The Lingua Spectrum
Cogito, ergo sum

Jurnal Oliy Attestatsiya Komissiyasi tomonidan 2025-yil 4-yanvar kuni 366/5-sonli qaror bilan tavsiya etilgan milliy ilmiy nashrlar ro'yxatiga kiritilgan

O‘zbek tili korpusi matnlarini lingvistik izohlash

Samatboyeva Madina To‘lqinjon qizi
samatboyevamadina@navoiy-uni.uz
Tayanch doktoranti,
Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti

Annotatsiya *Ushbu maqolada O‘zbek tili korpusini lingvistik jihatdan izohlashning nazariy va amaliy asoslari atroflicha yoritilgan. Korpus lingvistikasi doirasida matnlarni morfologik, sintaktik va semantik izohlash jarayonlari, ularning tilshunoslik tadqiqotlaridagi ahamiyati hamda korpus asosida olib boriladigan avtomatlashtirilgan til tahlillariga qo‘shgan hissasi tahlil qilinadi. Xususan, lingvistik izohlash komponentlari – morfologik tahlil (so‘zlarning shakl va grammatik xususiyatlarini aniqlash), sintaktik tahlil (so‘zlar orasidagi grammatik bog‘lanishlarni belgilash) hamda semantik izohlash (so‘z va iboralarning ma’no munosabatlarini ko‘rsatish) – ning ishlab chiqish usullari va ularni tajribadan o‘tkazishda qo‘llaniladigan zamonaviy texnologiyalar ko‘rib chiqiladi. Maqolada, shuningdek, xalqaro miqyosda tan olingan tabiiy tilni qayta ishlashga oid standartlashtirilgan ramkalar, xususan Universal Dependencies (UD) kabi formatlar asosida tahlil ishlari olib borilib, ularning o‘zbek tili uchun moslashuvi va amaliy qo‘llanilishi yuzasidan fikr yuritiladi. UD shakllarining o‘zbek tilida tatbiq etilishi misollar orqali ko‘rsatilib, izohlangan korpus matnlari asosida yaratilgan tahlil natijalari jadval shaklida taqdim etiladi. Mazkur maqola O‘zbek tilini raqamli til resurslari doirasida chuqurroq o‘rganish va tahlil qilishga xizmat qiladi hamda lingvistik izohlashning ilmiy-amaliy ahamiyatini yoritadi.*

Kalit so‘zlar *O‘zbek tili korpusi, lingvistik izohlash, lingvistik annotatsiya, morfologik izoh, sintaktik tahlil, avtomatik belgilash, UD*

Лингвистическая аннотация текстов корпуса узбекского языка

Саматбоева Мадина Тулкинжон кизи
samatboyevamadina@navoiy-uni.uz
Докторант,
Ташкентский государственный университет узбекского языка и литературы имени Алишера Навои

Аннотация *В данной статье подробно освещаются теоретические и практические основы лингвистической аннотации корпуса узбекского языка. В рамках корпусной лингвистики рассматриваются процессы морфологической, синтаксической и семантической аннотации текстов, их значимость в лингвистических исследованиях, а также вклад в автоматизированный языковой анализ на основе корпуса. В частности, анализируются компоненты лингвистической аннотации – морфологический разбор (определение форм и грамматических характеристик слов), синтаксический анализ (установление грамматических связей между словами), а также семантическая аннотация (отражение смысловых отношений между словами и выражениями). В статье также рассматривается использование*

международно признанных стандартов обработки естественного языка, в частности, формата Universal Dependencies (UD), а также обсуждается адаптация и практическое применение этих форматов к узбекскому языку. Применение формализма UD в узбекском языке иллюстрируется с помощью примеров, а результаты анализа, выполненного на аннотированных корпусных текстах, представлены в табличной форме. Настоящая работа направлена на углублённое изучение и анализ узбекского языка в рамках цифровых языковых ресурсов, а также раскрывает научно-практическую значимость лингвистической аннотации

