O‘ZBEK TILI KORPUSINI SINTAKTIK TEGLASH MASALASI

Elov Botir Boltayevich
Texnika fanlari bo‘yicha falsafa doktori PhD, dotsent
elov@navoiy-uni.uz
ToshDO‘TAU Kompyuter lingvistikasi va raqamli texnologiyalar kafedrasi mudiri
Abdullayeva Oqila Xolmo‘minovna
abdullayeva.oqila@navoiy-uni.uz
f.f.f.d., PhD, ToshDO‘TAU doktoranti

Annotatsiya. Bugungi kunda zamonaviy NLP modellarini yaratish lingvistik bilimlarning akustika va prosodika, fonetika, orfografiya, morfologiya, leksikologiya, sintaksis, semantika, pragmatika va diskurs kabi spetsifikatsiyasini talab qiladi. NLP bilan bog‘liq lingvistik bilimlar leksik, sintaktik, semantik va pragmatik xususiyatlarni o‘z ichiga oladi. Ushbu maqolada sintaktik teglashning asosiy amallari haqida fikrlar keltiriladi. Shuningdek, sintaktik teglash (parsing)da qo‘llanadigan shajara daraxti xususiyatlari tavsiflangan.

Abstract. Today, the creation of modern NLP models requires the specification of linguistic knowledge such as acoustics and prosody, phonetics, orthography, morphology, lexicology, syntax, semantics, pragmatics and discourse. Linguistic knowledge related to NLP includes lexical, syntactic, semantic and pragmatic features. This article provides an overview of the basic operations of syntactic tagging. Also, the features of the family tree used in syntactic tagging (parsing) are described.

Аннотация. Сегодня создание современных моделей НЛП требует уточнения таких лингвистических знаний, как акустика и просодия, фонетика, орфография, морфология, лексикология, синтаксис, семантика, прагматика и дискурс. Лингвистические знания, связанные с НЛП, включают лексические, синтаксические, семантические и прагматические особенности. В этой статье представлен обзор основных операций синтаксической разметки. Также описаны особенности генеалогического древа, используемые при синтаксической разметке (парсинге).

Kalit so‘zlar: *til korpusi, lingvistik teglash, morfologik tahlil, tokenlash, lemmalash, POS teglash, sintaktik tahlil.*

Kirish

Korpus matnlarini lingvistik teglash dastlab lingvistik nazariyalarni yoki bugungi kunda ma’lum bo‘lganidek, til korpuslarni ishlab chiqish va tahlil qilish uchun ma’lumot berish uchun amalga oshirildi. Matnlarni “qo‘lda” teglash uchun ko‘p vaqt va kuch talab qilinadi. Biroq so‘nggi o‘ttiz yil ichida hisoblash texnikalarning quvvati va katta hajmdagu ma’lumotlarni saqlash imkoniyati hosil

bo‘lgach, katta hajmdagi til kopruslarini ishlab chiqish va avtomatik teglash sohasidagi bir qator yutuqlarga erishildi. Matnlardan iborat katta hajmli til korpuslari ustida turli lingvistik va amaliy tadqiqotlar olib borildi. Tabiiy tilni yangi texnologiyalar asosda tadqiq qilish va, eng muhimi, ishonchli statistik modellarni ishlab chiqish uchun lingvistik teglangan matn va til korpusiga asoslangan tabiiy tilni qayta ishlash (NLP) sohasiga muhim xizmat qiladi.

