Garside, 1987*] va B.B.Greenlarning POSteglash bo‘yicha ilmiy tadqiqotlarida foydalanilgan [*Greene, Rubin, 1971:23*]. Brown korpusi singari 70-80-yillarda ishlab chiqilgan korpuslar odatda POS teglangan, ammo samarali

avtomatik usullarning yo‘qligi va qo‘lda teglashning murakkabligi boshqa til hodisalari uchun teglarni o‘z ichiga olgan yetarlicha katta hajmdagi korpuslarni ishlab chiqishga imkon bermagan.

1980-yillarning oxirida yangi katta hajmli til ma’lumotlarining mavjudligi lingvistik teglash tizimlarining ko‘payishiga olib keldi. Asosan, POS yoki morfo-sintaktik teglashlashga asoslangan va avtomatik teglash uchun statistik usullar ishlab chiqildi. Ushbu turdagi birinchi yirik sa’y-harakatlar ingliz tilidagi bir million so‘zdan iborat **Lancaster-Oslo-Bergen (LOB) korpusi** morfo-sintaktik va sintaktik teglar asosida teglangan[*Beale, 1985*].

Ushbu ilmiy tadqiqotga asoslanib, Penn Treebank loyihasi asosida Wall Street Journal maqolalarining bir million so‘zdan iborat korpus ishlab chiqilgan [*Marcus, Santorini, Marcinkiewicz, 1993*] va POS hamda sintaktik teglangan [*Marcus, Kim, Marcinkiewicz, MacIntyre, Bies, Ferguson, Katz, Schasberger, 1994*]. 1990-yillardagi avtomatik teglangan korpuslarga 1994-yilda ishlab chiqilgan 100 million so‘zli British National Corpus [*Erjaveç, 1998*]; MULTEXT ko‘p tilli korpusi[*Ide, Véronis, 1994: 588-592*]; PAROLE va SIMPLE korpuslari [*Kolodnytsky, Bernsen, Dybkjær, 2004*] o‘n to‘rtta Yevropa tilidagi ma’lumotlar va ularning POS teglarini o‘z ichiga olgan.

Matn alohida tokenlarga ajratilgandan so‘ng har bir token yoki tokenlar to‘plamini **so‘z turkumi** (ot, fe’l, olmosh va boshqalar) bilan belgilash mumkin. Bu vazifa NLPda **so‘z turkumini aniqlash va teglash** yoki **PoS teglash**dir. PoS teglash morfologik tahlil sanaladi, chunki o‘zakka qo‘shilgan grammatik shakl orqali so‘z turkumini aniqlashga yordam beradi. Misol uchun, morfologik teglash jarayoni *-gan, -di* qo‘shimchalarining fe’llar bilan qo‘llaniladigan morfologik birlik ekanligini anglatadi (istisnolar ham mavjud). Biroq morfologik (grammatik) shaklni olmagan so‘zlar ham teglanadi (2-jadval). Masalan, *mergan* – **ot**, *bergan* – **fe’l**. Bu so‘zlardagi *-gan* alohida birinchi so‘zda morfologik birlik emas, balki so‘zning bir bo‘g‘inidir. Bu kabi shakllar o‘zbek tilida uchrab turadi. Quyidagi 1-jadvalda shunday birliklarga misol keltiramiz:

1-jadval. O‘zbek tilidagi so‘zlarni POS teglash namunasi

№	So‘z	O‘zak	POS teg	Izoh
1.	mergan	mergan	Ot (N)	"mergan" alohida leskema, tarkibiy qismlarga bo‘linmaydi
2.	bergan	ber	Fe’l (VB)	"ber"+gan; *-gan* o‘tgan zamon yoki sifatdosh qo‘shimchasi; leksema + grammatik shakl
3.	olov	olov	Ot (N)	"olov" alohida leskema, tarkibiy qismlarga bo‘linmaydi
4.	ulov	ula	Fe’l (VB)	"ula"+v; *-v* so‘z yasovchi qo‘shimcha; leksema + grammatik shakl
5.	ming	ming	Ot (N) | Son (Num)	"ming" alohida leskema, tarkibiy qismlarga bo‘linmaydi