Ключевые слова *Корпус узбекского языка, лингвистическая аннотация, морфологическая разметка, синтаксический анализ, автоматическая аннотация, Universal Dependencies (UD)*

Linguistic annotation of texts in the Uzbek Language Corpora

Samatboeva Madina Tulkinjon kizi
samatboyevamadina@navoiy-uni.uz
PhD Student,
Alisher Navoiy Tashkent State University of Uzbek Language and Literature

Annotation *This article provides a comprehensive overview of the theoretical and practical foundations of linguistic annotation of the Uzbek language corpus. Within the framework of corpus linguistics, it analyzes the processes of morphological, syntactic, and semantic annotation of texts, their significance for linguistic research, and their contribution to automated language analysis based on corpora. In particular, the development methods of linguistic annotation components – morphological analysis (identifying word forms and grammatical features), syntactic analysis (determining grammatical relationships between words), and semantic annotation (indicating the semantic relationships of words and phrases) – as well as the modern technologies used in their implementation are discussed.*

The article also examines standardized frameworks widely recognized in the field of natural language processing, especially formats such as Universal Dependencies (UD), analyzing their adaptation and practical application for the Uzbek language. The implementation of the UD formalism in Uzbek is illustrated through examples, and the results of analyses conducted on annotated corpus texts are presented in form. This study contributes to the in-depth exploration and analysis of the Uzbek language as a digital linguistic resource and highlights the scientific and practical significance of linguistic annotation.

Key words *Uzbek language corpus, linguistic annotation, morphological tagging, syntactic analysis, automatic annotation, Universal Dependencies (UD)*

Kirish

Zamonaviy tilshunoslikda til materiallarini raqamli shaklda jamlash va ularni tizimli o‘rganish ehtiyoji ortib bormoqda. Korpus lingvistikasi – ana shu ehtiyojga javob beradigan yo‘nalish bo‘lib, u lingvistik tadqiqotlar, mashinali tarjima, matnni avtomatik tahlil qilish, til o‘qitish va boshqa ko‘plab sohalarda foydalanilmoqda. O‘zbek tilida bu yo‘nalish hali to‘liq shakllanmagan bo‘lsa-da, muhim qadamlar qo‘yilmoqda.

Tabiiy tilni qayta ishlash – bu sun’iy intellektning kichik sohasi bo‘lib, u mashinalarga inson tilini tushunish va qayta ishlashga yordam beradi. Tabiiy tilni qayta ishlash (natural language processing, NLP) vazifalarining aksariyati uchun eng asosiy qadam tabiiy tildagi shablon (qolip)larni tushunish va dekodlash uchun so‘zlarni raqamlarga aylantirishdir. NLPda bu bosqich matnli ko‘rinish (text representation) deb yuritiladi (Naseem, 2021; 35., Chai, 2023; 45., Probierz, 2023; 2846).

Mavzuga oid adabiyotlar tahlili

Til korpusi tuzish tamoyillari bilan shug‘ullangan olim borki, matnlarni lingvistik izohlash, annotatsiya bosqichlarini albatta tadqiq qiladi. Jahon olimlari ham bu borada o‘z ilmiy faoliyatida bir qancha samarali ishlarni amalga oshirishgan va yuqori natijalarni fanga ma’lum qilishgan.

Christopher D. Manning “The Stanford Part-of-Speech Tagger” maqolasida ingliz tilida avtomatik POS tagging (so‘z turkumi belgilash) algoritmini ishlab chiqqan. Naive Bayes, HMM (Hidden Markov Models), CRF (Conditional Random Fields) kabi statistik modellar asosida ishlovchi Stanford POS Tagger tizimini yaratgan. Manningning yondashuvi o‘zbek tiliga moslashtirish uchun asos bo‘la oladi, ayniqsa CRF modeli kontekstni inobatga olgan holatda so‘z turkumini aniqlash imkonini beradi. Aynan mana shu metodologiya o‘zbek tilida ham natijador bo‘lishi mumkin, agar korpus etarlicha katta va xilma-xil bo‘lsa (Manning, 2003; 8-12).

