MIRZO ULUG‘BEK NOMIDAGI O‘ZBEKISTON MILLIY UNIVERSITETI HUZURIDAGI ILMIY DARAJALAR BERUVCHI DSc.03/25.08.2021.Fil.01.16 RAQAMLI ILMIY KENGASH

O‘ZBEKISTON MILLIY UNIVERSITETI

RAXIMBOYEVA XULKAR G‘AYRATOVNA

O‘ZBEK TILIDA GAPLARNING IYERARXIK TAHLILI KORPUSINI YARATISH

10.00.11 – Til nazariyasi. Amaliy va kompyuter lingvistikasi.

FILOLOGIYA FANLARI BO‘YICHA FALSAFA DOKTORI (PhD) DISSERTATSIYASI AVTOREFERATI

Toshkent – 2025

UO‘K: 81’322:81’366:811.512.133

Filologiya fanlari bo‘yicha falsafa doktori (PhD) dissertatsiyasi avtoreferati mundarijasi

Оглавление автореферата диссертации доктора философии (PhD) по филологическим наукам

Contents of dissertation abstract of Doctor of Philosophy (PhD) on philological sciences

MIRZO ULUG‘BEK NOMIDAGI O‘ZBEKISTON MILLIY UNIVERSITETI HUZURIDAGI ILMIY DARAJALAR BERUVCHI DSc.03/25.08.2021.Fil.01.16 RAQAMLI ILMIY KENGASH

O‘ZBEKISTON MILLIY UNIVERSITETI

RAXIMBOYEVA XULKAR G‘AYRATOVNA

O‘ZBEK TILIDA GAPLARNING IYERARXIK TAHLILI KORPUSINI YARATISH

10.00.11 – Til nazariyasi. Amaliy va kompyuter lingvistikasi.

FILOLOGIYA FANLARI BO‘YICHA FALSAFA DOKTORI (PhD) DISSERTATSIYASI AVTOREFERATI

Toshkent – 2025

Filologiya fanlari bo‘yicha falsafa doktori (PhD) dissertatsiyasi mavzusi O‘zbekiston Respublikasi Oliy ta'lim, fan va innovatsiyalar vazirligi huzuridagi Oliy attestatsiya komissiyasida B2022.2.PhD/Fil2702 raqam bilan ro‘yxatga olingan.

Dissertatsiya Mirzo Ulug‘bek nomidagi O‘zbekiston Milliy universitetida bajarilgan.

Dissertatsiya avtoreferati uch tilda (o‘zbek, ingliz, rus (rezyume)) Imiy Kengashning veb-sahifasida (www.nuu.uz) va “ZiyoNET” Axborot ta’lim portalida (www.ziyonet.uz) joylashtirilgan.

Ilmiy rahbar:	**Sadullayeva Nilufar Azimovna** filologiya fanlari doktori, professor
Rasmiy opponentlar:	**Abduraxmonova Nilufar Zaynobiddin qizi** filologiya fanlari doktori, professor
	Toirova Guli Ibragimovna filologiya fanlari doktori, professor
Yetakchi tashkilot:	**Urganch davlat Universiteti**

Dissertatsiya himoyasi Mirzo Ulug‘bek nomidagi O‘zbekiton Milliy universiteti huzuridagi ilmiy darajalar beruvchi **DSc.03/25.08.2021.Fil.01.16** raqamli bir martalik Ilmiy kengashning 2025-yil “____” _____kuni soat _____dagi majlisida bo‘lib o‘tadi (Manzil: 100174, Toshkent shahar, Olmazor tumani, Universitet ko‘chasi, 4-uy. Tel.: (99871) 246-02-24; faks: (99871) 246-02-24; e-mail: nauka@nuu.uz. O‘zbekiston Milliy universiteti, O‘zbek filologiyasi fakulteti, 1-qavat, 108-xona).

Dissertatsiya bilan Mirzo Ulug‘bek nomidagi O‘zbekiton Milliy universitetining axborot-resurs markazida tanishish mumkin (___raqami bilan ro‘yxatga olingan). Manzil: 100174, Toshkent shahar, Olmazor tumani, Universitet ko‘chasi, 4-uy. Tel.: (99871) 246-02-24; faks: (99871) 246-02-24; e-mail: nauka@nuu.uz. O‘zbekiston Milliy universiteti, O‘zMU ma’muriy binosi, 2-qavat, 4-xona.

Dissertatsiya avtoreferati 2025-yil “____” _____kuni tarqatildi.

(2025-yil “____” _____ dagi _____ raqamli reyestr bayonnomasi).

N.A.Rahmanov
Ilmiy darajalar beruvchi bir martalik ilmiy kengash raisi, filol.f.d., professor

M.B.Xujamkulova
Ilmiy darajalar beruvchi bir martalik ilmiy kengash kotibi, PhD, dotsent

A.E.Mamatov
Ilmiy darajalar beruvchi ilmiy kengash qoshidagi bir martalik ilmiy seminar raisi, filol.f.d., professor

KIRISH

Dissertatsiya mavzusining dolzarbligi va zarurati. Jahon mamlakatlarining turli sohalari axborot olamining bevosita ta’siri ostida ekanligi, muloqot va axborotning ahamiyati juda kattaligi hammaga ma’lum. Shu bilan birga, kompyuter texnologiyalari yangi davr maqsadlariga erishishda muhim o‘rin tutib, hayotning barcha jabhalariga kirib bordi. Globallashuv sharoitida mashina tili va inson tili o‘rtasidagi farq bilan bog‘liq ma’lum qiyinchiliklar mavjud, bundan tashqari, inson tilining mavjud o‘ziga xos xususiyatlari bu muammoni hal qilish zaruratini tobora ko‘proq belgilamoqda. Shu nuqtayi nazardan kompyuter lingvistikasi muhim ahamiyat kasb etadi.

Dunyo tilshunosligida iyerarxik tahlil korpusi matnlarning sintaktik tuzilishini chuqur tahlil qilishda muhim ahamiyatga egadir. Til ko‘nikmalarini faollashtirish va uni dunyoga tanitish uchun uning nazariy hamda ilmiy tomondan tatbiq etilishi yetarli hisoblanmaydi, balki o‘sha tilni hozirda dolzarb hisoblanuvchi xalqaro tabiiy tillarni qayta ishlash (NLP – Natural Language Processing) resurslari orqali dunyoga tatbiq etish muhim ahamiyat kasb etadi. Kompyuter lingvistikasi tilni sun’iy intellekt modellari yordamida o‘rganadi, shuningdek, katta hajmdagi matnlar to‘plamini (korpusni) tahlil qilishni o‘z ichiga oladi.

Bugungi kunda mamlakatimizda ham korpus lingvistikasi yoki sun’iy intellekt kabi kompyuter sohalari ko‘lami rivojlanish bosqichida bo‘lib, ko‘rib chiqilmagan jihatlari yetarlicha ko‘p hisoblanadi, bu sohada lingvistik va tabiiy tillarni qayta ishlash resurslari sezilarli darajada kamligicha qolmoqda. Yetarli hajmdagi gaplarni qamrovchi iyerarxik (sintaktik, morfologik tahlil qilingan) tizimlangan korpus yaratish orqali o‘zbek tilining lingvistik tadqiqotlarini rivojlantirish, shuningdek, xalqaro ilmiy hamjamiyatda tan olinish imkoniyati oshadi.

Mazkur tadqiqot ish O‘zbekiston Respublikasi Prezidentining 2020-yil 5-oktyabrdagi PF-6079-son “Raqamli O‘zbekiston – 2030” strategiyasini tasdiqlash va uni samarali amalga oshirish chora-tadbirlari to‘g‘risida”gi farmoni, 2021-yil 17-fevraldagi PQ-4996-son “Sun’iy intellekt texnologiyalarini jadal joriy etish uchun shart-sharoitlar yaratish chora-tadbirlari to‘g‘risida”gi Qarori, 2019-yil 21-oktyabrdagi PF-5850-son “O‘zbek tilining davlat tili sifatidagi nufuzi va mavqeini tubdan oshirish chora-tadbirlari to‘g‘risida”gi farmoni, 2020-yil 20-oktyabrdagi PF-6084-son “Mamlakatimizda o‘zbek tilini yanada rivojlantirish va til siyosatini takomillashtirish chora-tadbirlari to‘g‘risida”gi farmonlarda va mazkur sohaga tegishli me’yoriy-huquqiy xujjatlarda belgilangan vazifalarni amalga oshirishda xizmat qiladi. Prezidentimiz aytganidek, “O‘zbek tilining o‘ziga xos xususiyatlari, shevalari, tarixiy taraqqiyoti, uning istiqboli bilan bog‘liq ilmiy tadqiqotlar samarasini oshirish”[1] mamlakatimizdagi kundalik talablardan biri sanaladi.

Tadqiqotning respublika fan va texnologiyalari rivojlanishining ustuvor yo‘nalishlariga mosligi. Tadqiqot respublika fan va texnologiyalar

[1] O‘zbekiston Respublikasi Prezidenti Sh.M.Mirziyoyevning 2019-yil 21-oktyabrdagi “Milliy o‘zligimiz va mustaqil davlatchiligimiz timsoli” mavzusida o‘zbek tiliga davlat tili maqomi berilganining o‘ttiz yilligiga bag‘ishlangan tantanali marosimdagi nutqi // Xalq so‘zi, 2019-yil, 22-oktyabr.

rivojlanishining “Axborotlashgan jamiyat va demokratik davlatni ijtimoiy, huquqiy, iqtisodiy, madaniy, ma’naviy-ma’rifiy rivojlantirish, innovatsion iqtisodiyotni rivojlantirish” ustuvor yo‘nalishiga muvofiq bajarilgan.

Muammoning o‘rganilganlik darajasi. Jahon tilshunosligida universal sintaktik bog‘liqlik mavzusining yoritilishida J.Nivre, D.Zeman, de Marneffe, D.Manning, De Marneffe, Donald Knut, Yevgeniy Karniak Daniel Jurafsky [2] kabi bir qancha olimlar tabiiy tillarni qayta ishlash jarayoni hamda korpus yaratish bo‘yicha bir qator tadqiqotlarga hissa qo‘shganlar.

Respublikamizda o‘zbek kompyuter lingvistikasi bo‘yicha A.Q.Po‘latov, M.M.Aripov, A.E.Mamatov, N.Z.Abduraxmanova, M.M.Kurbanova, M.M.Musayev, N.A.Sadullayeva, Sh.A.Nazirov, M.H.Hakimov, O‘.Hamdamov, N.A.Ignatev va G‘.R.Matlatipovlar[3] rahbarligida olib borilgan ilmiy tadqiqotlar o‘zbek tili grammatikasining formal modelini yaratish, o‘zbek, turk va qoraqalpoq tillaridagi gap tuzilmalarini formallashtirish, turkiy tillar uchun lug‘at yaratish, nutqli signallarning shakllanish jarayonlarini modellashtirish, nutq sintezatorlarini yaratish, ko‘p tilli modellashgan mashina tarjimasiga oid til obyektlarining mantiqiy-lingvistik va matematik modellarini yaratish masalalariga qaratilgan.

O‘zbekistonlik mashhur olim, fizika-matematika fanlari doktori, professor Abdumajid Po‘latov tomonidan uzoq yillar davomida tilshunoslikka oid “Ingliz tili”, “Dunyoviy o‘zbek tili. O‘zbek tilida fe’l shakllari va ularning rus, ingliz tillaridagi ko‘rinishlari”, “Kompyuter lingvistikasi”, “Ingliz tili – mustaqil o‘rganuvchilar uchun” va boshqa kitoblari [4], O‘.Sharipov, I.Yo‘ldoshevlar tomonidan “Tilshunoslik asoslari” [5] nomli o‘quv qo‘llanma, A.Rahimov tomonidan “Kompyuter lingvistikasi asoslari” [6] nomli o‘quv qo‘llanma, G‘.R.Matlatipov tomonidan “Mashina tarjimasi texnologiyalari” nomli o‘quv

---

[2]Nivre J., Zeman D. Universal Dependencies: A Cross-Linguistic Perspective on Grammar and Lexicon // Proceedings of the Workshop on Grammar and Lexicon: Interactions and Interfaces (GramLex 2016), 2016. – P. 8-17. URL: https://aclanthology.org/W16-3806.pdf.

