ISSN 2181-922X

TIL VA MADANIYAT

UZBEKISTAN

LANGUAGE & CULTURE

O‘ZBEKISTON

2025 Vol. 1

tsuull.uz
uzlc.tsuull.uz

ISSN 2181-922X

O‘ZBEKISTON:

TIL VA MADANIYAT

UZBEKISTAN:

LANGUAGE AND CULTURE

2025 Vol. 1

tsuull.uz

uzlc.tsuull.uz

Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti

Bosh muharrir: Shuhrat Sirojiddinov

Bosh muharrir o‘rinbosari: Nozliya Normurodova

Mas’ul kotiblar: Ozoda Tojiboyeva

Sanjar Mavlyanov

“O‘zbekiston: til va madaniyat” jurnali – lingvistika, tarix, adabiyot, tarjimashunoslik, san’at, etnografiya, falsafa, antropologiya va ijtimoiy tadqiqotlarni o‘rganish kabi sohalarni qamrab olgan akademik jurnal.

Jurnal bir yilda to‘rt marta chop etiladi.

Jurnalning maqsadi – ko‘rsatilgan sohalarga oid dolzarb mavzulardagi bahs-munozaraga undaydigan, yangi, innovatsion g'oyalarga boy, o‘z konsepsiyasiga ega bo‘lgan tadqiqotlarni nashr etishdir.

Ingliz, rus va o‘zbek tillaridagi, shuningdek, boshqa turkiy tillarda yozilgan maqolalar qabul qilinadi. Iqtisodiy tahlillar hamda siyosatga oid maqolalar e’lon qilinmaydi.

Jurnalda kitoblarga yozilgan taqrizlar, adabiyotlar sharhi, konferensiyalar hisobotlari va tadqiqot loyihalari natijalari ham e’lon qilinadi. Mualliflar fikri tahririyat nuqtayi nazaridan farq qilishi mumkin.

Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti.

O‘zbekiston, Toshkent, Yakkasaroy tumani, Yusuf Xos Hojib ko‘chasi, 103.

Email: uzlangcult@gmail.com
Website: www.uzlc.tsuull.uz

"Uzbekistan: Language and Culture" is an academic journal that publishes works in the field of linguistics, history, literature, translation studies, arts, ethnography, philosophy, anthropology and social studies.

The journal is published four times a year.

The purpose of the journal is to publish the results of the latest research that are rich in new, innovative ideas and has its own concept, which stimulates debate on topical issues in these areas.

The language of articles can be English, Russian and Uzbek. Other Turkic languages are also welcome. We do not publish economic analyses or political articles.

In addition to research articles, the journal announces book and literary work reviews, conference reports and research project results.

The authors' ideas may differ from those of the editors'.

Alisher Navoʻi Tashkent State University of the Uzbek Language and Literature.
103, Yusuf Khos Hojib, Yakkasaray, Tashkent, Uzbekistan.

Email: uzlangcult@gmail.com
Website: www.uzlc.tsuull.uz

MUNDARIJA

Lingvistika

Adabiyotshunoslik

Tarix. Manbashunoslik

San'at

CONTENT

Linguistics

Literature

History. Source studies

Art

ISSN 2181-922X 6—48

LINGVISTIKA

LINGUISTICS

O'zbek tilida tabiiy tilni qayta ishlashdagi morfologik, sintaktik va semantik tahlil metodlari

Botir Elov[1]
Oqila Abdullayeva[2]
Mastura Primova[3]

Abstrakt

Mazkur maqolada o‘zbek tilida tabiiy tilni qayta ishlashda qo‘llaniladigan morfologik, sintaktik va semantik tahlil metodlari muhokama qilingan. O‘zbek tilining lingvistik xususiyatlari – murakkab morfologiya, erkin so‘z tartibi va resurslarning cheklangani mazkur metodlarni qo‘llashda alohida yondashuv va tadqiqotlar zarurligini anglatadi [Senuma, Aizawa 2017, 100-109]. Tadqiqot doirasida morfologik tahlil usullari, so‘ng sintaktik va semantik tahlil metodlari ilmiy manbalar asosida ko‘rib chiqilgan. Har bir qismda mavjud afzallik va kamchiliklar, o‘zbek tilidagi qo‘llanilish tajribalari, shuningdek, xorijiy tillar bilan qiyosiy tahlil keltirilgan. O‘zbek tilida morfologik tahlilning qoidalarga asoslangan metodlar, statistik modellar (HMM, CRF va boshqalar), Neyron tarmoqlarga asoslangan yondashuvlari (BiLSTM-CRF, seq2seq) muhokama qilinib, natija ko‘rsatkichlari misol va foizlarda berilgan. Sintaktik parsing dependency va constituency parsing tahlil usullari orqali amalga oshirilishi ko‘rsatilgan. SOV tartibiga ega o‘zbek tili uchun UD treebankini qurish masalasi ko‘rib chiqilgan. Gaplardagi murakkab morfologik tuzilma va erkin so‘z

[1]*Elov Botir Boltayevich* – texnika fanlari falsafa doktori (PhD), dotsent, Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti.
E-pochta: elov@navoiy-uni.uz
ORCID: 0000-0001-5032-6648
[2]*Abdullayeva Oqila Xolmo‘minovna* – filologiya fanlari bo‘yicha falsafa doktori (PhD), Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti doktoranti.
E-pochta: abdullayeva.oqila@navoiy-uni.uz
ORCID: 0000-0002-2524-4832
[3]*Primova Mastura Hakim qizi* – Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti Kompyuter lingvistikasi va raqamli texnologiyalar kafedrasi o‘qituvchisi.
E-pochta: primovamastura@navoiy-uni.uz
ORCID: 0000-0002-0241-4659
Iqtibos uchun: Elov, B., Abdullayeva, O., Primova, M. 2025. “O'zbek tilida tabiiy tilni qayta ishlashdagi morfologik, sintaktik va semantik tahlil metodlari”. *O‘zbekiston: til va madaniyat* 1: 6 – 48.

tartibi parserlarni qurishdagi ta'siri yoritilgan. O'rganilgan yondashuvlar natijasida gibrid parserlarni qurish, morfologik analiz bilan integratsiya qilib, so'zlarning grammatik kategoriyalarini parserga berish masalasi ko'tarilgan. Shuningdek, neyron tarmoqlar asosida neural constituency parserlarni ishlab chiqish va undan olinadigan natijalar samaradorligi tahlil qilingan. NLP bilan bog'liq keyingi tahlil bosqichi semantik va sentiment tahlilda so'z ma'nolarini vektor shaklida ifodalovchi Word2Vec, FastText kabi modellardan tortib, BERT, mBERT, UzBERT kabi kontekstga moslangan transformator modellari ham muhokama qilingan. Matnlarda yuqori darajadagi semantik rol belgilash va WordNet, FrameNet kabi semantik tarmoqlarni yaratish masalalari ham ko'rib chiqilgan. O'zbek tili uchun NLP yondashuvlarini baholashda o'zbek tili boshqa tillar: ingliz, turk va rus tillari bilan qiyoslanib, tillarda qoidaviy, statistik va neyron metodlar qo'llanilishi, tahlil natijalari ko'rsatilgan.

Kalit so'zlar: *Pos teglash, parser, sintaktik parsing, semantik parsing, sentiment analiz, tabiiy tilga ishlov berish.*

Kirish

Bugungi kunda korpus matnlarini lingvistik teglashni qo'lda yoki yarim avtomatik tarzda amalga oshirish mumkin. Zamonaviy NLP modellarini yaratish lingvistik bilimlarning quyidagi to'qqizta spetsifikatsiyasini talab qiladi:

1-jadval. Lingvistik tavsif darajalari.

№	Lingvistik qatlam	Tavsif
1.	Akustika va prosodiya (Acoustics and prosody)	til ritmi, intonatsiyasi va ohangini tasvirlash, nutqda tovushlarning hosil bo'lishini tushuntirish uchun zarur bo'lgan bilimlar
2.	Fonetika (Phonology)	fonemalar birikmasi va fonemalarning morfemalarni hosil qilish jarayonini tasvirlash uchun zarur bilimlar. Bunda asosiy birlik tovushdir (og'zaki nutq)
3.	Orfografiya	so'zlar (yoki so'z birikmalari) orasidagi tuzilish qoidalarini tasvirlash va so'zlarning (yoki iboralarning) qanday qilib jumlalar hosil qilishini tushuntirish uchun zarur bo'lgan bilimlar (yozma nutq)
4.	Morfologiya (Morphology)	morfemalar birikmasini tavsiflash; morfemalarning so'z yasashini tushuntirish uchun zarur bo'lgan bilimlar
5.	Leksikologiya (Lexicology)	leksik tizim qoidalarini tavsiflash; so'zlarning o'ziga xos semantik xususiyatlarini tushuntirish uchun zarur bo'lgan bilimlar
6.	Sintaksis (Syntax)	so'z tartibi va gap tuzilishi, ularning grammatik xususiyatlarini tushuntirish uchun zarur bo'lgan bilimlar

7.	Semantika (Semantics)	gap tarkibiy qismlari orasidagi semantik munosabatlarni tavsiflash, gapni tashkil etuvchi komponentlardan butun gapning ma'nosini qanday chiqarishni tushuntirish uchun zarur bo'lgan bilimlar
8.	Pragmatika (Pragmatics)	gap ma'nosini uning kontekstiga bog'liq holda tavsiflash, so'zning ma'lum kontekstdagi vazifa (ma'no) si, u sodir bo'lgan kontekstga bog'liq bo'lgan nutqiy harakatning turli ma'nolarining chiqarishni tushuntirish uchun zarur bo'lgan bilimlar
9.	Diskurs (Discourse analysis)	gaplar orasidagi sintaktik qoidalarni tavsiflash; matn yoki dialogdagi uyg'unlik; gaplarning nutq yoki dialogni qanday tashkil etishini tushuntirish uchun zarur bo'lgan bilimlar.

NLP bilan bog'liq lingvistik bilimlar **leksik, sintaktik, semantik** va **pragmatik** xususiyatlarni o'z ichiga olishi kerak. Har bir xususiyat lingvistik ma'lumotni farqli yo'llar bilan uzatadi. Masalan, leksik xususiyat so'z darajasidagi asosiy tarkibiy qismlar (masalan, morfema) va uning flektiv shakllari haqidagi bilimlarni qamrab olishi mumkin; sintaktik xususiyat ma'lum bir tildagi so'z yoki so'z birikmalarining gap hosil qilishini o'z ichiga oladi; semantika ma'lum so'zlar yoki gaplarga qanday ma'no berishni; pragmatika suhbatda nutq markazidagi o'zgarishlarni, berilgan kontekstdagi gap ma'nosini izohlash haqida bilimlarni o'z ichiga oladi. Biroq lingvistik bilimga qo'shimcha ravishda NLP *axborot texnologiyalari, matematika, psixologiya, falsafa, statistika* va *biologiya* sohalardagi boshqa bilimlarni ham o'z ichiga olishi mumkin.

Tilshunoslikda lisoniy hodisalarni tahlil qilish va tavsiflash, odatda, bir necha alohida bosqichlarda amalga oshiriladi. Tilning turli tovushlari fonetikada tavsiflansa, yozuv tizimi orfografiyada o'rganiladi. Morfologiya so'zning shakllanishi, o'zgarishini, sintaksis so'zlarning joylashuvi, ularning so'z birikmasi va gapga aylanishini aniqlaydi. Semantika so'z *(leksik semantika)*, birikma hamda gapning ma'nosini *(kompozitsion semantika)* tahlil qiladi. So'z va so'z birikmalarining turli sharoit va vaziyatda o'ziga xos ma'no anglatishini pragmatik tahlil natijasida o'rganish mumkin. Shaxs va narsalar qanday qilib mavzu sifatida kiritilishi va keyinchalik qanday tilga olinishi diskurs tahlilning predmeti sanaladi.

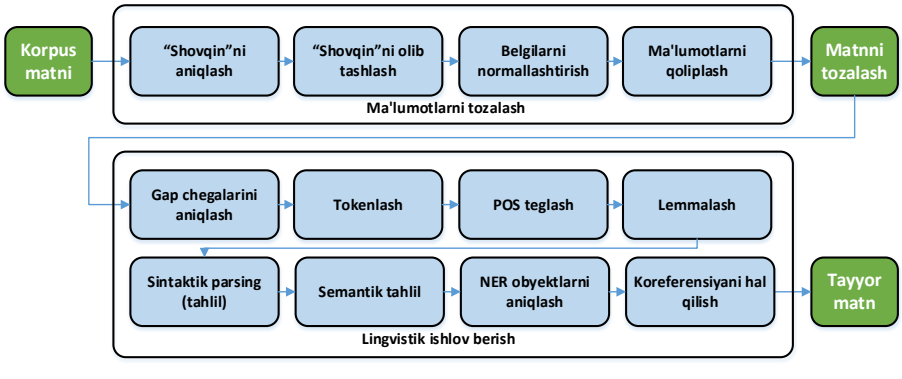
Fonetika va orfografiya eng kichik birliklar (alohida tovush va harf) bilan shug'ullanadi. Morfologiya, sintaksis va semantika o'rta hajmdagi birlik (so'z, so'z birikmasi va gap)larni o'rganadi. Diskurs va pragmatikaning predmeti yuuqorida keltirilganidek eng katta birliklar (*paragraf* va *dialog*)dir.

Bugungi kunda tilshunoslikda matnni teglashning zamonaviy usullari, shuningdek, turli xil teglash vazifalarini quyidagi – 2-jadval va 1-rasmda ko‘rsatilganidek, o‘xshash qatlamlar to‘plamiga joylashtirish mumkin bo‘lgan turli bosqichlarga ajratish mumkin.

2-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.



Biroq til korpuslarini teglashda bajariladigan vazifalar bilan tilshunoslik nazariyasidagi tavsif bosqichlari o‘rtasida faqat umumiy muvofiqlik mavjud.

Morfologik tahlil metodlari

Morfologik tahlil so‘z tarkibini – uning o‘zagi, affikslari va

grammatik ma'nolarini avtomatik aniqlashga xizmat qiladi. Bu vazifa agglyutinativ tillarda, jumladan o'zbek tilida murakkab: bir so'z ko'plab qo'shimchalar qabul qilishi va turli grammatik ma'nolarni ifodalashi mumkin. Masalan, *"kutubxona-lar-i-ga-mi"* so'zi tarkibida **lemma (kutubxona)**, *ko'plik -lar, 3-shaxs egalik -i, jo'nalish kelishigi -ga va so'roq qo'shimchasi -mi* mavjud [Abduraxmonova 2019]. Bunday boy morfologiyani tahlil qilish uchun tilshunoslik qoidalari hamda statistik modellardan foydalaniladi. Quyida morfologik tahlilning asosiy yondashuvlari – **qoidalarga asolangan tizimlar, statistik modellar** va **neyron tarmoqli modellar** – hamda maxsus **chegaralangan holatli mashina (FST)** texnikasi tafsilotlari bilan ko'rib chiqiladi.

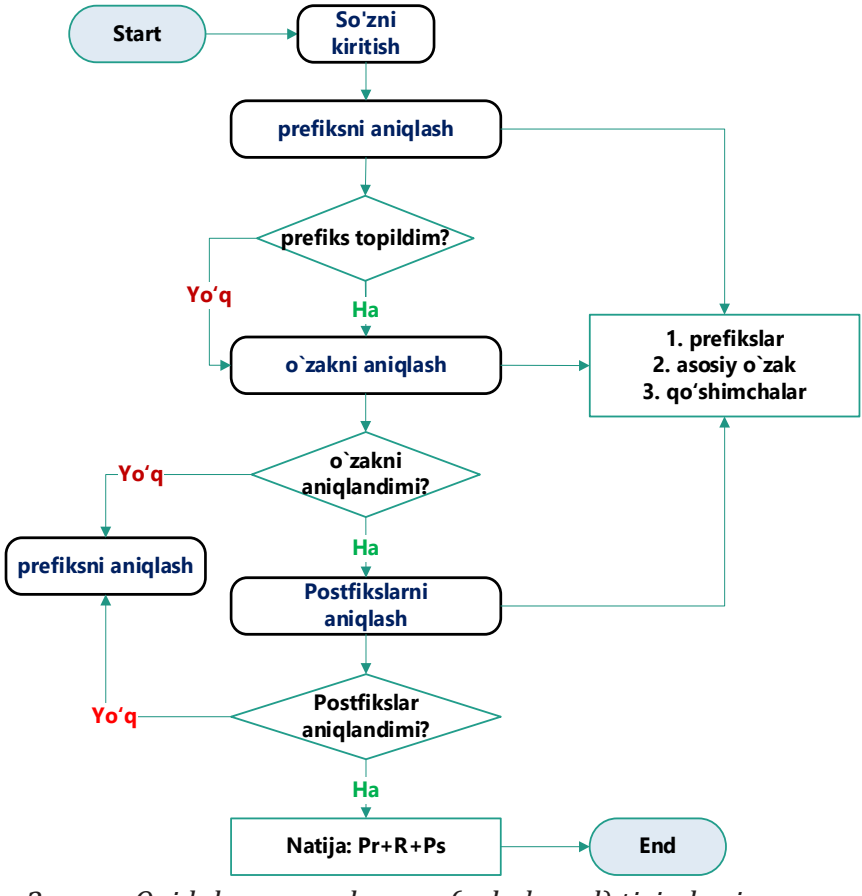
Qoidalarga asoslangan (rule-based) tizimlar

Qoidaga asoslangan yondashuvlarda lingvistlar tuzgan **qoida va lug'atlar** asosida so'zning morfologik tarkibi tahlil qilinadi. Masalan, o'zbek tilida so'z turkumlari va ularning birikishi mumkin bo'lgan affikslar ro'yxati, affikslarning ketma-ketligi qoidalari tuziladi [Hamroyeva 2021, 395-400]. Qoidalarga asoslangan morfotahlilda aksar hollarda **lug'at** (o'zak shakllar ro'yxati) va **qo'shimchalar jamlanmasi** yig'ilib, ularning birikish qoidalari belgilanadi.

