O‘ZBEK TILI KORPUSI UCHUN TEGLASH TIZIMINI ISHLAB CHIQISH MASALASI

Abdullayeva Oqila Xolmo‘minovna,
Filologiya fanlari falsafa doktori (PhD)
abdullayevaoqila@gmail.com
ToshDO‘TAU doktoranti DSc

Annotatsiya. Tabiiy tilni qayta ishlash (NLP)da matnlarni morfologik va sintaktik teglash va teglangan korpuslarni yaratish eng muhim vazifalardan biri hisoblanadi. Bugungi kunda universal POS teglar tizimidan ko‘plab tillarda tegsetlar tizimining asosi sifatida foydalanishib, keyinchalik har bir tilga xos xususiyatlarni ifodalovchi qo‘shimcha teglar bilan kengaytirilgan. Ushbu universal POS teglar tizimidan ko‘p tilli teglangan NLP ma’lumot platformalaridan biri hisoblangan Universal Dependencies (UD) loyihasi ham ishlab chiqildi. Ammo barcha tillar o‘z sintaksisi, morfologiyasi va fonetikasi bilan bir-biridan farqlanadi, bu esa POS teg va sintaktik teglar tizimini ishlab chiqish zaruriyatini keltirib chiqaradi. Mazkur maqolada tegsetlar tizimining, teglash modellarining yaratilishi bilan bog‘liq masalalar muhokama qilingan.

Abstract. In natural language processing (NLP), one of the most crucial tasks is the morphological and syntactic tagging of texts and the creation of tagged corpora. Currently, the universal POS (Part-of-Speech) tag system is widely used as the foundation for tagging systems in many languages, which has been subsequently expanded with additional tags to represent language-specific features. The Universal Dependencies (UD) project, one of the multilingual tagged NLP data platforms, was also developed based on this universal POS tag system. However, all languages differ in their syntax, morphology, and phonetics, necessitating the development of specialized POS tag and syntactic tag systems. This article discusses issues related to the creation of tagset systems and tagging models.

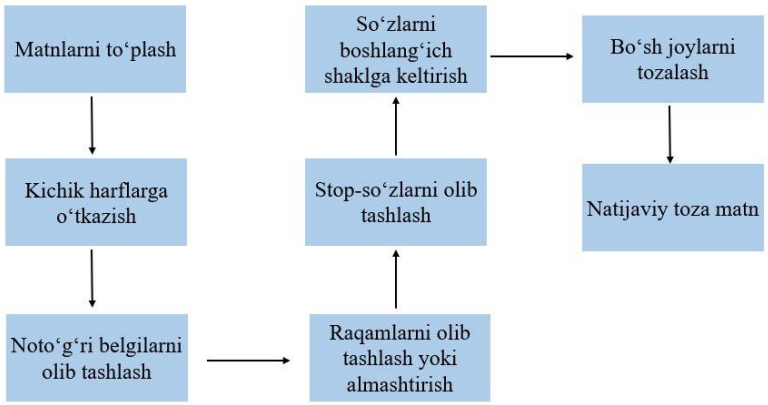
Аннотация. В области обработки естественного языка (NLP) одной из важнейших задач является морфологическая и синтаксическая разметка текстов и создание размеченных корпусов. В настоящее время универсальная система тегов частей речи (POS) широко используется в качестве основы для систем разметки во многих языках, которая впоследствии была расширена дополнительными тегами для отражения специфических языковых особенностей. Проект Universal Dependencies (UD), одна из многоязычных платформ размеченных данных для ОЕЯ, также был разработан на основе этой универсальной системы POS-тегов. Однако все языки различаются по своему синтаксису, морфологии и фонетике, что требует разработки специализированных систем POS-тегов и синтаксических тегов. В данной

статье рассматриваются вопросы, связанные с созданием систем наборов тегов и моделей разметки.