6.	ahvoling	ahvol	Ot (N)	"ahvol"+ing; *-ing* egalik qo‘shimchasi: leksema + grammatik shakl
7.	boisi	boisi	Ot (N)	"boisi" alohida leskema, tarkibiy qismlarga bo‘linmaydi
8.	ukasi	uka	Ot (N)	"uka"+si; *-si* egalik qo‘shimchasi

Mazkur jadvaldagi *mergan, olov, ming, boisi* kabi birliklar alohida leksema sanalib, tarkibiy qismlarga bo‘linmaydi, bu so‘zlardagi qismlar qo‘shimcha emas, o‘zakning bir qismi, boshqa holatda o‘zakka qo‘shilgan grammatik shakl sanaladi. Matn gaplarga bo‘lingandan so‘ng har bir gapdan *otli birikma* va *fe’lli birikma* kabi tarkibiy so‘z birikmalari aniqlanadi. Bu tahlil sintaktik tahlil bo‘lib, **lingvistik daraja (sintaksis)** va **annotatsiya vazifasi (sintaktik parsing)** o‘rtasida aniq muvofiqlik mavjud. Bugungi kunda lingvistik tahlil bosqichi bo‘yicha ko‘plab lingvistik nazariyalar mavjud bo‘lib, o‘zbek tili uchun hozirda keng qo‘llaniladigan **so‘z birikmalari modeli** yondashuvidan foydalanish maqsadga muvofiq.

POS teglash vazifasida matndagi har bir token yoki tokenlar ketma-ketligi POS yorlig‘i bilan belgilanadi. O‘zbek tilida 12 ta so‘z turkumi mavjud va ularning POS-tegi belgilangan (qarang: 2-jadval).

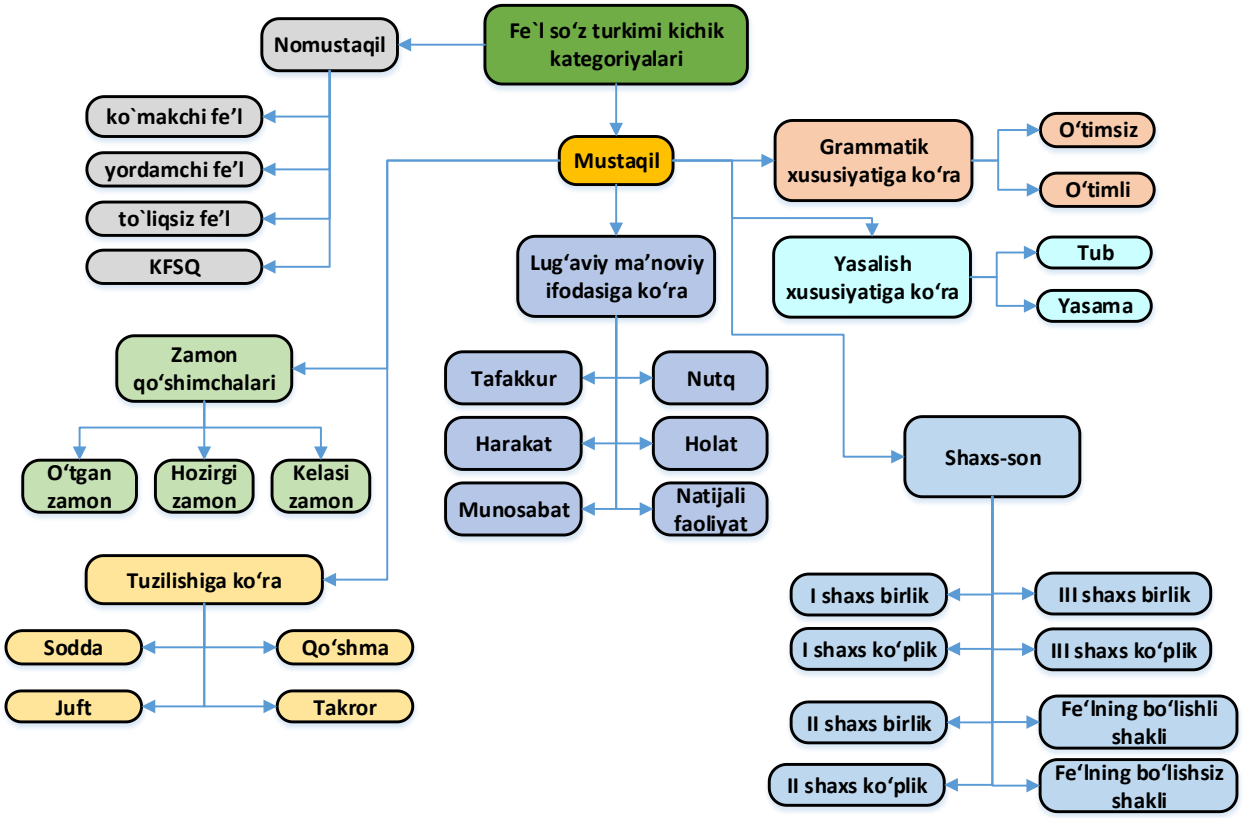
2-jadval. O‘zbek tilidagi POS teglar ro‘yxati (qisqa variant)