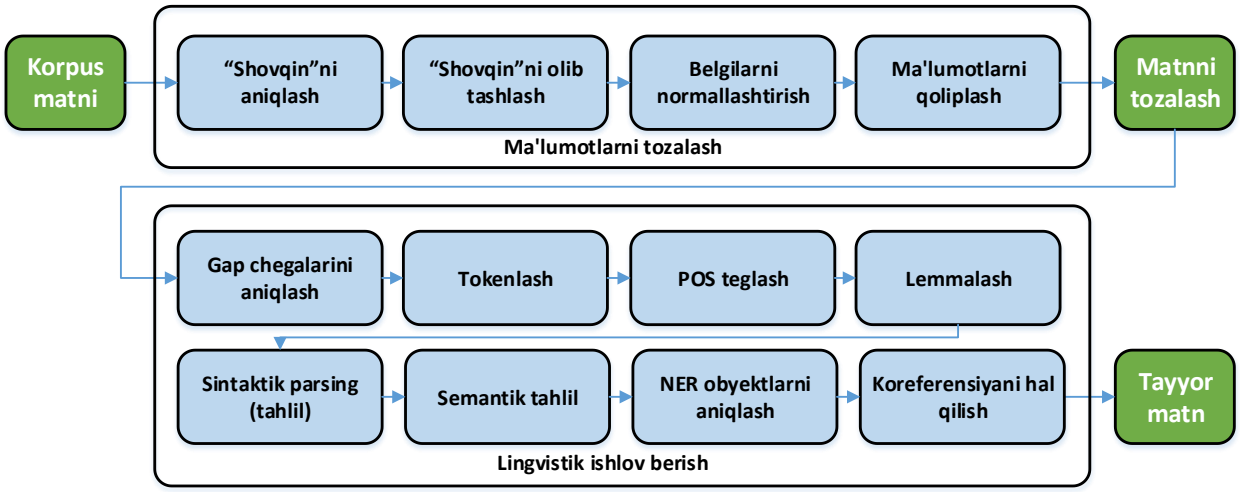
Bugungi kunga qadar o‘zbek tili korpusi matnlarini teglash uchun standartlashtirilgan tokenizatsiya amallari, lemmalash metodlari, POS teglashdagi noaniqliklarni bartaraf qilish usullari ishlab chiqilmagan. Bugungi kunda korpus matnlarini lingvistik teglashni qo‘lda yoki yarim avtomatik tarzda amalga oshirish mumkin.

NLP bilan bog‘liq lingvistik bilimlarning jihatlari haqida turli xil qarashlar mavjud. Umuman olganda, NLP tadqiqotchilarining aksariyati NLP bilan bog‘liq lingvistik bilimlar, hech bo‘lmaganda, **leksik, sintaktik, semantik** va **pragmatik** xususiyatlarni o‘z ichiga olishi kerak, deb hisoblashadi. Har bir xususiyat lingvistik ma’lumotni farqli yo‘llar bilan ko‘rsatadi. Masalan, sintaktik xususiyat ma’lum bir tildagi so‘z yoki so‘z birikmalarining gap hosil qilishini o‘z ichiga oladi.

Bugungi kunda tilshunoslikda matnni teglashning zamonaviy usullari, shuningdek, turli xil teglash vazifalarini quyidagi – 1-jadval va 1-rasmda ko‘rsatilganidek, o‘xshash qatlamlar to‘plamiga joylashtirish mumkin bo‘lgan turli bosqichlarga ajratish mumkin.

1-jadval. Lingvistik teglash bosqichlari

№	Teglash usuli	Tavsif
1.	Gap chegalarini aniqlash (sentence boundaries)	matnni gaplarga ajratish
2.	Tokenlash (tokenization)	matnni so‘zlarga ajratish
3.	Lemmalash	so‘zshaklni lemma (lug‘atdagi shakli)ga keltirish
4.	POS teglash (part-of-speech tagging)	so‘zlarning turkumlari bilan aniqlash va belgilash
5.	Sintaktik parsing (tahlil)	gapdagi tarkibiy so‘z birikmalarini tahlil qilish
6.	Semantik parsing (tahlil)	predikat-argument munosabatlarini belgilash (labeling predicate-argument relations)
7.	NER obyektlarini aniqlash (named entity recognition)	atoqli otlarni aniqlash va belgilash
8.	Koreferensiyani hal qilish (coreference resolution)	matndagi bir xil obyektlarga havolalarni bog‘lash



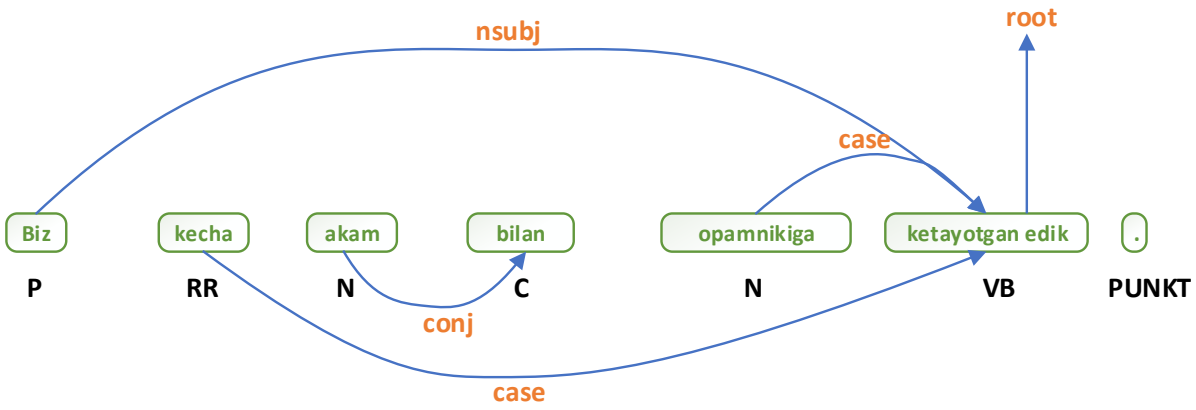
1-rasm. Korpus matnlarini teglashning zamonaviy usullari