Kalit so‘zlar: *teg, POS teg, sintaktik parsing, parser dasturlar, treebank.*

O‘zbek korpus lingvistikasi sohasidagi so‘nggi tadqiqotlarda o‘zbek tili Milliy korpusini ishlab chiqish, turli maxsus korpus turlarini shakllantirish, korpusda til birliklarini teglash masalasi ko‘rib chiqilmoqda. Korpuslarning yaratilishi bilan bog‘liq tadqiqotlar tahlil qilinganda, ilk morfologik va sintaktik teglash tizimi ishlab chiqilganini kuzatish mumkin. Bu tadqiqotlar XX asrning 50-70-yillariga borib taqaladi. Ilk yirik hajmli til korpusi Broun korpusi ustida amalga oshirilgan va korpus uchun tegsetlar ishlab chiqilgan [1]. Mazkur tadqiqotlar yanada mukammal tegsetlar tizimining, teglash modellarining yaratilishiga zamin bo‘lgan. Keyingi tadqiqotlar tillarning milliy korpuslari yaratilishi va tizimli teglar to‘plamining ishlab chiqilishi bilan bog‘liq bo‘ldi. Shunday tadqiqotlar sirasiga BNC uchun teglashda asos bo‘lib xizmat qilgan CLAWS POS-teglari [2] va statistik hamda qoidaga asoslangan yondashuvlarni birlashtirgan model asosida qurilgan Brill teglarining [3] ustida amalga oshirilgan tadqiqotlar edi. POS teggerlardan keyingi ilmiy izlanishlar morfologik va sintaktik teglarni standartlashtirish va universal teglovchi dasturlarni ishlab chiqish bo‘ldi. Sintaktik teglarni taklif qilgan yirik tizim UD Universal dependency loyihasi bugungi kunda 150 dan ortiq tillar uchun xizmat qilmoqda [4]. So‘nggi tadqiqotlar UD standarti, Universal dependenciesning tahlil qilish modellari, boshqa tillarga moslashuvi, UD grammatikasini ko‘p tilli modellarga integratsiya qilish mavzularida olib borilgan. Agglyutinativ tillarda ham dependency parsing tizimlari ishlab chiqilgan.

Sintaktik parsing va uning algoritmlari bilan ishlashda oddiy matnlar bilan cheklanishning iloji yo‘q. Tabiiy tilga ishlov beruvchi dasturlarni yaratishda xatoliklardan xoli, to‘g‘ri tuzilgan gaplardan iborat toza matnlar to‘plamiga ehtiyoj seziladi. Lekin o‘zbek tilida ishlov berilgan matnlar resursi kam hisoblanadi. Shuning uchun avvalo matnlarni xatoliklardan tozalash va keraksiz belgi hamda kodlarni olib tashlash kerak. Umuman yig‘ilgan matnni tozalash tabiiy tilni qayta ishlashning eng muhim birlamchi jarayoni hisoblanadi. Quyida dasturlar uchun toza matn to‘plamini olishdagi amallarni ko‘rish mumkin.



1-chizma. Matnni dastlabki qayta ishlash bosqichlari

Sintaktik parsing algoritmlarini ishlab chiqishda toza matn bilan cheklanib bo‘lmaydi. Yana qo‘shimcha annotatsiyalash zarur. Parser dasturining maqsadi gap tuzilishini aniqlashdir. Sintaktik tahlil algoritmlari bitta gap ustida amalga oshiriladi. Shuning uchun sintaktik tahlildan oldin bajarilishi kerak bo‘lgan birinchi muhim bosqich – bu gap chegaralarini aniqlashdir. Undan keyingi bosqich sintaktik tuzilmani so‘z yoki tokenlarga (so‘z va tinish belgilari) ajratishdir. Keyingi qayta ishlash – morfologik tahlildir, ya’ni har bir so‘zga asosiy shakl (lemma) va so‘z turkumi, jins, kelishik kabi grammatik ma’lumotlarini belgilovchi morfologik teg qo‘shiladi. Morfologik teg noaniq yoki aniq bo‘lishi mumkin – ikkinchi holatda bu jarayon teglash (tagging) deb nomlanadi. Morfologik tahlil haqidagi ma’lumotlar zamonaviy tahlil vositalari uchun juda muhim hisoblanadi [5]. Biz o‘zbek tili uchun, asosan, https://uznatcorpara.uz/ ga kiritilgan 14 ta POS teglari orqali morfologik tahlil qilmoqdamiz. 12 ta so‘z turkumi yuzasidan POS teglarni kengaytirish va qayta ishlash ustida tadqiqotlar olib borilmoqda. Morfologik va sintaktik teglash tizimi bo‘yicha ko‘plab tadqiqot ishlari olib borilgan. 22 ta tilning o‘zaro o‘xshash turkum va jihatlarini teglovchi universal so‘z turkumlari (Part-of-Speech, POS) to‘plamini ishlab chiqilib, takli qilindi [6]. Bunda aksariyat tillarda mavjud 12 ta so‘z turkumi POS teglari belgilab berilgan. Bunday universal teglar nazoratsiz va tillararo teglash tizimlari ishlab chiqish, turli analizatorlarni (parsers) qurish hamda zarur natijalarni olishda foydalidir. Ayniqsa bu umumiy belgilash tizimiga ega korpuslar mavjud bo‘lmaganda standart yondashuv sifatida qo‘llanilishi va qo‘lda teglanishi mumkin. Mazkur universal tegsetlarning yaratilishi haqida gap ketganda, kelajakdagi sintaktik tuzilmani nazoratsiz o‘rganish bo‘yicha tadqiqotlarni osonlashtirish va eng yaxshi amaliy natijalarni standartlashtirish maqsadida o‘n ikki universal so‘z turkumidan iborat bo‘lgan belgilash tizimi (tagset) taklif qilingani aytiladi. Ushbu belgilash tizimiga qo‘shimcha ravishda, 25 xil daraxtbanki (treebank) belgilash tizimlarini ushbu universal to‘plamga moslashtirish xaritasi ishlab chiqilgan. Natijada, daraxtbanki ma’lumotlari bilan birgalikda ushbu universal belgilash tizimi

