

O'ZBEKISTON RESPUBLIKASI
OLIIY TA'LIM, FAN VA INNOVATSIYALAR VAZIRLIGI
FARG'ONA DAVLAT UNIVERSITETI

**FarDU.
ILMIY
XABARLAR-**

1995-yildan nashr etiladi
Yilda 6 marta chiqadi

**3-2025
FILOLOGIYA**

**НАУЧНЫЙ
ВЕСТНИК.
ФерГУ**

Издаётся с 1995 года
Выходит 6 раз в год

ADABIYOTSHUNOSLIK

Sh.Amil, A.G.Sabirdinov	
XX asr o'zbek lirikasi va Oybek she'riyati	6
N.P.Farmonov	
Realistik va psixologik tasvir: "Girdob" romani qahramonlari misolida	14
N.A.Abdullayeva	
Gender leksik birliklarining lingvomadaniy ko'rinishlari	18
A.Eshniyazova	
Muallif konsepsiyasining ilmiy-nazariy asoslari.....	22
Z.Q.Eshanova	
Istiqlol davri o'zbek she'riyatida xalq dostonlari sintezi	27
J.B.Sayidolimov, I.B.Xabibullayev	
"Qisasi rabg'uziy" ning struktur tuzilishi	33
I.B.Xabibullayev	
Badiiy adabiyotda oila mavzusining shakllanishi va rivoji	42
E.G'.Qurbonova	
Lutfiy lirikasida badiiy mazmun va irfoniy talqin uyg'unligi	47
U.S.Ochilov	
Stiven Kingning «It» va «Green mile» romanlarida magik realizmning badiiy talqini.....	52
G.M.Kirgizova	
Iqbol Mirzo she'rlarida band tizimlari	58
N.N.Sharipova	
Xamsanavislik an'anasi va unda Toshlijali Yahyobey "Gulshan ul-anvor" dostonining o'rni	63
S.A.Olimjonov	
"Uch og'ayni" romanida "yo'qotilgan avlod" tasviri	69
I.M.Jurayev, H.A.Jo'rayev	
"Temur tuzuklari" hamda "Boburnoma"da an'ana va vorisiylik.....	74

TILSHUNOSLIK

Sh.M.Iskandarova, K.T.Israilova	
Ijtimoiy tarmoqlar tilini lingvistik o'rganishning nazariy asoslari.....	79
Sh.Sh.Uralova	
Ingliz va o'zbek tillarida oshxona anjomlari nomlarining suffikslar yordamida hosil bo'lishi va ularning so'z yasash modellari.....	83
I.F.Porubay, F.R.Nishanova	
Zamonaviy global jamiyat sharoitlarida lingvomadaniy birliklarning rivojlanish omillarining tahlili...	90
M.S.Sayidazimova	
O'zbek va ingliz tillaridagi fe'lining bo'lishsizlik shaklini yasovchi ayrim vositalar	96
X.G'.Raximboyeva, M.G'.Matlatipova	
O'zbek tili iyerarxiya korpusini yaratishda ma'lumotlar bazasini yig'ish, gaplarni annotatsiyalash va tokenlash jarayoni.....	100
Л.Мамедзаде	
Приемы рифмообразования в паремнологических единицах турецкого и азербайджанского языков.....	107
N.A.To'xtasinova	
Pretsedent nom, jumla va matnlarning mediamatn birliklari sifatidagi tadqiqi	113
Z.I.Akbarova	
Alisher Navoiy asarlarida keltirilgan "Ahl" komponentli lug'aviy birliklarning nutq uslublari doirasida qo'llanilishi	118
Ё.М.Алтишева	
Из истории изучения религиозных терминов в отечественной терминологии	121



UO'K: 811.8

**O'ZBEK TILI IYERARXIK KORPUSINI YARATISHDA MA'LUMOTLAR BAZASINI YIG'ISH,
GAPLARNI ANNOTATSIYALASH VA TOKENLASH JARAYONI****THE PROCESS OF COLLECTING A DATABASE, ANNOTATING SENTENCES, AND
TOKENIZING IN THE CREATION OF A DEPENDENCY PARSING OF THE UZBEK LANGUAGE****ПРОЦЕСС СБОРА БАЗЫ ДАННЫХ, АННОТИРОВАНИЯ ПРЕДЛОЖЕНИЙ И
ТОКЕНИЗАЦИИ ПРИ СОЗДАНИИ СИНТАКСИЧЕСКОГО АНАЛИЗА ЗАВИСИМОСТЕЙ
УЗБЕКСКОГО ЯЗЫКА****Raximboyeva Xulkar G'ayratovna¹** ¹O'zbekiston Milliy universiteti tayanch doktoranti**Matlatipova Muslima G'ayrat qizi²** ²Urganch davlat universiteti talabasi**Annotatsiya**

Ushbu maqolada o'zbek tili uchun iyerarxik tahlil korpusini yaratish bosqichlari keng yoritilgan. Xalqaro tajribada biror tilning iyerarxik tahlil korpusini yaratish uchun qo'llaniladigan beshta asosiy soddalashtirilgan bosqich mavjud bo'lib, ular batafsil ko'rib chiqiladi. Ushbu bosqichlar quyidagilar: matnlarni tanlash; oldindan ishlov berish vositalari va manbalarini tanlash; annotatsiyalash; o'zbek tiliga xos qo'llanmani hujjatlashtirish va tilning universal bo'lmagan xususiyatlarini aks ettirish; transliteratsiya bosqichidir..

