



UDK: 1751

Oqila ABDULLAYEVA,
ToshDO'TAU doktoranti
E-mail: abdullayevaokila@gmail.com

ToshDO'TAU professori v.b., f.f.d Sh.Hamroyeva taqrizi ostida

O'ZBEK TILIDA SINTAKTIK PARSING MASALASI: MUAMMO VA TAKLIFLAR

Аннотация

Tabiiy tilni qayta ishlashda (NLP) parsing, treebank usullari gaplarning sintaktik tahlili, gaplardagi birliklarning sintaktik aloqalari va gap tuzilmasini belgilash bilan shug'ullanadi. O'zbek tili morfoanalizatorida 35 000 ta turli uzunlikdagi gaplar POS teglangan bo'lib, keyingi bosqich esa o'zbek tilida sintaktik teglash tizimini ishlab chiqish, teglarni taklif etish hisoblanadi. Maqolada o'zbek tilida sintaktik parsing masalasi, gap bo'laklarini aniqlash modellari va o'zbek tili matnlarini dependency parsing qilish muammolari, sintaktik teglar masalasi tadqiq qilindi. O'zbek tilida sintaktik parsing va treebank masalasi boshqa agglutinativ tillar bilan solishtirilgan holda tahlil qilindi.

Kalit so'zlar: Parsing, dependency parsing, treebanklar, sintaktik teg, korpus.

SYNTACTIC PARSING IN THE UZBEK LANGUAGE: PROBLEMS AND PROPOSALS

Annotation

In natural language processing (NLP), parsing and treebank methods deal with the syntactic analysis of sentences, determining syntactic relationships between units in sentences, and identifying sentence structures. In the Uzbek language morphological analyzer, 35,000 sentences of varying lengths have been POS tagged. The next stage involves developing a syntactic tagging system for the Uzbek language and proposing appropriate tags. This article examines the issue of syntactic parsing in the Uzbek language, models for identifying sentence components, challenges in dependency parsing of Uzbek texts, and the matter of syntactic tags. The topic of syntactic parsing and treebanks in the Uzbek language is analyzed in comparison with other agglutinative languages.

Key words: Parsing, dependency parsing, treebanks, syntactic tag, corpus.

СИНТАКСИЧЕСКИЙ АНАЛИЗ В УЗБЕКСКОМ ЯЗЫКЕ: ЗАДАЧИ И ПРЕДЛОЖЕНИЯ

Аннотация

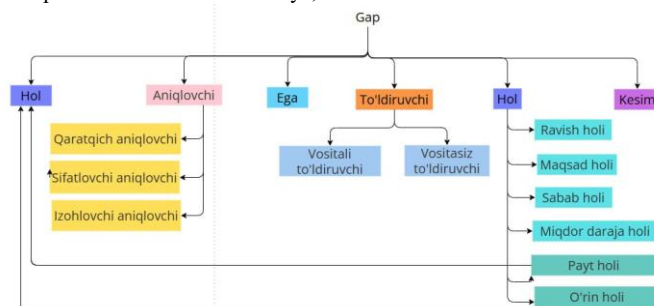
В области обработки естественного языка (NLP) методы синтаксического анализа и создания древовидных структур занимаются синтаксическим разбором предложений, определением синтаксических отношений между единицами в предложениях и выявлением структур предложений. В морфологическом анализаторе узбекского языка 35 000 предложений различной длины размечены по частям речи (POS). Следующим этапом является разработка системы синтаксической разметки для узбекского языка и предложение соответствующих тегов. В данной статье рассматриваются вопросы синтаксического анализа в узбекском языке, модели выявления компонентов предложения, проблемы анализа зависимостей в узбекских текстах и вопрос синтаксических тегов. Тема синтаксического анализа и создания древовидных структур в узбекском языке анализируется в сравнении с другими агглютинативными языками.