Bunday qo'lda yaratilgan tizimlar afzalligi shundaki, ular **aniq va izchil** ishlaydi, lingvistik jihatdan to'g'ri yechimlarni ta'minlaydi. Turkiy tillar uchun, jumladan, o'zbek tili uchun avtomatik morfologik tahlilda **chegaralangan holat mashinasi (ya'ni FST)** metodidan foydalanish samarali ekanligi tadqiqotchilar tomonidan ta'kidlangan [Adali 2020]. Darhaqiqat, o'zbek tili uchun morfologik tahlil vositalaridan biri Prolog'da qoidalarga tayangan parser bo'lib, murakkab so'z shakllari uchun yetarli imkoniyat bermagan.

Qoidalarga asoslangan tizimlar kamchiligi – **yangi yoki lug'atda yo'q so'zlarni** (masalan, yangi terminlar, nomlar) tahlil qila olmasligi va **ko'p ma'noli (omonim)** shakllarda kontekstni inobatga olmasligidir. O'zbek tilida, masalan, *"qilish"* yordamchi fe'li bilan birikmalar ko'p uchraydi *("yordam qil-di"*, *"ta'bir qil-moq"* kabi). Qoidalarga asoslangan analiz bunday birikmalarni alohida so'zlar sifatida tahlil qilsa-da, ularning birgalikdagi ma'nosini anglata olmaydi. Xuddi shuningdek, qoidaviy POS tagger (so'z turkumi belgilovchi) faqat qo'shimchalar va qo'shni so'zlarga asoslanib, kontekstni chuqur anglay olmaydi, bu esa ayrim hollarda xatolikka olib keladi [Sharipov va boshq. 2023]. Shunga qaramay, bu tur yondashuv kichik korpusli yoki cheklangan resursli tillarda dastlabki

bosqichda muhim ahamiyatga ega bo‘lib, ko‘pincha boshqa metodlar uchun asosiy annotatsiyalarni tayyorlashda qo‘llaniladi (masalan, qoida asosida tahlil qilingan matnlar keyinchalik statistik modellarni o‘qitish uchun ishlatiladi).



2-rasm. Qoidalarga asoslangan (rule-based) tizimlar jarayonlar ketma-ketligi.

Statistik modellar (HMM, CRF va boshqalar)

Statistik yoki **ma’lumotga asoslangan modellar** morfologik tahlil va teglashni ko‘p miqdordagi belgilangan matnlardan o‘rganish orqali amalga oshiradi. Bunda **yashirin Markov modeli (HMM)** va **shartli tasodifiy maydonlar (CRF)** kabi modellar keng qo‘llanilgan. Ilmiy manbalarda keltirilishicha, HMM [Kupiec 1992, 225-242] va CRF [Awasthi va boshq. 2006] modellarining joriy etilishi **POS teggerlar aniqligini sezilarli oshirgan** [Sharipov va boshq. 2023]. HMM modellarida so‘zlar ketma-ketligi orqali ehtimolliklar hisoblanib, har bir so‘z uchun eng ehtimolli grammatik teg topilsa, CRF modellarida butun ketma-ketlik uchun shartli ehtimollik mak-

simallashtiriladi.

Statistik modellar **annotatsiyalangan korpus** mavjud bo‘lganda ularning statistik xususiyatlarini hisoblab, qoidasiz avtomatik qaror qabul qiladi. Masalan, agar teglangan matnlar bo‘lsa, CRF modeli so‘z oxiridagi harflar, qo‘shimchalar mavjudligi, qo‘shni so‘zlar kabi xususiyatlarga asoslanib, so‘z turkumini belgilashni o‘rganishi mumkin. Bunday yondashuv o‘zbek tilida ham sinovdan o‘tkazilgan: Elov va boshqalar (2023) [Elov va boshq., 2023, 63-68] kichik hajmdagi o‘zbek matnlariga HMM modelini qo‘llab, ishchi prototip natijalarini ko‘rsatganlar. Statistik modellar qoidalarga qaraganda moslashuvchan, ya’ni yangi ma’lumot qo‘shilsa, tez o‘rgatish mumkin, inson mehnatini kamaytiradi.

Shu bilan birga, ularning chegarasi – **katta va muvozanatli ma’lumotga ehtiyoj**. O‘zbek tilida katta hajmli teglangan korpuslar yaqindagina paydo bo‘lib, ilgari bunday ma’lumot yetarli emas edi. Natijada, statistik modellar yomon ishlaydi. Ammo mavjud tadqiqotlar ko‘rsatmoqdaki, hatto cheklangan hajmdagi korpuslar uchun ham statistik yondashuvlar qoidalarga asoslangan tizimlar darajasiga yetib bormoqda. Masalan, Murat va Ali [Murat, Ali 2024] taklif qilgan ko‘p kanalli diqqat (multi-head attention) modelli POS tegger avvalgi BiLSTM, CNN va CRF kabi modellar natijasini 4.1% ga oshirib, 79.74% aniqlikka erishgan. Bu ko‘rsatkich qoidaviy dasturdan ham yuqori bo‘lsa-da, hali ham chuqur o‘rganuvchi modellar darajasiga yetmagan.

3-rasm. Statistik modellar (HMM, CRF va boshqalar) asosida morfologik tahlil qilish jarayonlar ketma-ketligi.

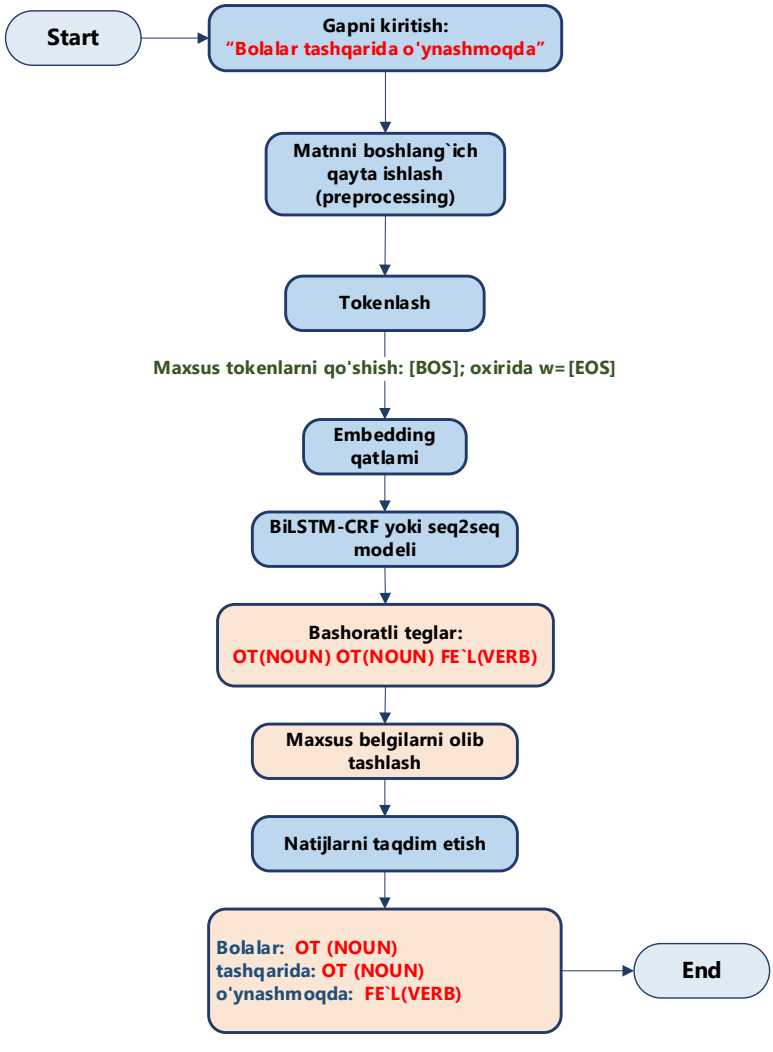
Statistik modellar morfologik tahlilning boshqa jihatlarida ham qo‘llaniladi. Masalan, **morfologik segmentatsiya (so‘zni tarkibiy qismlarga ajratish)** uchun HMM asosidagi *Morfessor* kabi vositalar tillararo muvaffaqiyatli sinovdan o‘tgan. Shuningdek,

lemmatizatsiya va morfologik disambiguation (bir necha mumkin bo‘lgan tahlillardan to‘g‘risini tanlash) vazifalarida ham CRF yondashuvi qo‘llanilib, yuqori natijalar qayd etilgan (turk tilida CRF disambiguator 95% aniqlikka erishgani ma’lum) [O‘zer, Korkmaz 2022, 1897-1913].

Neyron tarmoqlarga asoslangan yondashuvlar (BiLSTM-CRF, seq2seq)

So‘nggi yillarda chuqur o‘rganish usullari tabiiy tilni qayta ishlashda yangi yutuqlarga erishdi va bu morfologik tahlilda ham aks etdi [Meftah va boshq. 2018]. Neyron tarmoqlar katta hajmdagi ma’lumotdan **mustaqil xususiyatlarni o‘rganib**, yanada yuqori aniqlikdagi modellar yaratishga imkon beradi. Xususan, **ikki yo‘nalishli LSTM + CRF (BiLSTM-CRF)** arxitekturasi ketma-ketliklarni teglash (sequence labeling) muammolarida, jumladan **so‘z turkumlarini aniqlash** va **morfologik teglashda** eng yuqori natijalarni ko‘rsatdi [Huang va boshq. 2015]. BiLSTM so‘zning **kontekstdagi chap va o‘ng tomondan ma’no** izlarini hisobga olib, ichki holat vektorlarini shakllantiradi, oxiriga ulanadigan CRF qatlami esa butun gap bo‘yicha eng mos teglar ketma-ketligini tanlaydi. Bunday model ingliz tilida 97% aniqlikka yetgan bo‘lsa [Marcus va boshq. 1993, 313-330], o‘zbek tilida ham chuqur model qo‘llash katta sakrash bo‘ldi. Masalan, monolingual BERT modeliga asoslangan POS tegger o‘rtacha 97.8% aniqlikka erishdi [Elov va boshq. 2023, 57-62].

Neyron tarmoqlarning yana bir turi – **encoder-decoder arxitekturali ketma-ketlikdan ketma-ketlikka (seq2seq) modellari** – murakkab morfologik muammolarni yechishda qo‘llanilmoqda. Seq2seq modelida bir tarmoq (encoder) kirish so‘zni yoki gapni ichki vektorga kodlasa, ikkinchi tarmoq (decoder) undan chiqish ketma-ketligini hosil qiladi. Morfologiya uchun bu usul **so‘z yasash** va **so‘z turlantirish (inflection)** masalalarida ayniqsa muvaffaqiyatli bo‘ldi [Senuma, Aizawa 2017, 100-109]. Misol uchun, SIGMORPHON-2016 musobaqasida Kann va Schütze [Kann, Schutze 2016] seq2seq modelini qo‘llab, bir nechta tillarda lemmani berilgan grammatik kategoriya bo‘yicha to‘g‘ri shaklga o‘tkazishda birinchi o‘rinni egallashdi. Keyingi yili universal morfologik turlantirish musobaqasida ham yuqori resursli tillarda seq2seq tizimlari 91% aniqlik ko‘rsatdi. Bu shuni anglatadiki, neyron modellar yetarlicha ma’lumotga ega bo‘lganda morfologik jarayonlarni deyarli inson darajasida model qilishi mumkin.



4-rasm. Neyron tarmoqlarga asoslangan yondashuvlar (BiLSTM-CRF, seq2seq) asosidagi morfologik tahlil qilish jarayonlar ketma-ketligi.

O‘zbek tilida ham chuqur o‘rganuvchi modellar tatbiq etila boshlandi. Xususan, Elov va boshqalar [Elov va boshq. 2024, 126-130] tomonidan yaratilgan UzbMorphAnalyzer tizimi stemming, lemmatizatsiya va POS taglashni birlashtirib, o‘zbek tilida kompleks yechim taklif qilgan. Bu tizimda ham an’anaviy FST komponenti bilan birga ehtimoliy modellar integratsiya qilingan. Kelajakda esa yanada sof neyron yondashuvlar – masalan, so‘zning belgilar ketma-ketligidan to‘g‘ridan-to‘g‘ri morfologik teglar ketma-ketligini chiqaruvchi transformer modellar – paydo bo‘lishi kutilmoqda.

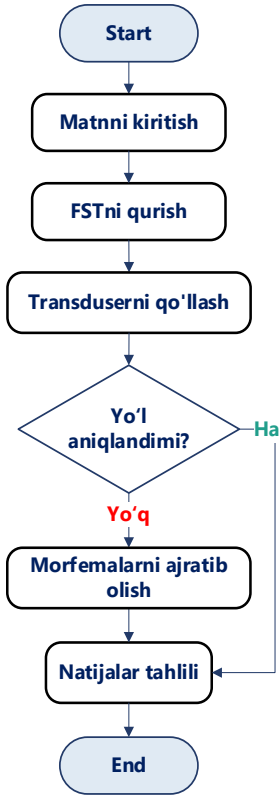
Chekli avtomatlar (FST) modeli va o‘zbek tilida qo‘llanishi

Chekli avtomatlar va transduserlar morfologik tahlilning klassik va samarali usullaridan biri bo‘lib, ayniqsa, agglyutinativ tillar uchun juda mos keladi [Hamroyeva 2021, 395-400]. FST modeli so‘z tarkibini holatlar va o‘tishlar orqali modellashtiradi: boshlang‘ich holatdan boshlab har bir affiks yoki so‘z qismi uchun yangi holatga o‘tiladi va so‘zning oxiriga yetganda to‘liq tahlil hosil qilinadi. **XFST, Foma, HFST** kabi vositalar yordamida grammatik qoidalarni avtomat shaklida kompilyatsiya qilish mumkin. Natijada, FST asosidagi analizator kiritilgan so‘z uchun barcha mumkin bo‘lgan “o‘zak+qo‘shimcha” zanjirlarini va ularning grammatik ma’nolarini chiqarib beradi.

O‘zbek tilida FST yondashuvi tadqiqotchilar e’tiborida bo‘lib keldi. Masalan, Hamroyeva [Hamroyeva 2021, 395-400] o‘zbek tili uchun morfologik analizator yaratishda FST modelini ishlab chiqib, unda **o‘zaklar bazasi, affikslar majmuasi** va **imlo qoidalarini** qat’iy tartibda qo‘llash kerakligini ta’kidlagan. Bunday tizim uchun *“daraxtlar”* so‘zi tahlili misol tariqasida ko‘rsatilgan: **daraxt (o‘zak)** + **lar (ko‘plik grammatik formasi)** tarzida ajratiladi. FST asosidagi analizatorlar har bir qo‘shimchaning qaysi turdagi so‘zga qo‘shilishi mumkinligini **allomorf qoidalar** orqali tekshiradi, bu esa grammatik xatolarga yo‘l qo‘ymaydi. O‘zbek tilida ham shunday qoida: masalan, sifatdosh va kelishik qo‘shimchalari ketma-ketligi qat’iy tartibga ega bo‘lib, FST bu tartib buzilishiga yo‘l qo‘ymaydi.