Abstract

This article provides an in-depth overview of the stages involved in building a syntactic treebank for the Uzbek language. It outlines five simplified but essential steps widely used in international practice when constructing a hierarchical corpus for any language. These steps include: text selection; pre-processing (including the selection of tools and resources); annotation; documentation of language-specific guidelines and addressing non-universal linguistic features; and finally, transliteration.

Аннотация

В данной статье представлен подробный обзор этапов создания дерева зависимостей (treebank) для узбекского языка. Рассматриваются пять упрощённых, но ключевых этапов, которые широко применяются в международной практике при построении иерархического корпуса для любого языка. Эти этапы включают: выбор текстов; предварительная обработка (включая выбор инструментов и ресурсов); аннотирование; документирование языковых особенностей и описание специфических, неуниверсальных черт языка; и, наконец, этап транслитерации.

Kalit so'zlar: Annotatsiya, bosqich, tokenlash, lemmalash, matn tanlovi, hujjatlashtirish, yo'riqnoma, natija, jarayon.

Key words: Annotation, steps, tokenization, lemmatization, text selection, documentation, guideline, result, process.

Ключевые слова: аннотация, этапы, токенизация, лемматизация, выбор текстов, документация, руководство, результат, процесс.

KIRISH

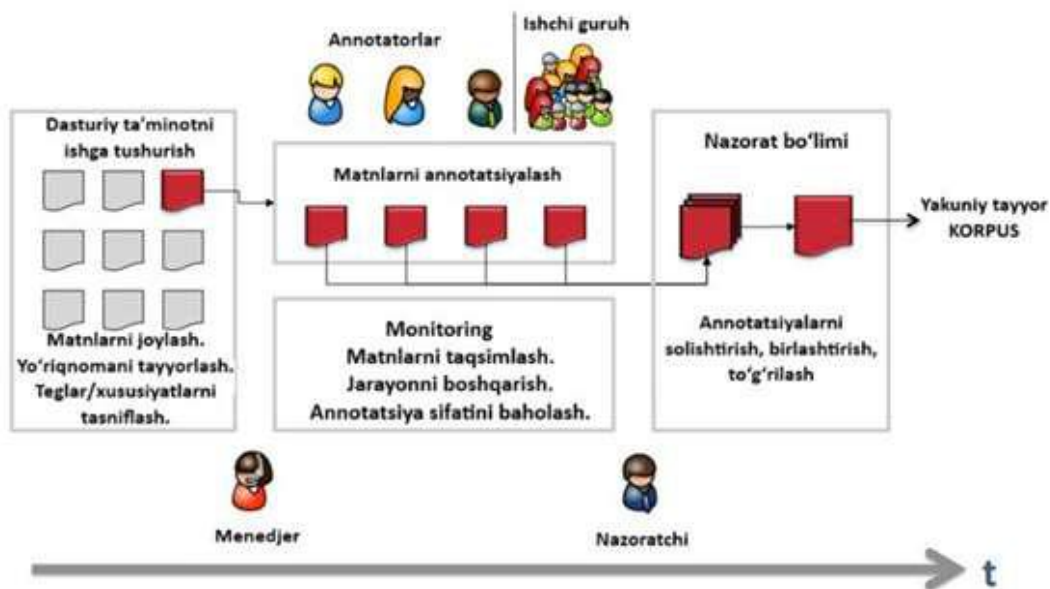
O'zbek tili uchun iyerarxik tahlil korpusi (dependency treebank)ni yaratish bosqichlari batafsil tavsiflanadi. Korpus uchun kerakli bo'lgan til materiallarini yig'ish, ularni maxsus vositalar yordamida annotatsiyalash (teglar bilan belgilash) va tokenlash (so'zlarni alohida tokenlarga ajratish) jarayonlari izchil bayon qilinadi. Korpus tuzish jarayonida ochiq manbadagi onlayn yangiliklar matnlari hamda mualliflik huquqi bo'yicha foydalanishga ruxsat etilgan badiiy adabiyotlardan olingan jumlarlar asos bo'ldi. Annotatsiyalash ishlari Germaniyaning Zaarland universitetida ishlab chiqilgan INCEption dasturiy platformasi yordamida amalga oshirildi. Jarayonga jami 5 nafar annotator (izohlovchi) jalb etildi, ularning har biri o'zbek tilining sintaksis va

TILSHUNOSLIK

morfologiyasini yaxshi biluvchi mutaxassislardir. Annotatorlar tomonidan matnning turli qismlarini belgilash ishlari amalga oshirildi. Korpus yaratilishida har bir qadamda sifatni ta'minlash uchun bir necha bosqichli nazorat va tahrir jarayonlari tashkil etildi.