Ключевые слова: Синтаксический анализ, парсинг, синтаксический анализ зависимостей, treebank, синтаксический тег, корпус.

Kirish. O'zbek tili boy va murakkab morfologik tuzilishga ega tillardan biri hisoblanadi. Tipik agglutinativ til sifatida o'zbek tili sintaktik xususiyatlariga ko'ra flektiv tillarga nisbatan farqli jihatlariga ega. O'zbek tili sintaktik tahlilida gaplarda SOV (Subyekt-Obyekt-Fe'l) so'z tartibiga ega bo'lib, erkin so'z tartibli til hisoblanadi. Bundan tashqari, o'zbek tili matnlarida asosiy gap bo'lagi ega va kesim hisoblansa-da, zarur hollarda ega gapda tushib qoldirilishi mumkin, bunday vaziyatda ham gap qurilishi va mazmuni buzilmaydi, kesim tarkibidagi grammatik qo'shimchalar orqali tiklanishi mumkin.

Mavzuga oid adabiyotlar tahlili. Tillarda matnlarni qayta ishlash bosqichlari sifatida ko'p holatlarda tokenizatsiya,

lemmatizatsiya, POS tegging, sintaktik va semantik parsing bosqichlari ko'rsatiladi. Biz ham o'zbek tili korpusi uchun kelajakda sintaktik analizator qurishni, matnlarni sintaktik tahlil qiluvchi, gap bo'laklarini aniqlovchi parser dasturlarni yaratishni maqsad qilganmiz. O'zbek tilida gaplarni sintaktik tahlil qilish – gapning qurilishini aniqlash, gapdagi birliklarning vazifasini va tartibini belgilashdan iborat. O'zbek tilida gap tuzilishi modeli $\downarrow < \text{Aniqlovchi} > \oplus < \downarrow \text{Ega} > \downarrow \oplus < \text{To'ldiruvchi} > \downarrow \oplus < \text{Hol} > \oplus < \text{Kesim} >$ shaklida bo'ladi. Quyidagi chizma orqali aniqroq tasavvur qilish mumkin.



1-chizma. O'zbek tilida gap qurilishi

O'zbek tilida gap bo'laklarining so'z tartibi mavjud bo'lsa-da, ammo tilda so'zlarning erkin joylashuviga ko'ra yuqoridagi chizmadan farqli qurilishlarni kuzatish mumkin. O'zbek tilida so'z tartibida moslashuvchanlik tufayli gapni SOV tartibidan OSV tartibiga ham o'zgartirish mumkin. Masalan, Universitetga ko'pchilik talabalar kelishmadi kabi. Gapdagi so'zning tartibi o'zgarsa-da, sintaktik bo'lakning vazifasi o'zgarishsiz qoldi. Chunki birlikning sintaktik vazifasini ko'rsatib turuvchi morfologik shaklga ega bo'lishidir. Ba'zi bo'laklarning

tartibi o'zgarganda sintaktik vazifasi ham o'zgaradi, bu o'rinda so'z tartibi muhim rol o'ynaydi.

Parser dasturlarini qurishda bevosita POS teglash jarayoni muhim hisoblanadi. Chunki o'zbek tilida so'zlar orasidagi sintaktik munosabatlarni aniqlashda morfologik tuzilma muhim rol o'ynaydi. Chunki gap bo'laklari muayyan grammatik shakllarda keladi. Shuning uchun o'zbek tili ta'limiy korpusi uchun yaratilgan morfoanalizatorida barcha matnlar morfologik tegladi. Hozirda 14 ta so'z turkumi (POS) tegi mavjud bo'lib, ular 1-jadvalda ko'rsatilgan.

