O‘ZBEK TILIDAGI SO‘ZLARNI POS TEGLASHNING ZAMONAVIY YONDASHUVLARI

Xudayberganov Nizomaddin Uktambay o‘g‘li
ToshDO‘TAU o‘qituvchisi
nizomaddin@navoiy-uni.uz

Annotatsiya. Ushbu maqola O‘zbek tilidagi so‘zlarni grammatik turkumlar (Part-of-Speech, POS) bo‘yicha teglashning zamonaviy usullarini tizimli tahlil qilishga bag‘ishlangan. O‘zbek tilining agglyutinativ tuzilishi, boy shakl yasovchi qo‘shimchalar tizimi va raqamli resurslarning cheklanganligi POS teglash jarayonini murakkablashtiradi. Tadqiqotda qoidaviy yondashuv, statistik modellar (Hidden Markov Model, Conditional Random Fields) va chuqur o‘rganish (Deep Learning) asosidagi usullar o‘zaro taqqoslanadi. Xususan, HMM asosidagi teglash, BiLSTM arxitekturasiga asoslangan neyron modellar va so‘nggi yillarda ishlab chiqilgan BERT asosidagi tillararo va monolingval modellar tahlil qilinadi. Shuningdek, O‘zbek tili uchun yaratilgan Universal Dependencies treebank va boshqa teglangan korpuslar imkoniyatlari baholanadi. Tadqiqot natijalari shuni ko‘rsatadiki, neyron tarmoq asosidagi modellar (91% gacha aniqlik) an’anaviy statistik va qoidaviy usullardan (86% atrofida) ustunlik qiladi, biroq resurs cheklangan sharoitda HMM modellari ham samarali alternativ bo‘la oladi.

Kalit so‘zlar: *POS teglash, O‘zbek tili, tabiiy tilni qayta ishlash, NLP, yashirin Markov modeli, HMM, chuqur o‘rganish, BiLSTM, Universal Dependencies, BERT.*

Annatation. This article is devoted to a systematic analysis of modern methods for part-of-speech (POS) tagging of words in the Uzbek language. The agglutinative structure of Uzbek, its rich system of derivational and inflectional affixes, and the limited availability of digital resources make the POS tagging process more complex. The study compares rule-based approaches, statistical models (Hidden Markov Model, Conditional Random Fields), and deep learning-based methods. In particular, HMM-based tagging, neural models based on the BiLSTM architecture, and recent cross-lingual and monolingual models based on BERT are analyzed. In addition, the capabilities of the Universal Dependencies treebank and other annotated corpora developed for the Uzbek language are evaluated. The results show that neural network-based models (up to 91% accuracy) outperform traditional statistical and rule-based methods (around 86%), although HMM models can still serve as an effective alternative in resource-constrained settings.

Keywords: *POS tagging, Uzbek language, natural language processing, NLP, Hidden Markov Model, HMM, deep learning, BiLSTM, Universal Dependencies, BERT.*

Аннотация. Данная статья посвящена системному анализу современных методов теггирования частей речи (POS) в узбекском языке. Агглютинативная структура узбекского языка, богатая система словообразовательных и формообразующих аффиксов, а также ограниченность цифровых ресурсов усложняют процесс POS-теггирования. В исследовании проводится сравнение правиловых подходов, статистических моделей (Hidden Markov Model, Conditional Random Fields) и методов, основанных на глубоком обучении. В частности, анализируются модели теггирования на основе HMM, нейронные модели с архитектурой BiLSTM, а также современные кросс-языковые и моноязыковые модели на основе BERT. Кроме того, оцениваются возможности Universal Dependencies treebank и других размеченных корпусов, созданных для узбекского языка. Результаты исследования показывают, что модели на основе нейронных сетей (точность до 91%) превосходят традиционные статистические и правиловые методы (около 86%), однако модели HMM также могут служить эффективной альтернативой в условиях ограниченных ресурсов.

Ключевые слова: *POS-теггирование, узбекский язык, обработка естественного языка, NLP, скрытая марковская модель, HMM, глубокое обучение, BiLSTM, Universal Dependencies, BERT.*

Kirish.