va moslashtirish xaritasi 22 ta til uchun umumiy so‘z turkumlarini o‘z ichiga olgan ma'lumotlar to‘plamini hosil qiladi. Ushbu resursdan foydalanishni uchta tajriba orqali namoyish etishgan: turli tillarda belgilash aniqliklarini taqqoslash, standart so‘z turkumlaridan foydalanmaydigan nazoratsiz grammatika induksiya yondashuvini taqdim etish va universal belgilash tizimidan foydalanib, tillarda tobelik munosabatini aniqlovchi analizatorlarni (dependency parsers) yaratish, bu esa zarur natijalarga erishishga imkon berishi ta’kidlangan [6]. Shunday bo‘lsa-da har bir til xususiy va grammatik jihatdan bir-biridan farqliligi sababli universal teglash tizimidan foydalanish doim ham foydali bo‘lmaydi.

O‘zbek tili korpuslari uchun oxirgi tadqiqotlarda teg tizimi taklif qilinganini kuzatish mumkin. Bunda UD (universal dependencies) teglaridan foydalanish holatlari va maxsus tegsetlar ham ishlab chiqilganini kuzatish mumkin. N.Abduraxmonova Protégé texnologiyasi asosida o‘zbek tilidagi sintaktik strukturalarni teglash uchun 18 ta universal teglar tizimi taklif etilgan [7].

1-jadval. Protégé texnologiyasi asosidagi o‘zbek tili sintaktik teglar tizimi

Tag	Name_English	Tag	Name_English
WC	Word combination	**SLP**	Singular personal pronouns
COLC	Collocation	**PPL**	Plural personal pronouns
FP\FCOLC	Free phrase\ Free collocation	**ICN\CPC\T**	Interconnectedness\Complicity
NP	Noun Phrase	**S**	Simple sentence
NA	Noun Adjoinment	**Sub**	Subject
NG	Noun Government	**Obj**	Object
NCS	Noun Collateral subordination	**Attr**	Attributive
VP	Verb Phrase	**Mod**	Modifier
AGRM	Agreement	**Pre**	Predicate

Yana bir guruh olimlar gapning asosini tashkil etuvchi beshta asosiy bo‘lak va gap bo‘lagi bo‘lib kelmaydigan komponentlar uchun sintaktik tegset taklif qilishgan [8]. O‘zbek tilida gaplarda asosiy komponent hisoblanuvchi kesim, ega, aniqlovchi, to‘ldiruvchi, hol va ularning turlari uchun teglar taklif qilingan. Shuningdek, undov va kiritmalar uchun ham teglar mavjud. UD tegsetlari tillarda grammatikani (so‘z va ularning sintaktik bog‘lanishlari) bir xil tarzda belgilash uchun ishlatiladigan bir tizim bo‘lsa-da, ammo tilning o‘z ichki qonun-qoidalarini hisobga olgan holda morfologik va sintaktik teglarning ishlab chiqilishi tabiiydir.