Sintaktik parsing (tahlil)

Sintaktik tahlilning vazifasi gap qurilmalarini tahlil qilishdir. Ma’lum qurilmaning turi ishlatiladigan grammatik nazariyaga bog‘liq. Masalan, **bog‘liqlik grammatikasi (dependency grammars)** so‘zlar orasidagi bog‘liqlik munosabatlarini aniqlash uchun grafik qurilmalardan foydalansa, **so‘z birikmalari qurilmasi grammatikasi (phrase structure grammars)** esa so‘z birikmalari orasidagi tobelik munosabatini ko‘rsatish uchun daraxt qurilmalaridan foydalanadi.

Gap yoki boshqa matn ma’lumotlari berilganda sintaktik tahlil qilish va ma’lumotlardan ajratish daraxtini yaratish vazifasi **parsing** qilish deyiladi. Bu daraxt grammatikada rasmiy qoidalarga asoslangan gapning sintaktik tuzilishini ajratib ko‘rsatadi. So‘z birikmasi grammatikasi asosida sintaktik tahlil qilishda gapdagi so‘zlar o‘zaro birlashtiriladi; uzunroq so‘z birikmalari hamda gapning tarkibiy qismlarini tashkil qiladi. Sintaktik tahlil qilish uchun eng ko‘p ishlatiladigan ikkita – **tobe munosabatlar tahlili (Dependency Parsing)** va **tarkibiy qismlar tahlili (Constituency Parsing)** qo‘llanadi.

Turli xil so‘z turkumlariga asoslangan so‘z birikmalari turlicha bo‘lib, otli birikma (NP) otga, fe’lli so‘z birikma (VP) fe’lga asoslanadi. So‘z birikmasiga asos bo‘lgan so‘z birikmaning **bosh so‘zi** deyiladi. Dependency Parsing jarayonida “bosh so‘z” va boshqa barcha so‘zlar o‘rtasidagi munosabatlarni aniq belgilaydigan **tobelik daraxti (Dependency Tree)** deb nomlanuvchi grammatik struktura shakllantiriladi. Ushbu strukturani graf ko‘rinishida hosil qilganimizda so‘zlar *tugunlar (nodes)*, ular orasidagi bog‘liqliklar *qirralardir(edges)*ga ajratiladi. Ba’zi ilmiy tadqiqotlarda ushbu graf **tobelik strukturasi (Dependency Structure)** deb yuritiladi. Ushbu strukturada muayyan ildiz so‘z mavjud. Matn ma’lumotlarining boshlang‘ich nuqtasi sifatida ushbu so‘z orqali boshqa barcha so‘zlarga erishish mumkin. Bu so‘z gapning markazidir.



2-rasm. “Dependency Tree”ga namuna

Shuningdek, gapga mos tobelik daraxtida munosabatlar deb ataladigan yorliqlarni[8] shakllantirish mumkin. Ushbu munosabatlar tobelik turi haqida batafsil ma’lumot beradi.

Yuqoridagi 2-rasmdagi h → d munosabatda, h – bosh va d – tobe so‘z hisoblanadi[9]. **Bosh qism (head)** so‘z birikmasining eng muhim tugunidir, **ildiz (root)** esa gapning eng muhim tugunidir: u bevosita yoki bilvosita boshqa har bir tugunning boshidir.

O‘zbek tilidagi POS teglarning to‘liq ro‘yxatida ot so‘z turkumining turli pastki turkumlarini ajratib turadi: *birlik shakldagi ot (NN), ko‘plik shakldagi ot (NNS), ... (NNP), ... (NNPS)* va boshqalar (2-jadval). Otli birikma ot so‘z turkumidagi *NN), NNS, NNP, NNPS* kabi kategoriyalarga asoslangan bo‘lishi kerak. Xuddi shunday, fe’lli birikma ham fe’lning har xil kichik toifalariga asoslanishi mumkin: *VBP, VBZ, VBD, VBN, VBG* va boshqalar.

