

O‘ZBEK TILI KORPUSIGA MORFOLOGIK ISHLOV BERISH

Nizomaddin Xudayberganov

nizomaddin@navoiy-uni.uz

ToshDO‘TAU o‘qituvchisi

Annotatsiya. Til korpusini morfologik tahlil qilish uchun uni morfologik teglash eng muhim NLP vazifalaridan biri hisoblanadi, chunki u korpusni sintaktik, semantik va pragmatik tahlil qilish uchun muhim qadam hisoblanadi. O‘zbek tilidagi til birliklarining morfologik teglash qo‘lda va yarim avtomatik tarzda amalga oshiriladi. Korpus matnlarini teglash chatbotlar, virtual yordamchilar, qidiruv tizimlari va mashina tarjimai kabi tabiiy tilni qayta ishlash (NLP) ilovalarida asosiy amal hisoblanib, u sun‘iy intellekt modellariga inson tili va muloqotni yanada samaraliroq tushunishga yordam beradi. Korpus matnlarini annotatsiyalash yoki teglash orqali so‘z, ibora, so‘z birikmasi shu kabi leksik birliklarni izohlash kabi matnga teg (izoh) qo‘shish va ularni belgilashdan iborat. Mashinali o‘rgatishda matn annotatsiyasi – bu mashinali o‘rgatish modellariga inson tilini tushunish va qayta ishlashda yordam berish uchun matnga metama’lumotlar yozish jarayoni. Yuqori sifatli ma’lumotlar to‘plami sun‘iy intellektni o‘qitishda juda muhim, chunki ular AI modellarining ishlashi, aniqligi va ishonchliligiga bevosita ta’sir qiladi. Matnni teglash orqali kalit so‘zlarni ajratib olish evaziga qidiruv tizimi natijasini optimallashtirish mumkin. ToshDO‘TAU Kompyuter lingvistikasi va raqamli texnologiyalari kafedrasida olimlari tomonidan til korpusi uchun maxsus teglar to‘plami ishlab chiqilgan bo‘lib, o‘zbek tilidagi so‘zlardagi gap bo‘laklari va qo‘shimchalari birikmalarining eng keng tarqalgan affiksalarini aniqlanib, ularni belgilangan teglar asosida tahlil qilish, til korpusining yanada rivojlanishiga asos bo‘lib xizmat qilishi mumkin. Til korpusini morfologik teglash uchun 12 dan ortiq maxsus teglar tanlab olindi va korpusda nutq birliklari teglandi.

Abstract. Morphological tagging of a language corpus for morphological analysis is one of the most important NLP tasks, as it is an important step for syntactic, semantic and pragmatic analysis of the corpus. Morphological tagging of Uzbek language units is done manually and semi-automatically. Corpus text tagging is a key practice in natural language processing (NLP) applications such as chatbots, virtual assistants, search engines, and machine translation, where it helps AI models understand human language and communication more effectively. Annotating or tagging corpus texts consists of adding a tag (explanation) to the text and defining them, such as explaining words, phrases, phrases, and similar lexical units. Text annotation in machine learning is the process of writing metadata to text to help machine learning models understand and process human language. High-quality data sets are critical to AI training, as they directly affect the performance, accuracy, and reliability of AI models. By tagging text, you can optimize search engine results by extracting keywords. ToshDO‘TAU scientists of the Department of Computer

Linguistics and Digital Technologies have developed a set of special tags for the language corpus, identifying the most common affixes of phrases and adverbs in Uzbek words, analyzing them based on the specified tags, for the further development of the language corpus. can serve as a basis. More than 12 special tags were selected for morphological tagging of the language corpus and speech units were tagged in the corpus.

Аннотация. Морфологическая разметка языкового корпуса для морфологического анализа является одной из важнейших задач NLP, поскольку является важным этапом синтаксического, семантического и прагматического анализа корпуса. Морфологическая разметка единиц узбекского языка осуществляется вручную и в полуавтоматическом режиме. Тегирование корпуса текста — это ключевая практика в приложениях обработки естественного языка (NLP), таких как чат-боты, виртуальные помощники, поисковые системы и машинный перевод, где они помогают моделям AI более эффективно понимать человеческий язык и общение. Аннотирование или маркировка корпусных текстов заключается в добавлении к тексту тега (пояснения) и его определении, например пояснении слов, словосочетаний, словосочетаний и подобных лексических единиц. Текстовые аннотации в машинном обучении — это процесс записи метаданных в текст, который помогает моделям машинного обучения понимать и обрабатывать человеческий язык. Наборы высококачественных данных имеют решающее значение для обучения AI, поскольку они напрямую влияют на производительность, точность и надежность моделей AI. Помечая текст тегами, вы можете оптимизировать результаты поисковых систем, извлекая ключевые слова. Ученые ТошДО'ТАУ кафедры компьютерной лингвистики и цифровых технологий разработали набор специальных тегов для языкового корпуса, определяющих наиболее распространенные аффиксы словосочетаний и наречий в узбекских словах, анализируя их на основе указанных тегов, для дальнейшего развития языкового корпуса. может служить основой. Для морфологической разметки языкового корпуса было выбрано более 12 специальных тегов и в корпусе были размечены речевые единицы.