POS teg	Belgilanishi	POS teg	Belgilanishi
Ot	N	Ko'makchi	Il
Sifat	JJ	Yuklama	Prt
Fe'l	VB	Modal so'z	MD
Son	NUM	Undov so'z	UH
Ravish	RR	Taqid so'zlar	IM
Olmosh	P	Atogli otlar	NER
Bog'lovchi	C	Iboralar	IB

1-jadval. O'zbek tilidagi POS teglar ro'yxati (qisqa variant)

So'nggi yillarda turkiy tillarda tabiiy tilning sintaktik tahlili bo'yicha sezilarli yutuqlarga erishildi, asosan, gapdagi so'zlar orasidagi sintaktik munosabatlarni aniqlash orqali uning grammatik tuzilishini ko'rsatadigan bog'liqlik tahlili (dependency parsing) dasturlari yaratildi. Bu ma'lumotlarga asoslangan (data-driven) yoki qoidalarga asoslangan usullar orqali amalga oshirildi. Turkiy tillardan turk tilida METU sabanci, BOUN Treebank, BoAT annotatsiyalovchi dastur va BIMST Treebankini, uyg'ur tilida, qozoq tilida hamda qirg'iz tillarida yaratilgan UD formatidagi Treebanklarni misol sifatida keltirish mumkin.

Til	Treebank nomi	Tashkilot/ Manba	Izoh	Ma'lumotlari
Turk tili	METU-Sabanci Turkcha Treebank	ODTÜ (METU), Sabanci Universiteti	Eng katta Turk tili tahlil daraxtblaridan biri	Bu turk tilidagi birinchi dependency treebank bo'lib, 7,262 jumlani o'z ichiga oladi va morfologik hamda sintaktik teglashni o'z ichiga oladi. Universal Dependencies (UD) standartlari asosida.
Turk tili	BOUN Treebank	Bog'azichi Universiteti	UD standartiga moslashgan Turk tili treebanki	BOUN Treebank turk tilidagi qo'lda teglangan yangi treebank bo'lib, 9,761 jumlani o'z ichiga oladi. Bu treebankni yaratishda Universal Dependencies (UD) asos qilib olingan.
Turk tili	BoAT (Bogazici Annotation Tool)	Bog'azichi Universiteti	Turk tili sintaktik annotatsiyalovchi vosita	Tabiiy tilni qayta ishlash (NLP) uchun mo'ljallangan, sintaktik va semantik belgilash (annotatsiya) jarayonlarini soddalashtirish va avtomatlashtirishga yordam beradigan dasturdir.
Turk tili	BIMST (Bogazici-ITU-METU-Sabanci Treebank)	Bog'azichi, ITU, METU, Sabanci universitetlari	Katta hajmli birlashtirilgan treebank	Ushbu treebank IMST korpusining yangilangan versiyasi bo'lib, yangi teglash sxemasi asosida qo'lda qayta annotatsiyalangan 5,635 gapni o'z ichiga oladi.
Qirg'iz tili	UD Kyrgyz-KTMU	Qirg'iz-Turkiya Manas Universiteti (KTMU)	UD formatida annotatsiyalangan Qirg'iz tili treebanki	781 jumla (7.4K token)dan iborat. Lemmatizatsiya, morfologik teglash va sintaktik munosabatlarni qo'lda teglangan.
Qozoq tili	UD Kazakh KTB	L.N. Gumilyov nomidagi Yevrosiyo Milliy Universiteti (ENU)	Qozoq tilining UD formatidagi treebanki	UD loyihasi doirasida ishlab chiqilgan. Qozoq tilining morfologik va sintaktik xususiyatlarini aks ettiruvchi belgilangan (annotatsiyalangan) korpus.
Uyg'ur tili	UD Uyghur UDT	S.Mamitimin, M.Aili va boshqalar	Uyg'ur tili uchun UD formatidagi treebank	UD Uyghur UDT (Uyghur Universal Dependency Treebank) 3,456 ta gap va 40,236 ta token (so'z)dan iborat.