De Marneffe M.-C., Manning C.D., Nivre J., Zeman D. Universal Dependencies // Computational Linguistics, 2021. Volume 47. Issue 2. – P. 255-270. URL: https://direct.mit.edu/coli/article/47/2/255/98516/Universal-Dependencies; Jurafsky D., Martin J.H. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition: Prentice Hall, 2000.

[3] Abdurakhmonova N.Z., Urdishev K. Corpus Based Teaching Uzbek As A Foreign Language // Journal of Foreign Language Teaching and Applied Linguistics. – Tashkent, 2019 // Abdurakhmonova Nilufar, Ismailov Alisher, Sayfullayeva Ra’no. MorphUz: Morphological analyzer for the Uzbek language / 7th International Conference on Computer Science and Engineering (UBMK), 2022. – № 9/14. – P. 61-66; Abdurakhmonova N. Dependency Parsing Based On Uzbek Corpus / Proceedings of the International Conference on Language technology for all (LT4all), 2019; Abdurakhmonova N.Z., Alisher S., Ismailov A., Mengliev D. Developing NLP tool for linguistic analysis of Turkic languages / IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON), 2022; Abdurahmonova Nilufar. Kompyuter lingvistikasi. – Toshkent: Nodirabegim, 2021; Aripov Mirsaid, Razzoqova Bibigul, Sharibbek Altinbay, Abdurakhmonova Nilufar. Ontology of grammar rules as example of Noun the Uzbek language and Kazakh languages / VI International scientific conference “Modern problems of the applied mathematics and information technology – Al Khorezmiy”. – T., 2018. – P. 37-38; Aripov Mirsaid, Begalov Baxodir, Begimqulov Uzoqboy, Mamarajabov Mirsalim. Axborot texnologiyalari. – Toshkent, 2009; Po‘latov A., Muhammedova S. Kompyuter lingvistikasi. – T., 2008; Po‘latov A. Kompyuter lingvistikasi. – T., 2011; Rahimov A. Kompyuter lingvistikasi asoslari. – T., 2011.

[4] Po‘latov A. Kompyuter lingvistikasi. – T., 2011.

[5] Sharipov O‘,. Yo‘ldoshev I. Tilshunoslik asoslari. – Toshkent: Nizomiy nomidagi Toshkent Davlat pedagogika universiteti, 2006.

[6] Rahimov A. Kompyuter lingvistikasi asoslari. – T., 2011.

qo‘llanma[7], N.Z.Abduraxmanova tomonidan “Korpus lingvistikasi” nomli darslik chop etilgan.

Shuningdek, jahonda tabiiy tillarning kompyuterga yo‘naltirilgan modellarini ishlab chiqish bo‘yicha bir qator tadqiqotlar olib borilgan bo‘lib, jumladan, N.A.Xomskiy (AQSH), Christopher D.Manning (AQSH), D.Juraffskiy (AQSH), Carlos Gómez-Rodríguez (Ispaniya), Joakim Nivre, Miguel A.Alonso (Ispaniya), Marco Kuhlmann (Shvetsiya), A.Jeyms, G.Fant, R.Delmonti [8] va boshqalarning bu boradagi ishlari qiyosiy o‘rganib chiqildi. Shuningdek, Mustaqil davlatlar hamdo‘stligi davlatlarida kompyuter lingvistikasi sohasida faoliyat yuritayotgan Ye.I.Bolshakova, A.Sharipbay[9] kabi olimlarning ushbu sohadagi qilingan ilmiy va uslubiy ishlari o‘rganildi.

Tadqiqotning maqsadi. O‘zbek tilida elektron shaklda saqlangan matnli ma’lumotlarni qayta ishlash uchun iyerarxik bog‘liqlik korpusini qurish, sun’iy intellekt modeli orqali baholash va dasturiy ta’minotini yaratish.

Tadqiqotning vazifalari. Tadqiqot maqsadini amalga oshirish uchun quyidagi vazifalarni hal qilish zarur bo‘ladi:

o‘zbek tilidagi iyerarxik korpusini annotatsiyalash uchun negizlash, morfologik xususiyatlash, gap bo‘laklarini aniqlash va bog‘liqlik qoidalarini ishlab chiqish uchun yo‘riqnoma tayyorlash;

o‘zbek tilidagi murakkab va sodda gaplardan tashkil topgan korpus bazasini yig‘ish hamda unga ishlov berish (pre-processing);

o‘zbek tilidagi korpusga kiritilgan gaplar tarkibidagi har bir so‘zning negizini (lemmasi) aniqlash va morfologik xususiyatlarini tahlil qilgan holda, avtomatik parser yordamida xulosalar chiqarish, to‘liq iyerarxik bog‘liqlik daraxti tahlil korpusini taqdim etish;

o‘zbek tilidagi korpusga kiritilgan gaplar tarkibidagi so‘zlarning sintaktik xususiyatlarini aniqlab, iyerarxik tarzda bog‘lanishlarini qurish va tobelanishni sun’iy intellekt modeli orqali tahlil qilish.

Tadqiqotning obyekti sifatida Kun.uz va daryo.uz saytlaridan tanlab olingan matnlar korpusda tahlillash uchun olingan.

---

[7] Matlatipov G‘.R. Mashina tarjimasi texnologiyalari. O‘quv qo‘llanma. – T., 2024.

[8] Gómez-Rodríguez Carlos, Nivre Joakim. A transition-based parser for 2-planar dependency structures / Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010. – P. 1492-1501; Nivre Joakim, Rimell Laura, McDonald Ryan, Gomez-Rodriguez Carlos. Evaluation of dependency parsers on unbounded dependencies / Proceedings of the 23rd international conference on computational linguistics, 2010. – P. 833-841; D.Gonzalez-Franco Joan, E.Preciado-Velasco Jorge, E.Lozano-Rizk Jose, Rivera-Rodriguez Raul, Torres-Rodriguez Jorge, A.Alonso-Arevalo Miguel. Comparison of Supervised Learning Algorithms on a 5G Dataset Reduced via Principal Component Analysis (PCA) – Future Internet, 2023. – P. 335; Ralph Debusmann, Denys Duchier, Alexander Koller, Marco Kuhlmann, Gert Smolka, Stefan Thater., A relational syntax-semantics interface based on dependency grammar COLING 2004 / Proceedings of the 20th International Conference on Computational Linguistics, 2004. – P. 176-182; Dann Juraffskiy., James H.Martin. Speech and language processing. – New Jersey: Upper Saddle River, 2008. February; Manning Ch., Schutze H. Foundations of Statistical Natural language Processing, MIT Press. Cambridge, MA, 1999. May; Andrew Ng. Machine Learning yearning, technical strategy for AI Engineers in the Era of Deep Learning, 2023. August.

[9] Bol'shakova E.I., Vorontsov K.V., Efremova N.E., Klyshinskiy E.S., Lukashevich N.V., Sapin A.S. Avtomaticheskaya obrabotka tekstov na estestvennom yazyke i analiz dannykh [Automatic natural language text processing and data analysis]. – Moscow: NIU VShE Publ, 2017; Sharipbay Altynbek, Razakhova Bibigul, Mukanova Assel, Yergesh Banu, Yelibayeva Gaziza. Proceedings of the Second International Conference on Data Science, E-Learning and Information Systems. – Kazakhstan, 2019. – P. 1-5.

Tadqiqotning predmeti o‘zbek tilidagi gaplarning sintaktik tuzilishini iyerarxik tarzda izohlash orqali, so‘zlar orasidagi grammatik munosabatlarni aniqlash va tasniflashdir.

Tadqiqotning usullari. Tadqiqot mavzusini yoritishda tasniflash, tavsiflash, avtomatik parsing[10], yo‘riqnoma orqali annotatsiyalash, o‘tkazish (transfer), korpus tahlili, statistik usullardan foydalanildi.

Tadqiqotning ilmiy yangiligi.

o‘zbek tilida iyerarxik bog‘liqlik korpusi yaratildi;

o‘zbek tilidagi iyerarxik korpusini annotatsiyalash, negizlash (lemmalash) morfologik xususiyatlash, gap bo‘laklarini aniqlash va bog‘liqlik qoidalarini ishlab chiqish yo‘riqnomasi yaratildi;

o‘zbek tilidagi jumlalarning iyerarxik qoidalari, morfologik hamda sintaktik xususiyatlari korpus doirasida chuqur tahlil qilindi va annotatsiyalandi;

o‘zbek tilining xususiyatlarini inobatga olgan holda tabiiy tilni qayta ishlash jarayoni uchun sun’iy intellekt modeli qurilib, korpusning ishonchliligi sun’iy intellekt modeli orqali sintaktik tahlil qilgan holda baholandi.

Tadqiqotning amaliy natijalari quyidagilardan iborat:

tabiiy tilni qayta ishlash (NLP) va sun’iy intellekt sohasida o‘zbek tilida avtomatik tarjima, matnni tahlil qilish, savol-javob tizimlari va ovozli yordamchilarni yaratishda asosiy til resursi sifatida xizmat qiladi;

tilshunoslik tadqiqotlari uchun aniq sintaktik ma’lumotlar bazasini ta’minlab, o‘zbek tilining grammatik tizimini chuqur tahlil qilish imkonini beradi;

ta’lim jarayonida lingvistik modellar asosida interaktiv vositalar va o‘quv dasturlarini ishlab chiqishga yordam beradi.

Tadqiqot natijalarining ishonchliligi. UAS (unlabeled attachment score – annotatsiyalanmagan bog‘lanish balli) hamda LAS (Labeled attachment score – annotatsiyalangan bog‘lanish balli) matematik baholash metodlari orqali korpusning natijaviy ishonchliligi ko‘rsatilgan.

Tadqiqot natijalarining ilmiy va amaliy ahamiyati.

Ilmiy ahamiyati iyerarxik tahlil korpusi tilshunoslik va kompyuter lingvistikasi tadqiqotlari uchun o‘zbek tilining sintaktik tuzilishini raqamli shaklda ifodalovchi, sintaktik hodisalarni chuqurroq tahlil qilish hamda boshqa tillar bilan qiyosiy ilmiy izlanishlarni amalga oshirish imkonini beruvchi baza vazifasini bajaradi. O‘zbek tilining sintaktik tuzilmasini raqamli shaklda ifodalaydi, bu esa tahliliy tadqiqotlarni rivojlantiradi va til modellarini yaxshilashga hissa qo‘shadi.

Amaliy ahamiyati esa tabiiy tilni qayta ishlash dasturlarini rivojlantirish uchun zarur resurs vazifasini bajaradi, jumladan, avtomatik tarjima, chatbotlar, matn mazmunini tahlil qilish, matnlarni avtomatik umumlashtirish, sentiment tahlil kabi vazifalarda ishlatiladi. Bu o‘z o‘rnida, o‘zbek tilida ishlovchi aqlli tizimlar, qidiruv tizimlari va axborot olish tizimlarining sifatini oshiradi, bu esa mazkur texnologiyalarning kundalik hayotda qo‘llanishini kengaytiradi. Shuningdek, o‘zbek tilidagi raqamli xizmatlar va dasturiy ta’minotlarni ishlab chiqish uchun

[10] Parsing (o‘zbekchada “tahlil qilish”, “ajratish”, “tarkibiy qismlarga ajratish”) degani, berilgan jumlaning grammatik tuzilishini aniqlash, ya’ni so‘zlarning o‘zaro bog‘liqlik munosabatlarini topish jarayonidir.

tayyor resurs sifatida xizmat qiladi, bu esa mahalliy va global IT kompaniyalariga o‘z mahsulotlarini yanada sifatli hamda foydalanuvchi talablariga mos ravishda yaratishlariga imkon beradi.