Kalit soʻzlar: *Til korpusi, matnni qayta ishlash, lingvistik tahlil, mashinali oʻqitish modellari, korpus turlari, parallel korpus, teglangan korpuslar, POS teg.*

Kirish

Morfologik teglangan korpus ilmiy tadqiqotchilariga turli xil lingvistik tahlil va tabiiy tilni qayta ishlash (NLP)ni amalga oshirish imkonini beradi. Morfologik ishlov berilgan/teglangan korpusdan tabiiy tildagi soʻz shakllari, nutq qismlari (soʻz turkumlari) va grammatik xususiyatlarning chastotasi, tarqalishi va birikmasi haqidagi maʼlumotlarni aniqlash uchun foydalanish mumkin. Shuningdek, morfologik tahlil, generatsiya, segmentatsiya va tildagi noaniqlik kabi vazifalar

uchun modellar va algoritmlarni o'rgatish va sinab ko'rish uchun ham qo'llaniladi. Sintaktik, semantik yoki pragmatik kabi boshqa turdagi korpusni teglash usullari uchun boshlang'ich qadam hisoblanadi.

Korpusni morfologik teglash haqida (Ide, Pustejovsky, 2017) “The Handbook of Linguistic Annotation” nomli kitobda morfologiyaning tamoyillari, usullari va muammolari haqida umumiy ma'lumot bergan [Erjavec, 1998; Kyoto, 1994]. Universal Dependencies loyihasi (Nivre va boshq., 2020)da bog'liqlik sintaksisi va morfologiyasi uchun tillararo teglash sxemalari va resurslarini ishlab chiqish haqida fikr mulohaza yuritilgan [Nivre, 2020]. Shuningdek, Savary va boshq. (2023) PARSEME ilmiy tadqiqotida 18 tildagi murakkab fe'l shakllarining morfologik teglash usullarini keltirilgan [Savary, 2023]. Kurimo va boshq. (2010) Morpho Challenge tadqiqotida turli xil morfologik tahlil va segmentatsiya uchun nazoratsiz mashinali usullarni baholagan [Kurimo, 2010].

Korpusga morfologik ishlov berish

Morfologik ishlov berish korpusda ishlatiladigan so'zlarni avtomatik tahlil qilishdan iborat. Morfologik tahlil jarayonida gap(matn)dagi so'z, qo'shma so'zlar va so'z birikmalari qayta ishlanadi. Morfologik tahlilning asosiy maqsadi – matndagi **so'zlarni aniqlash, uni kontekst qurshovidan ajratib olish, morfemik tuzilishini tahlil qilish, asl ma'nosi va matndagi sintaktik rolini aniqlash**dir. Matnni qayta ishlash natijasida olingan ma'lumotlardan NLPning *so'z ma'nosini aniqlash, lug'at tuzish, tahlil qilish, til o'rganish* kabi vazifalarida foydalanish mumkin. Ingliz va boshqa tillar uchun morfologik tahlilni amalga oshirish bo'yicha Greene va Rubin (1971), Karttunen va Wittenburg (1983), Koskenniemi (1983), deHaan (1984), Garside (1987), Church (1991), deRose (1991), Merialdo (1994) kabi olimlar ilmiy tadqiqotlar olib borishgan.

O'zbek tilida morfologik tahlilni quyidagicha amalga oshirish mumkin:

1. **So'z turkumi.**
2. **Yasalishi.**
3. **Tuzilishi.**
4. **Grammatik ma'nolari.**

Shuningdek, har bir so'z turkumi o'ziga xos morfologik xususiyatlarga ega bo'lganligi sababli har bir so'z turkumining tahlil parametrlari turlicha bo'ladi. Masalan, olmosh so'z turkumi morfologik tahlilida quyidagilarga e'tibor berish maqsadga muvofiq:

Olmosh so'z turkumi

1. Ma'noviy guruhlar:

- 1) kishilik olmoshi = [kish.olm.];
- 2) ko'rsatish olmoshi = [ko'r.olm.];
- 3) o'zlik olmoshi = [o'z.olm.];
- 4) gumon olmoshi = [gum.olm.];
- 5) so'roq olmoshi = [so'r.olm.];
- 6) belgilash olmoshi = [belg.olm.];

7) bo'lishsizlik olmoshi = [bo'l.olm.].