Tabiiy tilni qayta ishlash (NLP) sohasida so‘z turkumlarini teglash (Part-of-Speech tagging) eng muhim boshlang‘ich bosqichlardan biri hisoblanadi. Bu jarayonda matndagi har bir leksik birlik (so‘z) o‘zining grammatik turkumiga (ot, sifat, fe’l, ravish va h.k.) ko‘ra teg bilan belgilanadi. POS teglash keyingi tahlil bosqichlari – sintaktik parsing (tahlil), semantik tahlil, obyektlarni aniqlash (NER) va mashina tarjimasi – uchun asos bo‘lib xizmat qiladi.

O‘zbek tili NLP nuqtai nazaridan qator o‘ziga xos xususiyatlarga ega. Birinchidan, u agglyutinativ tillar guruhiga kiradi, ya’ni so‘z yasash va so‘z o‘zgartirish qo‘shimchalarni ketma-ket ulash orqali amalga oshiriladi. Bu xususiyat so‘z shakllarining ko‘p variantliligini keltirib chiqaradi va ularni lug‘at shakliga (lemmaga) keltirishni qiyinlashtiradi. Ikkinchidan, O‘zbek tilida omonimlik (bir so‘z shaklining turli ma’no va turkumlarda qo‘llanilishi) keng tarqalgan bo‘lib, kontekstual tahlilni talab qiladi. Uchinchidan, O‘zbek tili uchun tegishli darajada katta hajmli va yuqori sifatli teglangan korpuslar (annotated corpora) hozirga qadar cheklangan miqdorda mavjud.

So‘nggi yillarda O‘zbek tilini POS teglash bo‘yicha bir qator tadqiqotlar olib borilmoqda. Dastlabki ishlar asosan **qoidaviy yondashuv** (rule-based approach) ga asoslangan bo‘lib, unda tilshunoslik qoidalari va qo‘shimchalar bazasi asosida so‘zlarga teg berilgan. Keyinchalik **statistik modellar**, xususan Yashirin Markov Modeli (Hidden Markov Model, HMM) asosidagi teglagichlar ishlab chiqilgan. So‘nggi yillarda esa **chuqur o‘rganish** (deep learning) texnologiyalarining rivojlanishi bilan BiLSTM (Bidirectional Long Short-Term Memory) va transformator (BERT) arxitekturalariga asoslangan modellar paydo bo‘ldi.

Metodlar

O‘zbek tilidagi so‘zlarni POS teglashda qo‘llaniladigan zamonaviy yondashuvlarni asosan uchta katta guruhga ajratish mumkin: qoidaviy (rule-based) yondashuv, statistik (statistical) modellar va mashinaviy o‘qitish (machine learning) asosidagi usullar. Quyida har bir yondashuvning asosiy tamoyillari va O‘zbek tili uchun ishlab chiqilgan **POS teglagichlar** tahlil qilinadi.

Qoidaviy yondashuv (Rule-based Approach)

Qoidaviy POS teglash tilshunoslar tomonidan ishlab chiqilgan grammatik qoidalar va lug‘atlar asosida ishlaydi. Bu yondashuvda odatda quyidagi bosqichlar amalga oshiriladi:

1. **Tokenizatsiya**: Matnni alohida so‘z va tinish belgilariga ajratish.
2. **Lug‘at bilan solishtirish**: Har bir so‘zning o‘zak shakli oldindan tuzilgan lug‘at (lexicon) orqali tekshiriladi.
3. **Qo‘shimchalarni tahlil qilish**: Agar so‘z lug‘atda topilmasa, uning qo‘shimchalari ketma-ket ajratiladi va har bir qo‘shimchaning grammatik vazifasi aniqlanadi.
4. **Qoidalarni qo‘llash**: Aniqlangan o‘zak va qo‘shimchalar asosida so‘z turkumiga teg beriladi.