1. Amaliy jihatdan, o‘zbek tilining hozirgi kunda mavjud **ochiq kodli FST analizatori** Apertium loyihasi doirasida yaratilgan. Apertium [https://github.com/apertium/apertium-uzb] uchun o‘zbek morfologik tahlil moduli ishlab chiqilib, matnlarni tahlil qilish imkonini beradi. Bu analizator grammatik jihatdan barqaror bo‘lsada, undan olingan natijalarni disambiguation (noaniqlikni bartaraf etish) bosqichidan o‘tkazish talab qilinadi, chunki ko‘p so‘zlar bir nechta y interpretyatsiyaga ega bo‘lishi mumkin. O‘zbek tilida, masalan, *“yetti”* so‘zi son 7 ma’nosida ham, fe’lning buyruq zamon shakli (*“yetmoq”* fe’lining shakli) sifatida ham kelishi mumkin; faqat kontekstdagina uning haqiqiy ma’nosi aniqlanadi.

FST bunday holatda ikkala tahlilni ham chiqaradi, lekin qaysi biri to‘g‘ri ekanini aniqlamaydi. FST modellari asosidagi tizimlarning **afzalligi – to‘liq qamrov va aniqlik:** agar lug‘at va qoidalar bazasi mukammal tuzilgan bo‘lsa, u har qanday kiritmani grammatik jihatdan to‘g‘ri tahlil qila oladi.



5-rasm. Chekli avtomatlar (FST) modeli asosidagi morfologik tahlil qilish jarayonlar ketma-ketligi.

Kamchiligi – ularga yangi so‘z qo‘shish yoki o‘zgarish kiritish uchun lingvistik bilim talab etiladi va bunday tizimlar katta mehnat evaziga yaratiladi. Ammo o‘zbek tilidagi murakkab so‘z yasash va shakllantirish jarayonlarini inobatga olsak, FST yondashuvi bu til uchun eng ishonchli va izchil natija beruvchi usul bo‘lib qolmoqda [Adali 2020]. Keyingi bosqichlarda FST va neyron modellar integratsiyasiga asoslangan gibrid tizimlar – masalan, FST barcha imkoniyatlarni chiqarib, neyron tarmoq kontekstga ko‘ra qo‘llanishi kutilmoqda.

3-jadval. Morfologik tahlil metodlari: qiyosiy tahlil

Metod	Afzalliklari	Cheklovlari
Qoidaviy tizimlar	Aniq, izchil, tilga xos qonuniyatlar asosida ishlaydi	Kontekstni inobatga olmaydi, yangi yoki lug‘atda yo‘q so‘zlarni tanimaydi
Statistik modellar (HMM/CRF)	Korpusga asoslangan, moslashuvchan, qayta o‘qitish mumkin	Ko‘p va muvozanatli ma’lumot talab qiladi, kichik korpusda aniqlik past
Chuqur o‘rganish (BiLSTM/seq2seq)	Kontekstni yaxshi tushunadi, yuqori aniqlikka ega	Ko‘p resurs va o‘qitish talab qiladi, GPU zarur, interpretatsiyasi qiyin

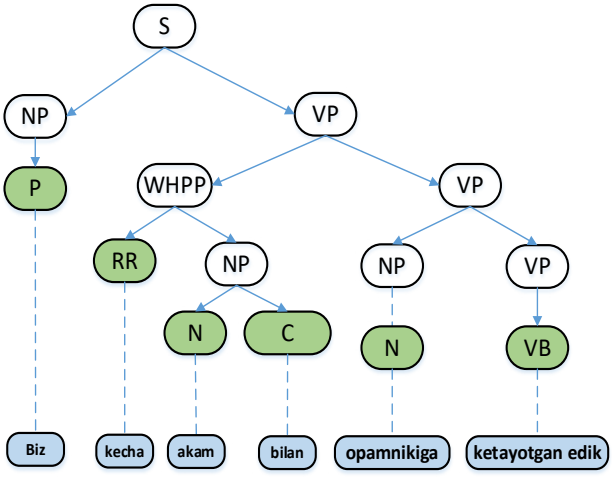
FST (Finite-State Transducer)	Grammatik jihatdan qat'iy, to'liq qamrovli, aniqligi yuqori	Tuzilishi mehnat talab qiladi, noaniqlikda tanlov bera olmaydi

Sintaktik tahlil

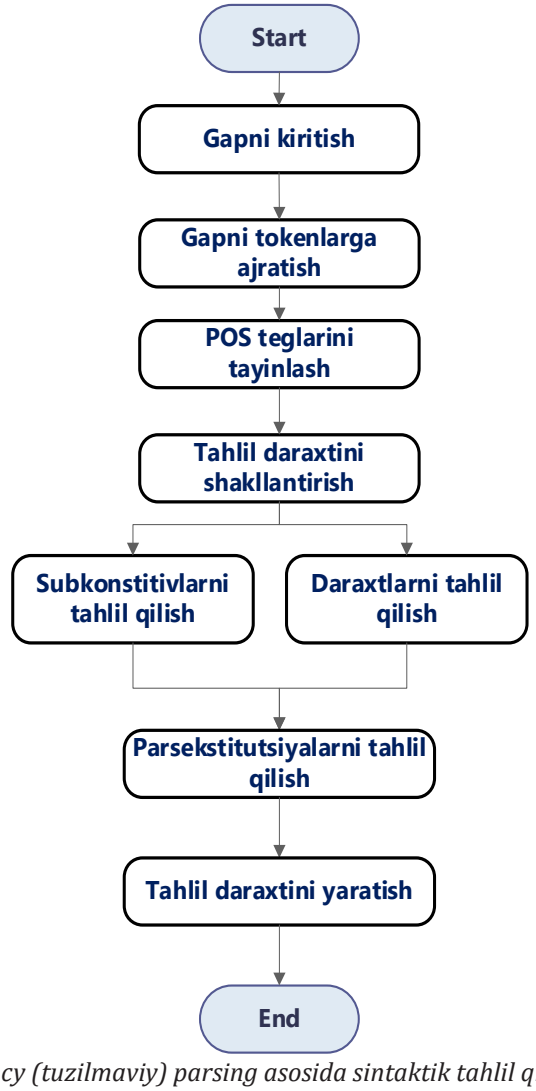
Sintaktik tahlil (parsing) natijasida gapdagi so'zlar o'rtasidagi grammatik bog'lanishlar aniqlanadi. Sintaksisni aniqlash uchun *ikki asosiy yondashuv mavjud*: **tarkibiy tuzilishga asoslangan parsing (constituency parsing)** va **bog'lanishga asoslangan parsing (dependency parsing)**. Tarkibiy tuzilma grammatik tahlilida **gap fraza tuzilishi daraxti** shaklida ifodalanadi – masalan, egaga birikib keluvchilar *"Ot guruhi (NP)"* hosil qilishi, kesimga birikadigan qism *"Fe'l guruhi (VP)"* tarkibiga kirishi kabi. Bog'lanishli tahlilda esa **har bir so'z boshqa bir so'zga tobe tarzda bog'lanib**, *"to'ldiruvchi->kesim"* kabi juftliklar orqali butun gapni qamrab oluvchi bog'liqlik grafi quriladi. Quyida ushbu usullar tafsiloti, Universal Dependencies standarti hamda o'zbek tilining **SOV** bazaviy so'z tartibi va erkin komponentlari sintaktik analizga qanday ta'sir qilishi ko'rib chiqiladi.

Bog'liqlik va tuzilma tahlil usullari

Constituency (tuzilmaviy) parsing maqsadi – gapni grammatik qoidalarga muvofiq ravishda ichma-ich uyushgan frazalarga ajratish. Bu yondashuv **anarxik grammatikalar** (masalan, **CFG, PCFG**) va tugunlardan iborat sintaktik daraxtlarga tayanadi. Masalan, ingliz tilida *"The quick brown fox jumps over the lazy dog"* gapining tuzilma daraxtida **NP (The quick brown fox)** va **VP (jumps over the lazy dog)** kabi tugunlar aniqlanadi. Bunday daraxt ichida har bir fraza o'z ichidagilarning tartibini ham aks ettiradi. Tuzilmaviy tahlil **Chomsky iyerarxiyasiga** binoan qat'iy qoidalarga tayangani bois, qat'iy so'z tartibli tillarda (masalan, ingliz tilida **SVO**) yaxshi ishlaydi – chunki grammatik munosabatlar asosan so'zlarning tartibi bilan ifodalanadi. Ammo **erkin so'z tartibli tillarda** bir fikrni turlicha tartibda ifodalash mumkin bo'lganligi tufayli, an'anaviy tuzilmaviy grammatikalar murakkablashadi.

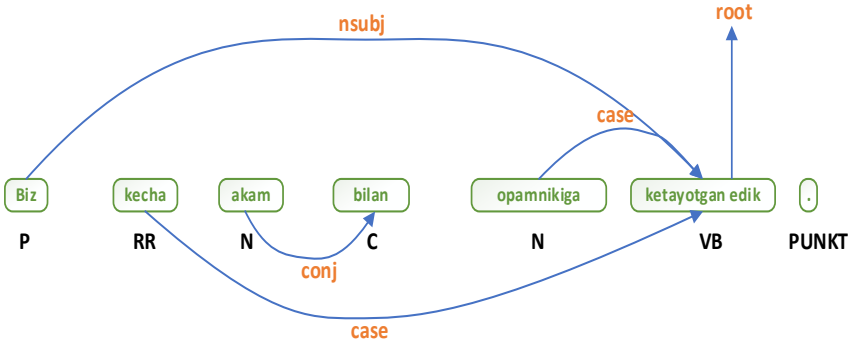


6-rasm. Tahlil daraxti diagrammasi.



7-rasm. Constituency (tuzilmaviy) parsing asosida sintaktik tahlil qilish jarayoni.

Dependency (bog‘liqlik) parsing esa **so‘zlar orasidagi to‘g‘ridan-to‘g‘ri munosabatlarga** e’tibor qaratadi. Bunda hech qanday maxsus fraza tugunlari kiritilmaydi, balki har bir so‘z boshqa bir *“so‘z”*ga bog‘lanadi. Masalan, *“Men kitobni o‘qidim”* gapida o‘qidim – **kesim** bosh so‘z, **Men** – unga bog‘langan **ega (nsubj)**, **kitobni – to‘ldiruvchi (obj)** sifatida bog‘lanadi. Bog‘liqlik tahlilida daraxtning har bir yoyiga grammatik rol (munosabat turi) tegi beriladi (masalan, nsubj, obj, advmod va hokazo). Bu usulning afzalligi – **tilga xos so‘z tartibiga unchalik bog‘liq emasligi**. Ega-kesim-to‘ldiruvchi munosabatlari qanday tartibda bo‘lishidan qat’iy nazar, dependency daraxtda baribir bir xil munosabatlar bilan bog‘lanaveradi. Shu sababli, bog‘lanishli parsing erkin tartibli, ko‘p qo‘shimchali tillarda barqarorroq natija beradi deb hisoblanadi [Abdurahmonova 2019].

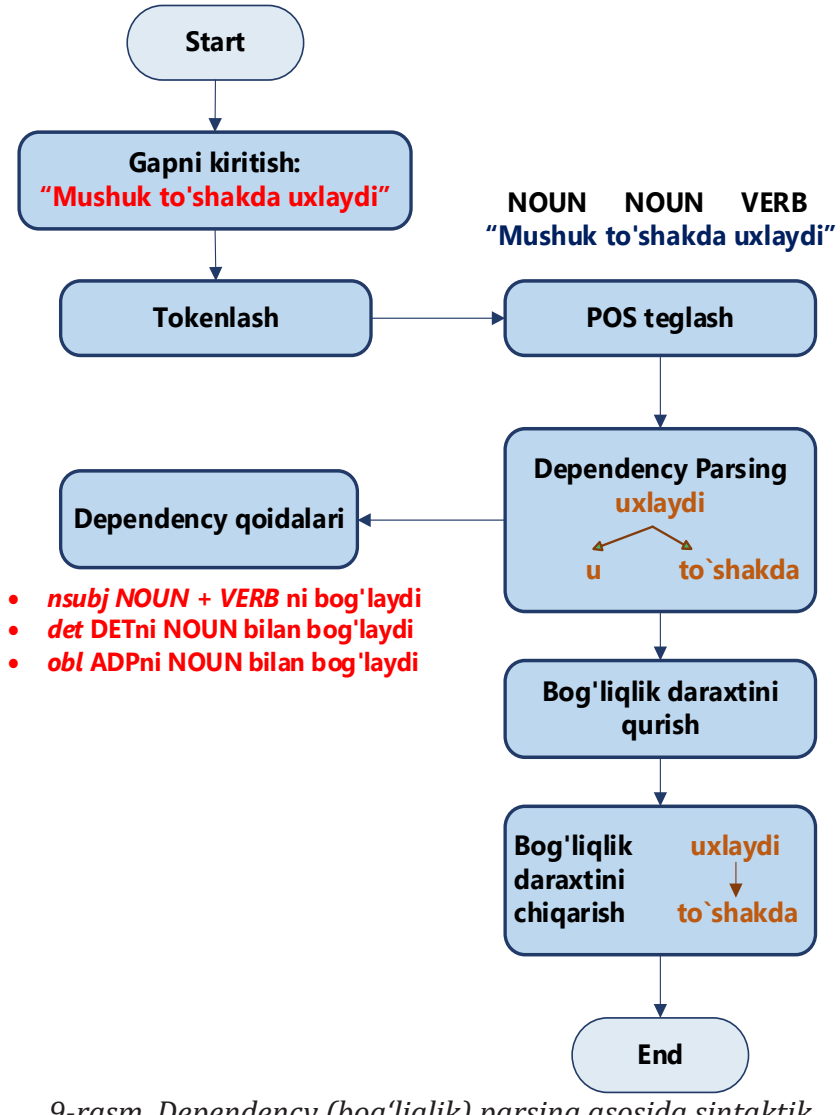


8-rasm. “Dependency Tree”ga namuna.

Amaliy jihatdan, bugungi kunda ko‘pchilik zamonaviy parserlar bog‘liqlik paradigmasidan foydalanadi yoki **dependency daraxtlarni** hosil qiladi. Bunga sabab – **Universal Dependencies** kabi standartlarning joriy qilinishi va bog‘liqlikning kross-lingvistik moslashuvchanligidir. Biroq, ba’zi vazifalarda constituency parsing ham qo‘l keladi: masalan, **matnning yuzaki tuzilmasini (treebank)** olish yoki **ayrim semantik tahlil** jarayonlarida birikma chegaralarini bilish kerak bo‘lsa, tuzilma daraxtlari ma’qul bo‘lishi mumkin. Zamonaviy yondashuvlar orasida hatto **gibrid parserlar** (masalan, birinchi bosqichda constituency, so‘ng dependency hosil qiluvchi) yoki bevosita **neural constituency parser** (Charniak va Johnson, 2005 yangilangan versiyalari, recent Span-CNN parser [Kahn va boshq. 2005, 233-240] va hokazo) ham mavjud.

O‘zbek tili sintaksisini ko‘rib chiqsak, an’anaviy grammatikada u mustaqil so‘zlar ergashuviga tayanadi va ko‘pincha kesim

gap oxirida keladi. Shu bois, **bog‘liqlik grammatikasi** o‘zbek tilini formalizatsiya qilishda qulay vosita bo‘lmoqda. Hozirgi kunda o‘zbek tilining to‘liq qoidalariga asoslangan constituency grammatikasi (masalan, CHELG yoki HPSG kabi) hali yaratilmagan, lekin dependency formatida kichik korpuslar belgilanishi boshlandi. Quyida UD formati va o‘zbek treebank misolida bu yondashuvga to‘xtalamiz.



9-rasm. Dependency (bog‘liqlik) parsing asosida sintaktik tahlil qilish jarayoni.

Universal Dependencies (UD) formati va o‘zbek UD Treebank

Universal Dependencies (UD) – bu turli tillar uchun yagona kelishilgan **bog‘liqlik annotatsiya formatini** taklif qiluvchi xalqa-

ro loyiha. UDda har bir so‘zga **universal turkum yorlig‘i (UPOS)**, morfologik xususiyatlar majmuasi va bosh so‘z bilan bog‘lovchi **bog‘lanish turi** beriladi. Masalan, UD da nsubj (ega), obj (to‘ldiruvchi), obl (hol) kabi munosabatlar barcha tillarda bir xil tamoyil bilan qo‘llanadi. UD, shuningdek, universal so‘z turkumlari (Ot, Fe’l, Sifat va boshq.) va birlik/ko‘plik, zamon, kelishik kabi morfologik kategoriyalar uchun qisqartmalarni belgilaydi. UD formatining maqsadi – har bir til uchun alohida formalizm o‘ylab topmasdan, barchasi uchun bir xil sxemada daraxtbanklar yaratish va shu orqali multilingvistik tadqiqotlarni yengillashtirish.