2. Tuzilishiga ko'ra turi:

- 1) sodda olmosh = [s.olm.];
- 2) qo'shma olmosh = [qo'sh.olm.];
- 3) juft olmosh = [j.olm.];
- 4) takror olmosh = [tak.olm.].

3. Yasalishi:

- 1) tub olmosh = [tub olm.];
- 2) yasama olmosh = [yas.olm.].

4. Olmoshlarda kelishik kategoriyasi:

- 1) [bosh kelishik] = [b.k.];
- 2) [qaratqich kelishigi] = [qar. k.];
- 3) [tushum kelishigi] = [tush. k.];
- 4) [jo'nalish kelishigi] = [j. k.];
- 5) [o'rin-payt kelishigi] = [o'.p. k.];
- 6) [chiqish kelishigi] = [ch. k.].

5. Son kategoriyasi:

- 1) [birlik son] = [birl. s.];
- 2) [ko'plik son] = [ko'p. s.].

6. Egalik kategoriyasi:

- 1) [I shaxs] = [I sh.];
- 2) [II shaxs] = [II sh.];
- 3) [III shaxs] = [III sh.].

Tabiiy til bilan shug'ullanadigan mutaxassislar tomonidan amalga oshirilgan morfologik tahlil samaradorligi yuqori hisoblanadi. Bugungi kunda mashinalar yordamida morfologik tahlilni amalga oshirishda 95% dan yuqori samaradorlikka erishilgan [Elov, Hamroyeva, Ahmedova, 2023].

POS teglash (Part-of-Speech Tagging)

Matndagi so'z yoki so'z birikmalarining muayyan xususiyatlarini ko'rsatish uchun ularga maxsus kodlar biriktiriladi va bu amal NLPda **lingvistik annotatsiya** yoki **teglash** deb nomlanadi. Xususiyatlarga tayinlangan kodlar **teg** deb nomlanadi.

POS teglash (so'z turkumini belgilash) – so'zning gapda qo'llanilgan turkumi asosida belgilash hisoblanadi. POS teglash amali uch bosqichda amalga oshiriladi [Elov, Hamroyeva, Abdullayeva, Uzakova, 2022].

- *boshlang'ich tahrirlash*;
- *avtomatik teglash*;
- *qo'lda tahrirlash*.

POS teglashning *boshlang'ich tahrirlash* bosqichda korpusdagi har bir so'z yoki so'z birikmasi POS teglash uchun mos formatga aylantiriladi. Leksik va grammatik omonimiya tufayli bitta so'z bir nechta POS tegiga ega bo'lishi mumkin.

Soʻzning turkumini (POS)ni dastlabki belgilashdan soʻng matnlardagi soʻzlarni ajratish uchun soʻzlarning teglari qoʻlda tuzatiladi.

Oʻzbek tili soʻz turkumlari uchun B.Elov va boshqalar tomonidan quyidagi teglar toʻplami ishlab chiqilgan [Elov;, Hamroyeva, Abdullayeva, Uzakova, 2022]:

1-jadval. Oʻzbek tilidagi POS teglar roʻyxati (qisqa variant)

№	POS teg	Belgilanishi
1.	Ot	N
2.	Sifat	JJ
3.	Feʼl	VB
4.	Son	NUM
5.	Ravish	RR
6.	Olmosh	P
7.	Bogʻlovchi	C
8.	Koʻmakchi	II
9.	Yuklama	Prt
10.	Modal soʻz	MD
11.	Undov soʻz	UH
12.	Taqlid soʻzlar	IM
13.	Atoqli otlar	NER
14.	Iboralar	IB

POS teglashning ikkinchi darajasi **grammatik teglash (grammatical tagging)** deb ataladi. Grammatik teglash jarayonida korpusdagi har bir soʻzga mos barcha grammatik maʼlumotlar belgilanadi. Ushbu jaryondan matnni tahlil qilishda, soʻzlarning grammatik maʼnosini aniqlash/farqlashda foydalaniladi. **Morfologik, leksik, semantik** va **sintaktik** tahlil bilan bogʻliq muammolarni hal qilishda matnda uchraydigan har bir soʻz uchun grammatik maʼlumotlarni oʻz ichiga olgan korpusga ehtiyoj mavjud.

Soʻz maʼnosini belgilash (Word Sense Tagging)

Soʻz maʼnosini teglash yoki **semantik teglash** uchun POS va grammatik teglangan matndan foydalaniladi, teglangan matnlar ustida amallar bajariladi. Matnni avtomatik semantik teglashdan soʻng har bir soʻz toʻgʻri semantik tasnifga ega boʻlishini taʼminlash maqsadida qoʻlda **post-tahrirlash** amalga oshiriladi.