O‘zbek tili uchun M.Sharipov va boshqalar tomonidan “**UzbekTagger”** nomli qoidaviy teglagich ishlab chiqilgan. Bu teglagich quyidagi xususiyatlarga ega:

- 12 ta asosiy so‘z turkumi uchun teglar to‘plami (tagset) ishlab chiqilgan.
- 20 xil soha (fan, texnika, adabiyot va h.k.) dan olingan matnlar asosida korpus tuzilgan.
- So‘z o‘zagini aniqlash uchun affikslarni ketma-ket ajratish (affix/suffix stripping) usuli qo‘llanilgan.
- Qo‘shimchalarning tasnifi va ularni ajratish uchun chekli holat mashinalari (Finite State Machine) dan foydalanilgan.

Ushbu yondashuvning afzalligi – katta hajmdagi teglangan korpus talab qilinmaydi va til qoidalariga asoslanganligi sababli izohlash oson. Biroq kamchiliklari – qoidalarni yaratish ko‘p vaqt talab qiladi, tilning murakkab va istisnoli holatlarini to‘liq qamrab ololmaydi va kontekstga sezgirlik darajasi past.

Statistik modellar (Statistical Models)

Statistik yondashuvda POS teglash ehtimollik modellari asosida amalga oshiriladi. Eng keng tarqalgan statistik usul bu Yashirin Markov Modeli (Hidden Markov Model, HMM) dir.

Yashirin Markov Modeli (HMM)

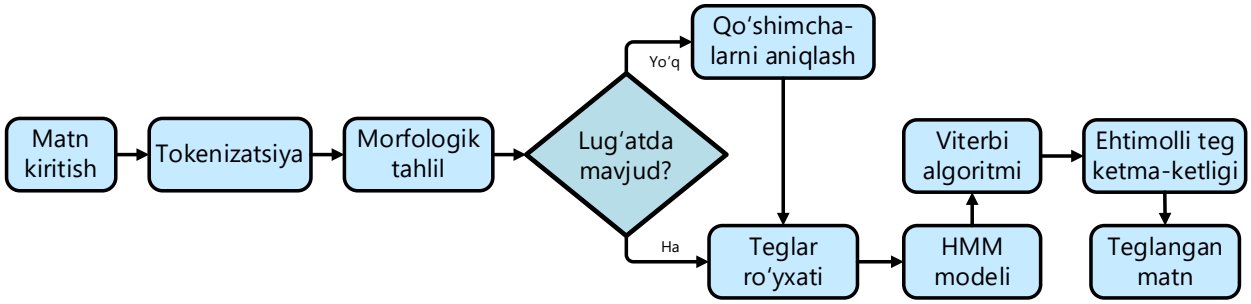
HMM asosidagi POS teglashda so‘zlar kuzatiladigan holatlar, teglar esa yashirin holatlar sifatida modellashtiriladi. Model ikkita asosiy ehtimollikni hisoblaydi:

- **O‘tish ehtimolligi (Transition probability)**: Bir tegdan ikkinchi tegga o‘tish ehtimoli, $P(t_i \mid t_{i-1})$.
- **Emissiya ehtimolligi (Emission probability)**: Berilgan teg ostida ma’lum bir so‘zning paydo bo‘lish ehtimoli, $P(w_i \mid t_i)$.

Teglanishi kerak bo‘lgan yangi gap uchun eng ehtimolli teglar ketma-ketligi Viterbi algoritmi orqali topiladi.

O‘zbek tili uchun HMM asosidagi teglagich bir qator tadqiqotlarda ishlab chiqilgan. Jumladan, O.Sobirov va boshqalarning ilmiy izlanishlari keltirilgan ishda:

- 13 ta POS teglar to‘plami ishlab chiqilgan.
- 30 xil soha bo‘yicha 48 079 so‘zdan iborat qo‘lda teglangan korpus tayyorlangan.
- Birinchi darajali HMM (first-order HMM) qo‘llanilgan.