ADABIYOTLAR TAHLILI VA METODOLOGIYA

Ma'lumotlarni yig'ish va tanlash. Dastlab, korpus uchun manbalar tanlab olindi. Ikki turdagi manbadan foydalanilgan: birinchisi, internetdagi ochiq axborot manbalaridan olingan yangiliklar maqolalari; ikkinchisi, jamoatchilik uchun ochiq e'lon qilingan yoki ruxsat olingan XX-XXI asrga oid o'zbek badiiy adabiyotlaridan olingan matnlar. Yangiliklar qismi asosan UzCrawl datasetidan olingan bo'lib (M. Mamasaidov va A. Shopulatov, 2023 yilgi loyiha) – ushbu dataset zamonaviy o'zbek tilidagi turli mavzulardagi yangilik saytlaridagi (masalan, Kun.uz, Daryo.uz kabi) maqolalarni jamlagan [7]. Badiiy adabiyotlar qismi esa internetda ochiq mavjud bo'lgan o'zbek adabiy asarlaridan tanlab olingan. Tanlash jarayonida materiallarning mavzuyi qamrovi keng bo'lishi, til uslubi va murakkabligi jihatdan xilma-xillik bo'lishiga e'tibor berildi. Shu tariqa, *jami 1200 ta jumla* tanlab olindi — shundan 750 tasi yangilik matnlaridan, 500 tasi badiiy asarlardan. Tanlangan jumlar korpusga kiritilishdan avval qo'lda tahrir qilinib, ortiqcha yoki takror bo'lib qolgan qismlar chiqarib tashlandi, imlo va tinish belgilari me'yorlari tekshirildi. Natijada, korpusga kiritilayotgan har bir jumla mustaqil va to'liq mazmunli bo'lishi ta'minlandi. Ushbu tanlama korpus kelgusida ham kengaytirilishi mumkin bo'lsa-da, hozircha u mazkur tadqiqot doirasida o'zbek tili sintaksisini qamrab oluvchi yetarlicha reprezentativ kichik hajmli korpus vazifasini o'tashga mo'ljallandi.



1-rasm. Korpusni izohlash (annotatsiya qilish) va yaratish jarayoni.

Annotatsiya vositasi va jamoaviy ishlash. Tanlab olingan matnlar INCEpTION dasturiga yuklanib, har bir jumlaning teglar bilan belgilash (annotatsiya qilish) jarayoni bajarildi. INCEpTION — bu matnni annotatsiya qilish uchun mo'ljallangan ochiq platforma bo'lib, unda turli xil annotatsiya qatlamlarini (masalan, morfologik teglar, sintaktik bog'lanishlar va h.k.) ishlab chiquvchi annotatorlar jamoaviy ishlashi mumkin [5]. Platforma bir vaqtda bir nechta annotatorning ishlashini qo'llab-quvvatlaydi, o'zgarishlarni kuzatish imkonini beradi. Bizning tadqiqot uchun INCEpTION'da maxsus Universal Dependencies (UD) sxemasiga mos keluvchi annotatsiya loyihasi yaratildi. Unda har bir so'z uchun tegishli *morfologik xususiyatlar* (masalan, asosiy so'z turkumi (UPOS), grammatik kategoriyalar, lemma) va sintaktik bog'lanish ma'lumotlarini kiritish uchun maydonlar belgilandi. Dastur interfeysida jumla so'zlari ro'yxat shaklida va daraxt ko'rinishida chiqib, annotatorlar har bir so'zning xususiyatlarini tanlashlari va uning gap ichida qaysi boshqa so'zga bog'lanishini ko'rsatishlari (bog'lovchi o'q orqali) lozim bo'ldi. Annotatorlar bir-birlarining ishi ustidan ham kuzatib, izohlar qoldirish orqali bahsli holatlarda umumiy muhokama orqali to'g'ri kelishuvga erishildi. Ayniqsa, murakkab sintaktik konstruksiyalar yoki noaniq ma'noli birikmalar uchraganda,

jamo'a a'zolari lingvistik adabiyotlarga tayangan holda birgalikda qaror qabul qildilar. Yakunda, har bir jumla bo'yicha annotatsiyalar yetakchi lingvist-annotator (korpus muallifi) tomonidan to'liq ko'rib chiqildi – bu bosqichda barcha xatolar tuzatildi, bir xil turdagi hodisalarning izchil belgilangani tekshirildi va korpus umumiy yagona uslubda annotatsiya qilinganiga ishonch hosil qilindi. Shunday qilib, jamoaviy mehnat va yakuniy tahrirlash natijasida korpusning annotatsiya sifati yuqori darajaga ko'tarildi.

Tokenlash va dastlabki avtomatik qayta ishlash. So'zlar odatda bo'sh joy bilan ajratiladi, lekin bunda quyidagi mustasnolar mavjud:

- Chiziqcha bilan ajratilgan qo'shma so'zlar va to'liq takrorlashlar (masalan, kuta-kuta, bora-bora), agar har bir so'zni bir so'z sifatida qabul qilinsa, ko'p so'zli tokenlar yaxlit holda shakllanadigan alohida semantik ifodadir.