Ya’ni bir korpus matnlari bilan tadqiqot olib borgan olim Adam Kilgarriff bo‘lib, uning ELRA jurnalida chop etilgan “Corpora and Lexicons: What’s in a Word?” maqolasi korpus matnlaridagi leksik birliklar tadqiqiga bag‘ishlangan. Kilgarriff korpuslarda leksik birliklarni avtomatik tarzda teglash va lemmatizatsiya qilishni tahlil qilgan. U korpus asosida so‘z ma’nolarini tahlil qilish uchun statistik yondashuvlarni taklif qilgan (Kilgarriff, 1997; 10-15). Uning ishlari leksik semantika bilan bog‘liq taglashda muhim asos hisoblanadi. O‘zbek tilida ko‘p ma’nolilik (polisemiya) muammosi aynan Kilgarriff yondashuvlari asosida hal etilishi mumkin.

Germaniyalik lingvist, xususan NLP mutaxasisi (kompyuter lingvisti) Helmut Schmiddir. U hozirda Ludwig-Maximilians-Universität München (LMU Munich) da “Centrum für Informations- und Sprachverarbeitung” (Axborot va Tilni Qayta Ishlash Markazi) markazida faoliyat yuritadi. Bu yerda u nutqni avtomatik tahlil qilish, POS-teg qo‘llash, morfologik tahlil va parserlash kabi sohalarda ilmiy-tadqiqot olib boradi. Asosiy tadqiqot ishlari kompyuter lingvistika texnologiyalari – masalan, “TreeTagger”, “RNNTagger”, “SFST”, “SMOR”, “BitPar” bilan bog‘liq bo‘lgan. Birgina “Treetagger – a tool for annotating text with part-of-speech and lemma information” maqolasi uni dunyoga tanitib yubordi (Schmid, 1995; 56-59). TreeTagger dasturi morfologik va lemmatik izohlash uchun eng mashhur vositalardan biri. Schmid modeli o‘z ichiga so‘zlar uchun ehtimollik asosidagi POS teglarni belgilaydi.

Yana bir kompyuter lingvistikasida tadqiqot olib boruvchi mashhur shvetsiyalik olim (kompyuter lingvistikasi professori-2008) – Joakim Nivre. U hozirda Shvetsiyaning Uppsala universitetida (Universitetet i Uppsala) faoliyat olib boradi. Universal Dependencies (UD) loyihasining asoschilaridan biri. Har xil tillar uchun yagona sintaktik annotatsiya modeli ishlab chiqqan. Bu modelda har bir til uchun umumiy belgilash sxemasi joriy qilingan.

U o'z ilmiy fikrlarini "Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection" maqolasida yoritib o'tgan (Nivre, 2020; 10).

UD loyihasi hozirda o'zbek tilini ham o'z ichiga olgan. Annotatsiyalarni xalqaro standartlarga mos ravishda ishlab chiqish uchun UD modeli asosiy tayanch bo'la oladi.

Yuqoridagi olimlar tomonidan yaratilgan teglash tizimlari va annotatsiya metodikasi mashina o'rganishi, statistik model, sintaktik bog'liqlik va lemmatizatsiya asosida shakllangan. O'zbek tili uchun esa bu tajribalardan foydalangan holda milliy til xususiyatlariga moslashtirilgan annotatsiya tizimi ishlab chiqish dolzarb vazifalardan biri hisoblanadi.