№	POS teg	Ingliz	Belgilanishi
1.	**Ot**	Noun	N
2.	**Sifat**	Adjective	JJ
3.	**Fe’l**	Verb	VB
4.	**Son**	Number	NUM
5.	**Ravish**	Adverb	RR
6.	**Olmosh**	Pronoun	P
7.	**Bog‘lovchi**	Conjunction	C
8.	**Ko‘makchi**	preposition	II
9.	**Yuklama**	particle	Prt
10.	**Modal so‘z**	modal words	MD
11.	**Undov so‘z**	exclamatory words	UH
12.	**Taqlid so‘zlar**	imitation words	IM

Turli tadqiqotchilar tomonidan boshqa ko‘plab muqobil tasniflar taklif qilingan. B.Elov va Sh.Hamroyevalarning “O‘zbek tilida POS tegging masalasi: muammo va takliflar” [*Elov, Hamroyeva, Abdullayeva, Uzakova, 2022: 51-68*] nomli maqolasida o‘zbek tili birliklarini POS teglash uchun teg tizimi ishlab chiqilgan keltirilgan, o‘zbek tili uchun POS teglarning to‘liq (kengaytirilgan varianti) “POS Tagging of uzbek texts using hidden Markov models (HMM) and Viterbi algorithm” [*Elov, Hamroyeva, Xudayberganov, Yodgorov, Yuldashev, 2023*] nomli maqolada o‘zbek tili birliklarini POS teglash uchun teg tizimidan foydalanish

tavsiflangan. Shuningdek, “Agglutinativ tillar uchun pos teglash va stemming masalasi (turk, uyg‘ur, o‘zbek tillari misolida)” [*Elov, Hamroyeva, Abdullayeva, Husainova, Xudayberganov, 2023: 6-39*] nomli maqolada agglutinativ tillarni teglashda turk, uyg‘ur va o‘zbek tillari teg timizi qiyoslangan. “O‘zbek tilidagi sodda gaplarning sintaktik teglangan bazasi orqali tahlil daraxtini qurish” [*Ramatova, 2023*] nomli maqolada o‘zbek tili birliklarini POS teglash uchun faqat mustaqil so‘z turkumlari POS teglari taklif qilingan. Quyida o‘zbek tili leksik birliklarini POS teglashda ahamiyatli bo‘lgan grammatik xususiyatlarni izohlaymiz.

Tilshunoslikda bu asosiy kategoriyalar yana kichik kategoriyalarga bo‘linadi. Masalan, fe’l so‘z turkumini teglashda, uning fe’l ekanligidan tashqari boshqa grammatik xususiyatlarini ham ko‘rsatish mumkin (3,4-rasm).