O‘zbek tili uchun ilk UD daraxtbank (treebank) yaqinda yaratildi. Akhundjanova va Talamo [Akhundjanova, Talamo 2025, 1-6] tomonidan Uzbek-UT deb nomlangan 500 gaplik treebank taqdim etildi. Ushbu treebank yangiliklar va badiiy matnlardan olingan 5,850 token (so‘z) uchun lemma, UPOS, morfologik xususiyatlar va bog‘lanish relatsiyalarini qo‘lda tekshirilgan holda beradi. Bu korpus UD me’yorlariga qat’iy amal qilgan va umumiy CoNLL-U formatida taqdim etilgan.

O‘zbek UD treebankini yaratishda mualliflar avval https://uznatcorpara.uz/uz/POSTag yordamida so‘zlarni ajratib, ularga avtomatik morfologik teglar berishgan, so‘ngra 150 ta gapni qo‘lda bog‘liqlik bilan belgilashgan. So‘ngra Stanford Stanza vositasining neyron parserini shu ma’lumotda dastlab o‘rgatib, qolgan gaplarga avtomatik bog‘lanishlar qo‘llashgan va ularni yana qo‘lda tuzatishgan. Natijada qisqa muddat ichida kichik hajmli bo‘lsa-da, to‘liq qo‘lda tekshirilgan birinchi **gold-standard treebank** hosil bo‘ldi.

Mazkur treebank ustida ishlash jarayonida o‘zbek tilining UD formatiga xos ba’zi qiyinchiliklari ham aniqlangan. Masalan, kesim+yordamchi fe’l konstruksiyalari (masalan, bor edi, qilgan edi) UD da bir nechta bog‘lanish orqali ifodalanadi va ularning qaysi so‘zni bosh, qaysi biri ergash so‘z bo‘lishi masalasi muhokama talab etgan. Shuningdek, kirish so‘zlar va modal konstruksiyalar (masalan, nazdimda, balki, kerak va hokazo) uchun UD relatsiyalarini aniqlashda muayyan ziddiyatlar yuzaga kelgan. Biroq, aynan shu muammolar UD hamjamiyati bilan hal etilib, o‘zbek tili sintaksisi UD doirasida ifodalash mumkinligi ko‘rsatildi.

O‘zbek UD treebank hajman kichik bo‘lsa-da, uning ahamiyati katta. U birinchidan, o‘zbek tilida statistik parserlar o‘qitish uchun data tayyorlaydi. Ikkinchidan, UD dagi 500 ta gap boshqa tillar bilan qiyosiy tahlil qilish imkonini beradi – masalan, turk yoki qozoq til-

lari treebanklari bilan strukturaviy o‘xshashlik va farqlarni ko‘rish mumkin bo‘ladi. Uchinchidan, bu treebank o‘zbek tilida lingvistik tadqiqotlarni (masalan, gap bo‘laklari tartibi, bog‘lanish munosabati) korpus asosida o‘tkazishga yo‘l ochadi.

SOV tartibi va erkin so‘z tartibining sintaktik tahlilga ta’siri

O‘zbek tilining **bazaviy so‘z tartibi – SOV (Subyekt–Obyekt–Predikat)** deya baholanadi. Ya’ni, odatiy gaplarda ega boshda, kesim esa gap oxirida keladi (*Men kitobni o‘qidim* – bu yerda “Men” – S, “kitobni” – O, “o‘qidim” – V). Biroq, amalda o‘zbek tili sintaksisi juda erkin: so‘zlarni gapda ma’lum ma’no urg‘usiga qarab qayta tartib- lash mumkin va bunda gap ma’nosi saqlanib qolaveradi (ko‘pincha faqat aktual bo‘laklar urg‘ulanadi). Masalan, *“Bugun men institutga bordim”* (oddiy tartib) va *“Men bugun institutga bordim”* o‘rtasida farq yo‘q; *“Institutga men bugun bordim”* shakli esa urg‘u yoki uslubiy maqsadga ko‘ra kelishi mumkin. Bunday **erkin so‘z tartibi morfologik ko‘rsatkichlar** orqali namoyon bo‘ladi – ot turkumlarining kelishik qo‘shimchalari ularning gapda qanday rol o‘ynayotganini (ega, to‘ldiruvchi yoki hol ekanini) ko‘rsatib turadi.

Erkin so‘z tartibi sintaktik tahlilga bir necha jihatdan ta’sir qiladi. Tuzilmaviy parsing nuqtayi nazaridan, o‘zbek tilida qat’iy qoidalar asosida birikmalarni aniqlash mushkul, chunki, masalan, ega har doim kesim oldidan kelmasligi mumkin, yoki aniqlovchi aniqlanmishdan keyin ham kelishi mumkin (kitob yaxshi, yaxshi kitob). Shuning uchun an’anaviy CFG grammatikasi juda ko‘p qoida va istisnolarga ega bo‘lib ketishi ehtimoli bor. Bog‘liqlik parsingi esa buni bevosita bog‘lanishlar orqali hal qiladi: masalan, *“Men kitobni o‘qidim”* gapida qanday tartibda bo‘lishidan qat’iy nazar, o‘qidim fe’liga Men subyekt sifatida bog‘lanadi (nsubj), kitobni esa obyekt (obj) bo‘ladi. Shu tariqa, bog‘liqlik strukturasida tartib o‘zgarishi grafiga ta’sir qilmaydi – faqat grafigi chizilganda so‘zlar chiziqli ketma-ketlikda boshqa o‘rinda tursa, yoylar chizilishi mumkin (hatto, no-proyektiv bog‘lanishlar paydo bo‘lishi ham mumkin).

O‘zbek tilida so‘z tartibi erkin bo‘lsa-da, **aktual bo‘laklar tartibi bor**: odatda gap boshida **tema (berilgan axborot), oxirida rema (yangi axborot)** joylashadi. Parserlar bunday pragmatik tartibni albatta bilmaydi, lekin ko‘pincha kesimning gap oxirida kelishi ustun holat bo‘lgani uchun, tartibni tahlil qilishi mumkin: **kesim bosh so‘z bo‘lib**, ega/to‘ldiruvchini bog‘laydi. Bu esa parser uchun **uzoq masofali bog‘lanishlarni topish vazifasini qo‘yadi**.

Ma'lumki, **transition-based (o'tishlarga asoslangan)** oddiy parserlar yaqin turuvchi so'zlarni bog'lashni oson, uzoqlarini mushkul o'rganadi. Shuning uchun o'zbek tilidagi parsing modelini tuzishda so'zlarning morfologik belgilarini (masalan, kelishiklarni) kirish sifatida berish foydali bo'ladi – chunki aynan shu belgilardan parser ega/to'ldiruvchini ajrata oladi.

Yana bir ta'sir jihati – **gap bo'laklarining tushirilishi (ellipsis).** O'zbek tilida ega yoki to'ldiruvchi tushib qolgan gaplar tez-tez uchraydi, chunki kesim orqali shaxs-son aniq bo'ladi (masalan, *"Keldim"* – egasi *"men"* tushirilgan gap). Bunday hollarda parsing jarayonida imPLICIT ravishda tushirilgan unsur bosh fe'lga bog'langan deb faraz qilinadi (UD da bu uchun Null tugunlar ishlatilmaydi, balki fe'l pro-drop xususiyati bilan belgilanishi mumkin). Parser esa buni konteksdan anglashga harakat qiladi yoki noto'g'ri tushunishi ham mumkin. Demak, erkin tartib aniqlikni pasaytiruvchi omillardir.

Hozircha o'zbek tili uchun sinov tariqasida o'tkazilgan parsing natijalariga nazar solinsa, kutilganidek, natijalar ingliz tiliga nisbatan pastroq. Masalan, 400 gaplik kichik o'quv to'plamida Stanza neyron parseri o'rgatilib, testda atigi **LAS 52% F1 baho** ko'rsatgan. Bu ko'rsatkichni ingliz tilidagi Penn Treebank'da o'qitilgan shunga o'xshash model 90% LAS ga erishishi bilan qiyoslash mumkin. Albatta, bu farqning katta qismi o'quv ma'lumotining kamligidan, qisman esa tilning o'ziga xos erkin tuzilishidadir. Shu bois, o'zbek tilida sintaktik tahlilni yaxshilash uchun:

1) ko'proq va turli soha matnlaridan iborat daraxtbank hajmini oshirish;

2) morfologik analiz bilan integratsiya – so'zlarning grammatik kategoriyalarini parserga berib borish;

3) multitil modellardan transfer learning qilish (turk, qozoq kabi tillar modellari yordamida) kabi choralar taklif etiladi.

Semantik tahlil metodlari

Semantik tahlil matndagi so'z va gaplar ma'nosini kompyuterga tushunarli tarzda ifodalash va interpretatsiya qilishni maqsad qiladi. Bu qatlam sintaksisdan keyingi bosqich bo'lib, lekin ko'pincha **leksik ma'no (so'z ma'nosi)** va **gap ma'nosi (predikat-argument tuzilmasi)** kabi qirralarga bo'linadi. So'nggi yillarda kompyuter lingvistikasida semantikani modellashtirish uchun ko'plab *vektorli metodlar* va *semantik lug'atlar* qo'llanilmoqda. Quyida so'z ma'nolarini vektor shaklida ifodalovchi **Word2Vec,**

FastText kabi modellardan tortib, **BERT, mBERT, UzBERT** kabi kontekstga moslangan transformator modellarigacha, shuningdek **sinonimiya/antonimiya, omonimlik** (ko‘p ma’nolilik)ni hisobga olish masalalari yoritiladi. Shuningdek, yuqori darajadagi **semantik rol belgilash** va **WordNet, FrameNet** kabi semantik tarmoqlarni yaratish masalalari ham ko‘rib chiqiladi.

Word2Vec, FastText asosidagi semantik vektorlar

Distributional Semantics tamoyiliga ko‘ra, *“so‘zlarning ma’nosi ularning kontekstlarida”* namoyon bo‘ladi [Firth, 1957]. Shu g‘oyaga asoslanib, **Word2Vec** [Mikolov va boshq. 2013 modeli katta matnlarda so‘zlarning qo‘llanish kontekstlarini o‘rganib, har bir so‘zga vektor moslashtiradi. Word2Vec metodining **Skip-gram** algoritmi tasodifiy bir so‘zdan uning atrofidagi so‘zlarni bashorat qilishga o‘rgansa, **CBOW** aksincha, kontekst bo‘yicha markaziy so‘zni topishga harakat qiladi. Natijada, o‘qitilgan model har bir so‘z uchun, aytaylik, 100 o‘lchamli vektor beradi va bu vektorlar semantik jihatdan yaqin so‘zlar uchun yaqin bo‘ladi. Klassik misol – *“king” - “man” + “woman” ~ “queen”* chiziqli vektor amali semantik mantiqqa ega bo‘lib chiqadi.

Word2Vec modellari o‘zbek tilida ham sinovdan o‘tgan. Katta hajmdagi o‘zbek matnlari (masalan, Vikipediya, yangiliklar korpusi) asosida Word2Vec o‘qitilganda, **so‘zlar orasidagi sinonimik yaqinlik** ancha yaxshi chiqadi. Masalan, model *“yigit”* vektori *“o‘g‘il”, “yosh odam”* vektorlariga yaqinlashishi yoki *“maktab”* so‘zi *“litsey”, “kollej”* so‘zlariga yaqin joylashishi kuzatiladi. Biroq, Word2Vec’ning cheklovi – har bir so‘zga bitta vektor biriktirishi. Ya’ni omonim so‘zlar (masalan, *“oy”* – [samoviy jism] va *“oy”* – [vaqt birligi]) uchun bitta vektor ikki ma’noni aralash ifodalaydi, natijada vektor kosmosida noaniqlik paydo bo‘ladi. Shuningdek, Word2Vec modelida **OOV (Out-Of-Vocabulary)**, ya’ni o‘qitish paytida uchramagan yangi so‘zlar uchun vektor hosil qilish imkoni yo‘q.

Shu muammo hal etish maqsadida **FastText** modeli [Bojanowski va boshq. 2017] taklif qilindi. FastText har bir so‘z vektorini uning belgilar n-grammalari vektorlari yig‘indisi sifatida ko‘radi. Masalan, *“kitob”* so‘zi 3-ta belgili qismaklar (ngrams) – *“kit”, “ito”, “tob”* kabilardan tashkil topib, ularning vektorlari yig‘ilib so‘z vektori hosil bo‘ladi. Bu yondashuv ayniqsa o‘zbek, turk, rus kabi ko‘p qo‘shimchali tillarda foydali: modomiki yangi so‘zning biror qismi modelda ko‘rilgan bo‘lsa, unga ham vektor tuzish mumkin. Grave va boshq. [Grave va boshq. 2018] tadqiqotida 157 til uchun

FastText vektor modellarini tayyorlab, ochiq e'lon qilgan – ulardan biri o'zbek tilidir. Ushbu model yordamida, masalan, *"boring"* va *"yo'q"* vektorlarini chiqarib, ular orasidagi kosinus o'xshashlikni hisoblasak, juda past chiqadi – bu tabiiy, chunki antonimlar (bor/yo'q) kontekstlarda kam uchrashadi birga. Aksincha, *"go'zal"* va *"chiroyli"* kabi sinonimlar ko'pincha bir xil kontekstlarda qatnashgani bois, ularning vektorlari juda yaqin bo'lib, model sinonimiyani *"his qilganini"* ko'ramiz.

O'zbek tilidagi **Word2Vec/FastText** modellari hali mukammal baholanmagan bo'lsa-da, **umumtil (general-language)**da yaxshi natijalar ko'rsatadi. Ular yordamida sinonimlar klasterlash, so'zlar orasidagi semantik masofani hisoblash, matnlardan mavzuga oid so'zlarni topish kabi masalalar hal qilindi. Biroq, so'z ko'p ma'noli bo'lsa, FastText ham uning barcha ma'nolarini bitta vektorga joylaydi, chunki model so'z kontekstini hisobga olmaydi – balki umumiy statistik pozitsiyasini oladi. Bu cheklovni bartaraf etish uchun kontekstga mos vektorlar zarur bo'ldi, bu esa BERT kabi modellar bilan amalga oshirildi (quyida tushuntiriladi).

BERT, mBERT, UzBERT va kontekstual modellar

BERT (Bidirectional Encoder Representations from Transformers) – 2018-yilda Google AI tomonidan taklif qilingan **transformator arxitekturasiga** asoslangan o'ta qudratli modeldir. BERT chuqur neyron tarmog'i matnni chapdan-o'ngga va o'ngdan-chapga bir vaqtda ko'rib, har bir so'zga uning butun gapdagi kontekstini e'tiborga olgan holda vektor (embedding) hosil qiladi. Muhimi, BERT modelida har bir so'z jonsiz emas, kontekstga bog'liq holda xarakterlanadi: masalan, *"oy"* so'zining *"Oy yulduzlar orasida eng yorqin jismdir"* gapidagi vektori bilan *"Oy davomida ko'plab ishlar qildik"* gapidagi vektori turlicha bo'ladi – model ularni turli ma'no farqlari bilan kodlaydi. BERT modellar **Masked Language Modeling** [Chung va boshq. 2021, 244-250] **(yashirin so'zni topish)** va **Next Sentence Prediction** [Shi, Demberg 2019] kabi vazifalarda ulkan korpuslarda o'qitilib, tilning ichki qoidalarini o'zlashtiradi.

Multilingual BERT (mBERT) – Google tomonidan 104 tilning Vikipediya matnlari ustida birgalikda o'qitilgan ko'p tilli model bo'lib, unga o'zbek tili ham kiradi. mBERT o'zbekcha matnni tushunishda ma'lum darajada qobiliyat ko'rsatadi, lekin u asosan lotin yozuvidagi Vikipediya bilan cheklangan (o'zbek tilining kirill alifbosi resurs sifatida qo'shilmagan edi). Shuningdek, mBERT alohida o'zbek tiliga moslashmagani bois ba'zan noaniq natijalar beri-

shi mumkin. Shu sabab, keyingi yillarda aynan o‘zbek tiliga mo‘ljallangan monolingual BERT modellar paydo bo‘ldi [Devlin va boshq. 2019].