2-jadval. Turkiy tillarda mavjud treebanklar va teglash dasturlari

Turk tilida yaratilgan Sintaktik tahlil treebanki - METU-Sabanci dasturida 13 ta POS teg va Universal Dependencies (UD) modeli asosida ishlab chiqilgan 20 dan oshiq yaqin sintaktik munosabatlarni belgilovchi teg mavjud [1]. BoAT (Bogazici Annotation Tool) [2] annotatsiyalovchi dasturida 37 ta, uyg'ur tilida [3] 23 ta dependency tegi, qozoq tilida [4] 40 ta, qirg'iz tilida [5] 25 ta sintaktik tegi mavjud. Teglarining tuzilishi o'rganilganda ko'pchiligi UD standart teglaridan olingani va ba'zilar moslashtirilganini ko'rishimiz mumkin. O'zbek tili uchun ham ba'zi teglar tavsiya qilingan. Protégé texnologiyasi asosida o'zbek tilidagi sintaktik strukturalarni teglash uchun 11 ta sintaktik teglar tizimi taklif etilgan [6].

Tadqiqot metodologiyasi. Tabiiy tilni qayta ishlash bosqichlarida "Meaning-Text Theory" (MTT) nazariyasi tamoyillarini qo'llaganda holda tahlil qilish anchayin keng tarqalgan. Ma'no-Matn Nazariyasi (MTT) turli tillarni tahlil qilish va ishlov berish uchun qo'llanilgan universal lingvistik yondashuvlardan bo'lib, ayniqsa morfologik va sintaktik jihatdan murakkab tillarga mos keladi. Bu nazariya asosida MTT

tomonidan taklif qilingan ko'p qatlamli til modeli - Sem $\Leftrightarrow$ DSynt $\Leftrightarrow$ SSynt $\Leftrightarrow$ DMorph $\Leftrightarrow$ SMorph $\Leftrightarrow$ DPhon $\Leftrightarrow$ SPhon hisoblanadi. MTT ning ko'p qatlamli tabiati uni turli xil morfologik va sintaktik tizimlarga ega tillarga moslashtirish imkonini beradi. Shu bois, u ko'plab tillar uchun qo'llanilgan va yangi tillarga tatbiq etish imkoniyati ham mavjud. O'zbek tilida esa hali to'liq ishlab chiqilmagan. Sem $\Leftrightarrow$ DSynt $\Leftrightarrow$ SSynt $\Leftrightarrow$ DMorph $\Leftrightarrow$ SMorph $\Leftrightarrow$ DPhon $\Leftrightarrow$ SPhon modeli ko'p darajali model bo'lib, har bir bosqichda oldingi darajadagi tahlillash keying bosqich uchun xizmat qiladi, ya'ni fonetik annotatsiyalashdan semantik annotatsiyalashgacha bo'lgan bosqichni o'z ichiga oladi. O'zbek tilida ham morfologik teglash amalga oshirilgan va keyingi bosqichda gap bo'laklarini aniqlashda muhim hisoblanadi. Tadqiqotimizning hozirgi holatida o'zbek tili matnlarini sintaktik parsing qilish uchun 18 ta sintaktik teglar taklif qilinmoqda. Quyidagi jadvalda gap bo'laklarini teglash uchun ishlab chiqilgan teglar ro'yxatini ko'rish mumkin.

POS teg	Belgilanishi	POS teg	Belgilanishi
Ega	E	Hol	
Kesim		Ravish holi	RH
Ot-ketim	OK	Payt holi	PH
Fe'l-ketim	FK	O'rin holi	O'H
Aniqlovchi		Sabab holi	SH
Sifatlovchi-anqiqlovchi	SA	Maqsad holi	MH
Izohlovchi-anqiqlovchi	IA	Miqdor-daraja holi	MDH
Qaratqich-anqiqlovchi	QA	Kirish so'z	KS
To'ldiruvchi		Kirish birikma	KB
Vositali-To'ldiruvchi	VT	Kirish gap	KG
Vositasiz-To'ldiruvchi	VST	Undalma	U