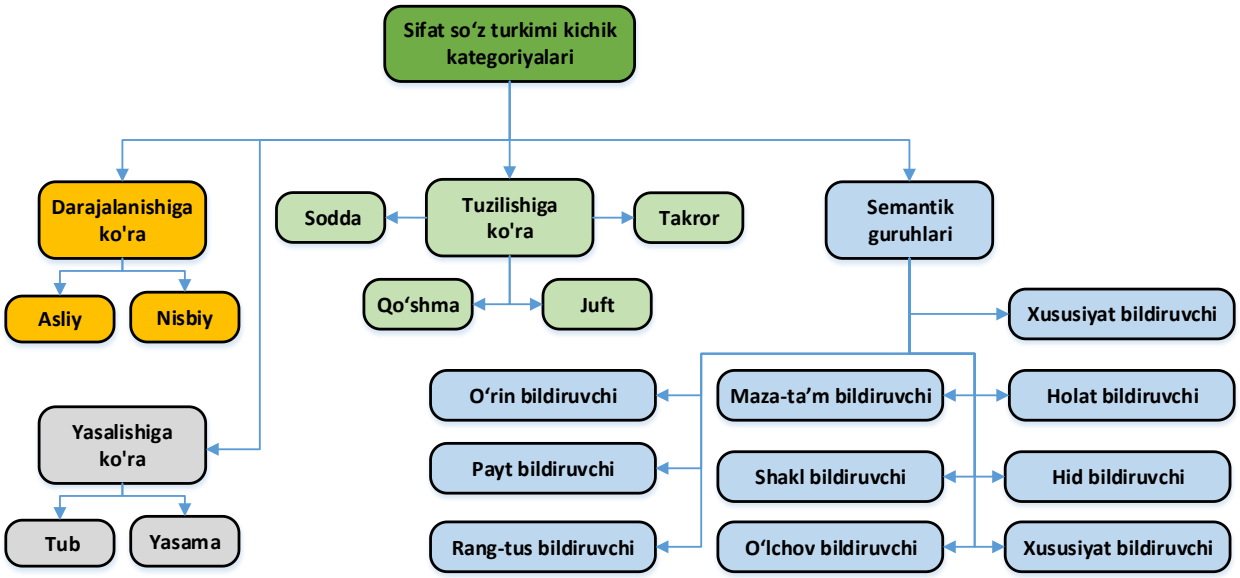


1-rasm. Fe’l so‘z turkimi grammatik kategoriyalari tasnifi

Fe’l so‘z turkumini morfoligik (POS) teglash quyidagicha amalga oshiriladi:

Og‘rib qoldi: *[fe’l], [mustaqil fe’l], [holat fe’li], [o‘timsiz fe’l], [bo‘lishli fe’l], [aniq nisbat], [sodda fe’l], [tub fe’l], [o‘tgan zamon], [III shaxs birlik], [har. t.], [yetakchi fe’l], [ko‘makchi fe’l].*

Bu yerda grammatik kategoriyalar to‘liq keltirildi, ammo teglash jarayonida kengaytirilgan teglar tizimidan foydalaniladi. Kengaytirilgan teglar tizimi “O‘zbek tilida POS tegging masalasi: muammo va takliflar” nomli maqolada batafsil keltirilgan [*Elov, Hamroyeva, Abdullayeva, Uzakova, 2022: 51-68*].



2-rasm. Sifat so‘z turkimi grammatik kategoriyalari tasnifi

Sifatni teglash umumiy grammatik ma’noni belgilashdan boshlanadi. Sifat belgining xususiyatini ifodalashiga ko‘ra ikkiga bo‘lingan: 1) asliy sifat; 2) nisbiy sifat. Sifat turkumida sintaktik kategoriya keng bo‘lmagan voqelanishga ega. Sifat turkumi egalik kategoriyasi UGMsini “keyingi sifatni oldingi so‘zga bog‘lash, mansublik, xoslik ma’nosini ifodalash” tarzida xususiylashtiradi. Kesimlik kategoriyasi esa sifat turkumida o‘z mohiyatini cheklangan darajada namoyon qiladi. Ko‘p hollarda sifat turkumida kesimlik kategoriyasi bog‘lama vositasida yuzaga chiqadi.

Kengaytirilgan teglar tizimi asosida sifat so‘z turkumining teglanishi quyidagicha bo‘ladi:

- **o‘lik:** *[sifat], [nisbiy sifat], [oddiy daraja], [yasama sifat], [sodda sifat], [jismoniy harakat LMG].*
- **tuzuk:** *[sifat], [asliy sifat], [oddiy daraja], [tub sifat], [yasama sifat], [xususiyat LMG].*
- **yiroq:** *[sifat], [asliy sifat], [oddiy daraja], [sodda sifat], [tub sifat], [xususiyat LMG].*
- **yo‘g‘onroq:** *[sifat], [asliy sifat], [qiyosiy daraja], [sodda sifat], [tub sifat], [shakl LMG].*