UzBERT – o‘zbek tilining ilk monolingual BERT modelidir [Mansurov va boshq. 2021]. U katta hajmdagi o‘zbek matnlari (Wiki, siyosiy axborotlari va boshqalar) ustida o‘qitilib, natijada **mBERT-dan maskalangan so‘zni topish aniqligi** bo‘yicha sezilarli ustunlikka erishgan. Mualliflar hisobotiga ko‘ra, UzBERT multilingvistik BERTni MLM vazifasida katta farq bilan ustunlik qilgan. Bu shuni ko‘rsatadiki, tilga xos modellar resurs kam bo‘lsa ham, til xususiyatlarini yaxshiroq o‘zlashtira oladi. Keyinchalik yana bir nechta modelllar: **BERTurk (turkcha), XLM-R (ko‘p tilli RoBERTa)** singari, o‘zbek tiliga moslab chiqdi yoki chiqish arafasida. Masalan, 2023-yilda Kuriyozov va boshq. *“BERTbek”* nomli modelni va Davronov va Adilova boshqa bir BERT variantini tayyorlashgani haqida xabar berildi. Bu modellar turli arxitektura va hajmga ega, ba’zilari hatto rasmiy maqolasiz ham jamiyatda paydo bo‘lgan (masalan, *Zabarjad AI loyihasining UzBERT modeli, yoki HuggingFace dagi jamoaviy loyiha modellari*).

Kontekstual modellar afzalligi – **ko‘p ma’nolilikni hal etish va ingibitorli bog‘lanishlarni ko‘ra olishidir**. Masalan, *BERT asosidagi POS tagger* qo‘shimchalar bilan ifodalanadigan o‘zgarishlarni ham kontekstda *“sezadi”* va shu bois qoidalarga asoslangan tagger ko‘ra olmagan nozik farqlarni to‘g‘ri belgilaydi. Bobojonova va boshq. (2025) izlanishida UzBERT modelini POS taglashga moslashtirish natijasida **91% aniqlikka erishildi** va bu natija mBERT hamda qoidaviy taggerdan ancha yuqori ekanligi ko‘rsatildi [Bobojonova va boshq. 2025]. BERT modelining o‘zbek tilida kontekstni hisobga olishi tufayli, masalan *“kelmoqda”* fe’li bilan *“kelmoqchi”* fe’lining farqini atrofidagi so‘zlardan anglab, birinchisiga davomiy zamon, ikkinchisiga esa maqsad ma’nosi tegi qo‘yishi mumkin – qoidalilar esa ikkalasini ham aynan bir xil *“kelmoq”* fe’li deb belgilagan bo‘lardi.

Semantik tahlilda BERT va shunga o‘xshash modellar nafaqat so‘z turkumlarini aniqlash, balki **ma’noga oid vazifalarda** ham qo‘llanilmoqda. Masalan, *matn tasnifi, his-tuyg‘u analizi, savol-javob* tizimlarida BERT representation’lari juda samarali bo‘ldi. O‘zbek tilida ham 2022-yilda sentiment analiz korpusi tuzilib, ko‘p tilda o‘qittirilgan XLM-RoBERTa modeli bilan ijobiy-salbiy izohlar 88-90% aniqlikda ajratilgan [Matlatipov va boshq. 2024]. Bu shuni ko‘rsatadiki, kontekstual semantik modellar til chegaralarini buzib,

hatto resursi kam tillarda ham yuqori natija berishi mumkin, chunki ular transfer learning orqali bilimlarni ko‘chirish qobiliyatiga ega.

Sinonim-antonim munosabatlari va omonimlik

Sinonimiya va antonimiya – so‘zlar orasidagi asosiy leksik-semantik munosabatlardan bo‘lib, kompyuter uchun ularni aniqlash ma’noni tushunishda muhim ahamiyatga ega. **Sinonimlar** – ma’nodosh so‘zlar (masalan: *go‘zal – chiroyli – obod*), antonimlar esa ma’nosi qarama-qarshi so‘zlar (*issiq – sovuq, yaxshi – yomon*). O‘zbek tilida sinonimik qatorlar boy bo‘lib, ba’zan 5-6 tacha so‘z bir ma’noni turli usulda ifodalay oladi, antonimik juftliklar esa, asosan, sifatlarga xos (*katta-kichik, yangi-eski* kabi).

Distributsion modellar (*Word2Vec, FastText* va hokazo) sinonimiyani qisman *“ushlaydi”:* odatda sinonim so‘zlar o‘xshash kontekstlarda ishlatilgani bois, ularning vektorlarini yaqin joylashtiradi. Bu modellar antonimiyani esa yaxshi farqlay olmaydi, chunki antonimlar ham ko‘pincha bir xil mavzuda uchrashadi (masalan, *issiq havo* va *sovuq havo* bir xil havo so‘zi bilan ketadi, natijada issiq/sovuq vektorlar orasidagi masofa kichik bo‘lishi mumkin). Shu bois, kompyuter sinonim/antonimni ajratishi uchun maxsus resurslar yoki algoritmik yondashuvlar kerak. Masalan, WordNetda har bir synset (sinonimlar guruhi) uchun antonimlar ham berilgan bo‘ladi, yoki tezauruslarda sinonim-antonim aloqalari qo‘lda kiritiladi.

Omonimlik (polysemy) – bir so‘zning bir nechta o‘xshash yoki tamoman boshqa ma’nolarga ega bo‘lishidir. O‘zbek tilida omonimlik keng tarqalgan: *“tun”* (geometrik figura burchagi) va *“tun”* (kecha vaqti) kabi yoki *“qalam” (asbob)* va *“qalam” (qo‘zichoq)* kabi misollar talaygina. Klassik vektor modelar (Word2Vec kabi) omonimlarni bir vektorga joylab qo‘yganligi yuqorida aytildi. Kontekstual modellar (BERT) esa ushbu muammoni birmuncha hal etadi – har xil gapda *“tun”* so‘zi turlicha vektor oladi va bu vektorlar orasi masofa sezilarli farq qiladi. Demak, BERT singari model yordamida kompyuter kontekstga qarab so‘zning qaysi ma’noda ishlatilganini ajrata oladi. Bunga qo‘shimcha ravishda, so‘z ma’nolarini alohida identifikatsiya qilish uchun sense embedding deb ataluvchi yondashuvlar ham mavjud: misol uchun, Word2Vec modelini alohida ma’nolar uchun o‘rgatish. Lekin bunday usullar uchun avval so‘zning necha ma’nosi borligini bilish kerak, bu esa qo‘lda belgilashni talab qiladi.

Sinonim-antonim munosabatlarni kompyuter tushunishi

turli ilovalarda zarur: qidiruv tizimlari sinonimlarni bilsa, foydalanuvchi so‘rovidagi so‘zga o‘xshash ma’noli hujjatlarni ham topa oladi; mashina tarjimasi antonimlarni chalkashtirmasligi lozim (masalan, *“uncha yomon emas”* iborasini *“not bad”* deb tarjima qilish); dialog tizimlari sinonimik ifodalarni yagona ma’no vakiliga keltira olishi kerak (parafrazni tushunish). Bunda **sezgir lug‘atlar (thesaurus)** va **semantik tarmoqlarning** yordami katta bo‘ladi.

Semantik rol belgilash (SRL)

Semantik rol belgilash – gapdagi predikat (odatda fe’l) va unga bog‘liq argumentlarning semantik rollarini aniqlash vazifasidir. Bu, sodda aytganda, *“Kim nima qildi, kimga qildi, qanday, qachon, qayerda?”* kabi savollarga struktural javob berishdir. Masalan, *“Ali bog‘da do‘stiga kitob berdi”* gapida *“Ali”* – **Agent** (harakatni bajaruvchi), *“kitob”* – **Theme** (berilayotgan narsa), *“do‘stiga”* – **Recipient** (qabul qiluvchi), *“bog‘da”* – **Locative** (joy holati). SRL natijasida bu rollar formal ravishda belgilab chiqiladi.

Ingliz tilida SRL uchun PropBank va FrameNet kabi katta teglangan korpuslar mavjud bo‘lib, ularga tayangan modellar **90% F1 baho** atrofida aniqlikka erishadi. O‘zbek tilida esa hozircha SRL uchun teglangan korpus mavjud emas. Ammo o‘zbek tilining boy flektiv (qo‘shimchali) xususiyati SRLni biroz yengillashtirishi ham mumkin: ko‘pincha qaysi argument qaysi rolni bajarishini kelishik qo‘shimchalari ko‘rsatib turadi. Masalan, **-ga** qo‘shimchasi odatda Recipient rolini (kimga?) ifodalasa, **-ni** qo‘shimchasi bevosita Theme/Patient (nimani?) rolini beradi. Albatta, hamisha ham bunday emas – ba’zan **-ni** egalikni ham bildirishi mumkin yoki **-ga** boshqa maqsad holini ham (qaerga? nimaga?) ifodalaydi. Shunday bo‘lsa-da, morfologik belgilarga qarab SRLni qoidaviy amalga oshirish imkoniyati mavjud.

SRLning ikki yondashuvi bor: **Frame-based** (FrameNet usuli) va **Predicate-based** (PropBank usuli). FrameNet yondashuvida har bir fe’l (yoki ot, sifat) ma’lum bir **semantik frame** (ramka) ga tegishli deb olinadi va o‘sha frame uchun oldindan belgilangan rollar to‘ldiriladi. Masalan, *“bermoq”* fe’li Giving frameiga mansub bo‘lib, uning doimiy rollari **Donor** (beruvchi), **Theme** (narsa), ***Recipient*** (oluvchi). PropBank usulida esa har bir predikat uchun **Arg0, Arg1, Arg2...** kabi ro‘yxat belgilangan bo‘lib, ular taxminan **Agent, Patient, ecc.**ga mos keladi. Hozircha o‘zbek tilida FrameNet ham, PropBank ham yo‘q, lekin rus tili uchun masalan, Russian FrameBank loyihasi boshlangan, turk tilida ham ba’zi SRL tadqiqot-

lari o‘tkazilgan. Ushbu tajribalarni o‘zbek tiliga transfer qilish mumkin: masalan, parallel korpuslarda inglizcha SRLni o‘zbek gaplariga proyeksiya qilish orqali dastlabki belgilashni hosil qilish va keyin uni tozalab chiqish mumkin.

SRL bajarishda zamonaviy modellardan foydalanish – bu **transformer model + biror klassifikator arxitekturasi**dir. Dastlab, BERT kabi model yordamida har bir so‘z kontekstual vektorga o‘tkaziladi; so‘ngra maxsus sarlavha (classifier) har bir so‘z uchun u qaysi rolga tegishli (yoki rol emas) ekanini belgilaydi. Bunda, albatta, gapda bir nechta predikat bo‘lishi mumkin, shuning uchun ko‘p predikatli SRL masalasi alohida e’tibor talab qiladi. Ko‘p tadqiqotlarda SRLni ikki bosqichga ajratishadi: avval predikat (asosiy fe’l)larni aniqlash, so‘ng ularga argumentlarni bog‘lash. O‘zbek tilida fe’l shakllari ko‘p, shu jumladan, masdarlardan ham predikativ foydalanish hollari bor (*“Borishim kerak”* – bu yerda **-im** kerak tarkibi predikat), shuning uchun SRL komponenti bularni ham e’tiborga olishi lozim. Umuman, SRL matnni chuqur tushunish yo‘lida muhim qadam hisoblanadi. U yordamida avtomatik ravishda matndan *“kim-kim bilan, nima qildi”* kabi ma’lumotlar chiqib, ma’lumot bazalari shakllanishi mumkin. O‘zbek tilida SRL ustida hali ishlar to‘liq boshlanmagan bo‘lsa-da, uning uchun zarur bo‘lgan sintaktik va morfologik tahlillar endilikda yo‘lga qo‘yilmoqda. Kelgusida, o‘zbek tilida ham FrameNet singari resurslar paydo bo‘lsa, SRL belgilashni avtomatlashtirish ancha osonlashadi.

WordNet, FrameNet va o‘zbekcha semantik lug‘atlar

WordNet – so‘zlarning semantik tarmog‘i bo‘lib, unda so‘zlar synset (ma’nodoshlik guruhi)larga ajratiladi va guruhlar orasida turli munosabatlar o‘rnatiladi (hypernymy – umumiy ma’no, hyponymy – xususiy ma’no, meronymy – qism/butun, antonymy – qarama-qarshilik, va hokazo). Asl Princeton WordNet ingliz tilida Prof. Miller boshchiligida yaratilgan bo‘lib, hozirda ko‘plab tillar uchun WordNet analoglari tuzilgan (**RusWordNet, EuroWordNet, Bolgar WordNet, TurkWordNet** va boshqalar). WordNet lug‘at emas, balki ontologik tarmoq – ya’ni kompyuter uchun ma’lum bir so‘zlar dunyosining modeli. Misol uchun, WordNetda *“kitob”* so‘zi *“yozuv ashyosi”* hyperonimiga ega, uning hyponimlari qatoriga *“darslik”, “qo‘llanma”* kabi kiradi; *“kitob”*ning qismi (meronim) sifatida *“sahifa”* bor, antonimi yo‘q; sinonim synsetiga esa *“asar”* kiritilishi mumkin edi.

O‘zbek tilida ham WordNet yaratish bo‘yicha dastlabki

qadamlar tashlangan. Agostini va boshq. [Agostini va boshq. 2021] UzWordNet deb ataluvchi lexical-semantik bazani taqdim etdilar, u dastlab 28 ming synset, 64 ming sense (so‘z ma’nosi) va 20 ming lemmadan iborat bo‘lib, taxminiy aniqligi 75.98% ekanini qayd etishgan. UzWordNetni yaratishda ko‘p tilli lug‘atlar va mavjud WordNetlar (masalan, TurkWordNet) ma’lumotlari tarjima va moslashtirish yo‘li bilan foydalanilgan [Madatov va boshq. 2023]. Yana bir yondashuv sifatida, Madatov va hammualliflar (2022) turkcha WordNet asosida o‘zbek WordNet tuzish haqida maqola chop etishgan – unda turk tilidagi synsetlar tarjimasi orqali o‘zbekcha synsetlar bazasi qurilganini bayon qilishgan. Hozirgi kunda UzWordNet rivojlantirilmoqda va uning aniqligini oshirish, hamda qamrovini kengaytirish ustida ishlar davom etmoqda. Bu kabi resurslarning ahamiyati shundaki, **ma’lumotlar integratsiyasida** (knowledge graphs), **so‘z ma’nosini aniqlash** (word sense disambiguation), **ma’noli qidiruv** (semantic search) kabi vazifalarda kompyuterga lug‘aviy bilim beradi.

FrameNet – semantik frames nazariyasiga asoslangan lug‘aviy baza. Frame – bu muayyan **voqealar yoki konseptlar** uchun yaratilgan andoza bo‘lib, unda ishtirokchi rollar (participant roles) va misollar keltiriladi. Masalan, Buying frame: **Seller, Buyer, Goods, Payment** kabi rollarni o‘z ichiga oladi va *“X sotib oldi Y dan Z ni P ga”* shaklida ifodalanishi mumkin. Ingliz tilida FrameNet (Baker va boshq., 1998) mingdan ortiq frame va o‘n minglab leksik birliklarni qamrab oladi. FrameNetdan SRL va semantik tahlil vazifalarida foydalanish mumkin – gapni qaysi framega tegishli ekanini topish va rollarini aniqlash ko‘rinishida.

O‘zbek tilida hozircha FrameNet yaratilmagan. Biroq, bu yo‘nalishda imkoniyatlar mavjud: masalan, inglizcha FrameNetni va o‘zbekcha tarjimasini lug‘atlar orqali bog‘lab, dastlabki **“O‘zbek FrameNet”**ni avtomatik hosil qilish mumkin. Ba’zi frame’lar to‘g‘ridan-to‘g‘ri mos tushsa (masalan, Movement frame – Harakat frame), ba’zilarini madaniy yoki til xususiyatlarini hisobga olib yangidan tuzish kerak bo‘lishi mumkin. O‘zbek tilida o‘zgacha frames ham paydo bo‘lish ehtimoli bor (masalan, *“Choy ichish”, “Gap tashlash”* kabi madaniy konseptlar). FrameNet yaratish – juda katta mehnat talab etuvchi ish, lekin natijasida tilning ma’no ontologiyasi shakllanadi va bu tilga “kompyuter tushunishi” uchun zamin yaratadi.

Yuqorida zikr etilgan **UzWordNet** lohiyasi – o‘zbek tilida ilk semantik lug‘aviy resurs bo‘lsa, FrameNet hali yo‘qligi ma’lum.

Shuningdek, tezaurus (ma'nolar asosida tuzilgan lug'at) yo'nalishida ham ishlar kam. Bir paytlar O'zRFA Til va adabiyot institutida *"O'zbek tilining sinonimlar lug'ati", "Antonimlar lug'ati"* kabi asarlar yaratilgan, ammo ular elektron shaklga to'liq o'tkazilgani va NLP-da ishlatilayotgani haqida ma'lumot kam. Kelgusida, WordNet va FrameNet kabi tuzilmalarni shakllantirish o'zbek tilini chuqurroq semantik tahlil qilish, shu tilga mos sun'iy intellekt tizimlarini yaratish uchun mustahkam poydevor vazifasini o'taydi. Xususan, WordNet yordamida **matnlarda so'z ma'nolarini avtomatik ajratish (WSD)** va **ma'nolar orasidagi munosabatlarni tahlil qilish** kabi masalalar yaqqol yechim topadi.