- Bog'lovchi "bo'lib", "holda" hol belgilari so'zga qo'shilganda ko'p so'zli leksema sifatida qabul qilinadi. Bo'shliqli so'zlar o'zbek tilida uchrab turadi.

Korpusda uchta maxsus dastlabki ishlov berish bosqichlari amalga oshirildi:

- metama'lumotlarga qo'shish - bu har bir namunaviy jumlagi kuzatish uchun identifikatorni, shuningdek, har bir jumla uchun ID berishni o'z ichiga oladi, korpus ma'lumotlarida metama'lumotlar (#) raqam belgisi bilan ifodalanadi;

- xatolarni o'chirish - bu matn terish xatolarini tuzatish, bo'sh joylarni tekshirish, yetishmayotgan tinish belgilarini qo'shish va jumalarni formatlashni o'z ichiga oladi;

- imloviy standartlashtirish – bu turli til darajalari (rasmiy, norasmiy, jargon, dialektlar va boshqalar) uchun imlo qoidalarini bir xil standartga moslashtirish jarayonidir, tilni tahlil qilishda izchillikka olib keladi, mashina tarjimasini va NLP tizimlari uchun aniqroq ma'lumot bazasi quriladi. Biroq imloni standartlashtirish qat'iy bo'lishi shart emas — asosan, o'zbek tili so'zlovchilari uchun tabiiy bo'lgan ifodani saqlab qolish va imloni ortiqcha murakkablashtirmaslik o'rtasidagi muvozanatni ta'minlash maqsadida, iyerarxik tuzilmadagi lemmalarning o'zgarishi orqali amalga oshirilgan.

Imloviy standartlashtirish asosiy jihatlari quyidagicha:

birlashtirilgan imlo qoidalarini – korpusdagi barcha matnlar (ilmiy, adabiy, matbuot, internet-muloqot) umumiy imlo normalariga moslashtiriladi;

darajaviy farqlar – har bir til darajasi (masalan, rasmiy yoki norasmiy matnlar) uchun mos imlo variantlari belgilanadi;

dialektal va jargon so'zlar – mahalliy yoki jargon so'zlar standart imlodagi qanday yozilishi kerakligi aniqlanadi;

elektron qidiruv uchun optimallashtirish – so'zlar variantlari (masalan, "qilmoq" / "qilish") bir xil yozuvga keltiriladi.

NATIJA VA MUHOKAMA

Negizlash (lemmalash) bosqichida har bir so'zning lug'aviy shakli (lemma), ya'ni lug'atlarda uchraydigan asosiy shakli aniqlanadi. O'zbek tilida negizlash - so'zni morfologik tahlil qilish orqali uning asosiga (lug'at shakliga / negiziga) qisqartirishni o'z ichiga oladi. O'zbek tili agglutinatif til bo'lgani uchun, lemma sifatida odatda hech qanday grammatik qo'shimcha olmaydigan sof lug'aviy shakl tanlanadi (masalan, "boryapti" so'zining lemma shakli "bor"). Har bir so'zni asosiy yoki lug'at shakligacha qisqartirish jarayoni (jumalarni negizlash / lemmalash) ko'plab tabiiy tillarni qayta ishlash vositalarida asosiy jarayondir. O'zbek tili agglutinatif tillar oilasiga mansub bo'lib, ko'pincha zamon, mayl, shaxs va holni bildirish uchun qo'shimchalardan foydalanadi. Shuning uchun o'zbek tilidagi negizlash qo'shimchalarni olib tashlashni talab qiladi. Matnli korpusni shakllantirish maqsadida negizlash (lemmalash) jarayonida quyidagicha misollarni uchratish mumkin:

Gap: Men kitob o'qiyapman va ukam musiqa tinglayapti. (Lemmalar: Men, kitob, o'qi, va, uka, musiqa, tingla);

Gap: U dars tayyorlayapti, chunki ertaga imtihon bor. (Lemmalar: U, dars, tayyorla, chunki, ertaga, imtihon, bor);

Gap: Agar ob-havo yaxshi bo'lsa, biz sayrga chiqamiz. (Lemmalar: Agar, ob-havo, yaxshi, bo'l, biz, sayr, chiq).

Tadqiqot ishida ishlab chiqilgan o'zbek tili iyerarxik korpusida jumalarni negizlash amalga oshirilgan quyidagi misollarni keltirish mumkin:

TILSHUNOSLIK

bu xalq , ayniqsa qishloq aholi turmush daraja oshir xizmat qil .
 Bu xalqimiz , ayniqsa qishloq aholisining turmush darajasini oshirishga xizmat qilmoqda .

qush pat tozalay boshla , kun issiq kel .
 Qushlar patlarini tozalay boshlasalar , kun issiq keladi ...

bu xalq , ayniqsa qishloq aholi turmush daraja oshir xizmat qil .
 Bu xalqimiz , ayniqsa qishloq aholisining turmush darajasini oshirishga xizmat qilmoqda .

bu bepoyon cho' llar o' z bag' rida yana qancha sirlarni yashir ekan ?
 Bu bepoyon cho'llar o'z bag'rida yana qancha sirlarni yashirgan ekan ?