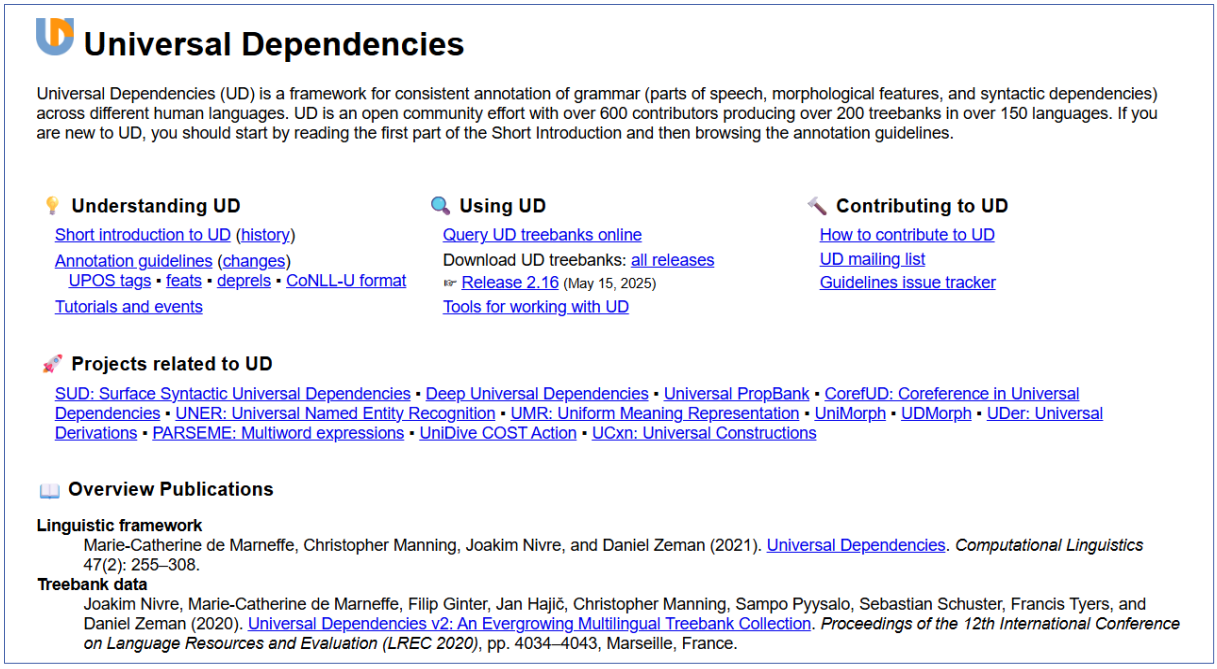
Tadqiqot metadologiyasi

Korpus lingvistikasi – til materiallarini kompyuterga asoslangan korpuslar orqali tahlil qilishni nazarda tutadi. Til korpuslari lingvistik izoh (annotatsiya) orqali ma'lumotga boyitiladi. Annotatsiya deganda, matndagi so'zlarning grammatik, sintaktik yoki semantik tavsiflari kiritilishi tushuniladi.

Morfologik annotatsiya – so'zning leksik shakli, grammatik kategoriyalari (masalan, ot, fe'l, birlik, ko'plik) kabi belgilar bilan izohlanadi. Sintaktik annotatsiya esa so'zlar orasidagi grammatik bog'liqliklarni (masalan, ega-kesim, aniqlovchi-to'ldiruvchi) ko'rsatadi. Chet tillarida *Universal Dependencies (UD)* loyihasi asosida bu jarayonlar ancha takomillashgan (Jurayev, 2021; 27).