Metodlar afzalliklari va kamchiliklari

Yuqorida batafsil ko'rib chiqilgan har bir metod va yondashuvning o'z afzallik va cheklovlari mavjud. Quyida umumiy xulosa tarzida ularni jamlaymiz va ayniqsa o'zbek tilining o'ziga xos jihatlari bilan bog'liq holda muhokama qilamiz.

1. **Qoidaviy (rule-based) metodlar: Afzalliklari** – lingvistik jihatdan izchil va aniq; kichik xatolarga yo'l qo'yadi va grammatik noodatiy holatlarga tayyor (chunki mutaxassislar qoida sifatida kiritgan bo'ladi). O'zbek tilida bunday yondashuv bilan, masalan, to'liqsiz gaplarni ham ma'lum darajada tahlil qilish mumkin (imlo va grammatik qoidalarga asoslangan tekshiruvlar kabi). **Kamchiligi** – qoidalarni tuzish juda ko'p inson mehnatini talab qiladi; yangi so'zlar va istisno holatlar paydo bo'lganda tizimni yangilab borish kerak; eng muhimi – **kontekstni his qila olmaslik**. Masalan, qoidaviy POS tagger faqat qo'shimchalarga qarab teg qo'yadi va to'liq kontekstni e'tiborga olmaydi, shu sababli u ko'p ma'noli yoki funksional so'zlar atrofida xatolarga yo'l qo'yadi. Bobojonova va boshq. qayd etishicha: mavjud qoidaviy teggerlar kontekstni his qilmaydi va shu sababli neyron modellar darajasiga yeta olmaydi.

2. **Statistik va klassik ML metodlari (HMM, CRF va boshq.): Afzalliklari** – ma'lumot bor joyda qoidalarni qo'lda tuzish o'rniga modelni o'zi o'rganadi, demak tezroq va osonroq moslashadi; murakkab chegaraviy holatlarni ehtimollar orqali hal qiladi (masalan, *"yuz"* so'zi sonmi yoki inson a'zosimi – korpusda qaysi hol ko'proq uchrasa, o'shanisiga moyillik beradi). **Kamchiligi** – ko'p ma'lumot talab etadi; agar korpus kichik bo'lsa, statistik model mukammal chiqmaydi yoki qoidalarga asoslangan metodlardan ko'p xato qiladi. O'zbek tilida erta bosqichda bunday modellar yetarli korpus bo'lmagani uchun sustroq edi. Misol uchun, HMM asosidagi

POS tegger 75% aniqlikka erisha olgan bo‘lsa, qoidaviy tegger esa o‘z ichki testida 90% ko‘rsatgan [Sharipov va boshq. 2023]. Shunga qaramay, korpuslar ortib, model parametrlarini to‘g‘ri tanlash bilan statistik usullar ham qoidalilarni quvib yetadi. CRFning afzalligi – turli xususiyatlarni qo‘shib bo‘lishi, masalan, so‘z oxiridagi 2 harf *“-ni”* bo‘lsa, ehtimol bu obj belgisidir degan qoidadan ham yaxshi ishlashidir, chunki ni ko‘plik shakli ham bo‘lishi mumkin (ko‘plik: lari va ni qoidalarini birga mulohaza qiladi).

3. **Chuqur o‘rganish modellari (BiLSTM, CNN, Transformer):** Afzalliklari – juda katta hajmdagi lingvistik bog‘lanishlarni avtomatik modellashtiradi, ya’ni qo‘lda kiritilmagan qoidalarni ham o‘z tahlilida aks ettiradi. Shu bois eng yuqori natijalarga erishadi: masalan, ingliz tilida POS teglashda inson darajasiga yaqin 97% aniqlikka chiqdi [Huang va boshq. 2015]. O‘zbek tilida ham BERT asosidagi modellar avvalgi yondashuvlardan sezilarli ustun chiqdi (teglashda 91% aniqlik, parsing yoki boshqa vazifalarda ham kutilmoqda). **Neyron modellar kontekstni to‘liq hisobga oladi**, masalan, oldingi va keyingi so‘zlarni, butun gap ma’nosini e’tiborga olib, mos kelmaydigan teglarni chiqarib tashlaydi – bu qoidaviy yoki an’anaviy statistik modellar uchun qiyin vazifa. Kamchiliklari – **ko‘p ma’lumot va resurs talab etadi**: o‘qitish uchun katta hajmdagi belgilangan korpus yoki o‘rgatilgan model (pretrained) kerak. Yana biri – tushuntirish qiyinligi: model nima sababdan bunday qaror qilganini izohlash mushkul (shaffoflik yo‘qligi). Amaliyotda yana bir muammo – agar ma’lumotda notekislik bo‘lsa (masalan, ba’zi konstruksiyalar juda kam uchrasa), model ularni yaxshi o‘rganmaydi.

4. **FST va qoidaviy morfologik tahlil:** Afzalliklari – to‘liq va aniqlik bilan barcha mumkin bo‘lgan tahlillarni beradi; morfologik jihatdan murakkab tillarda bu ajralmas vosita (ayniqsa, turk, fin, koreys kabi tillar uchun). O‘zbek tilida ham FST analizatori istalgan so‘zning grammatik tarkibini ajratib bera oladi (agar lug‘atda bo‘lsa, yoki so‘z yasalish qoidalariga asoslanib yangisini ham idrok eta oladi). Kamchiligi – noaniqlikni yechmaydi: bir so‘zning tahlilida shakliy o‘xshashlik bo‘lsa, barchasini chiqaradi, lekin qaysi biri to‘g‘ri – bu masalani keyingi bosqichga yuklaydi. Shuningdek, FST bazasi doimiy yangilanmasa, yangi jargon, sheva so‘zlarni tushunmaydi. Ammo bu kamchiliklar FSTni morfologik tahlildan voz kechishga sabab bo‘la olmaydi, chunki alternativini statistik modellar bilan yechish mumkin. Bu modelga cheksiz kombinatsiyadagi qo‘shimchalarni o‘rgatish amalda qiyin. Shu bois ko‘plab tizimlarda birlam-

chi tahlil FST bilan, so‘ng ranking (saralash) ML modellar bilan tahlil qilinadi.

5. **Dependency parsing va Constituency parsing:** Dependency parsing afzalligi – **mustaqil**; erkin so‘z tartibiga moslashuvchan; UD kabi standart mavjudligi uchun ko‘p tillar tajribasi jamlangan; semantik tahlilga bevosita o‘tishda qulay (predikat-argument tuzilmani osongina ko‘rish mumkin). Constituency parsing afzalligi – **an’anaviy sintaktik tahlilga mos** (masalan, aniqlovchi+ aniqlanmish bir fraza deb ko‘riladi, dependencyda esa ikkita alohida bog‘lanishlar); ba’zi lingvistik nazariya va qo‘llanmalar constituency asosida bo‘lgani uchun tushuntirishga qulay. O‘zbek tilida, yuqorida aytganimizdek, bog‘liqlik yondashuvi samarali va shu yo‘l tanlanishi maqsadga muvofiq. Ayniqsa, **SOV va erkin tartib bog‘liqlik** usulida yaxshi *“ko‘rinadi”* – chunki so‘zlarning roli tartibdan emas, bog‘lanish turidan bilinadi. Kamchiligi – UD treebank hajmi kichikligi natijasida hozircha parsing aniqligi past. Shuningdek, dependency formatida birikma chegaralari yo‘qligi ayrim qo‘llanmalar uchun noqulay: masalan, nutqning sintaktik ravonligini tekshirishda constituency daraxtdagi fraza balansi hisobga olinar edi, dependencyda bunday to‘g‘ridan-to‘g‘ri tushuncha yo‘q.

6. **Word2Vec/FastText statik embedding:** Afzalliklari – **o‘ta katta korpuslarda o‘rgatish oson**, hech qanday qo‘lda belgilangan ma’lumot talab etmaydi (unsupervised); vektor fazoda semantik o‘xshashliklarni oddiy evklid masofa yoki kosinus o‘xshashlik orqali aniqlash mumkin; ularni saqlash va ishlatish nisbatan yengil (har bir so‘zga 300 o‘lchamli vektor, matnga qarab tanlab olinadi). O‘zbek tilida ham shunday model bo‘lsa, hatto resurs kam vazifalarda (masalan, klasterlashda) foydalanish mumkin. Kamchiligi – **polisemiya hisobga olinmaydi**, **kontekstni farqlamaydi;** faqat so‘z darajasidagi reprezentatsiya beradi, butun gap yoki matn ma’nosini ifodalash uchun qo‘shimcha mexanizm kerak (masalan, average pooling yoki sifatsiz RNN). Shuningdek, **statik vektorlar antonimiyani farqlay olmaydi**, chunki statistik jihatdan sinonimga yaqin qo‘yishi mumkin – bu yuqorida aytilgan.

7. **BERT va kontekstual modellar:** Afzalliklari – **kontekstga bog‘liq nozik ma’no farqlarini ajrata oladi**; ko‘p funksiyani bitta model bajarishi mumkin (bir modeldan embedding olib, istalgan tasniflagichga bersa bo‘ladi); hozirda deyarli barcha SOTA natijalar transformer modellar tomonidan o‘rnatilmoqda. O‘zbek tilida ham bu trend kuzatilmoqda: masalan, POS taglashda BERT asosidagi model qoidaviydan ham, eski ML modellaridan ham ustun bo‘ldi;

matn tasnifida ko‘p tilli model 90% dan oshiq aniqlikka erishdi. Kamchiliklari – **hajmi kattaligi va hisoblash resursi**: BERT-base 110 million parametr, o‘zbek tilida uni noldan o‘qitish (pretraining) juda og‘ir va ko‘p GPU vaqtini talab qiladi; hatto o‘qitilgach, uni xizmatda qo‘llash ham sekin (har bir input uchun murakkab transformer hisoblari). Bu muammo yechimi uchun siqilgan yoki optimallashtirilgan versiyalar (DistilBERT, TinyBERT) paydo bo‘lgan, lekin ular asosan resurslari ko‘p tillarda tayyorlangan. Yana bir cheklov – BERT kontekstual bo‘lsa-da, u gap doirasida: agar matn katta bo‘lsa (masalan, paragraf), BERT odatda 512 token bilan cheklangan. Matn darajasidagi semantik tushunish uchun yuqoriroq tuzilma (maybe transformer-decoder modellar, GPT kabilar) kerak bo‘ladi. Ammo umumiy holda, kontekstual modellar semantik tahlil uchun hozir eng kuchli qurol bo‘lib, ularni cheklashdan ko‘ra moslashish maqsadga muvofiq.

8. **Leksik resurslar (WordNet, FrameNet):** Afzalliklari – **inson bilimini struktura** shaklida berib, kompyuterga ma’lum bir tilning ma’no olamini tushunish imkonini beradi; kichik ma’lumotlarda ham ishonchli ishlaydi (chunki qo‘lda tuzilgan qoidalar/aloqalar); natijalarni tushuntirish va izohlash oson (masalan, nega bu ikki so‘z o‘xshash – chunki WordNetda bir synsetda). Kamchiliklari – **yana resurs yaratish qiyinligi**: WordNet tuzish katta mehnat, FrameNet undan ham murakkab; yangi atamalar va sohalar qamrab olinmasligi mumkin (masalan, kompyuter sohasi jargonlari WordNetga kiritilmagan bo‘lishi mumkin); tilning ba’zi o‘ziga xosliklarini formal tuzilmaga solish qiyin (masalan, idiomatik iboralar). O‘zbek tilida bu resurslar endi yaratila boshlagan, ularning sifati inglizcha analoglardan past (UzWordNet 76% aniqlik deb baholangan). Shunday bo‘lsa-da, kichik korpusli tahlilda yoki qidiruvda WordNet katta manba bo‘lishi mumkin – masalan, foydalanuvchi *“yurt”* deb izlaganda *“vatan”, “xalq”, “o‘lka”* degan sinonimlarni ham qidirish uchun, WordNet’dan sinonimlar to‘plami olinishi mumkin. FrameNet va SRL resurslari esa tilni chuqurroq anglashga xizmat qiladi, lekin ularni tuzish eng qiyin ish bo‘lib qolmoqda.

9. **O‘zbek tilining o‘ziga xos muammolari:** Yuqorida muhokama qilinganidek, o‘zbek tili morfologik jihatdan murakkab va sintaktik jihatdan o‘zgaruvchan tuzilishga ega. Bu, bir tomondan, tahlilni qiyinlashtirsa (ko‘p qismlar, ko‘p variantlar), boshqa tomondan, ba’zi yechimlarni osonlashtiradi (masalan, kelishik orqali gap bo‘laklarini bilsa bo‘ladi). O‘zbek tilining yana bir muammosi – **kirill va lotin yozuvi muammosi**: matnlar ikki alifboda kelishi mum-

kin. Modellarni tuzishda buni e'tiborga olish kerak – masalan, Uz-BERT modellaridan birida kirill matnlarni ham qamrab olishi uchun maxsus versiya qilindi. Yana **resurslarning kamligi va bir joyga jamlanmagani** – masalan, katta korpuslar qurilmagani, treebank yaqinda paydo bo'ldi, semantik belgilangan korpus yo'q – shular ham jiddiy to'siq. Ammo bu muammolarni bosqichma-bosqich hal etish yo'lida ilmiy ishlar olib borilmoqda va ayni bo'limning yozilishi ham shu yo'nalishdagi izlanishlar natijasidir.

Eksperimentlar va natijalar: statistik ko'rsatkichlar

O'zbek tilida morfologik, sintaktik va semantik tahlil bo'yicha o'tkazilgan ilmiy eksperimentlar hali ko'p emas, ammo mavjudlari u yoki bu yondashuvning samaradorligini ko'rsatib bermoqda. Ushbu bo'limda ba'zi muhim natijalar va ko'rsatkichlar keltiriladi. Model va metodlarni taqqoslash uchun aniqlik (accuracy), F1-o'lchov (aniqlik va to'liqlik uyg'unligi), perpleksiya (til modellari uchun) va boshqalar ishlatiladi. Quyida ba'zi eksperimental natijalarni jadvalda keltirilgan:

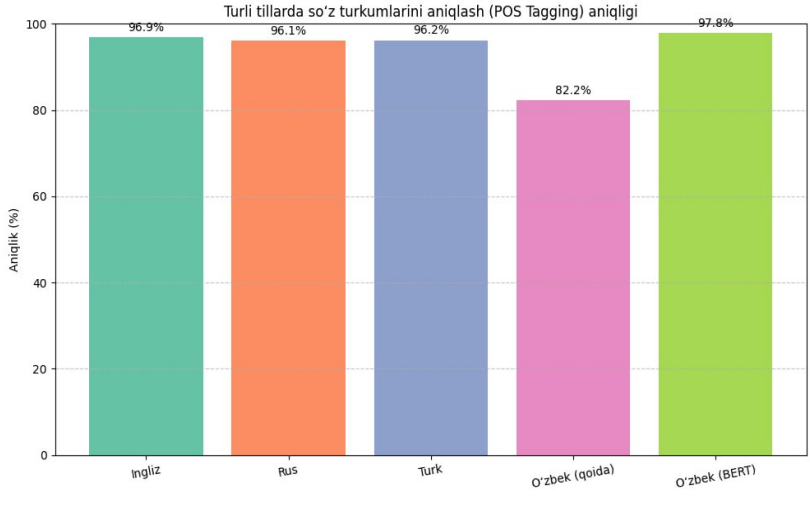
Vazifa/Resurs	Metod/Model	O'lchov (metrka)	Natija
So'z turkumlarini belgilash (POS, uz)	Qoidaviy yondashuv (UzbekTagger)	Aniqlik	82.2%
	HMM/CRF model (baseline)	Aniqlik	87.5%
	Neyron (BiLSTM+attention, 2024)	Aniqlik	89.7%
	BERT fine-tuning (Elov)	Aniqlik	97.8%
Bog'liqlik parsing (uz)	Stanford Stanza (UD 500 gap)	LAS (F1)	8.9
Til modeli (uz)	3-gram (statistik)	Perplexity (test)	3.12
UzWordNet semantik bazasi	Qo'lda + avto	Synsetlar soni (aniql.)	28140 synset (75.98%)

Yuqoridagi jadvaldan ko'rinadiki, o'zbek tilida **neyron yondashuvlar** an'anaviy usullarga nisbatan sezilarli ustunlikka ega bo'la boshladi: masalan, Elov tomonidan ishlab chiqilgan POS teglashda BERT modeli 97.8% aniqlikka erishib, qoidaviy dasturdan ancha yuqori natija berdi. Bunda, albatta, yangi belgilangan korpus (500 gap)ning hissasi katta bo'ldi – axir avval ma'lumot yo'qligi sababli qoidaviy yondashuvlar ustun turardi. Endi esa ilk bor baholash imkoniyati yaratilgach, kamchiliklar aniqlandi: masalan, qoidaviy tegger atrofdagi faqat yaqin **bir so'zga qarab qaror**

qiladi, BERT modeli esa **butun gapni e'tiborga olib** to'g'riroq teg qo'ymoqda.