Xulosa

Til korpusi matnlari ustida o‘tkazilgan turli tajribalar shuni ko‘rsatadiki, o‘zak ma‘lumotlarini sintaktik vazifa bilan birlashtirish morfologik jihatdan boy til uchun POS teglash natijasini yaxshilaydi, bu esa NLP vazifasining hal qilish samaradorligini oshirishga xizmat qiladi. O‘zbek tili korpusi matnlarini POS teglash vazifasining o‘ziga xos jihatlari, POS teglar kategoriyalari va kichik kategoriyalar ro‘yxati keltirildi. Jumladan fe’l va sifat so‘z turkumlari grammatik kategoriyalari tasnifi va namunalar izohlandi. Maqola keltirilgan POS teglash usullaridan o‘zbek

tili morfologik analizatorni ishlab chiqishda va boshqa murakkab NLP vazifalarni hal qilishda foydalanish mumkin.

Foydalanilgan adabiyotlar:

1. Maverick, G. v. (1969). Computational Analysis of Present-Day American English. Henry Kučera, W. Nelson Francis. *International Journal of American Linguistics*, *35*(1). https://doi.org/10.1086/465045

2. Church, K. W. (1989). Stochastic parts program and noun phrase parser for unrestricted text. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2*. https://doi.org/10.3115/974235.974260

3. Derose, S. J. (1988). GRAMMATICAL CATEGORY DISAMBIGUATION BY STATISTICAL OPTIMIZATION. *Computational Linguistics*, *14*(1).

4. Garside, R. (1987). The CLAWS word-tagging system. *The Computational Analysis of English: A Corpus-Based Approach*.

5. Greene, B.B., Rubin, G.M.: Automatic Grammatical Tagging of English. Brown University, Department of Linguistics (1971)

6. Beale, A. D. (1985). Grammatical analysis by computer of the Lancaster-oslo/bergen (lob) corpus of british English texts. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, *1985-July*. https://doi.org/10.3115/981210.981246

7. Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2). https://doi.org/10.1162/coli.2010.36.1.36100

8. Marcus, M., Kim, G., Marcinkiewicz, M.A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., Schasberger,B.: The penn treebank: annotating predicate argument structure. In: Proceedings of the Workshop on Human Language Technology, pp. 114–119. Association for Computational Linguistics, Stroudsburg, PA, USA (1994)

9. Erjaveç, T., Ide, N.: The MULTEXT-East corpus. In: Proceedings of First International Conference on Language Resources and Evaluation, pp. 971–974 (1998)

10. Ide, N., Véronis, J. MULTEXT: multilingual text tools and corpora. In: Proceedings of the 15th International Conference on Computational Linguistics (COLING 94), vol. I, pp. 588–592. Kyoto, Japan (1994)

11. Kolodnytsky, M., Bernsen, N. O., & Dybkjær, L. (2004). A visual interface for a multimodal interactivity annotation tool: Design issues and implementation solutions. Proceedings of the Workshop on Advanced Visual Interfaces AVI. https://doi.org/10.1145/989863.989937

12. Elov B., Hamroyeva Sh., Abdullayeva O., Uzakova M. O‘zbek tilida POS tegging masalasi: muammo va takliflar / Uzbekistan: Language and Culture. Applied philology. 2022/2(5). – 51-68-b.

13. Elov B., Hamroyeva Sh., Xudayberganov N., Yodgorov U., Yuldashev A. Pos tagging of Uzbek texts using hidden Markov models (HMM) and Viterbi algorithm. O‘zMU xabarlari. Mirzo Ulug‘bek nomidagi O‘zbekiston Milliy universiteti ilmiy jurnali. 2023 yil Maxsys son.

14. Elov, B.B., Hamroyeva, Sh.M., Abdullayeva, O.X., Husainova, Z.Y., Xudayberganov, N.U. 2023. “Agglutinativ tillar uchun pos teglash va stemming masalasi (turk, uyg‘ur, o‘zbek tillari misolida)”. O‘zbekiston: til va madaniyat 2: 6-39.

15. Ramatova M. O‘zbek tilidagi sodda gaplarning sintaktik teglangan bazasi orqali tahlil daraxtini qurish // Educational Research in Universal Sciences. VOLUME 2 | ISSUE 12 | 2023. // https://zenodo.org/records/10463799.