Universal Dependencies (UD) – bu tabiiy tillarni sintaktik va morfologik jihatdan izohlash uchun ishlab chiqilgan xalqaro standartlashtirilgan ramka (standart) hisoblanadi. U turli tillarda umumiy tahlil modelini yaratishga xizmat qiladi, shu bilan birga har bir tilning o'ziga xos xususiyatlarini ham hisobga oladi. UD – bu 150 dan ortiq tillarni qamrab olgan, 200 dan ortiq "daraxtbanklar" (treebank) yaratgan, 600 dan ortiq ishtirokchilardan iborat ochiq hamjamiyat tashabbusidir. (1-rasm)

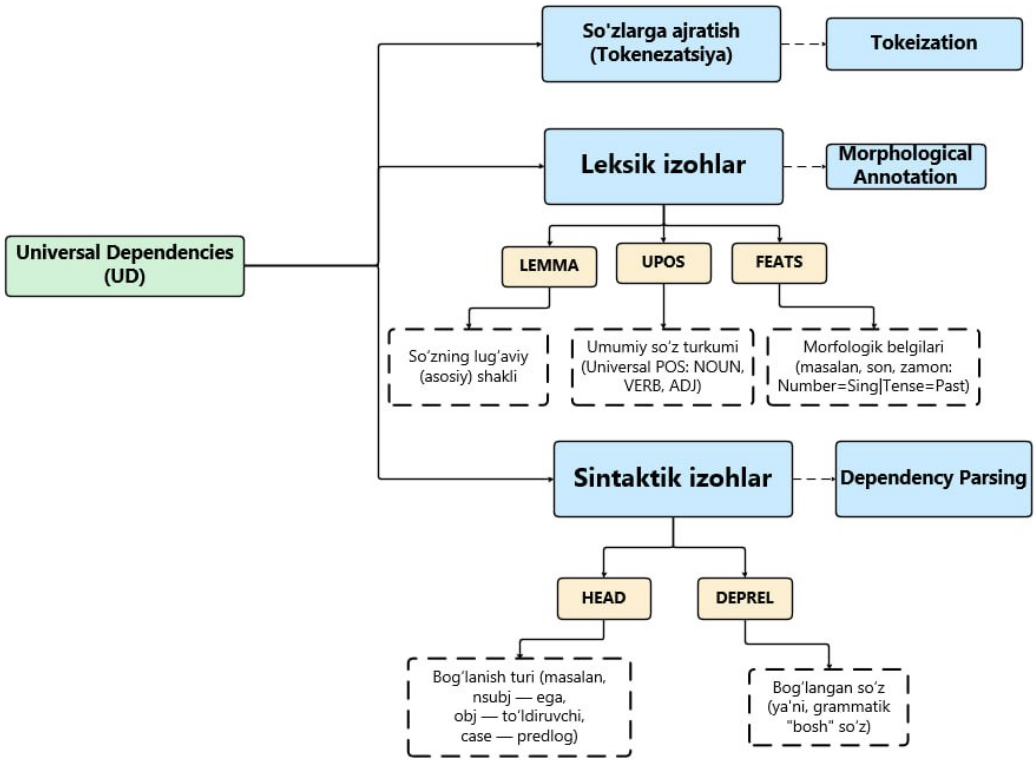
Universal Dependencies (UD)ni bir vaqtning o'zida *lingvistik izohlash ramkasi* deb ham atash mumkin. Chunki ushbu standart asosida butun dunyo tillari korpus yaratish faoliyatida foydalanishadi. Ya'ni bunda korpusdagi matnlar lingvistik izohlanadi (morfologik, sintaktik analiz qilinadi, maxsus teglash ishlab chiqilib, so'zlarga biriktiriladi).



1-rasm. *Universal Dependencies (UD) interfeysi (https://universaldependencies)*

Ushbu atama tarkibiy tarjimasi **Universal** – umumiy, xalqaro, barcha tillar uchun yagona bo'lgan → *Umumtil* (yoki "Umumiy tilga oid"), **Dependencies** – sintaktik bog'liqliklar (ya'ni, so'zlar orasidagi grammatik munosabatlar) → *Bog'liqliklar, Sintaktik munosabatlar* kabi ma'nolarni ifodalaydi. Ko'pgina ilmiy adabiyotlarda ushbu tushunchani turlicha shakllarda qo'llashadi. Masalan, **Umumtil sintaktik bog'liqliklar** *(eng keng tarqalgan va aniq ifoda)*, **Umumiy til bog'lanmalari** *(soddalashtirilgan lekin noaniqroq)*, **Xalqaro sintaktik izohlash standarti** *(ko'proq UD loyihasining mohiyatini ifodalaydi)*. Ular orasidan eng maqbul va mazmunan tog'ri variant sifatida **Universal Dependencies (UD) – Umumtil sintaktik bog'liqliklar** deb ifodalash maqsadga muvofiq.