Tadqiqot natijalarining joriy qilinishi. O‘zbekiston Respublikasi Innovatsion rivojlanish vazirligi huzurida tashkil etilgan Jahon bankining “O‘zbekiston milliy innovatsion tizimini modernizatsiya qilish” loyihasi tomonidan asos solingan REP-25112021/113 “UzUDT: O‘zbek tilida tabiiy tilni qayta ishlash uchun universal bog‘liqlik daraxti korpusi va uning semantik tahlili” granti doirasidagi amaliy tadqiqotlarda tayyorlangan “O‘zbek tilida gaplarning iyerarxik tahlili korpusini yaratish” mavzusidagi tadqiqot ilmiy natijalaridan foydalanildi. Jumladan, amaliy loyiha doirasida o‘zbek tili matnlaridagi so‘zlarni morfologik tahlil qilish yuzasidan negizlash (lemmalash)ni bajarish, iyerarxik sintaktik bog‘lanishlarni aniqlash, ularni teglash va baholash, matndagi so‘z turkumlarini annotatsiyalash va baholash, morfologik xususiyatlar bazasini ishlab chiqish masalalarida dissertatsiya ishining ilmiy natijalaridan foydalanildi;

Xorazm Ma’mun akademiyasida faoliyat olib borayotgan A-FA-2019-9 shifrli “Qadimiy yozma noyob qo‘lyozma va manbalarni tadqiq qilish, ularning raqamlashtirilgan bibliotekasini yaratish” amaliy loyihasida mazkur tadqiqotdan keng foydalanildi, olingan natija va xulosalarning amaliyotga tatbiq etilishi quyidagilarga xizmat qildi: dissertatsiya materiallari tarixiy asarlarda uchraydigan mumtoz adabiyot vakillarining ijod namunalarini tahlil qilishda; tarixiy asarlar matnida uchraydigan so‘zlar va gaplarni to‘g‘ri tabdil qilish, izohlashda; XIX asrga tegishli adabiy manbalar tilini lingvistik tahlil qilish, oldingi va keyingi davr til xususiyatlari bilan solishtirish, o‘ziga xos belgilarni aniqlashda yordam berdi;

Erasmus + dasturining Project №598340-EPP-1-2018-1-ES-EPPKA2-CBHE-JP University Cooperation Framework for Knowledge Transfer in Central Asia and China (UNICAC) grant loyihasida dissertatsiyadan olingan natijalardan foydalanildi: tillararo semantik qidiruv tizimlarini va ma’lumotlar bazalarini kengaytirish; ma’lumotlarga asoslangan tahlillar yordamida til o‘rganish va tillaridagi sintaktik o‘ziga xos jihatlarni ta’kidlaydigan ta’lim vositalarini yaratish; xalqaro loyihalarda foydalanish uchun lingvistik resurslarni standartlashtirishda foydalanildi.

Tadqiqot natijalarining aprobatsiyasi. Mazkur tadqiqotning muhim g‘oyalari va amaliy natijalari 3 ta xalqaro (shundan 1 ta Scopus), 6 ta respublika ilmiy-amaliy anjumanlarida aprobatsiyadan o‘tkazilgan.

Tadqiqot natijalarining e’lon qilinganligi. Dissertatsiya mavzusi bo‘yicha jami 14 ta ish e’lon qilingan, jumladan, O‘zbekiston Respublikasi Oliy attestatsiya komissiyasining doktorlik dissertatsiyalari asosiy ilmiy natijalarini chop etish tavsiya etilgan ilmiy nashrlarda 5 ta maqola, jumladan, 4 ta respublika va 1 ta xorijiy nashrlarda chop qilingan.

Dissertatsiyaning tuzilishi va hajmi. Dissertatsiya kirish, uchta bob, umumiy xulosa va tavsiyalar, foydalanilgan adabiyotlar ro‘yxatidan tashkil topgan bo‘lib, ishning umumiy hajmi 133 sahifadan iborat.

DISSERTATSIYANING ASOSIY MAZMUNI

Kirish qismida dissertatsiya mavzusining dolzarbligi asoslangan, dissertatsiya mavzusi bo‘yicha xorijiy ilmiy-tadqiqotlar sharhi va muammoning o‘rganilganlik darajasi bayon etilgan, tadqiqotning maqsadi va vazifalari, shuningdek, obyekti va predmeti aniqlangan, tadqiqot ishining fan va texnologiyalarni rivojlantirishning ustuvor yo‘nalishlariga mosligi ko‘rsatilgan hamda tadqiqotning ilmiy yangiligi, natijalarning ishonchliligi, nazariy va amaliy ahamiyati, natijalarning amaliyotga joriy etilishi, e’lon qilinganligi, ishning tuzilishi borasida ma’lumotlar berilgan.

Tadqiqotning “**Iyerarxik tahlil korpusining asosiy tushunchalari**” deb nomlangan birinchi bobida iyerarxik korpusi va uning sintaktik tahlili, o‘zbek tili texnologiyasi hamda iyerarxik korpus yaratish metodlari, izohlash (annotatsiyalash) sxemalari va foydalaniladigan elektron vositalar alohida tushuntirilib, misollar keltirilgan.

Ko‘p tilli tabiiy tilni qayta ishlash talabining ortib borishi bilan o‘zbek tiliga ham qiziqish ortib bormoqda. Ayniqsa, morfologik jihatdan boy bo‘lgan o‘zbek tili uchun uning iyerarxik bog‘liqliklarini va semantikasini aniqlash muhim ahamiyat kasb etadi. O‘zbek tilida erkin so‘z tartibi va agglyutinativ morfologiya keng tarqalgan bo‘lib, bog‘liqlikni aniqlash aniq va kontekstga mos tabiiy tillarni qayta ishlash vositalarini yaratishda zarurdir. Ushbu tadqiqot ishi o‘zbek tilidagi iyerarxik bog‘liqlikni aniqlashning ahamiyati, metodologiyasi, muammolari va qo‘llanilish sohalarini o‘rganadi hamda uning o‘zbek tilida so‘zlashuvchilar uchun tabiiy tillarni qayta ishlash sohasi rivojlantirishidagi transformatsiyaviy rolini ta’minlaydi.

Iyerarxik bog‘liqlik – bu gapning grammatik tuzilishini tahlil qilish jarayoni bo‘lib, unda so‘zlar o‘rtasidagi munosabatlar aniqlanadi. Bog‘liqlikni tahlil qilishda (parsing) har bir so‘z “bosh” so‘z yoki qo‘shimcha bilan bog‘lanadi, bu esa har bir so‘zning gapdagi sintaktik rolini ko‘rsatadigan “bog‘liqlik munosabatini” o‘rnatadi[11]. Masalan, “Talaba kitobni o‘qiyapti” jumlasida “kitobni” (kitob) so‘zi “o‘qiyapti” (o‘qimoqda) fe’liga obyekt sifatida bog‘lanadi, “Talaba” (talaba) esa subyekt sifatida xizmat qiladi.

Bog‘liqlikni aniqlash ko‘plab tabiiy tillarni qayta ishlash ilovalari, jumladan, mashina tarjimasi, sentiment tahlil qilish va savol-javob tizimlari uchun muhimdir. Iyerarxik bog‘liqlikni aniqlash batafsil sintaktik tuzilmalarni taqdim etish orqali gaplarni aniq talqin qilish, mantiqiy va kontekstga mos natijalarni ishlab chiqarishga imkon beradi.

O‘zbek tilidagi iyerarxik bog‘lanishli tahlil qilish o‘ziga xos hisoblanadi. Turkiy tillar oilasiga kiruvchi o‘zbek tili o‘zaro bog‘liqlikni tahlil qilishni ham qiyin, ham foydali qiladigan bir qancha xususiyatlarga ega.

Agglyutinativ morfologiya: O‘zbek tili so‘zlarni ildizga bir nechta suffikslarni qo‘shish orqali hosil qiladi. Masalan, “o‘qiyapman” (Men o‘qiyotganim) so‘zida “o‘q” negizi, vaqt hamda subyekt ma’lumotlarini ifodalovchi -iya, -p, va -man

[11] Shukurov M. O‘zbek tili va grammatikasi. – Toshkent: O‘zbekiston, 2010. – B. 86-87.

suffikslar mavjud. Bu tuzilmalarni aniq parserlash so‘zlar o‘rtasidagi sintaktik munosabatlarni tushunish uchun juda muhimdir[12].

*Erkin so‘z rartibi***:** O‘zbek tilida odatda Subyekt-Obyekt-Fe’l (SOV) tartibi kuzatiladi, ammo bu tuzilma yozma va og‘zaki shakllarda sezilarli darajada farq qilishi mumkin. Bu erkinlik bog‘liqlikni aniqlovchilar turli gap tuzilmalariga moslashuvchan bo‘lishi, shu bilan birga, sintaktik munosabatlarning aniqligini saqlab qolishi kerak.

Hol belgilari va morfologik-sintaktik moslik: O‘zbek tilida grammatik rollarni, masalan, subyekt, obyekt yoki egasi kimligini ko‘rsatish uchun hol belgilari ishlatiladi. O‘zbek tilidagi bog‘liqlikni aniqlashda ushbu belgilarni hisobga olishi kerak, chunki ular so‘zlar o‘rtasidagi munosabatlarga bevosita ta’sir qiladi.

O‘zbek tilidagi bog‘liqlikni aniqlashning asosiy tarkibiy qismlari: O‘zbek tilidagi iyerarxik bog‘liqlikni aniqlashda tilning o‘ziga xos tuzilmalarini hisobga olish, batafsil annotatsiyalar va lingvistik belgilashlar kerak bo‘ladi.

Quyida asosiy komponentlari keltirilgan:

So‘z turkumni belgilash: So‘z turkumlari har bir so‘zning sintaktik toifasini (masalan, ot, fe’l, sifat) aniqlaydi. O‘zbek tilida so‘z turkumlarini belgilash birinchi bosqichdir, chunki bu bog‘liqlik munosabatlarini o‘rnatish uchun asosiy ma’lumot qatlamini ochib beradi.

Jumlalarda bosh so‘zlarni tanlash va bog‘liqlik munosabatlari: O‘zbek tilida gapdagi so‘zlar “bosh” so‘zga, odatda asosiy fe’lga, bog‘liq bo‘ladi, u bog‘liqlik tuzilmasining ildizi sifatida xizmat qiladi. Dunyo amaliyotida universal bog‘liqlik munosabatlari mavjud va bu qoidalarni tatbiq qilish o‘zbek kopmyuter lingvistikasi uchun juda muhim.

O‘zbek tili uchun quyidagi bog‘liqlik munosabatlarini aniqlaymiz:

- *nsubj:* Fe’lning subyektini belgilaydi.
- *obj:* Fe’lning obyektini ko‘rsatadi.
- *obl:* Obyektiv argumentlarni belgilaydi, odatda, oldin so‘zlar yoki hol sifatli modifikatorlardir.
- *det:* Egasini bildiruvchi zamonlarni ifodalaydi, masalan, egasi zamonlari.
- *amod:* Otlarni modifikatsiya qiluvchi sifat modifikatorlari.

O‘zbekcha jumlani ko‘rib chiqamiz: *Mening yangi do‘stim maktabga bordi.*

1-jadval.

O‘zbek tilida iyerarxik tahlil qilish misoli.

So‘z	Lemma	POS	Bosh so‘z	Iyerarxik bog‘lanish
Mening	men	PRON	do‘stim	nmod
yangi	yangi	ADJ	do‘stim	amod
do‘stim	do‘st	NOUN	bordi	nsubj
maktabga	maktab	NOUN	bordi	obl
bordi	bormoq	VERB	ILDIZ (ROOT)	ILDIZ (ROOT)

[12] Nurmonova D. Hozirgi o‘zbek adabiy tili: sintaksis. Ma’ruzlar matni. – Toshkent, 2016. – B. 287.

Ushbu misolda:

- **Mening** (mening) – **do‘stim** (do‘stim) so‘zining egasi modifikatori (nmod).
- **Yangi** (yangi) – **do‘stim** (do‘stim) so‘zining sifat modifikatori (amod).
- **Do‘stim** (do‘stim) – **bordi** (bordi) asosiy fe’lning subyekti (nsubj).
- **Maktabga** (maktabga) – manzilni ko‘rsatuvchi obyektiv modifikator (obl).