Soʻz maʼnosini belgilashning quyidagi misoli Buyuk Britaniyaning Lankaster universiteti loyihasidan olingan. *I like a particular shade of lipstick.*

2-jadval. Soʻz maʼnosini belgilash

№	POS va grammatik teg	Soʻz	Semantik teg
1.	PPIS1	<i>I</i>	Z8
2.	VV0	<i>like</i>	E2+

3.	AT1	<i>a</i>	Z5
4.	JJ	<i>particular</i>	A4.2+
5.	NN1	<i>shade</i>	O4.3
6.	IO	<i>of</i>	Z5
7.	NN1	<i>lipstick</i>	B4

Bu jadvalda matn pastga qarab o'qiladi, chap tomonda POS va grammatik teglar, o'ng tomonda esa so'z ma'nosini anglatuvchi teglar keltirilgan.

Semantik teglar umumiy nutq maydonini ko'rsatadigan bosh harf, maydonning birinchi bo'linmasini ko'rsatadigan raqam, aniqroq bo'linmani ko'rsatish uchun o'nli nuqtadan keyin keyingi raqam, ko'rsatish uchun bir yoki bir nechta “ortiqcha” yoki “minyslardan” iborat semantik miqyosda **ijobiy** yoki **salbiy** pozitsiya kabi ma'lumotlardan tashkil topadi.

Masalan, **A4.2+** tegi “*umumiy va mavhum so'zlar*” (**A**), “*tasniflash*” (**A4**) toifasidagi, “*alohida va umumiy*” (**A4.2**) va “*alohida*” toifasidagi so'zni ko'rsatadi. “**umumiy**” (**A4.2+**). Xuddi shunday, **E2+** “*hissiy holatlar, harakatlar, hodisalar va jarayonlar*” (**E**), “*yoqtirish va yoqtirmaslik*” (**E2**) toifasiga kiradi va “*yoqmaslik*” (**E2+**) emas, balki “**yoqtirish**”ni anglatadi.

Xulosa

Tabiiy tilni qayta ishlash vazifalarining aksariyatida zamonaviy tizimlar odatda matematik modellarini yaratish uchun mashinali o'rganish usullariga tayanadi. Ushbu tizimlarning aksariyati nazorat ostida o'rganish strategiyalarini qo'llashini hisobga olsak, muammoli soha uchun teglangan korpus muhim ahamiyatga ega. Korpusga morfologik ishlov berishning hozirgi usuli bir nechta mutaxassislarni yollash va ularni qo'lda yoki yordamchi dastur yordamida korpusga teglashdan iborat. Biroq, bu usul qimmat va vaqt talab etadi. Ushbu maqolada biz yuqoridagi muammolarni hal qilishga qaratilgan yangi usul taklif qilindi.

Foydalanilgan adabiyotlar:

1. Erjaveç, T., Ide, N.: The MULTEXT-East corpus. In: Proceedings of First International Conference on Language Resources and Evaluation, pp. 971–974 (1998)
2. 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)
3. Nivre, J., de Marneffe, M. C., Ginter, F., Haji, J., Manning, C. D., Pyysalo, S., Schuster, S., & Zeman, F. T. D. (2020). Universal dependencies v2: An evergrowing multilingual treebank collection. LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings.
4. Savary, A., Khelil, C. ben, Ramisch, C., Giouli, V., Mititelu, V. B., Mohamed, N. H., Krstev, C., Liebeskind, C., Xu, H., Jiang, M., Stymne, S., Güngör,



T., Pickard, T., Guillaume, B., Bhatia, A., Butler, A., Candito, M., Gantar, A., Iñurrieta, U., ... Walsh, A. (2023). PARSEME Corpus Release 1.3. 19th Workshop on Multiword Expressions, MWE 2023 - Proceedings. <https://doi.org/10.18653/v1/2023.mwe-1.6>

5. Kurimo, M., Virpioja, S., Turunen, V., & Lagus, K. (2010). Morpho challenge competition 2005-2010: Evaluations and results. Proceedings of the Annual Meeting of the Association for Computational Linguistics.

6. Boltayevich, E. B., Mirdjonovna, H. S., & Ilxomovna, A. X. (2023). Methods for Creating a Morphological Analyzer. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 13741 LNCS. https://doi.org/10.1007/978-3-031-27199-1_4

7. 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.

8. 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 xabarлари. Mirzo Ulug'bek nomidagi O'zbekiston Milliy universiteti ilmiy jurnali. 2023 yil Maxsys son.

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