UD bilan lingvistik izohlash jarayoni bir qancha bosqichni qamrab oladi (2-rasm). Avvalo, matn tarkibidagi berilgan gaplar tokenlar(odatda so'z va tinish belgilariga)ga ajratiladi. Keyingi bosqich leksik izohlash(morfologik jarayon boshlanishi)ga o'tiladi. Har bir token uchun (lemma, upos, feats)morfologik xususiyatlar biriktiriladi. Oxirgi bosqich sintaktik izohlash bosqichida so'zlar orasidagi grammatik bog'liqliklar aniqlanadi. Bunda bog'langan so'z va bog'lanish turlari aniqlanadi.



2-rasm. *Universal Dependencies (UD – Umumtil sintaktik bog'liqliklar) tizimi*

Universal Dependencies (UD) tillararo tahlilni osonlashtirish, NLP (Natural Language Processing) ilovalari uchun standart sintaktik va morfologik belgilash, annotatsiyalangan korpuslar yaratish maqsadlarini ilgari suradi.

Shunday avtomatik annotatsiya tizimlari mavjudki, ulardan hozirgi kunga qadar tadqiqot ishlarida foydalanib kelishadi: (1-jadval).

№	Avtomatik annotatsiya tizimi	Ta’rifi
1	UDPipe	Universal Dependencies asosida grammatik izohlarni yaratish uchun.
2	Stanza (Stanford NLP)	Ko‘p tilli, shu jumladan o‘zbek tilini qisman qo‘llab-quvvatlovchi tizim.
3	SpaCy	Yuqori tezlikda ishlaydigan NLP vositasi.

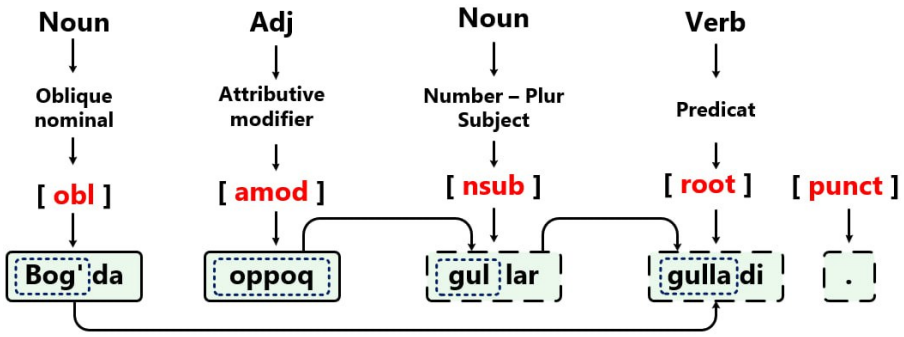
1-jadval. *Avtomatik annotatsiya tizimlari*

Tahlil va natijalar

Matnlarni lingvistik izohlash jarayoni bir qancha bosqichlarni o‘z ichiga oladi:

1. **Tokenizatsiya** – matnni so‘z birliklariga ajratish.
2. **Lemmatizatsiya** – har bir so‘zni lug‘aviy shakliga keltirish.
3. **POS tagging** – so‘z turkumini aniqlash.
4. **Dependency parsing** – sintaktik bog‘liqliklarni aniqlash.