3-jadval. O'zbek tilidagi sintaktik teglar ro'yxati (qisqa variant)

Tahlil va natijalar. O'zbek tili ham boshqa agglutinativ tillar qatori gapda asos va unga qo'shilgan qo'shimchalar to'plamidan iborat. Gaplar ma'lum sintaktik tuzilishga ega birikmalarni birlashtirish orqali emas, ko'proq grammatik shakllarni qo'shish va erkin so'z

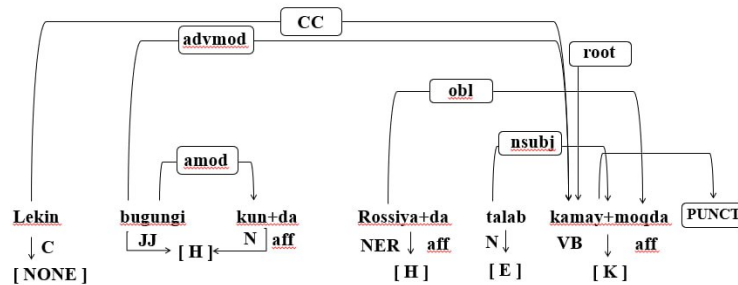
tartibi orqali shakllanadi. Masalan, lug'aviy shakl hosil qiluvchi qo'shimchalar so'zning lug'aviy ma'nosini o'zgartirmaydi, balki grammatik rolini aniqlaydi. Masalan: -ning qaratqich kelishidagi birliklar gapda qaratqich-aniqlovchi, -ni tushum kelishidagi so'zlar esa vositasiz to'ldiruvchi vazifasida keladi. Egalik qo'shimchasini olgan birliklar esa ega gap bo'lagi vazifasida keladi.

Grammatik qo'shimchalar	Misol	Gapdagi vazifasi
Wismalar+egalik qo'shimchalari	Maktab+im katta.	Ega
Wismalar+tushum kelishigi	Kitob+ni ber.	Vositasiz to'ldiruvchi
Wismalar +jo'nalish kelishigi	Imtihon+ga ruxsat berilmadi.	Vositali to'ldiruvchi
Wismalar +o'rin-payt kelishigi	Saytda xabarlar berilmagan.	Vositali to'ldiruvchi
Wismalar+qaratqich kelishigi	Talabalar+ning hammasi keldi.	Qaratqich aniqlovchi
Wismalar+ko'plik qo'shimchasi	Kelgan+lar baholandi.	Ega
Wfe'l+shaxs-son, zamon, mayl qo'shimchasi	Zavodda ishlashmayapti.	Kesim

4-jadval. Gaplarda gap bo'lagining uchrashi

Shuningdek, gapdagi birliklarning sintaktik munosabatlarini ham grammatik qo'shimchalar orqali aniqlash mumkin.

Masalan: "Lekin bugungi kunda Rossiyada talab kamaymoqda" gapini parsing qilib, gapdagi bo'laklarning vazifasini va sintaktik aloqalarini quyidagi chizma orqali tahlil qilish mumkin:



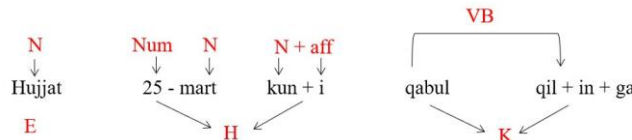
1-chizma. O'zbek tilida gapni dependency parsing qilish