1-rasm. HMM asosidagi POS teglash jarayonining blok-sxemasi

Conditional Random Fields (CRF)

HMM dan tashqari, Conditional Random Fields (CRF) ham statistik modellar sirasiga kiradi. CRF HMM dan farqli o‘laroq, butun gap kontekstini bir vaqtda

hisobga oladi va qo‘shimcha xususiyatlar (features) ni modelga kiritish imkonini beradi. O‘zbek tilidagi ravish so‘z turkumini teglashda CRF usuli qo‘llanilgan tadqiqotlar mavjud.

Statistik modellarning afzalligi – ular teglangan korpusdan avtomatik o‘rganadi va kontekstni ma'lum darajada hisobga oladi. Kamchiligi – katta hajmdagi teglangan korpus talab qiladi va qoidaviy usulga nisbatan kamroq izohlanuvchan (less interpretable).

Chuqur o‘rganish asosidagi yondashuvlar (Deep Learning Approaches)

So‘nggi yillarda neyron tarmoqlarga asoslangan modellar POS teglash vazifasida eng yuqori natijalarni ko‘rsatmoqda. O‘zbek tili uchun quyidagi chuqur o‘rganish modellari qo‘llanilgan:

BiLSTM asosidagi modellar

BiLSTM (Bidirectional Long Short-Term Memory) modellari matnni ikki yo‘nalishda (oldinga va orqaga) o‘qib, har bir so‘z uchun kontekstga bog‘liq vektor ifodalar (contextualized embeddings) hosil qiladi.

S.Matlatipov ning ilmiy tadqiqotlarida O‘zbek tili uchun **Universal Dependencies (UD) treebank** yaratilgan va BiLSTM asosidagi deep biaffine dependency parser o‘qitilgan. Ushbu korpus:

- 686 ta gap (7,800 token) dan iborat.
- Badiiy va ilmiy-ommabop matnlarni o‘z ichiga oladi.
- Lemmalar, POS teglar, morfologik xususiyatlar va bog‘liqlik munosabatlari qo‘lda teglangan.
- Annotatorlar kelishuvi (inter-annotator agreement) lemmatizatsiya va UPOS (Universal Part-of-Speech) uchun 95% dan yuqori.

Model Stanza pipeline'ida implementatsiya qilingan BiLSTM va deep biaffine attention mexanizmiga asoslangan.

Transformator (BERT) asosidagi modellar

Transformator arxitekturasi, xususan BERT (Bidirectional Encoder Representations from Transformers) modellari hozirgi kunda NLP vazifalarida eng ilg‘or (state-of-the-art) hisoblanadi.

L.Bobojonova, A.Akhundjanova va boshqalarning ilmiy izlanishlarida keltirilgan tadqiqotda O‘zbek tili uchun:

- Ikkita monolingval BERT modeli (o‘zbek matnlarida qo‘shimcha o‘qitilgan) sinovdan o‘tkazilgan.

- Birinchi marta ochiq foydalanish mumkin bo‘lgan UPOS-teglangan benchmark dataset yaratilgan.
- Modellar fine-tuning qilinib, o‘rtacha 91% aniqlikka erishilgan.

BERT modellari qoidaviy teglagichlardan farqli o‘laroq, qo‘shimchalar orqali yuzaga keladigan oraliq POS o‘zgarishlarini aniqlay oladi va kontekstga yuqori sezgirlik ko‘rsatadi.

Qo‘llanilgan korpuslar va teglar to‘plami

O‘zbek tili uchun POS teglashda foydalanilgan asosiy korpuslar 1-jadvalda keltirilgan.

1-jadval. O‘zbek tili uchun POS teglash korpuslari

Korpus nomi	Hajmi (so‘z)	Teglar soni	Annotatsiya turi
UzbekTagger korpusi	48 079	12	Qo‘lda, soha bo‘yicha muvozanatlangan
UzUDT (Universal Dependencies)	7 800 (686 gap)	17 (UPOS)	Qo‘lda, lemmalar va morfologik xususiyatlar bilan
Benchmark UPOS dataset		17 (UPOS)	Ochiq foydalanish uchun

2-jadvalda O‘zbek tili uchun ishlab chiqilgan POS teglar to‘plami (UzbekTagger asosida) va ularning Universal Dependencies (UPOS) dagi ekvivalentlari keltirilgan.