Dependency parsing bo'yicha hozircha natijalar past – 52% LAS, lekin bu faqat 400 ta gapda modelni o'qitish natijasi. Solishtirish uchun, ingliz tilida 40k gapli treebanklarda parserlar 90%dan yuqori LASga erishadi. Demak, to'plam kattalashsa, o'zbek tilida ham xatoliklar sezilarli kamayadi. Yana kuzatilgan bir holat: o'zbek parsing natijalarida label aniqligi (Qaysi munosabat turi ekanini topish) nisbatan baland, lekin head aniqligi (Qaysi so'zga bog'lanishi) pastroq bo'lgan. Bu shuni anglatadiki, parser odatda munosabat turini to'g'ri taxminlaydi (masalan, subyekt ekanini), lekin qaysi so'zga bog'lashda adashadi – bu katta ehtimol bilan so'z tartibining erkinligidan kelib chiqmoqda.

Til modellari uchun perplexity ko'rsatkichlari berildi: 3-gram model testda 8.9, LSTM model esa 3.12. Perpleksiyaning pasayishi – modelning matnni bashorat qilish qobiliyati oshganini bildiradi. Ko'rinadiki, neyron LSTM til modeli an'anaviy n-gram modelga nisbatan 2.8 barobar past perplexity ko'rsatdi, ya'ni aynan shu tadqiqotda [Mukhamadiyev va boshq., 2023] nutq tanib olish tizimining aniqligi oshganligi qayd etildi. Bu misol semantik model sifatida til modellari (masalan, GPT-2 kabi)ni ham o'zbek tilida o'qitish samarali bo'lishini ko'rsatadi.

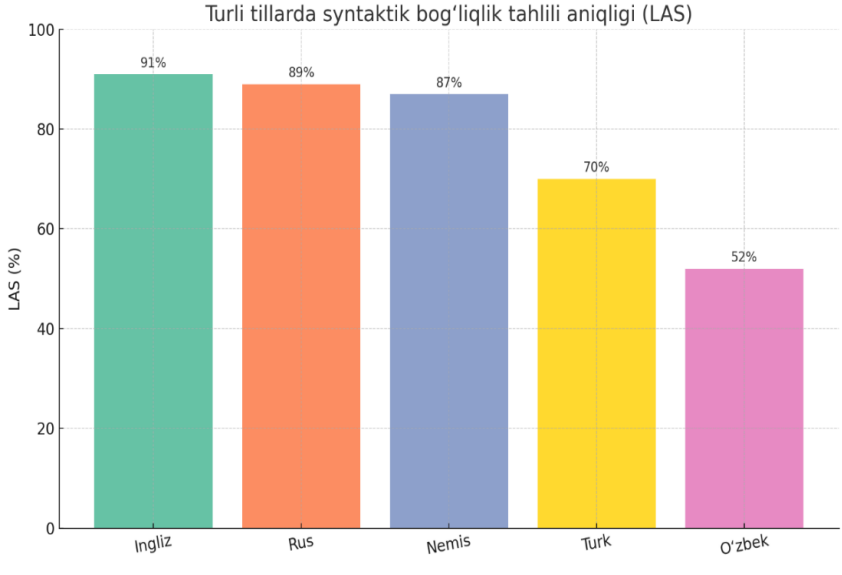


10-rasm. Turli tillarda so'z turkumlarini aniqlash (POS tagging) aniqligi.

Semantik resurslardan UzWordNet hozircha 28 ming synset bilan cheklangan, bu ingliz WordNetning 10% qismidir. Uning

aniqligi 76% atrofida baholanmoqda – demak, ayrim bog‘lanish va guruhlashlar to‘g‘ri emas. Bu tabiiy, chunki avtomatik tarjima va moslashtirish orqali tuzilgan. Ammo bu ham nihoyatda qimmatli ilk qadam: u orqali allaqachon ba’zi tadqiqotlarda foydalanish boshlandi (masalan, matndagi bir-biriga yaqin ma’noli so‘zlarni topishda). Kelgusida UzWordNet kengaytirilsa va to‘g‘rilansa, uning ko‘plab qo‘llanmalari paydo bo‘ladi – masalan, semantik izlash (foydalanuvchi so‘zining sinonimlarini ham izlash), avtomatik tarjima (so‘zning qaysi ma’noda ishlatilganini anglashi uchun), ilmiy bilimlar bazasi (ontologiyalar) qurish va hokazo.

Ingliz, rus va turk tillarida so‘z turkumlarini avtomatik aniqlash anchayin yuqori (96–97% atrofida) [Huang va boshq. 2015]. O‘zbek tilida esa avval qo‘llanilgan qoidaviy yondashuv mustaqil sinovlarda 82.2% aniqlikni ko‘rsatgan [Bobojonova va boshq. 2025]. Yangi o‘rgatilgan UzBERT asosidagi model esa 97.8%ga erishib, sezilarli yaxshilanishga olib keldi. Ushbu grafik o‘zbek tilida qo‘shimcha va kontekstni chuqur o‘rganuvchi modellar samaradorligini yaqqol ko‘rsatadi.



11-rasm. Xorijiy tillar va o‘zbek tilida syntaktik bog‘liqlik tahlili aniqligi (LAS).

Ingliz tilida katta treebanklar mavjudligi tufayli bog‘lanishli parsing 90% aniqlikka erishadi (StanfordNN parser, 2017). Rus tilida ham yirik SynTagRus korpusi asosida ~80% lar atrofida natija ko‘rsatilgan. Turk tili erkin so‘z tartibli bo‘lsa-da, uning uchun o‘rta hajmli treebank bor va 70% atrofida LAS kuzatilgan (CoNLL

2017 jamoaviy natijalari). O‘zbek tilida treebank juda kichik (500 gap) bo‘lgani bois hozircha 52% LAS, ya’ni juda past aniqlik kuzatilmoqda. Ushbu raqamlar resurs hajmi va sintaktik murakkablik parsing natijasiga katta ta’sir qilishini ko‘rsatadi: **erkin so‘z tartibli va kam korpusli tillarda aniq parsing modelini yaratish** hozircha qiyin, lekin korpus oshishi bilan sezilarli yaxshilanish kutilmoqda.

Yuqoridagi tahliliy ma’lumotlar va grafiklardan ko‘rinib turibdiki, o‘zbek tili uchun kompyuter lingvistikasi sohasida ish boshlangani samara bera boshladi, biroq hali talay vazifalar turibdi. Kelgusida korpuslar kengayishi, modellar takomillashishi bilan bu raqamlar oshib boradi. Ayni paytda xorijiy tillar bilan taqqoslanganda ham yutuqlar, ham orqada qolishlarimiz aniqlandi – bu haqida so‘nggi qismda batafsilroq to‘xtalamiz.

Xorijiy tillar bilan taqqoslash: ingliz, turk va rus tillari

O‘zbek tili uchun NLP yondashuvlarini baholashda uni **ingliz, turk** va **rus** tillari misolida qiyoslash foydali bo‘ladi. Ingliz tili – analitik til, sodda morfologiyaga ega va juda ko‘p resurslar jamlangan; turk tili – o‘zbek tiliga juda yaqin agglyutinativ til hisoblanadi, erkin so‘z tartibi unsurlariga ega va oxirgi yillarda resurslari oshgan; rus tili – flektiv til, o‘rta darajada boy morfologiyaga ega, so‘z tartibi nisbatan erkinroq va ayni paytda katta treebank va korpus resurslari mavjud.

Morfologik tahlil va teglash: Ingliz tilida morfologik analizda grammatik shakllari, asosan, fe’llarning zamon yoki otlarning ko‘plik shaklini aniqlash bilan cheklansa, turk va o‘zbek tillarida har bir so‘z ketma-ket qo‘shimchalar zanjiri bilan murakkablashishi mumkin. Shu sabab, ingliz tilida qoidaviy morfologik tahlil hatto sodda regex qoidalar bilan hal bo‘ladi, yoki umumiy lug‘at bilan (masalan, WordNet lemmas). Turk tilida esa Oflazer (1994) yaratgan FST analizator standarti hali ham asosiy vosita bo‘lib kelmoqda – u 98-99% kiritmalarni to‘g‘ri qamrab oladi deyiladi. O‘zbek tilida shunday FST yaqinda paydo bo‘ldi va qamrovi hali to‘liq baholanmagan, lekin nazariy jihatdan turkchaga yaqin bo‘lishi kerak. Rus tilida morfologik tahlil uchun, masalan, **pymorph2** kabi qoidaviy+statistik analizatorlar mavjud bo‘lib, so‘zni lug‘atdan topib, mos paradigmani beradi; uning aniqligi 95% dan yuqori. Shuningdek, rus tilida Neural morphological tagger (UDPipe modeli) 98% aniqlikni ko‘rsatgan (POS+feats). Demak, **resurs ko‘pligi va tilning soddaligi bo‘yicha ingliz > rus > turk > o‘zbek** tarzida morfologik

tahlil murakkablashadi. Biroq turk va o‘zbek tillari orasida tajriba almashish qulay – masalan, turk tilida CRF asosida morfologik disambiguation 96% bo‘lsa, o‘zbek tilida ham shunga erishish mumkin deb taxmin qilinadi [Özer, Korkmaz 2022].

So‘z turkumlari (POS) va morfologik teglash: Ingliz tilida POS tagging uchun millionlab tokenli Penn Treebank mavjud, **state-of-the-art modellar** 97.5% aniqlikka erishadi [Huang va boshq. 2015]. Hatto oddiy HMM modellar ham 95% dan oshadi. Rus tilida ham SynTagRus va boshqa korpuslar tufayli POS tagging 96-97% ga yetgan (StanfordNLPCore rus modeli 96% dan yuqori [https://github.com/stanfordnlp/CoreNLP/issues/480]. Turk tilida katta **Bosphorus treebank** va **IMST treebank** bor, ammo murakkab morfologiya tufayli POS emas, to‘liq morfologik teglar majmuasini aniqlash ko‘proq e’tiborli – unda so‘nggi natijalar 93-94% lar atrofida (masalan, transformer model bilan 96%). O‘zbek tilida hozircha POS uchun 500 gap teglandi, natijada birinchi neural tagger 97.8% ga yetdi, lekin to‘liq morfologik teglar (kelishik, son, egalik va b.) aniqligi hali pastroq (80%lar atrofida bo‘lishi mumkin, taxminan). Bu yerda ko‘rinadiki, **korpus hajmi hal qiluvchi omil**: ingliz/rusda yuz minglab so‘zlar, turkda o‘n minglab, o‘zbekda esa bir necha ming token bilan cheklanmoqda. Yaqin yillarda o‘zbek POS korpuslari 100k token chegarasidan oshsa, teggerlar aniqligi ham 95% ga yaqinlashishi mumkin.

Sintaktik parsing: Ingliz tilida hozirda constituency parsing ham 95% F1 (PTB, 2021), dependency parsing LAS 92% (stanfordnlp 2018). Rus tilida UD treebank (SynTagRus 50k gap) asosida LAS 88-90% ga chiqqan (neyron model, 2020). Turk tilida bir nechta treebank bor (10k gap), unda LAS 75-80% eng yuqori (2018-yili bilvosita o‘lchovlar). Bundan tashqari, **turk tilida morfologik bo‘linmalarni** so‘z o‘rniga asosiy birlik qilib parse qilinsa, aniqlik oshishi ko‘rsatilgan. O‘zbek tilida treebank endi ishlab chiqilyapti, hali ko‘p experiment amalga oshirilgani yo‘q, ammo yuqorida keltirilgan 52% – bu minimal model natijasi. Xullas, parsingda ham **ingliz va rus tillari** oldinda, sababi:

1) Korpus hajmi kattaligi;
2) So‘z tartibi qat’iyligi (inglizda);
3) Ko‘p tadqiqot tajriba yillari.

Turk va o‘zbek tillari orqada:

1) Korpus hajmi kichik;
2) Erkin so‘z tartibi va murakkab morfologiya modelni murakkablashtiradi.

Lekin e'tiborli jihati – turk tilidagi yondashuvlar aynan o'zbek tiliga ko'chirib sinab ko'rishga arziydi, chunki struktur va tipologik jihatdan juda yaqin. Masalan, turk treebankiga tayyorlangan parserni o'zbek matnida qo'llab, keyin o'zbekcha ohangga moslab fine-tune qilish mumkin (transfer learning). Xuddi shunday, turk tilidagi WordNet (BalkaNet) asosida o'zbek WordNet tuzish tajribasi bo'ldi.

Semantik tahlil va kontekstual modellar: Ingliz tilida BERT va undan keyingi GPT-3 kabi modellar matnni deyarli tushunish darajasida qayta ishlamoqda (masalan, suxandon-tarjima, dialog yarata olish). Rus tilida ham Sberbank AI kabi kompaniyalar RuBERT, RuGPT3 modellarini chiqardi. Turk tilida Germaniyada DBMDZ guruhi BERTurkni chiqargan (2019) va Turkiyada Istanbul Univ. BERTurkce kabi modellar paydo bo'lgan. Ular turk tilidagi ko'plab vazifalarda >90% aniqliklarni ta'minlamoqda (masalan, savol-javob tizimida EM/F1 = 79/87%, inglizcha ekvivalenti 85/92%). O'zbek tilida mBERT bilan cheklangan edik, endi UzBERT modellar chiqdi – lekin ularning parametrlari kichikroq va o'rgatish korpusi cheklangan. Shunga qaramay, BBPOS tadqiqoti ko'rsatdi: monolingual UzBERT mBERTdan 3-4% yuqori aniqlik berdi. Bu yaxshi dalil. Demak, agar resurs yaratilsa, o'zbek tilida ham mustaqil ChatGPT yoki BERT kabi model bo'lishi mumkin. Faqat katta hajmli tekst korpusi, hisoblash resurslari va vaqt kerak.

WordNet va lug'atlar: Ingliz WordNet 117k synset, RusWordNet 90k synset, TurkWordNet 15k synset (cheklangan); UzWordNet 28k synset. Ko'rinadiki, rus tili WordNetini Yandex va boshqalar ancha rivojlantirgan. TurkWordNet unchalik emas, ammo turk tilida katta dasturlar bor. O'zbek tilida esa ilk qadam boshqalardan yomon emas – 28k synset ozmuncha emas (eslatma: Uzbek lug'atlarda 60-80 ming atrofida so'z bor deb hisoblasak). Lekin sifat masalasi mavjud, ya'ni rus WordNeti filologlar tomonidan tekshirilgan.

Umumiy taqqos: Ingliz tili **resurs va yondashuvlar xilma-xilligi** bo'yicha yetakchi – har bir muammo uchun o'nlab modellar va tadqiqotlar mavjud. Shuning uchun ham ingliz tilida hozir fundamental yangi yechimlar kamroq – asosan modellarning sifatini oshirish, arxitektura yangiliklari bo'lyapti. Turk va rus tillari o'rta holatda – ba'zi sohalarda to'liq yechimlar bor (rus tilida mashinaviy tarjima, nutqni tanish kuchli; turk tilida tovushli assistentlar chiqmoqda). O'zbek tili hozircha resurslar yetishmovchiligi tufayli ortda. Biroq, yaqin qardosh tillar tajribasi va ko'p tilli modellar

bizga yordam berishi mumkin.

Xulosa sifatida, xorijiy tillar bilan taqqoslash shuni koʻrsatadiki:

1. Metodlar va yondashuvlar universalligi: hamma tillarda qoidaviy, statistik va neyron metodlar qoʻllaniladi. Faqat til xususiyatiga qarab implementatsiya va natijalar farq qiladi. Masalan, ingliz tilida morfologiya sodda – HMM ham yetadi; turk/oʻzbekda – FST kerak boʻladi. Lekin chuqur model hamma joyda yuqori natija bermoqda, faqat ma'lumot hajmi bilan cheklanadi.

2. Model sifati va resurs bogʻliqligi: korpus va lugʻatlar qancha katta va puxta boʻlsa, model shuncha yaxshi ishlaydi. Bu borada oʻzbek tiliga koʻproq e'tibor va sarmoya kerak. Xususan, millionlab tokenli matn korpusi, ona tilida belgilangan treebank (masalan, 10k gap), WordNet/FrameNet toʻliq shakllantirish – bular ustuvor vazifalar.