Mazkur misoldan ham ko‘rish mumkinki o‘zbek tilida iyerarxik bog‘liqlikni tahlil (izohlash) qilishda quyidagi turdagi muammolar kuzatiladi:

- *Ma’lumot yetishmovchiligi:* rivojlangan tillardagi kabi keng tarqalgan tillarga nisbatan olganda, o‘zbek tilida katta va annotatsiyalangan korpus deyarli mavjud emas, bu esa iyerarxik parserlarni yetarli ma’lumot bilan o‘qitishni qiyinlashtiradi. Cheklangan resurslar ko‘pincha murakkab jumla tuzilmalarida past parsing aniqligiga olib keladi.
- *Dialekt farqlari:* O‘zbek tilida bir nechta dialektlar mavjud bo‘lib, har biri o‘ziga xos leksik va sintaktik farqlarga ega. Standart o‘zbek tilida o‘qitilgan parser dialekt farqlarini qayta ishlashda qiyinchiliklarga duch kelishi mumkin, bu esa dialektga xos o‘qitish ma’lumotlariga yoki moslashuvchan modellarga ehtiyoj tug‘diradi.
- *Morfologik murakkablik:* O‘zbek tilining agglutinativ tabiati affikslar orqali keng so‘z hosil qilish imkonini beradi. Har bir shakl so‘zning jumladagi rolini sezilarli darajada o‘zgartirishi mumkin, bu esa parserlarning ushbu shakllarni to‘g‘ri tanib olish va qayta ishlashini muhim
- *Moslashuvchan sintaksis***:** Moslashuvchan so‘z tartiblarini parsing qilish qiyin, chunki parserlar ko‘pincha bog‘liqliklarni aniqlash uchun barqaror naqshlarga tayanadi. O‘zbek tilida, jumlalar har xil so‘z tartiblarida grammatik jihatdan to‘g‘ri bo‘lishi mumkin, bu esa sintaktik tahlilni murakkablashtiradi.

O‘zbek tili uchun iyerarxik parsingini rivojlantirish tilning o‘ziga xos xususiyatlariga e’tibor qaratadigan takomillashtirilgan resurslar va murakkab usullarga muhtoj. Dissertatsiyaning adabiyotlarida iyerarxik parsingning kelgusida kutiladigan yo‘nalishlari keltirilgan.

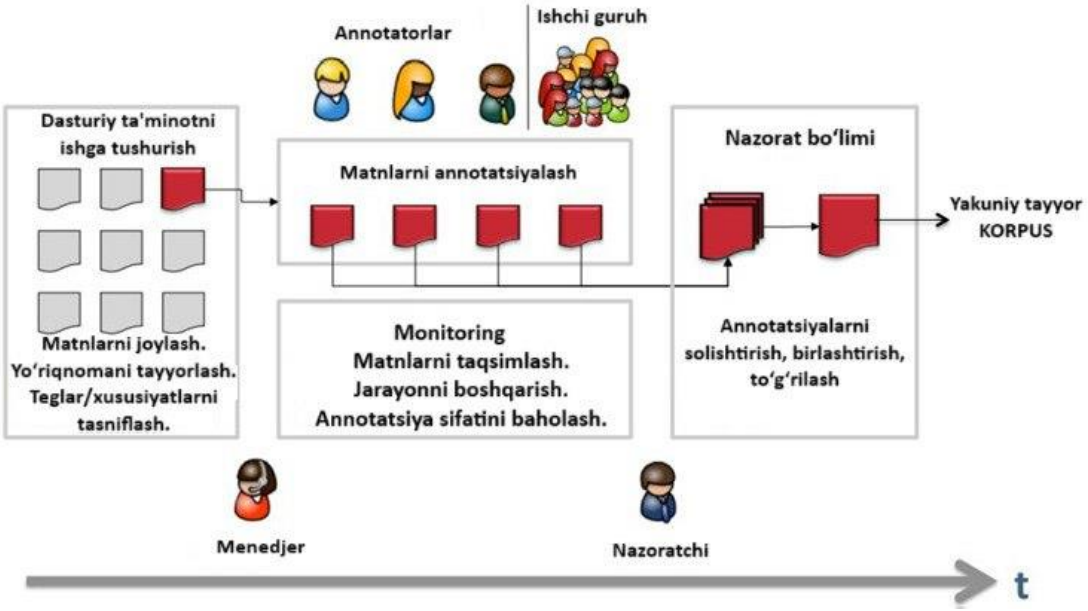
Tadqiqotning II bobi “**O‘zbek tilining ierarxik tahlili korpusini yaratish metodlari va uning amaliy-texnologik jihatlari**” deb nomlangan.

Korpusda uchta maxsus dastlabki ishlov berish bosqichi amalga oshirildi:

metama’lumotlarga qo‘shish – bu har bir misol jumlaga kuzatish, tushunarlilik va unikallik uchun ID-identifikatsiya belgisi berishdir, tarjimalar, shuningdek, ijtimoiy tarmoq va veb sahifalardan olingan jumlalar identifikatsiya qilinadi. Iyerarxik daraxt korpusi ma’lumotlari, metama’lumotlar raqamlash belgisi (#) bilan bog‘lanadi;

xatolarni o‘chirish – bu xatolarni tuzatish, bo‘shliklarni tekshirish va yetishmayotganlarni to‘ldirish, tinish belgilari mavjud gaplarni formatlashni o‘z ichiga oladi. Xabarlar jumlalarida xatolar va noqulay tarjimalar ham bo‘lishi mumkin bo‘lib ular tuzatiladi;

imloviy standartlashtirish – bu o‘zbek tili uchun zarur jarayon, chunki aniq imlo qoidalarisiz tilning qayta ishlanishi murakkablashadi[13]. Biroq imlo standartlashtirish ortiqcha qat’iy bo‘lmasligi kerak, balki muvozanatni saqlash muhimdir. Bu mahalliy o‘zbek tilida so‘zlashuvchilar uchun tabiiy bo‘lgan shakllarni saqlash va imlo tizimini ortiqcha murakkablashtirmaslik o‘rtasidagi muvozanatni ta’minlash uchun amalga oshiriladi. O‘zbek tilida dependency treebank yaratishda, daraxt tuzilishidagi lemmalarning o‘zgarishi ham ushbu muvozanatni saqlash uchun muhim ahamiyatga ega.



1-rasm. Ieyarxik daraxti korpusini ishlab chiqish jarayoni (ketma-ketligi).

Keyingi bosqichda gapni alohida so‘zlarga yoki belgilarga ajratiladi (tokenlash), masalan, “Mushuklar tez yugurmoqda” jumlasi [“Mushuklar”, “yugurmoqda”, “tez”]ga tokenlanadi. Tokenlash jarayoni tinish belgilarini, qisqarishlarni va so‘zlar orasidagi turli intervallarni boshqarishi mumkin. Murakkab so‘zlar va qisqartmalarning segmentatsiyasi ham alohida o‘rin egallaydi (masalan, “FA”, “Nyu York”, va h.k.)

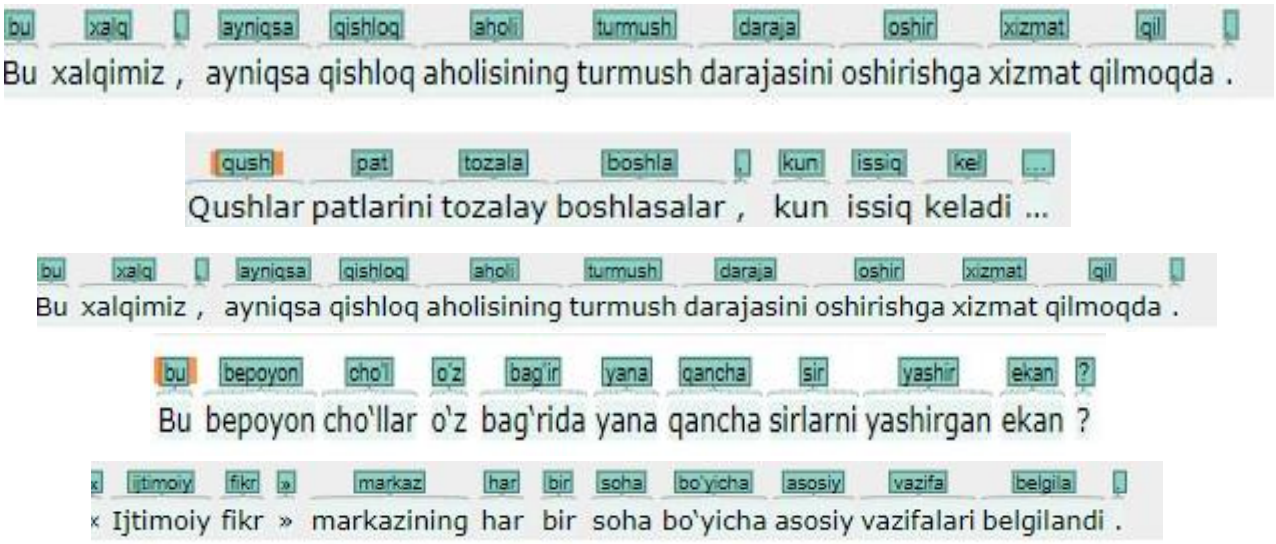
Negizlash (Lemmalash) bosqichida har bir so‘zning lug‘aviy shakli (lemma), ya’ni lug‘atlarda uchraydigan asosiy shakli aniqlanadi. O‘zbek tili agglyutinativ til bo‘lgani uchun, lemma sifatida odatda hech qanday grammatik qo‘shimcha olmaydigan sof lug‘aviy shakl tanlanadi.

Keyingi bosqichda jumladagi har bir so‘zni asosiy yoki lug‘at shakligacha qisqartirish jarayoni bo‘lib o’tadi (jumlalarni negizlash / lemmalash), u ko‘plab tabiiy tillarni qayta ishlash vazifalarida foydalidir. O‘zbek tilida negizlash – so‘zni morfologik tahlil qilish orqali uning asosiga (lug‘at shakliga / negiziga) qisqartirishni o‘z ichiga oladi. O‘zbek tili aglyutinativ til bo‘lib, u ko‘pincha

[13] Bernd Bohnet. Very high accuracy and fast dependency parsing is not a contradiction // University of Stuttgart. – Germany, 2010. – P. 93-94.

zamon, mayl, shaxs va holni bildirish uchun qo‘shimchalardan foydalanadi. Shuning uchun o‘zbek tilidagi negizlash o‘zakni olish uchun qo‘shimchalarni olib tashlashni talab qiladi. Matnli korpusni shakllantirish maqsadida negizlash (lemmalash)da quyidagicha misollarni uchratishimiz 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).

Bu yerda jumlalarni negizlash jarayonidan quyidagi misollarni keltirishimiz mumkin:



2-rasm. Yaxlit bir gapdagi so‘zlarning negizlarini (lemmalarini) belgilash.

So‘zlarda, asosan, asos so‘zlar shu shaklda olindi va tizimda avtomatik tarzda lemmalash amalga oshirildi. Gapdagi har bir so‘z, tinish belgilari alohida lemma qilib olindi.

So‘z turkumlarini aniqlash/teglash (POS). Har bir token uchun so‘z turkumlari aniqlanadi, chunki negizlash jarayonida aniq natijalarni ta’minlash so‘z turkumlarini aniqlanashga bog‘liq. so‘z turkumlarini aniqlanash jarayonida har bir so‘zga ot, fe’l, sifat va boshqa grammatik belgilar tayinlanadi. Python dasturlash tilidagi NLTK (Natural Language Toolkit) yoki spaCy kabi kutubxonalar so‘z turkumlari aniqlanash jarayonida POS teglarini taqdim etadi.

PoS (Part-of-Speech) teglarini quyidagi uchta katta guruhga ajratish mumkin:

Tokenlash (so‘zlarni segmentlash). So‘zlar odatda bir biridan bo‘sh joy orqali ajratiladi, lekin bunda quyidagi mustasnolar mavjud:

– chiziqcha belgisi (“-”) bilan ajratilgan qo‘shma yoki takroriy so‘zlar (masalan, kuta-kuta, bora-bora);

– “bo‘lib” “holda” kabi bog‘lovchilari hol belgilarining paydo bo‘lishi so‘zga qo‘shilgan bo‘lish ko‘p so‘zli leksema sifatida leksema qilinadi;

2-jadval.

So ‘z turkumlarining sinflari.