2-jadval. O‘zbek tilidagi sintaktik teglar ro‘yxati

№	Teg	Tavsif	Namuna
1.	**NP**	Otli birikma	Akamning moshinasi
2.	**ADVP**	Ravishli birikma	Juda ko‘p
3.	**INTJ**	Undov so‘zlar	Oh-ho
4.	**S**	Gap (darak gap)	Men keldim.
5.	**SINV**	Inversiyaga uchragan gap	Keldi bu bahorlar.
6.	**WHADJP**	So‘roq so‘zli birikma	Qayerda?
7.	**SBAR**	Murakkab gapdagi tobe qism	*Bahor kelsa*, gullar ochiladi.
8.	**NX**	Otli birikmaning boshi	Mening *ukam*
9.	**RRC**	Qisqartirilgan nisbiy bo‘laklar	Kelayotgan odam

[8] https://universaldependencies.org/u/dep/index.html
[9] https://universaldependencies.org/u/pos/all.html

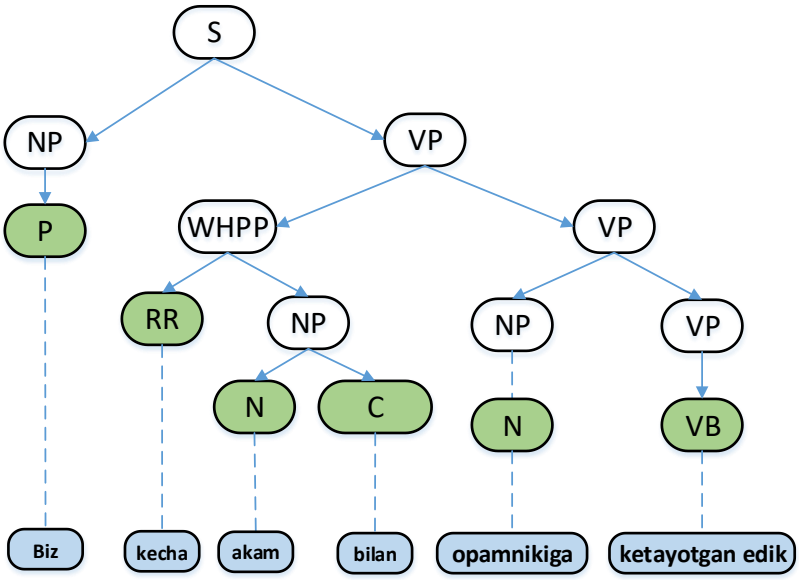
10.	**CONJP**	Bog‘lovchili birikma	Va, bilan, ba’zan…. ba’zan, ..
11.	**VP**	Fe’lli birikma	Men *ertaga maktabga boraman*.
12.	**FRAG**	Gaplardagi bo‘lak qismlar	*Yoqimli shabada!*
13.	**QP**	Miqdor ko‘rsatgichli birikma (otli birikmada)	*Ko‘p* bolalar
14.	**PRT**	Yuklama	*Faqat, axir, -mi?, -chi?*
15.	**ADJP**	Sifatlovchi ibora	*Qora* sochlar, *eng uzun* bino
16.	**UCP**	Gapdagi bo‘lak	U *ustozlari faxrlanadigan talaba* bo‘lgan.
17.	**TOP**	Daraxtning eng yuqori tuguni (root)	Men (men - root) bordim (bor-root)
18.	**PRN**	Qavs ichidagi izohlar	Alisher Navoiy (1441-1501) buyuk shoir, mutafakkir.
19.	**SBARQ**	So‘roq so‘zli birikmalar	nima?, nima uchun?
20.	**WHADVP**	Holli birikmalar	Qachon? Kecha
21.	**WHPP**	To‘ldiruvchi	Akam bilan, sen uchun
22.	**WHNP**	So‘roq so‘z (otli birikmada)	Kim?
23.	**NML**	Aniqlovchi (Otli birikmada)	O‘zbek olimlari, rasmli kitob

Tobe munosabatlar tahlili (Dependency Parsing) tobelik grammatikasiga asoslangan bo‘lsa, *Tarkibiy qismlar tahlili* (*Constituency Parsing)* **kontekstga bog‘liq bo‘lmagan grammatikaga (context-free grammar)** asoslanadi. Sintaktik tahlilning bu turi matn ma’lumotlaridagi so‘z birikmalari ustida amallar bajaradi. *Tarkibiy qismlar tahlili* jarayonida matnni grammatik toifaga asoslanib, quyi so‘z birikmalari, tarkibiy qismlarga ajratiladi. Ushbu qismlar asosan grammatik birliklardir. Grammatik birliklariga **otli birikma (NP)**, **fe’lli birikma (VP)** yoki ba’zan **old so‘z birikmasi (PP)** misol bo‘ladi.