Gapdagi har bir birlik grammatik shakliga ko'ra gap bo'lagi vazifasini bajarmoqda. Ammo gaplarda sintaktik aloqaga kirishmaydigan bo'laklar ham mavjud bo'ladi: -"lekin" bog'lovchisi [NONE] tegi bilan annotatsiyalandi, chunki bu birlik sintaktik aloqaga kirishmaydi, [CC] UD tegi bog'lovchilik munosabatini bildirgani uchun belgilangan; -"bugungi kunda" birikmasi gapda payt holi [H] gap bo'lagi bo'lib, "bugungi - kunda" {JJ va N} munosabati UD universal [amod] tegi bilan, "bugungi kunda - kamaymoqda" birikmasi dependency parsing tahliliga ko'ra tobe munosabatni ifodalab, [advmod] UD tegi bilan belgilangan; - "Rossiyada" o'rin-joy holi [H] gap bo'lagi bo'lib, "Rossiyada kamaymoqda" tobe munosabatli sintaktik aloqasi uchun [obl] UD tegi bilan belgilangan; -"talab kamaymoqda" birikmasida "talab" ega [E], "kamaymoqda" kesim [K] vazifasida kelgan, birgalikda ega-kesimlik tobe munosabatiga ko'ra [nsubj]

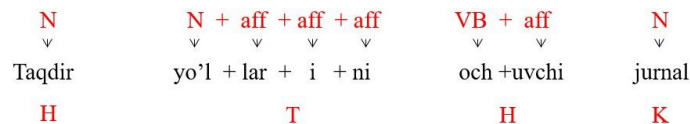
UD tegi bilan belgilangan; barcha tinish belgilari [PUNCT] UD tegi bilan belgilanganligi sababli, nuqta ham xuddi shu shaklda annotatsiyalangan.

O'zbek tili murakkab morfologik tuzilishga ega tillardan biri hisoblanadi. Gapdagi birliklarning morfologik shakllari o'sha birlikning vazifasini aniqlashga yordam beradi. Masalan, ot so'z turkumidagi so'zlar bosh kelishik yoki ko'plik shaklida doim ega gap bo'lagi vazifasida kelishi, birliklar shaxs-son, zamon, mayl qo'shimchalarini olsa kesim vazifasida kelishi aniq. Parsing dasturini tuzishda qoidalarga asoslangan usullar orqali avtomatik sintaktik tahlil qilishda aynan mana shu kabi grammatik qoidalar muhim hisoblanadi. Ammo yuqori natijaga erishish mumkinligi shubhali. Misollar orqali ko'ramiz. Masalan,

Hujjat 25-mart kuni qabul qilingan.



Taqdir yo'llarini ochuvchi jurnal.



Yuqoridagi misollar orqali ko'rish mumkinki, 1-gapdagi ot so'z turkumidagi bosh kelishikdagi birlik qoidaga ko'ra ega bo'lib, 2-gapda esa ot so'z turkumidagi bosh kelishik shaklida ko'ringani bilan belgisiz qaratqich kelishigini olganligi sababli aniqlovchi vazifasida kelmoqda. 2-gapdagi "jurnal" so'zi bosh kelishikda bo'lsa-da, kelish o'rni va vazifasiga ko'ra ot-kesimdir. O'zbek tili grammatik qoidalariga ko'ra gaplar SOV shaklida bo'lsa-da, bu ham qoidalarga asoslangan usul orqali tuzilgan

dasturda yuqori natijani bermaydi. Chunki gap tuzilishi ham muloqat maqsadiga ko'ra o'zgarishi mumkin. MTT tomonidan taklif qilingan Sem ⇔ DSynt ⇔ SSynt ⇔ DMorph ⇔ SMorph ⇔ DPhon ⇔ Sphon ko'p qatlamli til modelining SSynt (surface syntactic) tuzilmasi orqali gaplarning sintaktik bo'laklari aniqlandi.