Ochiq sinf so‘zlar (Open class words) *Bu sinfga kiruvchi so‘zlar doimiy ravishda yangi so‘zlar bilan to‘ldirilishi mumkin. Ochiq sinf so‘zlar jumlada asosiy mazmunni yetkazadi va quyidagilarni o‘z ichiga oladi. Masalan: Otlar (Nouns – NOUN, PROPN): kitob, talaba, O‘zbekiston. Fe’llar (Verbs – VERB): yozmoq, o‘qimoq, gapirdi.*	**Yopiq sinf so‘zlar (Closed class words)** *Bu sinfga kiruvchi so‘zlar miqdor jihatdan cheklangan, ular deyarli yangi so‘zlar bilan boyib bormaydi va ko‘pincha grammatik vazifalarni bajaradi. Masalan: Olmoshlar (Pronouns – PRON): men, sen, ular. Bog‘lovchilar (Conjunctions – CCONJ, SCONJ): va, yoki, ammo, chunki.*	**Boshqalar (other)** *Bu toifaga grammatik vazifasi jihatidan yuqoridagi ikki guruhga kirmaydigan boshqa toifalar kiradi, masalan: Tinish belgilar (Punctuation – PUNCT) nuqta (.), vergul (,), savol belgisi (?) Simvollar (Symbols – SYM): $, %, @, №.*
ADJ: sifat	ADP: ko‘makchi	PUNCT: tinish belgilari
ADV: ravish	AUX: yordamchi fe’l	SYM: simvollar
INTJ: undov so‘z(interjection)	CCONJ: bog‘lovchi	X: boshqalar
NOUN: ot	DET: aniqlovchi	
PROPN: atoqli ot	NUM: son	
VERB: fe’l	PRON: olmosh	

O‘zbek tili turkiy tillar oilasiga mansub bo‘lib, uning sintaktik tuzilmasi ko‘plab o‘ziga xos xususiyatlarga ega. O‘zbek sintaksisining asosiy xususiyatlari quyidagicha:

O‘zbek tilida gaplar odatda “Ega – To‘ldiruvchi – Kesim” (ETK) tartibida tuziladi (masalan, “Men kitob o‘qiyman” gapida). O‘zbek tilida so‘zlar ko‘pincha qo‘shimchalar yordamida o‘zgaradi, ushbu qo‘shimchalar so‘zning ma’nosini va sintaktik rolini aniqlaydi. O‘zbek tilida so‘zlar orasidagi bog‘lanishlar ham ko‘pincha qo‘shimchalar orqali amalga oshiriladi, shuning uchun so‘z tartibi nisbatan erkin. O‘zbek sintaksisini boshqa tillar bilan taqqoslaganda bir qator qiziqarli o‘xshashliklar va farqlarni ko‘rish mumkin. Ingliz tilida ham ETK tuzilmasi mavjud, lekin o‘zbek tilidan farqli o‘laroq, fe’lning o‘rni va shakli qat’iyroqdir. Ingliz tilida so‘zlar orasidagi bog‘lanish asosan yordamchi fe’llar va kelishiklar orqali amalga oshiriladi.

Korpusning rasmiy formati (CONLL-U)

Gaplar bir yoki bir nechta so‘z qatorlaridan tashkil topadi va so‘z qatorlari quyidagi maydonlarni o‘z ichiga oladi:

1. **ID:** So‘z indeksi, har bir yangi gap uchun 1 dan boshlanadigan butun son; ko‘p so‘zli birliklar uchun oraliq bo‘lishi mumkin; bo‘sh tugunlar uchun kasr son bo‘lishi mumkin (kasr sonlar 1 dan kichik bo‘lishi mumkin, lekin 0 dan katta bo‘lishi shart).

2. **FORM:** So‘z shakli yoki tinish belgisi.

3. **LEMMA:** So‘z shaklining lemmasi yoki o‘zagi.

4. **UPOS:** so‘z turkumi belgisi.

5. **XPOS:** Ixtiyoriy tilga xos (yoki istalgan iyerarxik daraxt modeliga xos) so‘z turkumi / morfologik belgi; mavjud bo‘lmasa, pastki chiziq qo‘yiladi.

6. **FEATS (XUSUSIYATLAR)**: Xususiyatlar ro‘yxatidan yoki belgilangan tilga xos kengaytmadan olingan morfologik xususiyatlar ro‘yxati; mavjud bo‘lmasa, pastki chiziq qo‘yiladi.

7. **HEAD(BOSH)**: Joriy so‘zning boshi, bu ID qiymati yoki nol (0) bo‘lishi mumkin.

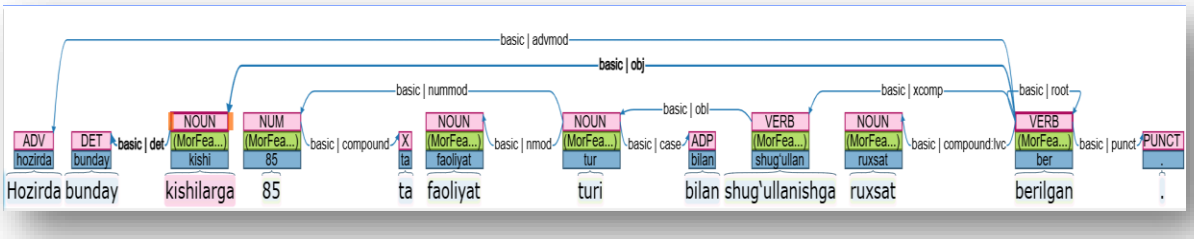
8. **DEPREL**: BOSH ga nisbatan bog‘liqlik munosabati (agar BOSH = 0 bo‘lsa, ildiz) yoki uning belgilangan tilga xos kichik turi.

9. **DEPS**: Bosh-bog‘liqlik juftliklari ro‘yxati ko‘rinishidagi kengaytirilgan bog‘liqlik grafigi.

10. **MISC** (BOSHQA): Har qanday boshqa izoh.

```
# text = Hozirda bunday kishilarga 85 ta faoliyat turi bilan shug‘ullanishga ruxsat berilgan.
1	Hozirda	hozirda	ADV	_	_	11	advmod	_	_
2	bunday	bunday	DET	_	_	3	det	_	_
3	kishilarga	kishi	NOUN	_	Case=Dat|Number=Plur	11	obj	_	_
4	85	85	NUM	_	NumType=Card	7	nummod	_	_
5	ta	ta	X	_	_	4	compound	_	_
6	faoliyat	faoliyat	NOUN	_	Case=Nom|Number=Sing	7	nmod	_	_
7	turi	tur	NOUN	_	Case=Nom|Number=Sing	9	obl	_	_
8	bilan	bilan	ADP	_	_	7	case	_	_
9	shug‘ullanishga	shug‘ullan	VERB	_	Case=Dat|VerbForm=Vnoun	11	xcomp	_	_
10	ruxsat	ruxsat	NOUN	_	Case=Nom|Number=Sing	11	compound:lvc	_	_
11	berilgan	ber	VERB	_	Tense=Past|VerbForm=Fin|Voice=Pass	0	root	_	SpaceAfter=No
12	.	PUNCT	_	_	11	punct	_	_
```

3-rasm. Bitta jumlaning CONLL-U rasmiy formatida tahlil jarayonida ko‘rinishi.

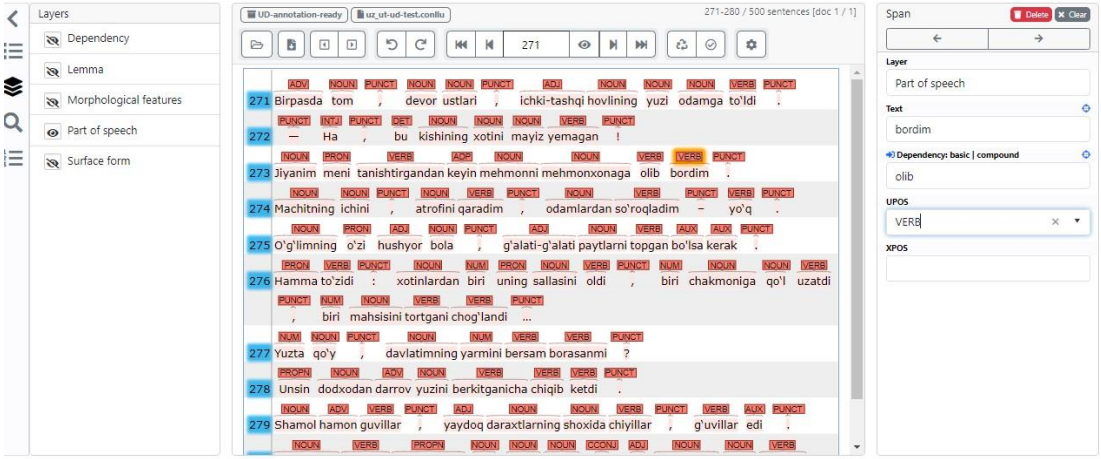


4-rasm. Jumlaning iyerarxik daraxti shaklidagi tahlil ko‘rinishi.

So‘z turkumlarini aniqlash va tizimga kiritish jarayoni ham mazkur bo‘limda yoritib berilgan. O‘zbek tili korpusida so‘z turkumlarini (POS) teglash matndagi har bir so‘zga grammatik toifalar yoki nutq qismlarini belgilash jarayonidir. Jumladagi har bir so‘z ot, fe‘l, sifat, qo‘shimcha va hokazo ekanligini ko‘rsatadigan teg bilan teglanadi[14]. POS belgilari tabiiy tilni qayta ishlashda

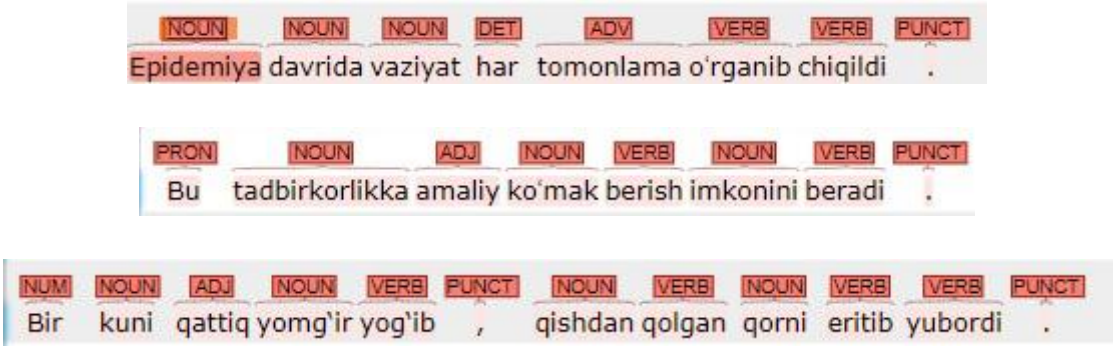
[14] Zeman D. Reusable Tagset Conversion Using Tagset Drivers / In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC ’08). In Marrakech. – Morocco, 2008. – P. 213-218.

jumlalarning tuzilishi va ma’nosi haqida tushuncha beradigan muhim dastlabki ishlov berish bosqichidir.



5-rasm. Iyerarxik korpusda gapdagi so‘zlarning so‘z turkumlari (POS)ni belgilab olinishi va korpusdagi umumiy ko‘rinishi.

So‘z turkumlari teggerlari yordamida POS teglarini tayinlash: POS teglari so‘z morfologiyasi, tarkibi va sintaksis asosida jumla qismlarini aniqlashga o‘rgatilgan maxsus belgilardir. Ba’zi umumiy POS teglari quyidagilarni o‘z ichiga oladi: Ot (Noun), fe’l (Verb), sifat (Adjective), olmosh (PRN), predlog (PREP) va boshqalar. POS teggerlari oldindan taxmin qilish uchun ko‘pincha lingvistik qoidalar, mavjud modellar yoki mashinani o‘rganishdan foydalanadi.



6-rasm. Gapdagi so‘zlarning turkumlarini teglar (POS tags) bilan ajratish.

So‘z turkumlarini teglashdagi usullar.

Qoidalarga asoslangan teglash: Kontekstga asoslangan teglarni belgilash uchun qo‘lda yaratilgan grammatik qoidalardan foydalanadi (masalan, agar so‘z maqoladan keyin bo‘lsa, u ot bo‘lishi mumkin); Stoxastik (statistik) teglash: Yashirin Markov modellari (HMMs) yoki shartli tasodifiy maydonlar (CRFs) kabi ehtimollik usullaridan foydalanib, so‘zning POS yorlig‘i ehtimolini uning kontekstiga asoslangan holda hisoblash orqali teglarni tayinlaydi;

Mashinali o‘rganish / chuqur o‘rganishga asoslangan teglash: Neyron tarmoqlar, ayniqsa, takroriy tarmoqlar va transformatorlar (masalan, BERT) POS

teglarini bashorat qilish uchun katta izohli korpuslarda o‘qitiladi. Ushbu modellar tildagi murakkab obyektlarni aniqlashni o‘rganadi va adekvat xulosalar chiqaradi.[15]

NLP kutubxonalari bilan POS tegiga misollar.