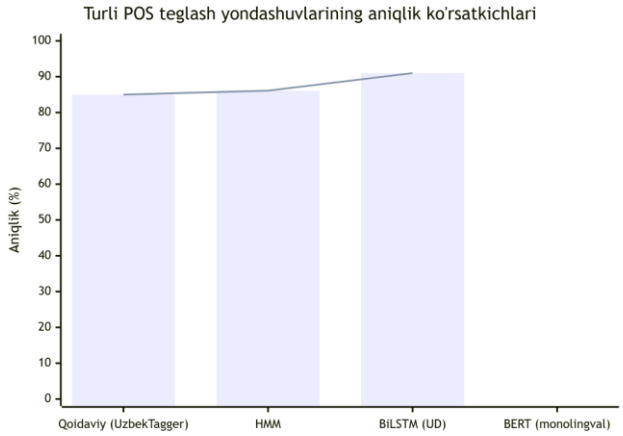
2-jadval. O‘zbek tili POS teglari va UPOS ekvivalentlari

Turkum	UzbekTagger tegi	UPOS tegi	Izoh
Ot	OT	NOUN	Shaxs, predmet, tushuncha nomlari
Sifat	SIF	ADJ	Belgini bildiruvchi so‘zlar
Son	SON	NUM	Sanoq va tartib sonlar
Olmosh	OLM	PRON	Ot o‘rnida qo‘llanuvchi so‘zlar
Fe’l	FEL	VERB	Harakat va holatni bildiruvchi so‘zlar
Ravish	RAV	ADV	Harakat belgisini bildiruvchi so‘zlar
Ko‘makchi	KOM	ADP	Ot va olmoshlar bilan qo‘llanuvchi yordamchi so‘zlar
Bog‘lovchi	BOG	CCONJ	So‘z va gaplarni bog‘lovchi so‘zlar
Yuklama	YUK	PART	Ma’no nozikliklarini ifodalovchi so‘zlar
Undov	UND	INTJ	His-hayajonni ifodalovchi so‘zlar

Modal so‘z	MOD	[yo‘q]	Munosabat bildiruvchi so‘zlar
Taqlid	TAQ	[yo‘q]	Tovush va holatga taqlid so‘zlar

Yondashuvlarning samaradorlik ko‘rsatkichlari

Turli yondashuvlar asosidagi POS teglagichlarning aniqlik (accuracy) ko‘rsatkichlari 2-rasm va 3-jadvalda keltirilgan.



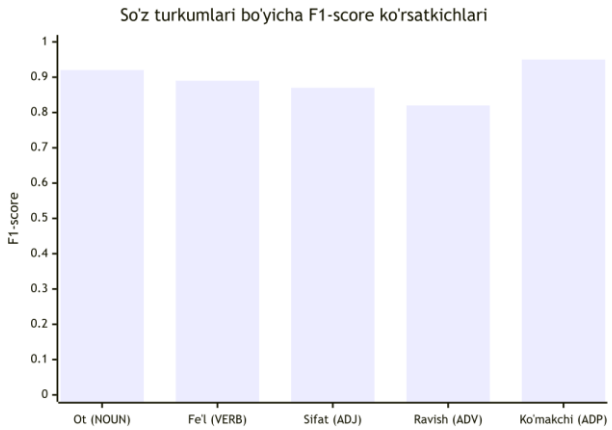
2-rasm. Turli POS teglash yondashuvlarining aniqlik taqqoslamasi

3-jadval. POS teglagichlarning qiyosiy tahlili

Yondashuv	Model/Usul	Aniqlik (Accuracy)	Afzalliklari	Kamchiliklari
Qoidaviy	UzbekTagger (rule-based)	≈ 85%	Korpus talab qilmaydi, izohlash oson	Kontekstga past sezgirlik, qoidalarni yaratish murakkab
Statistik	HMM (first-order)	≈ 86.10%	Kontekstni hisobga oladi, matematik asoslangan	Katta korpus talab qiladi
Neyron (BiLSTM)	Deep Biaffine (Stanza)	86.10% (UPOS)	Kontekstga yuqori sezgirlik, qo‘shimcha xususiyatlarni o‘rganish	Juda katta korpus talab qiladi, hisoblash resurslari talabchan
Neyron (Transformator)	Monolingval BERT	≈ 92%	Kontekstni chuqur tushunish, qo‘shimcha o‘zgarishlarni aniqlash	Juda katta resurs talab qiladi, interpretatsiya qiyin