2-jadval. M.Sharipov va boshqa mualliflar tomonidan taklif qilingan sintaktik teglar ro‘yxati

Nomi	Sintaktik teg	Ta’rifi
SUBJECT (ega)	EG	Subject
PREDICATE (kesim)	OK	Noun Predicate
	FK	Verb Predicate

ATTRIBUTE (aniqlovchi)	QA	Genetive Attribute
	SA	Adjectival Attribute
OBJECT (to‘ldiruvchi)	VL	Indirect Object
	VS	Direct Object
	VH	Condition Modifiers
	PH	Time Modifier
MODIFIERS (hol)	OH	Place Modifier
	SH	The Reason Modifier
	MH	The Aim Modifier
EXCLAMATION (undov)	UN	A person or object that is focused on speech
THE ENTRY WORD (kiritma)	KR	The entry word

Yuqorida taklif qilingan tegsetlar biz tadqiqotimiz uchun ishlab chiqqan tegsetlar tizimiga o‘xshash bo‘lsa-da, ammo ba’zi teglarning belgilanishi farqlidir. Qachonki umumiy kelishuvga kelinganda, O‘zbek tili teggeri ishlab chiqilishi mumkin. Quyida o‘zbek tili ta’limiy korpusi uchun ishlab chiqilgan sintaktik teglar tizimi bilan tanishish mumkin.

3-jadval. O‘zbek tili sintaktik analizatori uchun taklif qilingan sintaktik teglar ro‘yxati

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

Agglyutinativ va boy morfologiyaga ega til sifatida o‘zbek tilining parser dasturlarini ishlab chiqishda ba’zi qiyinchiliklarga sabab bo‘ladi. Ma’lumki, har bir tilning gap qurilishi va qoidalari aniq belgilangan. Masalan, ingliz tili va unga o‘xshash tillarda gap qurilishi aniq va so‘zlar tartibi qat’iydir. O‘zbek tili uchun gap qurilishi va so‘z tartibi erkin hisoblanadi. Qoidalarga asoslangan parsing usullarini ishlab chiqish qat’iy grammatik qoidalar va lug‘at ma’lumotlaridan foydalanishga asoslanadi. Ko‘pincha bunda formal grammatika qoidalar to‘plamidan, ya’ni CFG (Context free grammar)dan foydalaniladi. Kontekstsiz grammatikada parse

daraxtidan foydalanadi. Parse daraxtini ko‘rsatish yoki grafik tasvirlanishi parserni tasvirlashning qulay usulidir.

Agglyutinativ tillarda gaplardagi so‘z tartibining erkinligi, o‘ziga xos sintaktik bog‘lanishlar va morfologiyasining murakkabligi sababli qoidalarga asoslangan parser dasturlarini natijadorligini pasaytiradi. **Qoidalarga asoslangan va statistik yondashuvlarni** birlashtirgan holda qoidalarga asoslangan parser va mashinali o‘rganishga asoslangan modellarni kombinatsiya qilish orqali yuqori natijalarni olish mumkin.

Morfologiya va sintaksis NLPdagi eng muhim lingvistik qatlam hisoblanadi. Grammatik morfemalar turkumlarning shaklini o‘zgartirishi bilan birga sintaktik tahlilda ham rol o‘ynaydi. Chunki o‘zbek tilida grammatik shakllar orqali gap bo‘laklarining formal modellarini qurish mumkin. Modellar orqali esa gapning komponentlarini aniqlashga erishiladi. POS teglangan korpus orqali matnlarni sintaktik teglash mumkin. Buni korpusda tahlil qilish mumkin:

Tokzorni[[N]] eslayman[[VB]].
Tokzorni[[T]] eslayman[[K]].

U[[P]] qo‘rqib ketgandi[[VB]].
U[[E]] qo‘rqib ketgandi[[K]].

Bundan[[P]] kimga[[P]] foyda[[N]]?
Bundan[[T]] kimga[[T]] foyda[[K]]?

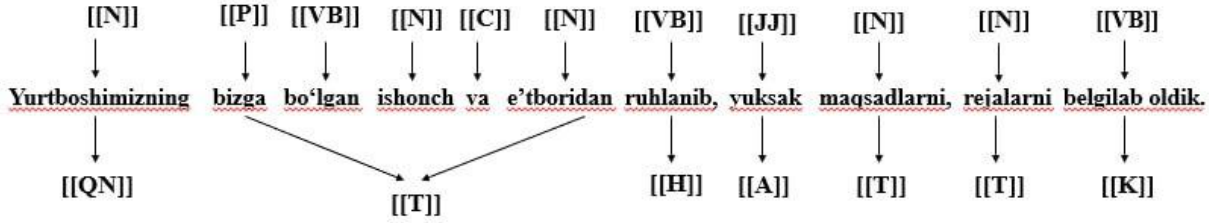
Yana[[RR]] bir[[NUM]] masala[[N]] bor[[MD]].
Yana bir[[A]] masala[[E]] bor[[K]].

Albatta[[MD]], siyosatchilar[[N]] quloq solmadi[[IB]].
Albatta[[KS]], siyosatchilar[[E]] quloq solmadi[[K]].