3. Turkiy tillar orasida hamkorlik: turk, oʻzbek, qozoq, uygʻur tillari oʻxshash tuzilishga ega. Agar shu tillar lingvistlari resurslarni baham koʻrib, birgalikda modellar yaratsa, hammaga foyda boʻladi. Masalan, turk tilidagi BERTni koʻpqirrali qilib, unga ozroq oʻzbek va qozoq matnlarini ham qoʻshib oʻrgatsa, mintaqaviy koʻp tilli model paydo boʻladi. Bu kelajak uchun istiqbolli yoʻnalish.

4. Kirill-Lotin muammosi ham hududiy xususiyat: rus tilida bunday masala yoʻq (faqat kirill). Turk va ingliz tilida lotin. Oʻzbek tilida ikki xil, qozoq tilida esa uch xil (!) (kirill, lotin, arab). Demak, biz modellarni shunga moslashimiz kerak. Transliteratsiya vositalari mavjudligi yaxshi natija beradi. Mansurov (2021) lotin-kirill oʻzbek teksti uchun aralash pretraining oʻtkazgan.

Xulosa

Amalga oshirilgan tajribalar shuni koʻrsatadiki, til texnologiyalarini rivojlantirishda dastlab katta mehnat va resurs talab etiladi, lekin bir marta oʻzlashtirilgach, koʻp natijalarni qisqa vaqt ichida qoʻlga kiritish mumkin. Turli tillarda ishlab chiqilgan metodlardan foydalanish til xususiyatiga koʻra natijalari farq qiladi.

Oʻzbek tili bilan ingliz, rus va turk tillari qiyoslanganda, POS teglash aniqlik darajasi gaplarda soʻz tartibining qat'iyligi, katta hajmli korpuslarning mavjudligi, uzoq yillik tajribalar, sun'iy intellektga asoslangan modellarning (masalan, UDPipe, RuBERT, RuGPT3 kabi modellar) ishlab chiqilganligi, koʻp tarmoqli tahlillangan lugʻatlar, WordNetlarning mavjudligiga koʻra foizlarda farqli tahlil natijalarini beradi. Oʻzbek tili borasidagi bugungi tadqiqotlar

bosqichma-bosqich chuqur gibrid modellarni ishlab chiqish yo‘lidan bormoqda.

O‘zbek tilida neyron yondashuvlar an’anaviy usullarga nisbatan sezilarli ustunlikka ega bo‘lishini tahlil natijalarida kuzatish mumkin. POS teglashda BERT modeli orqali 97.8% aniqlikka erishib, qoidaviy dasturdan ancha yuqori natija berganini tadqiqot natijasida oldik. O‘zbek tili o‘ziga xos tarixiy taraqqiyotga ega, murakkab tuzilishli va boy tillardan biri hisoblanadi. O‘zbek tilida mavjud kirill va lotin alifbosiga asoslangan o‘zbek tilida mavjud ma’lumotlar bazasini korpusga jamlash, to‘plangan matnlar ustida dastlabki qayta ishlash bosqichlarini bajarish bilan NLPdagi masalalarning yechilishiga xizmat qiladi. Yaqin kelajakda ingliz va rus tillari qatori, o‘zbek tilida ham matnlarni erkin tushunuvchi, analiz va sintez qila oluvchi sun’iy intellekt tizimlarini ko‘rishimizga umid qilsa bo‘ladi. Biz tahlil qilgan morfologik, sintaktik va semantik metodlar esa ana shu tizimlarning poydevoridir.

Adabiyotlar

Abdurakhmonova, N. 2019. “Dependency parsing based on Uzbek Corpus”. *In Proceedings of the International Conference on Language Technologies for All* (LT4All).

Adalı, E. 2020. *Türkçe Doğal Dil İşleme.* Akçağ Yayınları.

Agostini, A., Usmanov, T., Khamdamov, U., Abdurakhmonova, N., & Mamasaidov, M. 2021, January. Uzwordnet: “A lexical-semantic database for the uzbek language”. *In Proceedings of the 11th Global Wordnet conference,* 8-19.

Arofat Akhundjanova, Luigi Talamo. 2025. “Universal Dependencies Treebank for Uzbek”. *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains* (RESOURCEFUL 2025), 1–6.

Awasthi, P., Rao, D., & Ravindran, B. 2006. “Part of speech tagging and chunking with hmm and crf”. *Proceedings of NLP Association of India (NLPAI) Machine Learning Contest* 2006.

Bobojonova, L., Akhundjanova, A., Ostheimer, P., & Fellenz, S. (2025). *BBPOS: BERT-based Part-of-Speech Tagging for Uzbek. arXiv preprint* arXiv:2501.10107.

Boltayeviç, E. B., Yuldashevna, X. Z., Mamurjonovna, U. S., Ermamatovich, N. S., Kızı, A. Ş. A., & Shavkatjon, M. 2024. “Algorithms for Parsing Roots and Stems of Words in Uzbek Language”. *In 2024 9th International Conference on Computer Science and Engineering* (UBMK), 126-130. IEEE.

Boltayevich, E. B., Adalı, E., Mirdjonovna, K. S., Xolmo’Minovna, A. O., Yuldashevna, X. Z., & Uktamboy O’g’li, X. N. 2023. “The Prob-

lem of Pos Tagging and Stemming for Agglutinative Languages (Turkish, Uyghur, Uzbek Languages)". *In 2023 8th International Conference on Computer Science and Engineering (UBMK),* 57-62. IEEE.

Boltayevich, E. B., Samariddinovich, S. S., Mirdjonovna, K. S., Adalı, E., & Yuldashevna, X. Z. 2023. "POS taging of Uzbek text using hidden markov model". *In 2023 8th International Conference on Computer Science and Engineering (UBMK*), 63-68. IEEE.

Chung, Y. A., Zhang, Y., Han, W., Chiu, C. C., Qin, J., Pang, R., & Wu, Y. 2021. "W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training". *In 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU),* 244-250. IEEE.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. 2019. "Bert: Pre-training of deep bidirectional transformers for language understanding". *In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers),* 4171-4186.

https://github.com/apertium/apertium-uzb

https://github.com/stanfordnlp/CoreNLP/issues/480

Huang, Z., Xu, W., & Yu, K. 2015. *Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv*:1508.01991.

Huang, Z., Xu, W., & Yu, K. 2015. *Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:*1508.01991.

Kahn, J. G., Lease, M., Charniak, E., Johnson, M., & Ostendorf, M. 2005. "Effective use of prosody in parsing conversational speech". *In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 233-240.

Kann, K., & Schütze, H. 2016. *Single-model encoder-decoder with explicit morphological representation for reinflection.* arXiv preprint arXiv:1606.00589.

Kupiec, J. 1992. "Robust part-of-speech tagging using a hidden Markov model". *Computer speech & language,* 6(3): 225-242.

Madatov, K., Bekchanov, S., & Vičič, J. 2023. *Uzbek text summarization based on TF-IDF*. arXiv preprint arXiv:2303.00461.

Mansurov, B., & Mansurov, A. 2021. *UzBERT: pretraining a BERT model for Uzbek.* arXiv preprint arXiv:2108.09814.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. 1993. "Building a large annotated corpus of English: The Penn Treebank". *Computational linguistics,* 19(2): 313-330.

Matlatipov, S. G., Rajabov, J., Kuriyozov, E., & Aripov, M. 2024. "UzABSA: Aspect-Based Sentiment Analysis for the Uzbek Language".

In Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages@ LREC-COLING 2024, 394-403.

Meftah, S., Semmar, N., & Sadat, F. 2018. "A neural network model for part-of-speech tagging of social media texts". *In LREC 2018-Eleventh International Conference on Language Resources and Evaluation.*

Mirdjanovna, K. S. 2021. "Finite State Machine Model for Uzbek Language Morphological Analyzer". *In 2021 6th International Conference on Computer Science and Engineering (UBMK),* 395-400. IEEE.

Murat, A., & Ali, S. 2024. *Low-resource POS tagging with deep affix representation and multi-head attention*. IEEE Access.

Özer, H., & Korkmaz, E. E. 2022. "Transmorph: a transformer based morphological disambiguator for Turkish". Turkish Journal of Electrical Engineering and Computer Sciences, 30(5): 1897-1913.

Senuma, H., & Aizawa, A. 2017. "Seq2seq for morphological reinflection: When deep learning fails". *In Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection,* 100-109.

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

Shi, W., & Demberg, V. 2019. "Next sentence prediction helps implicit discourse relation classification within and across domains". *In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP),* 5790-5796.

Morphological, Syntactic, and Semantic Analysis Methods in Natural Language Processing in Uzbek

Botir Elov[1],
Oqila Abdullayeva[2],
Mastura Primova[3]

Abstract

This article discusses the methods of morphological, syntactic, and semantic analysis used in natural language processing for the Uzbek language. The linguistic features of Uzbek – complex agglutinative morphology, free word order, and limited resources – necessitate a specialized approach and research in applying these methods [Senuma, Aizawa 2017, 100-109]. Within the framework of this study, morphological analysis methods, followed by syntactic and semantic analysis methods, were examined based on scientific sources. Each section presents the existing advantages and disadvantages, experiences in applying these methods to the Uzbek language, as well as comparative analyses with foreign languages. For morphological analysis in Uzbek, rule-based methods, statistical models (HMM, CRF, etc.), and neural network-based approaches (BiLSTM-CRF, seq2seq) are discussed, with results provided in examples and percentages. It is demonstrated that syntactic parsing is carried out using dependency and constituency parsing methods. The issue of constructing a UD treebank for the Uzbek language, which follows the SOV word order, has been examined. The impact of complex morphological structure and free word order in sentences on parser construction is highlighted. As a result of the studied approaches, the issue of building hybrid

[1]*Elov Botir Boltayevich* – doctor of philosophy of technical sciences (PhD), associate professor. Tashkent State University of Uzbek Language and Literature named after Alisher Navo'i.
E-mail: elov@navoiy-uni.uz
ORCID: 0000-0001-5032-6648
[2] *Abdullayeva Oqila Xolmo'minovna* – PhD, post-doctorate student. Alisher Navo'i Tashkent State University of Uzbek Language and Literature.
E-mail: abdullayeva.oqila@navoiy-uni.uz
ORCID: 0000-0002-2524-4832
[3]*Primova Mastura Hakim qizi* – Teacher of Tashkent State University of Uzbek Language and Literature named after Alisher Navo'i.
E-mail: primovamastura@navoiy-uni.uz
ORCID: 0000-0002-0241-4659
For citation: Elov, B., Abdullayeva, O., Primova, M. 2025. "Morphological, Syntactic, and Semantic Analysis Methods in Natural Language Processing in Uzbek". *Uzbekistan: Lamguage and Culture* 1: 6 – 48.

parsers, integrating them with morphological analysis, and feeding grammatical categories of words into the parser has been raised. Additionally, the development of neural constituency parsers based on neural networks and the effectiveness of their results were analyzed. For the subsequent analysis stage related to NLP, particularly in semantic and sentiment analysis, models ranging from Word2Vec and FastText, which represent word meanings in vector form, to context-adapted transformer models such as BERT, mBERT, and UzBERT were also discussed. The text also considers issues of high-level semantic role labeling and the creation of semantic networks such as WordNet and FrameNet. In evaluating NLP approaches for the Uzbek language, Uzbek is compared with other languages: English, Turkish, and Russian. The application of rule-based, statistical, and neural methods in these languages is examined, and the results of the analysis are presented.

Key words: *POS tagging, parser, syntactic parsing, semantic parsing, sentiment analysis, natural language processing.*

References

Abdurakhmonova, N. 2019. "Dependency parsing based on Uzbek Corpus". *In Proceedings of the International Conference on Language Technologies for All* (LT4All).

Adalı, E. 2020. *Türkçe Doğal Dil İşleme.* Akçağ Yayınları.

Agostini, A., Usmanov, T., Khamdamov, U., Abdurakhmonova, N., & Mamasaidov, M. 2021, January. Uzwordnet: "A lexical-semantic database for the uzbek language". *In Proceedings of the 11th Global Wordnet conference,* 8-19.

Arofat Akhundjanova, Luigi Talamo. 2025. "Universal Dependencies Treebank for Uzbek". *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains* (RESOURCEFUL 2025), 1–6.

Awasthi, P., Rao, D., & Ravindran, B. 2006. "Part of speech tagging and chunking with hmm and crf". *Proceedings of NLP Association of India (NLPAI) Machine Learning Contest* 2006.

Bobojonova, L., Akhundjanova, A., Ostheimer, P., & Fellenz, S. (2025). *BB-POS: BERT-based Part-of-Speech Tagging for Uzbek. arXiv preprint* arXiv:2501.10107.

Boltayeviç, E. B., Yuldashevna, X. Z., Mamurjonovna, U. S., Ermamatovich, N. S., Kızı, A. Ş. A., & Shavkatjon, M. 2024. "Algorithms for Parsing Roots and Stems of Words in Uzbek Language". *In 2024 9th International Conference on Computer Science and Engineering* (UBMK), 126-130. IEEE.

Boltayevich, E. B., Adalı, E., Mirdjonovna, K. S., Xolmo'Minovna, A. O., Yuldashevna, X. Z., & Uktamboy O'g'li, X. N. 2023. "The Problem of Pos Tagging and Stemming for Agglutinative Languages

(Turkish, Uyghur, Uzbek Languages)". *In 2023 8th International Conference on Computer Science and Engineering (UBMK),* 57-62. IEEE.

Boltayevich, E. B., Samariddinovich, S. S., Mirdjonovna, K. S., Adalı, E., & Yuldashevna, X. Z. 2023. "POS taging of Uzbek text using hidden markov model". *In 2023 8th International Conference on Computer Science and Engineering (UBMK*), 63-68. IEEE.

Chung, Y. A., Zhang, Y., Han, W., Chiu, C. C., Qin, J., Pang, R., & Wu, Y. 2021. "W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training". *In 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU),* 244-250. IEEE.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. 2019. "Bert: Pre-training of deep bidirectional transformers for language understanding". *In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers),* 4171-4186.

https://github.com/apertium/apertium-uzb

https://github.com/stanfordnlp/CoreNLP/issues/480

Huang, Z., Xu, W., & Yu, K. 2015. *Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv*:1508.01991.

Huang, Z., Xu, W., & Yu, K. 2015. *Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:*1508.01991.

Kahn, J. G., Lease, M., Charniak, E., Johnson, M., & Ostendorf, M. 2005. "Effective use of prosody in parsing conversational speech". *In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 233-240.

Kann, K., & Schütze, H. 2016. *Single-model encoder-decoder with explicit morphological representation for reinflection.* arXiv preprint arXiv:1606.00589.

Kupiec, J. 1992. "Robust part-of-speech tagging using a hidden Markov model". *Computer speech & language,* 6(3): 225-242.

Madatov, K., Bekchanov, S., & Vičič, J. 2023. *Uzbek text summarization based on TF-IDF*. arXiv preprint arXiv:2303.00461.

Mansurov, B., & Mansurov, A. 2021. *UzBERT: pretraining a BERT model for Uzbek.* arXiv preprint arXiv:2108.09814.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. 1993. "Building a large annotated corpus of English: The Penn Treebank". *Computational linguistics,* 19(2): 313-330.

Matlatipov, S. G., Rajabov, J., Kuriyozov, E., & Aripov, M. 2024. "UzABSA: Aspect-Based Sentiment Analysis for the Uzbek Language". *In Proceedings of the 3rd Annual Meeting of the Special Interest*

Group on Under-resourced Languages@ LREC-COLING 2024, 394-403.

Meftah, S., Semmar, N., & Sadat, F. 2018. "A neural network model for part-of-speech tagging of social media texts". *In LREC 2018-Eleventh International Conference on Language Resources and Evaluation.*

Mirdjanovna, K. S. 2021. "Finite State Machine Model for Uzbek Language Morphological Analyzer". *In 2021 6th International Conference on Computer Science and Engineering (UBMK),* 395-400. IEEE.

Murat, A., & Ali, S. 2024. *Low-resource POS tagging with deep affix representation and multi-head attention*. IEEE Access.

Özer, H., & Korkmaz, E. E. 2022. "Transmorph: a transformer based morphological disambiguator for Turkish". Turkish Journal of Electrical Engineering and Computer Sciences, 30(5): 1897-1913.

Senuma, H., & Aizawa, A. 2017. "Seq2seq for morphological reinflection: When deep learning fails". *In Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection,* 100-109.

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

Shi, W., & Demberg, V. 2019. "Next sentence prediction helps implicit discourse relation classification within and across domains". *In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP),* 5790-5796.