So‘z turkumlari bo‘yicha natijalar tahlili

Turli so‘z turkumlarini teglash aniqligi turlicha bo‘lishi mumkin. Masalan, ravish va sifat so‘z turkumlari o‘ziga xos qiyinchiliklarga ega. 3-rasmda ayrim so‘z turkumlari bo‘yicha teglash aniqligi (BiLSTM modeli misolida) keltirilgan.



3-rasm. So‘z turkumlari bo‘yicha teglash aniqligi (BiLSTM modeli)

Ko‘makchi (ADP) va ot (NOUN) kabi tez-tez uchraydigan va aniq morfologik belgilarga ega turkumlar yuqori aniqlikda teglanadi. Ravish (ADV) esa ma’no xilma-xilligi va shakliy o‘zgaruvchanligi sababli nisbatan past aniqlik ko‘rsatadi.

Xatolik tahlili (Error Analysis)

POS teglash jarayonida uchraydigan asosiy xatolik turlari quyidagilardan iborat:

1. **Omonimlik (Homonymy) muammosi**: Bir so‘z shakli turli kontekstlarda turli turkumlarda qo‘llanilishi mumkin. Masalan, “ish” so‘zi “yaxshi ish” (ot) va “ish boshlamoq” (fe’l o‘zagi) ma’nolarida kelishi mumkin. HMM va neyron modellar kontekst orqali omonimiyani aniqlashda nisbatan yaxshi natija ko‘rsatadi.
2. **Yangi va lug‘atdan tashqari so‘zlar (OOV so‘zlari)**: Korpusda uchramagan yangi so‘zlar (xususan, yangi qo‘shma so‘zlar yoki o‘zlashmalar) to‘g‘ri teglanmasligi mumkin. Qoidaviy usullar morfologik tahlil orqali bunday so‘zlarni teglash imkoniyatiga ega.
3. **Omonamiya holatlar**: Ayrim so‘z shakllari bir vaqtning o‘zida bir necha turkumga xos xususiyatlarni namoyon qilishi mumkin. Masalan, sifatdosh fe’l va sifat xususiyatlarini birlashtiradi.

Umumiy muhokama

Tahlil natijalari shuni ko‘rsatadiki, O‘zbek tilidagi so‘zlarni POS teglashda chuqur o‘rganish asosidagi modellar, xususan BERT, eng yuqori aniqlikka (91%) erishmoqda. Bu modellar kontekstni chuqur tushunish va qo‘shimchalar orqali

yuzaga keladigan murakkab morfologik o‘zgarishlarni modellashtirish qobiliyati bilan ajralib turadi.

Ammo shuni ta’kidlash kerakki, neyron modellarning yuqori samaradorligi **katta hajmdagi teglangan korpuslar** va **yuqori hisoblash quvvatiga** bog‘liq. Resurslar cheklangan bo‘lgan holatlarda (masalan, maxsus soha korpuslari yoki kam resursli tillar) **HMM asosidagi modellar** ham munosib alternativ bo‘la oladi. O.Sobirov va boshqalarning ilmiy izlanishlarida keltirilgan HMM teglagichi 48,079 so‘zlik korpusda 86.10% aniqlikka erishgan bo‘lib, bu nisbatan kichik korpus uchun yaxshi ko‘rsatkichdir.

Qoidaviy yondashuv esa o‘zining soddaligi va izohlanish osonligi bilan ajralib turadi. UzbekTagger M.Sharipov, E.Kuriyozov va boshqalarining izlanishlaridagi kabi qoidaviy teglagichlar til qoidalariga asoslanganligi sababli, ularning xatolarini tahlil qilish va tuzatish oson. Biroq ular kontekstual o‘zgarishlarni to‘liq qamrab ololmaydi.