Sintaktik daraxtlar tabiiy tilda gap yoki so‘z birtikmasining sintaktik tuzilishi grafik tasvirini ifodalaydi. Bu daraxtlar gapdagi so‘z va birikmalar qanday iyerarxik tarzda tuzilganligi, ular orasidagi grammatik munosabatlarni ko‘rsatadi. Sintaktik daraxt lingvistik tahlilning asosiy tushunchasi bo‘lib, turli xil tabiiy tillarni qayta ishlash vazifalarida muhim rol o‘ynaydi. Sintaktik daraxtlarining asosiy elementlari:

Sintaktik daraxt lingvistik tahlil uchun muhim ahamiyat kasb etadi va turli NLP ilovalarida, masalan, *tahlil qilish, imloni tekshirish* va *mashina tarjimasida* qo‘llaniladi. U gapning grammatik tuzilishini vizual tarzda taqdim etadi, bu inson va kompyuter uchun tilni tushunish va boshqarishni osonlashtiradi.

So‘z birikmasi tuzilishi qoidalari bosh so‘zning qanday qilib boshqa so‘zlar bilan birikma hosil qilishi mumkinligini belgilaydi. Berilgan gapdan so‘z birikmalarini hosil qilish usuli ko‘pincha quyidagi 3-rasmdagi kabi tahlil daraxti diagrammasi orqali ko‘rsatiladi.



3-rasm. Tahlil daraxti diagrammasi

Ba’zi ilmiy tadqiqotlarda sintaktik tahlil qilingan gapga mos ko‘rinishlardan bir qavs ichidagi gapdir. Yuqoridagi 3-rasmga mos ifoda quyidagi ko‘rinishga ega:

(**S** (**NP** (**P** *Biz*)) (**VP** (**WHPP** (**RR** *kecha*) (**NP** (**N** *akam*) (**C** *bilan*))) (**VP** (**NP** (**N** *opamnikiga*)) (**VP** (**VB** *ketayotgan edik*)))))

Matnni sintaktik teglash uchun turli xil so‘z birikmalari teglari (*NP*, *VP*, *PP*, *S* va boshqalar) **sintaktik tarkibiy teglar** deb ataladigan teglar to‘plamidir.

Ushbu maqolada matnni sintaktik teglash va uning usullari haqida ma’lumot keltirildi va yechimlari o‘zbek tili leksik birliklari misolida keltirildi.

Foydalanilgan adabiyotlar:

1. Maverick, G. v. (1969). Computational Analysis of Present-Day American English. Henry Kučera, W. Nelson Francis. International Journal of American Linguistics, 35(1). https://doi.org/10.1086/465045

2. Demirşahin, I., & Zeyrek, D. (2017). Pair Annotation as a Novel Annotation Procedure: The Case of Turkish Discourse Bank. In Handbook of Linguistic Annotation. https://doi.org/10.1007/978-94-024-0881-2_46

3. Dickinson, M., & Tufiş, D. (2017). Iterative Enhancement. In Handbook of Linguistic Annotation. https://doi.org/10.1007/978-94-024-0881-2_9

4. Core, M., Ishizaki, M., Moore, J., & Nakatani, C. H. (1999). The report of the Third Workshop of the Discourse Resource Initiative. Chiba University and Kazusa Academia Hall.

5. Cunningham, H. (2002). GATE, a general architecture for text engineering. Computers and the Humanities, 36(2). https://doi.org/10.1023/A:1014348124664

6. Day, D., Aberdeen, J., Hirschman, L., Kozierok, R., Robinson, P., & Vilain, M. (1997). Mixed-initiative development of language processing systems. 5th Conference on Applied Natural Language Processing, ANLP 1997 - Proceedings. https://doi.org/10.3115/974557.974608

7. Bird, S., Day, D., Garofolo, J., Henderson, J., Laprun, C., & Liberman, M. (2000). ATLAS: A flexible and extensible architecture for linguistic annotation. 2nd International Conference on Language Resources and Evaluation, LREC 2000.

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 xabarlari. 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. “Agglutinativ tillar uchun pos teglash va stemming masalasi (turk, uyg‘ur, o‘zbek tillari misolida)”. O‘zbekiston: til va madaniyat 2: 6-39.

10. Ramatova M. O‘zbek tilidagi sodda gaplarning sintaktik teglangan bazasi orqali tahlil daraxtini qurish // Educational Research in Universal Sciences. VOLUME 2 | ISSUE 12 | 2023. // https://zenodo.org/records/10463799.