Tokzorni eslayman.	[[T]] [[K]]
Yaradorlar kasalxonaga yotqizilgan.	[[E]] [[H]] [[K]]
Yog'ochidan yaxshi mebel chiqarkan.	[[T]] [[A]] [[K]]
Biroq hech qachon xazina topilmagan.	[[H]] [[E]] [[K]]
Bu haqda EFE agentligi ma'lum qildi.	[[T]] [[E]] [[K]]
Hojatxonalar yo'qligi bitta masala.	[[A]] [[E]] [[A]] [[K]]
Rasmiylar namoyishchilar bilan kurashish uchun 13 ming politsiyachini safarbar qilgan.	[[E]] [[T]] [[A]] [[T]] [[K]]

5-jadval. Sodda gaplarning sintaktik tuzilishiga misollar

Yuqoridagi jadvalda sodda gaplarning sintaktik tuzilishi gap bo'laklarini parsing qilish orqali aniqlanganini ko'rish mumkin. Katta masshtabli matnlar yig'ilgan korpus orqali gaplarning turli tuzilmalarini tahlil qilish mumkin. Bu esa qoidalarga asoslangan $\downarrow < \text{Aniqllovchi} > \oplus < \downarrow \text{Ega} > \downarrow \oplus < \text{To'ldiruvchi} > \downarrow \oplus < \text{Hol} > \oplus < \text{Kesim} >$ modeli orqali tahlil dasturini qurish har doim ham o'zini oqlamasligini ko'rsatadi.

Xulosa va takliflar. O'zbek tilida parser dasturini yaratish uchun kamida 20 000 ta gapli korpus bo'lishi va bu

gaplar sintaktik analizatorida trening qilinishi kerak. Parser dasturida gap bo'laklarini belgilovchi 18 ta teg taklif qilinmoqda. Dependency parsing uchun UD universal teglarini moslashtirganda holda qabul qilinishi maqsadga muvofiq. MTT tomonidan taklif qilingan $\text{Sem} \Leftrightarrow \text{DSynt} \Leftrightarrow \text{SSynt} \Leftrightarrow \text{DMorph} \Leftrightarrow \text{SMorph} \Leftrightarrow \text{DPhon} \Leftrightarrow \text{Sphon}$ ko'p qatlamli til modelining SSynt (surface syntactic) tuzilmasi orqali gaplarni sintaktik tahlil qilish, modellarini ishlab chiqish zarur hisoblanadi. Keyingi tadqiqotlarda o'zbek tilida sintaktik teglarni tanlash yuzasidan lingvistik tahlillar o'tkazilishi zarur.

ADABIYOTLAR

1. Atalay B. Nart, ets. All, Oflazer Kemal, Say Bilge. The Annotation Process in the Turkish Treebank. Proceedings of 4th International Workshop on Linguistically Interpreted Corpora (LINC-03) at EACL 2003.
2. Salih Furkan Akkurt, Bushra Marshan, Susan Uskudarli. BoAT v2 - A Web-Based Dependency Annotation Tool with Focus on Agglutinative Languages. ALTNP2022: The International Conference on Agglutinative Language Technologies as a challenge of Natural Language Processing, June 6-8, 2022, Koper, Slovenia.
3. Mairehaba Aili va boshq. Universal dependencies for Uyghur. Proceedings of WLSI/OIAF4HLT, pages 44–50, Osaka, Japan, December 12 2016.
4. Aibek Makazhanov, Aitolkyn Sultangazina, Olzhas Makhambetov, Zhandos Yessenbayev. Syntactic annotation of Kazakh: following the universal dependencies guidelines. a report. 3rd International Conference on Computer Processing in Turkic Languages. Kazan, Tatarstan, 2015.
5. Benli Ibrahim, Sharshembayev Bakyt. Dependency Parsing Based Treebank for Kyrgyz Language. <http://ymerdigital.com/>
6. Abduraxmonova N. O'zbek tili elektron korpusining kompyuter modellari. Monografiya. – Toshkent, 2021.
7. Sayfullayeva R. va boshq. Hozirgi o'zbek adabiy tili. – Toshkent, 2009.
8. Mahmudov N., Nurmonov A. O'zbek tilining nazariy grammatikasi (Sintaksis). – Toshkent, 1995.