2-jadval.

Python-ning NLTK kutubxonasidan foydalanishda namuna

```
import nltk
nltk.download('averaged_perceptron_tagger')
jumla = “Chaqqon, jigarrang tulki dangasa itning ustiga sakraydi”
tokens = nltk.word_tokenize(sentence)
pos_tags = nltk.pos_tag(tokens) print(pos_tags)
   # Output: ('chaqqon', 'JJ'), ('jigarrang, 'JJ'), ('tulki', 'NN'), ('sakraydi', 'VBZ'),
                    ('ustiga', 'IN'), ('dangasa', 'JJ'), ('it', 'NN')]
```

Tadqiqotning III bobi “**O‘zbek tilidagi gaplarning iyerarxik tahlili korpusining samaradorligi**” deb nomlangan. Ushbu bo‘limda o‘zbek tilidagi jumlalarning iyerarxik tahlilini o‘zida mujassamlashtiruvchi va namoyish etuvchi korpusning yakuniy ko‘rsatkichlari, ishlashi, sifati va qamrovi baholangan.

O‘zbek tili uchun tadqiq qilingan iyerarxik korpusi 1200 ta sintaktik jihatdan belgilangan sodda-murakkab aralash gaplarni o‘z ichiga oladi. Korpus tarkibida online yangiliklardan iborat vebsite ma’lumotlari va badiiy adabiyot matnlari namunalari mavjud bo‘lib, ular tegishli ravishda 700 va 500 gapdan iborat. 3-jadvalda korpusning asosiy hajm ko‘rsatkichlari keltirilgan.

3-jadval.

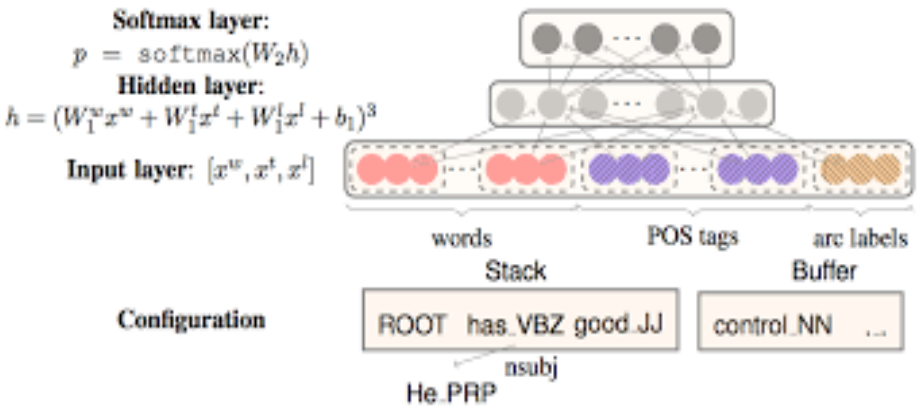
Yaratilgan korpusdagi gaplaring jami soni va ularning tahlil ko‘rsatkichlari.

Ko‘rsatkich	Qiymat
Jami gaplar soni	1200 (yangilik: 700; badiiy: 500)
Jami tokenlar soni	18 000
O‘rtacha gap uzunligi	15 ta token
UD sintaktik bog‘lanish turlari	33 (to‘liq qamrov)
UD so‘z turkumlari (POS)	17 (barchasi mavjud)

O‘tishga-asoslangan (Transition-based) iyerarxik bog‘liqlik parsing usuli – matnni tahlil qilishda so‘zlar orasidagi sintaktik bog‘lanishlarni ketma-ket amallar (shift, reduce, arc-left, arc-right) orqali aniqlovchi sun’iy intellekt modellar oilasiga mansub bo‘lib (7-rasm), u quyidagicha jarayonda ishlaydi:

Boshlang‘ich stek (Stack), bufer (Buffer) va o‘tish (Transition) amallari orqali bajariladi. Har bir qadamda sintaktik bog‘liqliklar ketma-ket ravishda aniqlanadi.

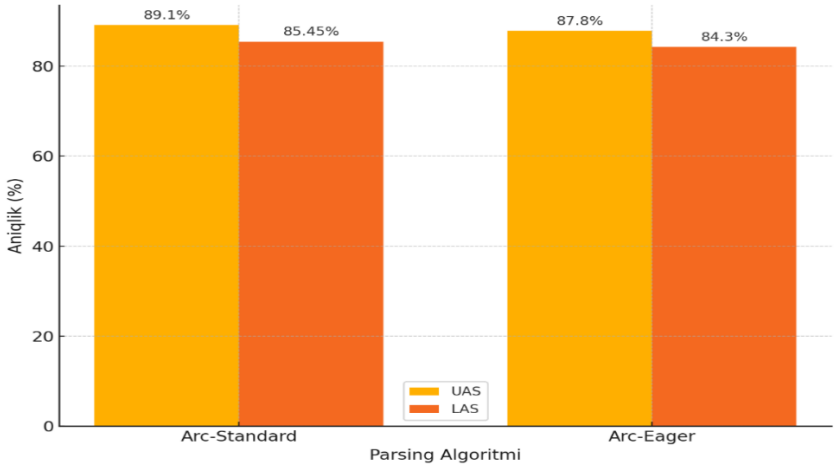
[15] Zimbra D., Abbasi A. and Zeng D. The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation, ACM Transactions on Management Information Systems, 2018.Vol. XX, No. X, 9.



7-rasm. O‘tishga-asoslangan(Transition-based) iyerarxik bog‘liqlik parsing sun’iy intellekt modeli.

Asosiy baholash mezonlari quyidagilardan tashkil topgan:

- **UAS** *(Unlabeled Attachment Score)* – To‘g‘ri aniqlangan bog‘lanishlar ulushi (teglarsiz).
- **LAS** *(Labeled Attachment Score)* – Yorliq bilan to‘g‘ri aniqlangan bog‘lanishlar foizi.



8-rasm. Sun’iy intellekt modelining Arc Standard (Standart yoy) va Arc Eager (Yoysimon Shoshqin model) UAS/LAS baholash natijalari.

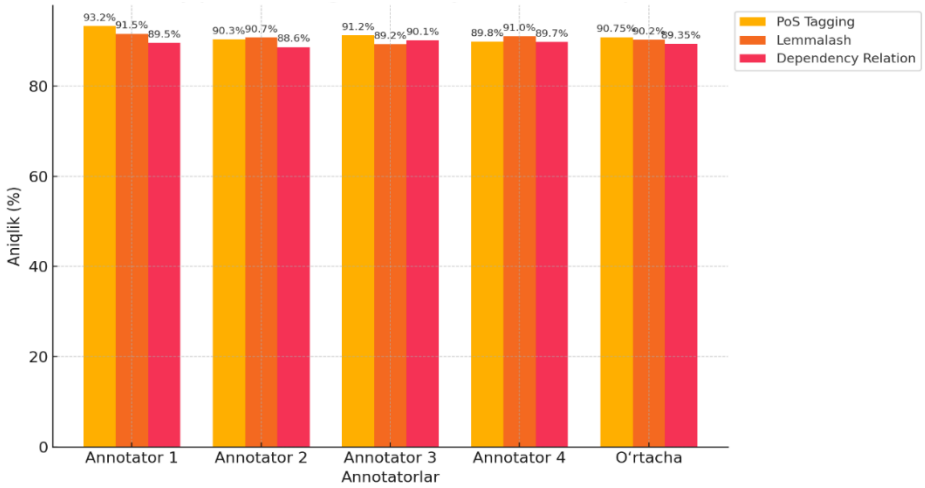
Yuqoridagi rasmda O‘tishga-asoslangan (Transition-based dependency parsing) algoritmlaridan Arc-Standard va Arc-Eager ning samaradorligi **UAS** va **LAS** mezonlari orqali solishtirilmoqda.

UAS (Unlabeled Attachment Score). Bog‘liqlikni turini hisobga olmay, faqat to‘g‘ri yoki noto‘g‘ri ekanligini aniqlaydi. Bu ko‘rsatkich bo‘yicha Arc-Standard algoritmi (89.10 %) Arc-Eager algoritmidan (87.80 %) yuqoriroq natija ko‘rsatmoqda.

LAS (Labeled Attachment Score). Bog‘liqlikni aniqlash bilan birgalikda, uning turini ham aniqlaydi. Arc-Standard algoritmi bu ko‘rsatkich bo‘yicha ham (85.45 %) Arc-Eager algoritmidan (84.30 %) yaxshiroq natijani berdi.

Arc-Standard algoritmi. Arc-Standard stek (stack) asosli algoritm bo‘lib, SHIFT, LEFT-ARC, RIGHT-ARC kabi o'tish (transition) amallaridan foydalanadi. Har bir amal faqat stekdagi yuqori elementlar bilan bajariladi. Arc-Standard, odatda, aniq va yuqori sifatli parsing natijalarini beradi.

Arc-Eager algoritmi. Arc-Eager algoritmi ham o'tish (transition) asosli parsing algoritmi hisoblanadi. Arc-Standard dan farqi, Arc-Eager algoritmi bog‘liqliklarni imkon qadar erta aniqlashga harakat qiladi va LEFT-ARC hamda RIGHT-ARC amallarini boshqacharoq amalga oshiradi. Arc-Eager, odatda, tezroq ishlaydi, ammo ba’zi holatlarda aniqligi Arc-Standard ga nisbatan biroz pastroq bo‘ladi.



9-rasm. Annotatorlarning ish samaradorligi baholash.

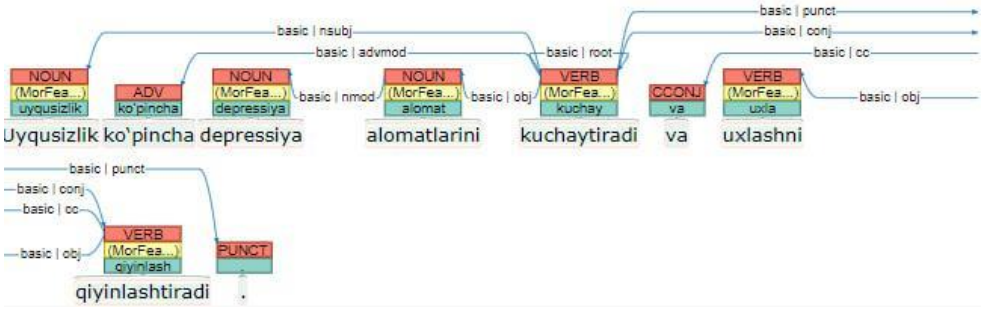
Annotatsiya jarayonining sifatini tekshirish uchun annotatorlarning ish samaradorligi quyidagi asosiy ko‘rsatkichlar orqali baholandi:

PoS Tagging (So‘z turkumini aniqlash) Annotatorlar tomonidan so‘zlarning turkumini aniqlashdagi umumiy o‘rtacha aniqlik **90.75 %** bo‘ldi. Bu annotatorlarning so‘z turkumlarini aniqlashda yuqori darajada bir xil fikrda ekanligini ko‘rsatadi.

Lemmalash (So‘zlarning lug‘aviy shaklini aniqlash) Annotatorlarning so‘zlarni lemmalashdagi umumiy aniqligi **90.20 %**ni tashkil qildi. Bu annotatorlar orasida so‘zlarning bazaviy shakllarini aniqlashda yaxshi kelishuv mavjudligini bildiradi.

Bog‘liqlik munosabatlarini aniqlashda o‘rtacha aniqlik darajasi **89.35 %** bo‘ldi. Ushbu natija annotatorlarning sintaktik bog‘lanishlarni belgilashda aniq va yuqori sifatli annotatsiya qilish imkoniyatiga ega ekanligini ko‘rsatmoqda.

Annotatsiya jarayonida foydalanilgan dasturiy platforma orqali olingan bu yuqori natijalar annotatsiya sifatini va annotatsiya jarayonining muvaffaqiyatli o'tganligini tasdiqlaydi. Bu o‘zbek tili uchun yaratilayotgan dependency treebankning kelgusida ishonchli foydalanishga tayyorligini anglatadi.



10-rasm. Yaratilgan korpusda iyerarxik tahlil qilingan gaplarning alohida va yaxlit holdagi parsing ko‘rinishi.