« ijtimoiy fikr » markaz har bir soha bo' yicha asosiy vazifa belgila .
 « Ijtimoiy fikr » markazining har bir soha bo'yicha asosiy vazifalari belgilandi .

korxona jamoa yalpi mahsulot kam 80 foiz eksport chiqar mo'ljalla .
 Korxona jamoasi yalpi mahsulotning kamida 80 foizini eksportga chiqarishni mo'ljallamoqda .

ha , olam uzra sochilib ketgan aziz kishilarning sanog'ini Xudodan o'zga hech kim bilolmaydi .
 Ha , olam uzra sochilib ketgan aziz kishilarning sanog'ini Xudodan o'zga hech kim bilolmaydi .

2-rasm. O'zbek tili iyerarxik korpusida gapdagi so'zlarning negizlarini (lemmalarini) belgilash.

So'zlarning asos qismlari shu shaklda olindi va tizimda avtomatik tarzda lemmalash amalga oshirildi. Gapdagi har bir so'z, tinish belgilari alohida lemma qilib olindi.

Lemmalash qoidalarini qo'llash: korpusda so'zning asosiy shaklini topish uchun lemmalash funksiyasidan foydalaniladi, bunda kirish sifatida so'z va uning natijasi sifatida POS tegi chiqariladi.

Misol uchun, agar *Pythondan* foydalanilganda, NLTK-dagi WordNetLemmatizer funksiyasi so'zlarni POS teglari asosida lemmalashi mumkin:

```
from nltk.stem import WordNetLemmatizer
from nltk.corpus import wordnet
lemmatizer = WordNetLemmatizer()
def get_wordnet_pos(word):
    tag = nltk.pos_tag([word])[0][1][0].upper()
tag_dict = {"J": wordnet.ADJ, "N": wordnet.NOUN, "V": wordnet.VERB, "R": wordnet.ADV} return
tag_dict.get(tag, wordnet.NOUN)
lemmatizer.lemmatize("yugurish", get_wordnet_pos("yugur"))
```

Lemmalarni asl so'zlarga xaritalash. Har bir asl so'zni lemmasi bilan almashtiriladi. Misol uchun, "yugurish"ning lemmalashtirilgan shakli "yugur" bilan almashtiriladi.

Annotatsiya jarayonini samarali tashkil etish uchun, avvalo, barcha jumlar tokenlarga ajratildi (ya'ni har bir so'z alohida satrga chiqarildi). Bu bosqichda mavjud tayyor vositadan – NLTK tokenizatori yordamidan foydalanildi [6]. NLTK (Natural Language Toolkit) – Python dasturlash tilidagi tilni qayta ishlash kutubxonasi bo'lib, undagi standart tokenizator satrni bo'sh joy va tinish belgilari bo'yicha so'zlarga ajratadi. NLTK yordamida 1200 ta jumla 9800 ta tokenga ajratildi, ya'ni korpus umumiy hajmi 9800 so'zni tashkil etdi. Tokenlash jarayonida quyidagi qoidalar va maxsus holatlar inobatga olindi:

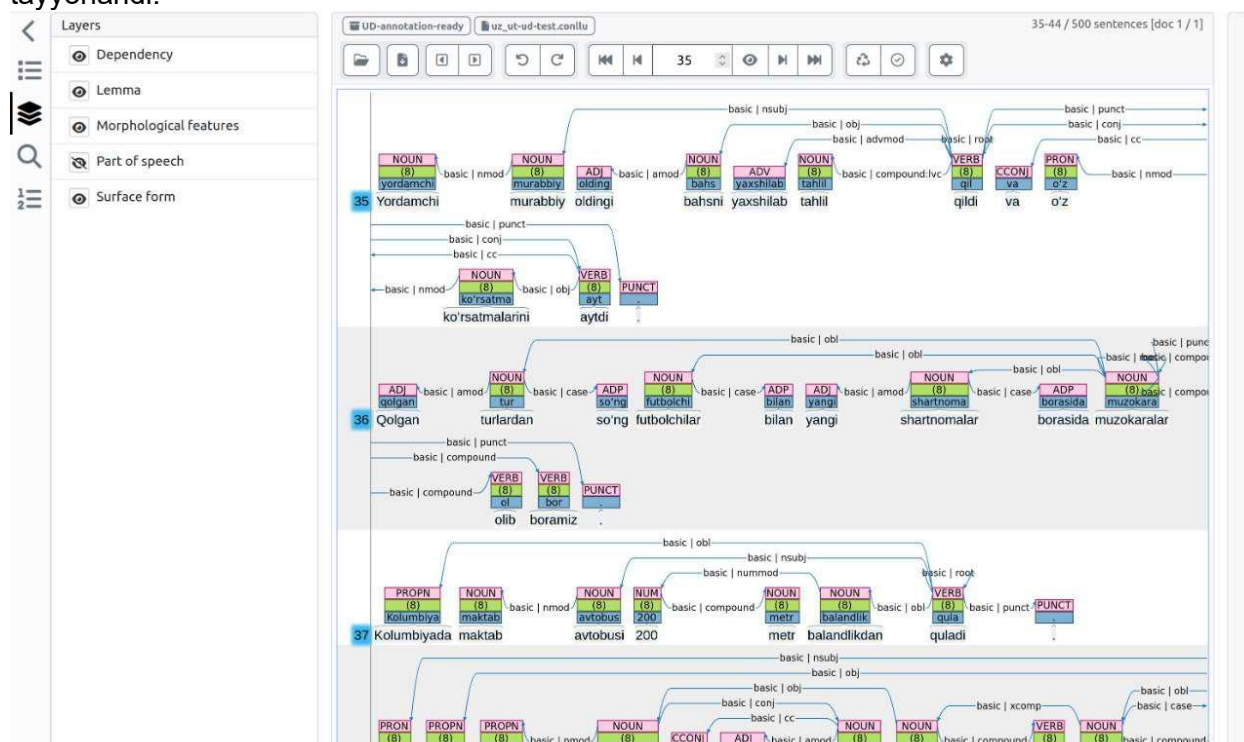
Oddiy holatda, gapdagi har bir so'z yoki birikma bo'shliq bilan ajralgan bo'lsa, alohida token sifatida olindi. Masalan, "Talaba kitob o'qidi" jumlasida uchta token – "Talaba", "kitob", "o'qidi" ajratiladi.

Tinish belgilaridan so'ng ham token chegarasi qo'yildi. Masalan, vergul, nuqta, so'roq-belgisi kabi tinish belgilari alohida token sifatida belgilandi. "Keldi, ko'rdi, g'alaba qozondi." kabi gapda vergul va nuqta ham alohida token bo'ladi.

Keltirib chiqarilmagan qo'shimchalar: Ba'zi tinish belgilari so'zga qo'shilib yozilgan bo'lsa, ularni ajratish kerak bo'ladi. Masalan, "kitobni." oxirida nuqta bor – tokenlash jarayonida "kitobni" va "." alohida tokenlar sifatida olinadi.

Qisqartmalar: matnda nuqta bilan tugaydigan qisqartma soʻzlar boʻlsa, ular ichidagi nuqta tokenlash vaqtida ajratib yuborilmasdan, soʻzning tarkibida saqlab qolindi. Chunki UD formatida nuqta abbreviatura tarkibida kelsa, u soʻzning bir qismi deb qaraladi. Masalan, “mln.” (million qisqartmasi) bir butun token sifatida olindi. Xuddi shuningdek, “A. Navoiy” kabi ism familiya qisqartma holatida – “A.” alohida token (nuqta bilan birga) va “Navoiy” alohida token tarzida koʻrib chiqildi.

Koʻp soʻzli birikmalar (multi-word expressions): UD qoidalariga koʻra, token ichida boʻshliq boʻlishiga yoʻl qoʻyilmaydi, yaʼni ikki yoki undan ortiq soʻzdan tashkil topgan birikma bitta token sifatida birlashtirilmaydi. Buning oʻrniga, agar mazmunan yaxlit boʻlgan birikma bir necha soʻzdan iborat boʻlsa, ular alohida tokenlar sifatida qoldiriladi va keyinchalik maxsus sintaktik bogʻlanish bilan oʻzaro bogʻlanadi. Masalan, “Togʻli Qorabogʻ” geonomik nomi ikki soʻzdan iborat boʻlsa ham, UD tizimida u *flat* munosabat bilan bogʻlangan ikkita alohida token sifatida beriladi. Yaʼni, “Togʻli” va “Qorabogʻ” soʻzlari har biri token, va annotatsiyada ular *flat* teg orqali birikkan bir nom sifatida tasvirlanadi. Bunday yondashuv, ayniqsa, ism-familiya, murakkab joy nomlari, birikma feʼllar va soʻz birikmalarini belgilashda qoʻllanildi. Natijada, tokenlash qoidalariga mos holda ishlab chiqilib, barcha jumlar bir xil mezonda soʻzlarga ajratildi va keyingi annotatsiya bosqichiga tayyorlandi.



3-rasm. INCEpTION dasturida Oʻzbeki tili iyerarxik korpusi uchun annotatsiya jarayoni bosqichlari va tokenlash qoidalarining qoʻllanilishi: chap tomonda matn boʻlaklarga (tokenlarga) ajratilib, har biriga ID raqami berilgan; oʻng tomonda esa, shu tokenlarga tegishli morfologik teglar va sintaktik bogʻlanishlar koʻrsatilgan (bogʻlovchi oʻqlar bilan).

Bosqichma-bosqich annotatsiyalash jarayoni. Korpus uchun tokenlar tayyor boʻlgach, har bir jumlaning toʻliq belgilash bir necha bosqichda amalga oshirildi.