Xulosa

Ushbu maqolada O‘zbek tilidagi so‘zlarni POS teglashning zamonaviy yondashuvlari – qoidaviy, statistik (HMM) va chuqur o‘rganish (BiLSTM, BERT) asosidagi usullar – tizimli tahlil qilindi. Tadqiqot natijalari quyidagi asosiy xulosalarni chiqarish imkonini beradi:

- **Ierarxik samaradorlik**: POS teglash aniqligi bo‘yicha transformator asosidagi modellar (BERT, ≈ 91%) neyron tarmoq asosidagi modellar (BiLSTM, ≈ 86.10%) statistik modellar (HMM, ≈ 86.10%) qoidaviy modellar (≈ 85%) tartibi kuzatiladi.
- **Resurslar va samaradorlik o‘rtasidagi muvozanat**: Katta hajmdagi teglangan korpus va hisoblash resurslari mavjud bo‘lganda, BERT asosidagi modellar eng yaxshi tanlov hisoblanadi. Resurslar cheklangan bo‘lsa, HMM asosidagi modellar munosib alternativ bo‘la oladi.
- **Kontekstning ahamiyati**: O‘zbek tilidagi omonimiya va polifunksionallik muammolarini hal qilishda kontekstni hisobga oluvchi modellar (HMM, BiLSTM, BERT) qoidaviy usullardan ustun turadi.

Foydalanilgan adabiyotlar

1. Boltayevich, E. B., Mirdjonovna, H. S., & Ilxomovna, A. X. (2022, October). Methods for creating a morphological analyzer. In *International Conference on Intelligent Human Computer Interaction* (pp. 27-38). Cham: Springer Nature Switzerland.

2. Xudayberganov, N., & Karimova, Z. (2025). RAVISH SO‘Z TURKUMINI GRAMMATIK POS TEGLASH ALGORITMLARI: NAZARIY YONDASHUV,

AMALIY TATBIQ VA KELGUSIDAGI YO‘NALISHLAR. *COMPUTER LINGUISTICS: PROBLEMS, SOLUTIONS, PROSPECTS*, *1*(1).

3. Matlatipov, S. A. (2025). O ‘ZBEK TILI UCHUN UNIVERSAL BOG‘LIQLIK DARAXTI KORPUSI ASOSIDA CHUQUR BI-AFFIN TOBELIK TAHLILINING NEYRON MODELI. *«ACTA NUUz»*, *2*(2.2. 1), 85-93.

4. Najmiddin qizi, M. S. (2025). Sifat so‘z turkumini darajalanishiga ko‘ra POS teglash. *Kokand SPI ilmiy xabarlari*, 5(5), 2169–2173.

5. Bobojonova, L., Akhundjanova, A., Ostheimer, P. S., & Fellenz, S. (2025, January). BBPOS: BERT-based part-of-speech tagging for Uzbek. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages* (pp. 287-293).

6. Sharipov, M., Abjalova, M., & Sobirov, O. (2025, September). Design and Implementation of a Markov Model-Based POS Tagger for the Uzbek Language. In *2025 10th International Conference on Computer Science and Engineering (UBMK)* (pp. 343-347). IEEE.

7. Abjalova, M., & Adilova, M. (2025). So‘z turkumlarini teglash usullari. *Uzbekistan: Language and Culture*, *1*(01).

8. Sharipov, M., Kuriyozov, E., Yuldashev, O., & Sobirov, O. (2023). UzbekTagger: The rule-based POS tagger for Uzbek language. *arXiv preprint arXiv:2301.12711*.

9. Xusainova, Z. (2022). NLP: tokenizatsiya, stemming, lemmatizatsiya va nutq qismlarini teglash. *Prospects of Uzbek applied philology*, *1*(1).

10. Sharipov, M., & Sobirov, O. (2022). Development of a rule-based lemmatization algorithm through Finite State Machine for Uzbek language. *arXiv preprint arXiv:2210.16006*.