Root (asosiy bog‘liqlik nuqtasi) sifatida “*qildi*” fe’li tanlangan. Bu jumlaning asosiy harakati hisoblanadi va barcha boshqa bo‘laklar shu so‘zga bog‘lanadi.

Asosiy bog‘liqliklar quyidagicha:

- *qildi* (root, VERB) – fe’l;
- *murabbiy* (nsubj, NOUN) – fe’lning bajaruvchisi (subyekti);
- *yordamchi* (nmod, NOUN) “*murabbiy*” so‘ziga bog‘langan aniqlovchi (qanday murabbiy? – *yordamchi* murabbiy);
- *bahsni* (obj, NOUN) – fe’lning obyekti (harakat nimaga yo‘naltirilgan?);
- *oldingi* (amod, ADJ) “*bahsni*” so‘ziga bog‘langan sifat (qaysi bahs? – *oldingi* bahs);
- *yaxshilab* (advmod, ADV) – fe’lga bog‘langan ravish (qanday tahlil qildi? – *yaxshilab*);
- *tahlil* (compound: lvc, NOUN) – fe’l bilan birgalikda qo‘shma fe’l hosil qiladi (*tahlil qildi*);
- *aytdi* (conj, VERB) – asosiy fe’lga (*qildi*) teng bog‘langan ikkinchi harakat (*qildi* va *aytdi*);
- *va* (cc, CCONJ) **–** bog‘lovchi so‘z sifatida (*qildi* va *aytdi*);
- *ko‘rsatmalarini* (obj, NOUN) "aytdi" fe’lining obyekti (nimani aytdi? – ko‘rsatmalarini);
- *o‘z* (nmod, PRON) "ko‘rsatmalarini" so‘ziga tegishlilikni bildiradi (kimning ko‘rsatmalari? – o‘zining);
- . (punct, PUNCT) – gap yakunini ko‘rsatuvchi tinish belgisi.

XULOSA

1. Morfo-sintaktik xususiyatlarning aniqlanishi**:** O‘zbek tili morfologik jihatdan boy bo‘lgani uchun korpusni tuzishda turli morfologik va sintaktik belgilarni, masalan, fe’lning zamon, shaxs, holatlar kabi belgilarini to‘g‘ri aniqlash muhim ahamiyat kasab etdi. Iyerarxik tahlil jarayonida, asosan, ot, sifat, fe’l, va boshqa asosiy so‘z turkumlariga oid xususiyatlar batafsil belgilanib, ularning sintaktik tuzilmadagi o‘rni aniqlab olindi.

2. Korpusda ishlatilgan asosiy universal so‘z turkumlari teglarining tanlovi: O‘zbek tili iyerarxik darxti korpusida Universal bog‘liqliklar UD) standartlariga asoslangan teglar (labels) qo‘llanildi. Masalan, *nsubj* (nominal subyekt), *obj* (obyekt), *advmod* **(**adverbial modifier) kabi umumiy sintaktik munosabatlarni ifodalovchi teglar ishlatildi. Bunda turkiy tillarga xos ayrim tuzilmalarning universal bog‘liqlik tizimida qanday ifodalanishi sinovdan o‘tkazildi va natijalar hujjatlashtirildi.

3. O‘zbek tilining erkin so‘z tartibini aks ettirish: O‘zbek tilida so‘z tartibi nisbatan erkin bo‘lganligi sababli daraxtsimon tahlilda aynan so‘z tartibidagi o‘zgarishlar sintaktik bog‘lanishlarga qanday ta’sir ko‘rsatishi kuzatildi. Buning natijasida negiz va unga bog‘liq so‘zlar o‘rtasidagi sintaktik munosabatlarni to‘g‘ri aks ettirishga erishildi.

4. Annotatsiya jarayonida yuzaga kelgan qiyinchiliklar: O‘zbek tili uchun daraxtsimon tahlil korpusini tuzishda ba’zi murakkab tuzilmalarda (masalan, kesim va hol birliklari, yoki mustaqil so‘zlashuv birliklari) annotatsiya qilish qiyinchilik tug‘dirdi. Ayniqsa, turli morfologik shakllarning turlicha semantik vazifalari mavjud bo‘lib, ularni to‘g‘ri belgilash uchun ba’zi holatlarda qo‘shimcha ishlanmalar kiritishi talab qilindi.

5. Qo‘shimcha sifatida leksik xususiyatlarni aniqlash: O‘zbek tilidagi ayrim o‘ziga xos morfologik va sintaktik shakllarni to‘g‘ri aks ettirish uchun leksik xususiyatlar, masalan, yordamchi fe’llar (auxiliary verbs) va qaratqichli so‘zlar maxsus qoidalar asosida qayta tahlil qilindi. Buning natijasida korpusdagi ayrim so‘zlarning funksional roli yanada aniqroq aks ettirildi.

6. Yuqori sifatli iyerarxik daraxti korpusining ahamiyati: O‘zbek tili uchun yuqori sifatli daraxtsimon tahlil korpusini yaratish o‘zbek tilining sintaktik tuzilmasini yanada chuqur o‘rganish imkoniyatini beradi. Ushbu korpus boshqa turkiy tillarni o‘rganishda yoki turkiy tillar orasidagi umumiy sintaktik qoidalarni aniqlashda ham foydali bo‘lishi mumkin.

7. Kelgusida takomillashtirish imkoniyatlari: O‘zbek tili iyerarxik korpusini yanada kengaytirish va boshqa soha yoki uslubdagi matnlar bilan boyitish kelgusida ushbu korpusning qamrovini yanada oshiradi. Shu bilan birga, korpusda mavjud bo‘lmagan yoki hozirgi tahlil (annotatsiya) qoidalariga mos kelmaydigan strukturalarni kiritish imkoniyati ham mavjud.

8. Tabiiiy tillar jarayonida morfologik xususiyatlarni teglash belgilaridan 17 ta standartlashtirilgan teg belgilari mavjud bo‘lib, ulardan asosan 15 tasi ishlatilgan. Teg belgilar: ADJ, ADP, ADV, CCONJ, DET, INTJ, NOUN, NUM, PART, PRON, PROPN va SCONJ ni o‘z ichiga olgan. Shu bilan birga, AUX (yordamchi), SYM (belgi) va X (boshqa) belgilarini ham ayrim o‘rinlarda uchratish mumkin.

9. Tabiiy tillar jarayonida standartlashtirilan 16 ta teglash belgilari hujjatlashtirilgan. Ushbu iyerarxik daraxt korpusida 15 ta oldindan belgilangan morfologik xususiyatlardan foydalanilgan: Aspect, Kelishik (Case), Clusivity, Degree, Deixis, Mood, Neutral, Number, PartType, Person, Polarity, PronType va Voice. Bundan tashqari, ikkita teglar o‘zbek tili uchun o‘ziga xos bo‘lganligi sababli kiritilgan, xususan, aniqlovchi suffiks va neytral ko‘rsatkich suffiksi.

10. O‘zbek tili korpusini annotatsiya qilishni avtomatlashtirish imkoniyatlarini, masalan, qoidalar asosida ishlaydigan algoritmni o‘rganish muhim ahamiyat kasb etadi. Qoidalar asosidagi ushbu algoritm mazkur o‘zbek tili iyerarxik tahlil korpusi uchun ko‘proq mos kelishi mumkin, chunki u statistik algoritmlar bilan solishtirganda kamroq miqdordagi ma’lumot talab qiladi.

SCIENTIFIC COUNCIL DSc.03/25.08.2021.Fil.01.16 ON AWARDING SCIENTIFIC DEGREES AT NATIONAL UNIVERSITY OF UZBEKISTAN NAMED AFTER MIRZO ULUGBEK

NATIONAL UNIVERSITY OF UZBEKISTAN

RAXIMBOYEVA XULKAR G‘AYRATOVNA

CREATING DEPENDENCY PARSING TREEBANK FOR UZBEK LANGUAGE

10.00.11 – Theory of language. Applied and computational linguistics

ABSTRACT OF DISSERTATION OF THE DOCTOR OF PHILOSOPHY (PhD) IN PHILOLOGICAL SCIENCES

Tashkent – 2025

The theme of the dissertation of Doctor of Philosophy (PhD) was registered at the Supreme Attestation Commission under the Ministry of Higher Education, Science and Innovation of Republic of Uzbekistan under the number B2022.2.PhD/Fil2702.

The dissertation has been carried out at the National Universuty of Uzbekistan named after Mirzo Ulugbek.

The abstract of the dissertation is posted in three languages (Uzbek, English, Russian (resume)) on the Scientific Council's website (www.nuu.uz) and on the information and educational portal "ZiyoNet" (www.ziyonet.uz).

Scientific supervisor:	**Sadullayeva Nilufar Azimovna** Doctor of philological Sciences, Professor
Official opponents:	**Abdurakhmonova Nilufar Zaynobiddin kizi** Doctor of philological Sciences, Professor
	Toirova Guli Ibragimovna Doctor of philological Sciences, Professor
Leading organization:	**Urgench State University**

The defense will take place on _____, 2025 at _____ at the meeting of the Scientific Council awarding Scientific Degrees DSc.03/25.08.2021.Fil.01.16 at National University of Uzbekistan. (Address: 100174, Tashkent city, Almazar district, University Street, 4. the faculty of Uzbek philology, 1st floor. Phone: (99871) 246-02-24; fax: (99871) 246-02-24; e-mail: nauka@nuu.uz.).

The dissertation can be reviewed at the Information Resource Centre of National University of Uzbekistan (registered under the number ___). Address: 100174, Tashkent city, University Street, 4. Phone: (99871) 246-02-24; fax: (99871) 246-02-24.

The abstract of the dissertation was distributed on _____, 2025.

(Protocol of the register № _____ on _____, 2025).

N.A.Rahmanov
Chairman of the one-time scientific council for awarding Scientific degrees, Doctor of philological sciences, professor

M.B.Khujamkulova
Scientific secretary of the one-time scientific council for awarding Scientific degrees, Doctor of philosophy on philological sciences, associate professor

A.E.Mamatov
Chairman of the one-time scientific seminar at the council for awarding Scientific degrees, Doctor of philological sciences, professor

INTRODUCTION (abstract of PhD dissertation)

Relevance and necessity of the dissertation topic. It is known that various sectors in countries around the world are directly influenced by the information age, and the significance of communication and information is extremely high. At the same time, computer technologies have penetrated all spheres of life, playing a crucial role in achieving the goals of the new era. In the context of globalization, certain difficulties arise from the differences between machine language and human language. Moreover, the unique characteristics of human language increasingly necessitate addressing these issues. From this perspective, computational linguistics holds great importance.

In global linguistics, hierarchical analysis corpora are essential for conducting in-depth analysis of the syntactic structure of texts. To promote language skills and present a language to the world, its theoretical and scientific application alone is not sufficient – it is also vital to introduce the language through modern and relevant natural language processing (NLP) resources. Computational linguistics studies language using artificial intelligence models and involves the analysis of large text collections (corpora).

Today, in our country as well, fields such as corpus linguistics and artificial intelligence within the realm of computer science are still in the development stage, with many aspects yet to be explored. Linguistic and natural language processing resources in this area remain significantly limited. By creating a structured dependency parsing treebank (with syntactic and morphological analysis) that covers a sufficient number of sentences, it becomes possible to advance linguistic research on the Uzbek language and increase its recognition within the international scientific community.

This dissertation contributes to the implementation of tasks outlined in the following decrees and resolutions of the Republic of Uzbekistan: Presidential Decree No. PF-6079 (October 5, 2020) – On the Approval and Effective Implementation of the "Digital Uzbekistan – 2030" Strategy; Presidential Resolution No. PQ-4996 (February 17, 2021) – On Measures to Create Conditions for the Rapid Implementation of Artificial Intelligence Technologies; Presidential Decree No. PF-5850 (October 21, 2019) – On Measures to Fundamentally Enhance the Prestige and Status of the Uzbek Language as the State Language; Presidential Decree No. PF-6084 (October 20, 2020) – On Measures for the Further Development of the Uzbek Language and the Improvement of Language Policy. This research also aligns with other relevant regulatory and legal documents in this field: as the president of Uzbekistan has stated, "Enhancing scientific research on the unique characteristics, dialects, historical development, and future prospects of the Uzbek language" [16] is one of the country's current priorities.