Bu jarayonni bitta misol orqali ko‘rib chiqamiz. “Bog‘da oppoq gullar gulladi”. Ushbu gapda ikkita ot (Noun - N) (bog‘da, gullar), bitta sifat (Adjective - Adj) (oppoq), gulladi (fe’l) (Verb - V) mavjud. Sintaktik tahlilda esa bog‘da – oblique nominal (obl) ravish, oppoq – attributive modifier (amod) aniqlovchi, gullar – subject (subj) ega, Gulladi – predicat(gapning asosi) (root) kesim. Nuqta – tinish belgisi esa punctuation (punct).



ID	So‘z	Lemma	UPOS	XPOS	Feat	Head	DepRel
1	Bog'da	bog‘	NOUN	NOUN	Case=Loc	4	obl
2	oppoq	oppoq	ADJ	ADJ	Degree=Pos	3	amod
3	gullar	gul	NOUN	NOUN	Number=Plur	4	nsubj
4	gulladi	gullamoq	VERB	VERB	Mood=Ind	Tense=Past	Person=3
5	.	.	PUNCT	PUNCT	_	4	punct

- **Bog‘da** – joyni bildiruvchi so‘z, *bog‘* so‘zining joy (locative) holidagi shakli. Bu yerda *obl* (oblique nominal) sifatida predikatga bog‘langan.
- **Oppoq** – sifat, *amod* (attributive modifier), ya’ni “gullar” so‘zini aniqlovchi so‘z.
- **Gullar** – ko‘plikda ot, gapning bosh ishtirokchisi (*nsubj*).
- **Gulladi** – asosiy fe’l (predikat), gapning ildiz so‘zi (*root*).
- **.** – nuqta, tinish belgisi (*punct*).

Xulosa va takliflar

O‘zbek tilidagi matnlarni lingvistik izohlash korpus lingvistikasi doirasida dolzarb masala hisoblanadi. Raqamli korpuslar va avtomatik annotatsiya tizimlari bu yo‘nalishda muhim vosita bo‘lib xizmat qiladi. Annotatsiya sifati va aniqligini oshirish uchun tilga xos leksik va grammatik xususiyatlar chuqur tahlil qilinishi, zamonaviy texnologiyalar bilan uyg‘unlashtirilishi kerak.

Ko‘p hollarda, mavjud annotatsiya tizimlari qo‘lda amalga oshirilgan bo‘lib, bu jarayon samaradorligini pasaytiradi. Annotatsiyalash jarayonini avtomatlashtirish esa vaqt va mehnat resurslarini tejash bilan birga, aniqlikni oshiradi.

Adabiyotlar ro‘yxati:

1. Naseem, U., Razzak, I., Khan, S. K., & Prasad, M. (2021). A Comprehensive Survey on Word Representation Models: From Classical to State-of-the-Art Word Representation Language Models. ACM Transactions on Asian and Low-Resource. 2021 Language Information Processing, 20(5). https://doi.org/10.1145/3434237
2. Chai, C. P. (2023). Comparison of text preprocessing methods. *Natural Language Engineering, 29(*3). 2023 https://doi.org/10.1017/S135132492200021
3. Probierz, B., Hrabia, A., & Kozak, J. (2023). A New Method for Graph-Based Representation of Text in Natural Language Processing. *Electronics, 12*(13). 2023 https://doi.org/10.3390/electronics1213284
4. Christopher D. Manning. The Stanford Part-of-Speech Tagger. Stanford University NLP Group. 2003
5. Adam Kilgarriff. Corpora and Lexicons: What’s in a Word?. ELRA Journal. 1997
6. Helmut Schmid. Treetagger — a tool for annotating text with part-of-speech and lemma information. Institut für Maschinelle Sprachverarbeitung, University of Stuttgart. 1995
7. Joakim Nivre. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. *Proceedings of the Twelfth Language Resources and Evaluation Conference* (LREC 2020). European Language Resources Association publisher. 2020, may
8. Jurayev A. Korpus lingvistikasi asoslari. Toshkent: Fan, 2021.
9. Universal Dependencies: https://universaldependencies.org/