Quyida annotatsiyalashning asosiy bosqichlari ketma-ketligi va har bosqichda bajarilgan ishlarning qisqacha bayonini keltiramiz:

Lemmalash (asosiy shaklni aniqlash): har bir token uchun uning lugʻaviy asosi (lemma) avtomatik tarzda aniqlab chiqildi. Buning uchun mavjud dasturiy vositalardan foydalanildi – xususan, oʻzbek tilining UzbekLemmatizer modelidan foydalanildi. Avtomatik lemmatizator yordamida har bir soʻzning boshlangʻich shakli (masalan, feʼlning infinitiv yoki otning birlik-bosh kelishikdagi shakli) belgilanib, annotatsiya tizimiga kiritildi. Avtomatik jarayon taxminan 85-90% aniqlik bilan lemmalarni toʻgʻri topib berdi, biroq ayrim hollarda xatolar kuzatildi – masalan, maʼnosi bir-biridan farq qiluvchi, lekin yozilishi bir xil boʻlgan soʻzlarda (omonimlarda) lemmalash jarayonida notoʻgʻri

TILSHUNOSLIK

tanlov qilishi kuzatildi. Shu sababli, keyingi bosqichlarda annotatorlar har bir soʻzning lemmalarini yana qoʻlda tekshirib chiqdilar va zarur joylarda toʻgʻrilashdi. Masalan, “och” soʻzi gapga qarab “ochmoq” (yopiq narsani ochish) yoki “och” (toʻqlikning antonimi) lemmasiga ega boʻlishi mumkin – bunday noaniq holatlarda tegishli lemmani tanlash annotatorning qaroriga havola qilindi.

Morfologik teglash (soʻz turkumlari va xususiyatlar): keyingi qadam – har bir soʻzga uning grammatik soʻz turkumi va batafsil morfologik xususiyatlarini teg tarzida belgilashdir. Bu jarayon qisman avtomatik, qisman qoʻlda bajarildi. Avvalo, maxsus UzMorphAnalyzer dasturi (Ulugʻbek Salaev, 2023) yordamida har bir soʻzning barcha mumkin boʻlgan grammatik tavsiflari ishlab chiqildi [4]. UzMorphAnalyzer oʻzbek tilidagi soʻzlarni tahlil qilib, ularning soʻz turkumini (masalan, ot, feʼl, sifat va h.k.) hamda shaxs-son, zamon, kelishik kabi grammatik kategoriyalarini aniqlaydi. Ushbu dastur natijalari asosida dastlabki Universal POS (UPOS) teglari avtomatik tarzda berildi: masalan, UzMorphAnalyzer natijasida “kitobni” soʻzi uchun NOUN (ot) deb belgilandi, “oʻqidi” soʻzi uchun VERB (feʼl) va hokazo. Biroq avtomatik belgilash toʻliq ishonchli chiqmagani bois, barcha UPOS teglar 5 nafar annotator tomonidan birma-bir tekshirilib, zarur tuzatishlar kiritildi. Shu jarayonda anʼanaviy oʻzbek tili grammatikasidagi toifalar bilan UD standartidagi 17 ta soʻz turkumi orasidagi mapping (mos kelish) jadvali ham tuzildi. Masalan, anʼanaviy grammatikada “ravishdosh” deb ataluvchi turkum UD tizimida VERB (feʼl) sifatida belgilandi va unga qoʻshimcha VerbForm=Conv xususiyati berildi.

TADQIQOT MUHOKAMASI.

Morfologik xususiyatlar (FEATS) uchun UD tomonidan belgilangan 42 xil universal xususiyatlardan mos keladiganlari tanlandi. Bular qatoriga Kelishik (Case), Son (Number), Egallik olmoshlari (Possessive), Daraja (Degree), Zamon (Tense), Holat (Mood), Shaxs (Person), Niqob (Polarity) va Daraja (Voice) kabi turli kategoriyalar kiradi. Har bir soʻzga uning gapdagi shakliga mos keluvchi xususiyatlar toʻplami birlashtirildi. Masalan, “kitobni” soʻziga Case=Acc (tushum kelishigi) va Number=Sing (birlik) xususiyatlari, “oʻqidi” feʼliga esa Tense=Past (oʻtgan zamon), Person=3, Number=Sing va Polarity=Pos kabi belgi qiymatlari qoʻyildi. Boshida 150 ta jumla shu tarzda toʻliq qoʻlda morfologik teglab chiqildi va Stanza (tabiiy til tahlili Python paketi) vositasi [3]. yordamida mashinani oʻqitish orqali qolgan jummalarga avtomatik teg qoʻyishga harakat qilindi. Stanford Stanza – bu turli tillar uchun modelni oʻqitib, analiz qiluvchi zamonaviy kutubxona (Qi va boshq., 2020) [2,15]. Biz 150 ta qoʻlda teglangan jummalarni Stanza modeliga oʻrgatib, boshqa jumlar uchun soʻz turkumlari va xususiyatlarini avtomatik topish amaliyotini ham bajardik. Natijada baʼzi oddiy kategoriyalar (masalan, Number yoki Case singari) deyarli xatosiz topildi, murakkabroq joylarda esa xatolar kuzatildi. Shuning uchun yakunda barcha 1200 ta jumlaning morfologik teglari yana qoʻlda koʻrib chiqildi va avtomatik bosqichda yuzaga kelgan xatolar tuzatildi. Bu bosqich yakunida korpusning har bir soʻzi uchun toʻgʻri LEMMA, UPOS va FEATS toʻplami aniqlandi.