Yuqoridagi gaplarning POS teg va sintaktik teglashidan kuzatish mumkinki, o‘zbek tilida qoidalarga asoslangan parser qurish orqali gap bo‘laklarini aniqlash mumkin. Bunda har bir so‘z turkumining qanday grammatik morfemalarni olganda, qaysi gap bo‘lagi vazifalarida kelish mumkinligi aniqlanishi, formal modellari ishlab chiqilishi kerak. Murakkab tuzilishli gaplarda to‘g‘ri natijaga ega bo‘lmasligimiz mumkin. Chunki gaplarda sintaktik teglashda bitta lemma bitta bo‘lak bo‘lmasligi yoki ot so‘z turkumidagi so‘z doim ham ega yoki fe’l kesim bo‘lmasligi mumkin. Buni parser qanday aniqlaydi? Masalan:

Yurtboshimizning bizga bo‘lgan ishonch va e’tiboridan ruhlanib, yuksak maqsadlarni, rejalarni belgilab oldik gapini POS teglash va sintaktik teglaganimizda lemmalar soni bir xilmi? Tahlilni quyida ko‘ramiz.

yurtboshimizning – qaratqich aniqlovchi
bizga bo‘lgan ishonch va e’tiboridan – kengaytirilgan to‘ldiruvchi
ruhlanib – hol
yuksak – aniqlovchi
maqsadlarni – to‘ldiruvchi
rejalarni – to‘ldiruvchi
belgilab oldik – fe’l-kesim vazifasida kelmoqda.



Yuqoridagi tahlilda gap POS va sintaktik teglandi. Gapda 12 ta token va 11 ta lemma mavjud. Demak, 11 ta POS teglangan lemma, lekin 7 ta gap bo‘lagi mavjud. Demak, gaplardagi lemma va gap bo‘lagi mos emas. Har bir POS teg bitta gap bo‘lagi bo‘lolmaydi.

Xulosa. Xulosa qilib aytganda, o‘zbek tilida morfologik va sintaktik annotatsiyalangan korpuslarni yaratish NLP sohasidagi eng muhim talablardan biridir. Buning uchun o‘zbek tili matnlarida sintaktik munosabatlarni chuqurroq annotatsiya qilish imkoniyatini o‘rganishimiz, sintaktik teglovchi dasturlar uchun model va algoritmlarni ishlab chiqishimiz zarur.

Turkiy tillarning grammatik xususiyatlari va gap tuzilmalari umumiy jihatdan o‘xshash bo‘lgani uchun kelgusidagi ishlarimizda turkiy tillarda yaratilgan parserlarni tahlil qilishimiz, o‘zbek tilida treebank qurishda tillarda mavjud modellar asosida tahlillarning aniqlik darajasini kuzatishimiz zarur hisoblanadi. Bu kabi tadqiqot va izlanishlar natijasi o‘zbek tilida tabiiy tilga ishlov berish vazifalarini samarali bajarishda dolzarb hisoblanadi.

Foydalanilgan adabiyotlar:

1. Francis W. Nelson, Henry Kucera. Computational Analysis of Present-Day American English. Providence, RI: Brown University Press., 1967.; Greene B.B., Rubin G. M. Automatic grammatical tagging of English. Department of Linguistics, Brown University, Providence, Rhode Island, 1971.

2. Garside R. The CLAWS word-tagging system. In Garside, R., Leech, G., and Sampson, G. (Eds.), The Computational Analysis of English, 30–41. Longman, 1987.

3. Brill Eric. "A simple rule-based part of speech tagger." In Proceedings, Third Conference on Applied Natural Language Processing, ACL, Trento, Italy, 1992.

4. https://universaldependencies.org/

5. Vojtech Kovář. Automatic Syntactic Analysis for Real-World Applications. PhD thesis. Brno, Spring 2014.

6. Slav Petrov, Dipanjan Das, Ryan McDonald. A Universal Part-of-Speech Tagset. // In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), pages 2089–2096, Istanbul, Turkey. European Language Resources Association (ELRA), 2012.

7. Abduraxmonova N. O‘zbek tili elektron korpusining kompyuter modellari. Monografiya. – Toshkent, 2021.

8. Sharipov M., Mattiev J., Sobirov J., Baltayev R. Creating a morphological and syntactic tagged corpus for the Uzbek language / The International Conference and Workshop on Agglutinative Language Technologies as a challenge of Natural Language Processing (ALTNLP). - Koper, Slovenia, June 7-8, 2022.