Sintaktik bogʻlanishlarni belgilash (iyerarxik tahlil): korpus yaratish jarayonidagi eng muhim bosqich – bu gapdagi soʻzlar orasidagi sintaktik munosabatlarni, yaʼni iyerarxik bogʻlanishlarni belgilashdir. Universal Dependencies doirasida har bir bogʻlanish HEAD (bosh soʻz) va DEPREL (bogʻlanish turi) parametrlari bilan ifodalanadi.

XULOSA

Ushbu maqolada korpusimizda UD tavsiya etgan 37 ta asosiy sintaktik munosabatlardan 32 tasini qoʻllangani aks etadi (qolganlari bizning kiritgan jumlarimizda uchramadi). Sintaktik bogʻlanishlarni belgilash qoʻl mehnatini talab qiluvchi, murakkab vazifa boʻlgani bois, uni ham qisman avtomatlashtirishga harakat qildik. Dastlab, yuqorida morfologik teglar bilan boyitilgan 150 ta jumla toʻgʻridan-toʻgʻri qoʻlda sintaktik bogʻlanishlar qoʻyildi. Bunda maxsus grafik vositalardan – Grew (Graph Rewriting tool, Guillaume, 2021) [1,168]. – foydalandik; Grew yordamida UD qoidalariga zid kelmaydigan daraxt tuzilmasini tekshirib borish mumkin. Shuningdek, sintaktik tahlil modelini mashinada oʻqitishda yordam berish uchun fastText dan olingan oʻzbekcha soʻz vektorlaridan foydalanildi. 150 ta jumla qoʻlda bogʻlanib boʻlgach, Stanza tarkibidagi parser modelini shu maʼlumotlar asosida dastlabki oʻqitdik. Ushbu dastlabki model keyin yana bir qancha jumlani avtomatik tahlil qilib berdi, yaʼni har bir soʻzga HEAD va DEPREL qiymatlarini belgilab chiqdi. Albatta, bu avtomatik natijalar toʻliq ishonchli boʻlmagan – annotatorlar tomonidan ushbu jumlalar har biri diqqat bilan tekshirilib, avtomatik tahlil xatolari tuzatildi. Tuzatilgan

jumlalar qo'shimcha tarzda dasturga qayta o'rgatilib, sintaktik analizning ikkinchi modeli hosil qilindi. Nihoyat, qolgan jumlar ham shu ikkinchi model yordamida avtomatik tahlil qilindi va ular ham mutaxassis tomonidan qo'lda tekshirilib, zaruriy joylarda tuzatildi. Jarayon yakunida sintaktik bog'lanishlarni belgilovchi model uchinchi bor takomillashtirilib o'qitildi (1-bosqichdan 3-bosqichgacha model sifat ko'rsatkichlari oshib bordi — F1 o'lchovi birinchi yurgizishda 46%, oxirgi yurgizishda 52% gacha ko'tarildi). Oxir-oqibat, korpusdagi barcha jumalarning izchil sintaktik daraxtlari (bog'lanishlari) tayyor bo'ldi.

Annotatsiya jarayonining ushbu bosqichma-bosqich yondashuvi jamoaga ishni ma'lum darajada avtomatlashtirib, vaqt va kuchni tejashga imkon berdi. Yakuniy bosqichda esa barcha qatlamlarning mutanosibligini yana bir bor sinchiklab tekshirish, yagona formatga keltirish va CoNLL-U fayllariga eksport qilish amalga oshirildi.

ADABIYOTLAR RO'YXATI

1. Bruno Guillaume. 2021. Graph Matching and Graph Rewriting: GREW tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 168–175, Online. Association for Computational Linguistics.
2. P Qi, Y Zhang, Y Zhang, J Bolton, Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. CD Manning. Association of Computational Linguistics (ACL) System Demonstrations.
3. <https://stanfordnlp.github.io/stanza/>
4. Salaev U. UzMorphAnalyser: A morphological analysis model for the Uzbek language using inflectional endings //AIP Conference Proceedings. – AIP Publishing, 2024. – T. 3244. – №. 1
5. Stefanie Dipper, Cora Haiber, Anna Maria Schröter, Alexandra Wiemann, and Maike Brinkschulte. 2024. Universal Dependencies: Extensions for Modern and Historical German. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17101–17111, Torino, Italia. ELRA and ICCL.
6. <https://www.nltk.org/api/nltk.tokenize.html>
7. <https://huggingface.co/datasets/tahrirchi/uz-crawl>