Relevance of the research to the priority areas of science and technology development in Republic. The research corresponds to the priority directions for the

[16] O‘zbekiston Respublikasi Prezidenti Sh.M.Mirziyoyevning 2019 yil 21 oktyabrdagi “Milliy o‘zligimiz va mustaqil davlatchiligimiz timsoli” mavzusida o‘zbek tiliga davlat tili maqomi berilganining o‘ttiz yilligiga bag‘ishlangan tantanali marosimdagi nutqi // Xalq so‘zi, 2019 yil, 22 oktyabr.

development of science and technology in the Republic of Uzbekistan, specifically aligning with the priority area of "Developing an Information Society and Advancing a Democratic State through Social, Legal, Economic, Cultural, Educational, and Spiritual Development, as well as Fostering an Innovative Economy."

Level of study of the problem. In world linguistics, the coverage of the topic of Universal Dependency is associated with the names of J.Nivre, D.Zeman, de Marneffe, D.Manning, De Marneffe, and Daniel Jurafsky[17]

Scientific research on Uzbek computer linguistics in our republic, led by A.Q. Pulatov, M.M. Aripov, A.E. Mamatov, N.Z. Abdurakhmanova, M.M. Kurbanova, M.M. Musaev, N.A. Sadullaeva, Sh.A. Nazirov, M.H. Hakimov, O. Hamdamov, N.A. Ignatev and G.R. Matlatipov[18] is focused on creating a formal model of Uzbek grammar, formalizing sentence structures in Uzbek, Turkish and Karakalpak languages, creating dictionaries for Turkic languages, modeling the processes of speech signal formation, creating speech synthesizers, and creating logical-linguistic and mathematical models of language objects for multilingual modeled machine translation.

The famous Uzbek scientist, Doctor of Physical and Mathematical Sciences, Professor Abdumajid Pulatov has published for many years the following books on linguistics: "English Language", "World Uzbek Language. Verb forms in Uzbek and their manifestations in Russian and English", "Computer Linguistics", "English Language - for Independent Learners" and others [19],, the textbook "Fundamentals of Linguistics"[20] by Sharipov O., Yuldoshev I., the textbook "Fundamentals of Computer Linguistics"[21] by A. Rahimov, the textbook "Technologies of Machine Translation" [22], by G.R. Matlatipov, and the textbook "Corpus Linguistics" by N.Z. Abdurakhmanova.

---

[17] De Marneffe, Christopher D. Manning, Joakim Nivre, Daniel Zeman. Universal Dependencies. Computational Linguistics: Volume 47, Issue 2, 2021. – P. 255. – URL: https://direct.mit.edu/coli/article/47/2/255/98516/ Universal-Dependencies. 59 Nivre J., Zeman D.;
Nivre J., Zeman D. Universal Dependencies: A Cross-Linguistic Perspective on Grammar and Lexicon // Proceedings of the Workshop on Grammar and Lexicon: Interactions and Interfaces (GramLex 2016). – 2016. – P. 8–17. – URL: https://aclanthology.org/W16-3806.pdf.
D. Jurafsky, JH. Martin. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition: Prentice Hall, 2000

[18] N.Z.Abdurakhmonova, K Urdishev., Corpus Based Teaching Uzbek As A Foreign Language.-Tashkent: Journal of Foreign Language Teaching and Applied Linguistics.2019.// Nilufar Abdurakhmonova, Ismailov Alisher, Rano Sayfulleyeva., MorphUz: Morphological analyzer for the Uzbek language.- 2022 7th International Conference on Computer Science and Engineering (UBMK). 2022/9/14.pp. 61-66.// N Abdurakhmonova., Dependency Parsing Based On Uzbek Corpus, Proceedings of the International Conference on Language technology for all (LT4all), 2019.// Nilufar Z Abdurakhmonova, Alisher S Ismailov, Davlatyor Mengliev., Developing NLP tool for linguistic analysis of Turkic languages, 2022 IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON).// Nilufar Abdurahmonova., Kompyuter lingvistikasi. Nodirabegim. Toshkent, 2021.
Mirsaid Aripov, Bibigul Razzoqova Altinbay Sharibbek, Nilufar Abdurakhmonova., Ontology of grammar rules as example of Noun the Uzbek language and Kazakh languages, VI international scientific conference "Modern problems of the applied mathematics and information technology - Al Khorezmiy -2018". Pp37-38// Mirsaid Aripov, Baxodir Begalov, Uzoqboy Begimqulov, Mirsalim Mamarajabov., Axborot texnologiyalari, oshkent-2009..
A.Polatov, S.Muhammedova. Kompyuter lingvistikasi. -T.,2008; A.Po'latov. Kompyuter lingvistikasi. -T., 2011: A.Rahimov. Kompyuter lingvistikasi asoslari. –T.,2011,

[19] A.Po'latov. Kompyuter lingvistikasi. -T., 2011:

[20] Sharipov O'.. Yo'ldoshev I. Tilshunoslik asoslari,-Toshkent: Nizomiy nomidagi Toshkent Davlat pedagogika universiteti, 2006.

[21] A.Rahimov. Kompyuter lingvistikasi asoslari. –T.,2011,

[22] G'.R.Matlatipov, "Mashina tarjimasi texnologiyalari". O'quv qo'llanma, 2024.

Moreover, a number of studies have been conducted around the world on the development of computer-oriented models of natural languages, including the work of N.A. Chomsky (USA), Christopher D. Manning (USA), D. Juraffsky (USA), Carlos Gómez-Rodríguez (Spain), Joakim Nivre, Miguel A. Alonso (Spain), Marco Kuhlmann (Sweden), A. James, G. Fant, R. Delmonti [23] and others. Also, the scientific and methodological work of scientists working in the field of computer linguistics in the countries of the Commonwealth of Independent States, such as Ye.I. Bolshakova and A. Sharipbay, [24] was studied.

The purpose of the dissertation is to build a treebank relationship tree like corpus for processing text data stored electronically in the Uzbek language, evaluate it using an artificial intelligence model, and create software.

The tasks of the research consists of the following:

to prepare a guideline for annotating the dependency parsing treebank of the Uzbek language, including lemmatization, morphological tagging, identification of sentence constituents, and development of dependency rules;

to collect a corpus database composed of simple and complex sentences in the Uzbek language and performing pre-processing;

to identify the lemma of each word in the sentences included in the Uzbek corpus, analyzing their morphological features, and using an automatic parser to generate results and present a complete dependency parsing analysis corpus;

to determine the syntactic properties of words in the sentences of the Uzbek corpus, constructing their dependency parsing, and analyzing subordination using an artificial intelligence model.

The object of the research. Texts selected from the websites *Kun.uz* and *Daryo.uz.*

The subject of the research. Creating a treebank (hierarchical) analysis corpus of the Uzbek language, the content, principles, methods, forms, grammatical connections, analysis algorithms, and annotation processes.

[23] Carlos Gómez-Rodríguez, Joakim Nivre., A transition-based parser for 2-planar dependency structures, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010. PP1492-1501.
Joakim Nivre, Laura Rimell, Ryan McDonald, Carlos Gomez-Rodriguez., Evaluation of dependency parsers on unbounded dependencies. Proceedings of the 23rd international conference on computational linguistics. 2010. PP833-841.;Joan D Gonzalez-Franco, Jorge E Preciado-Velasco, Jose E Lozano-Rizk, Raul Rivera-Rodriguez, Jorge Torres-Rodriguez, Miguel A Alonso-Arevalo., Comparison of Supervised Learning Algorithms on a 5G Dataset Reduced via Principal Component Analysis (PCA)- Future Internet. 2023. P335.;Ralph Debusmann, Denys Duchier, Alexander Koller, Marco Kuhlmann, Gert Smolka, Stefan Thater., A relational syntax-semantics interface based on dependency grammar, COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics. 2004.PP 176-182.; Dann Juraffskiy., James H.Martin. Speech and language processing. -New Jersey: Upper Saddle River, February 2008.; Ch. Manning., H. Schutze. Foundations of Statistical Natural language Processing, MIT Press. Cambridge, MA: May 1999.; Andrew Ng. Machine Learning yearning, technical strategy for AI Engineers in the Era of Deep Learning, August 2023.

[24] EI Bol'shakova, KV Vorontsov, NE Efremova, ES Klyshinskiy, NV Lukashevich, AS Sapin. Avtomaticheskaya obrabotka tekstov na estestvennom yazyke i analiz dannykh [Automatic natural language text processing and data analysis].- Moscow: Izd-vo NIU VShE Publ, 2017.; Altynbek Sharipbay, Bibigul Razakhova, Assel Mukanova, Banu Yergesh, Gaziza Yelibayeva., Proceedings of the Second International Conference on Data Science, E-Learning and Information Systems.-Kazakhstan,2019.-1-5p.;

Research methods. To address the research topic, methods such as classification, description, automatic parsing[25] , annotation using guidelines, statistics, transfer and corpus analysis were applied

The scientific novelty of the research is as follows:

a dependency treebank was created in the Uzbek language;

a guideline was created for the annotation, lemmatization, morphological characterization, identification of parts of speech, and development of relational rules for the dependency treebank corpus in the Uzbek language;

the dependency rules and morphological features of the given sentences in the Uzbek language were deeply analyzed and annotated within the corpus;

an artificial intelligence model was built for natural language processing, taking into account the specific features of the Uzbek language, and the reliability of the corpus was evaluated through syntactic analysis performed by the AI model.

The practical result of the research

serves as a key language resource for automatic translation, text analysis, question-answering systems, and voice assistants in the field of natural language processing (NLP) and artificial intelligence in the Uzbek language

provides a precise syntactic database for linguistic research, enabling a deep analysis of the grammatical system of the Uzbek language.

helps in developing interactive tools and educational programs based on linguistic models in the educational process.

Reliability of research results. The obtained results and statements demonstrate the resulting reliability of the corpus through the mathematical evaluation methods UAS (unlabeled attachment score) and LAS (Labeled attachment score).

Scientific and practical significance of research results. Scientific significance – the dependency treebank serves as a basis for linguistics and computational linguistics research, representing the syntactic structure of the Uzbek language in digital form, allowing for a deeper analysis of syntactic phenomena and comparative scientific research with other languages. It represents the syntactic structure of the Uzbek language in digital form, which advances analytical research and contributes to the improvement of language models.

Practical significance – it serves as a necessary resource for the development of natural language processing programs, including automatic translation, chat bots, text content analysis, automatic text summarization, sentiment analysis. This, in turn, increases the quality of intelligent systems, search engines and information retrieval systems working in the Uzbek language, which expands the application of these technologies in everyday life. It also serves as a ready-made resource for developing digital services and software in the Uzbek language, allowing local and global IT companies to create their products with higher quality and user-friendliness.

Implementation of research results. The scientific results of the dissertation on the topic "Creating a dependency parsing treebank for Uzbek language" prepared in the framework of the practical research within the framework of the grant REP-

[25] **Parsing** (in Uzbek: *tahlil qilish*, *ajratish*, *tarkibiy qismlarga ajratish*) refers to the process of identifying the grammatical structure of a given sentence – that is, determining the syntactic relationships and dependencies between the words within it.

25112021/113 - "UzUDT: “Universal dependency treebank corpus and Its Semantic analysis for natural language processing in the Uzbek language” funded by the World Bank project "Modernization of the National Innovation System of Uzbekistan" organized under the Ministry of Innovative Development of the Republic of Uzbekistan were used. In particular, the scientific results of the dissertation were used in the morphological analysis of words in Uzbek texts within the framework of the practical project to perform their lemma, identify hierarchical syntactic connections, tag and evaluate them, annotate and evaluate word groups in the text, and develop a database of morphological features.

In the practical project A-FA-2019-9, titled "Researching Ancient Written Manuscripts and Sources and Creating Their Digital Library," which is being carried out at the Khorezm Mamun Academy, this dissertation was extensively utilized. The findings and conclusions derived from the dissertation were applied in the following areas: Analyzing the creative works of classical literature representatives found in historical texts; Accurately translating and interpreting words and sentences encountered in historical works; Conducting linguistic analysis of literary sources from the 19th century, comparing the linguistic features of earlier and later periods, and identifying distinctive characteristics; Utilizing ancient written manuscripts and rare sources, as well as contributing to the development of a digital library for such materials; Assisting in the creation of the electronic portal and the website ms.mamun.uz for the practical project, including the development of a special interface, search system, and website design.

The results of the dissertation were widely utilized in the Erasmus+ program's grant project №598340-EPP-1-2018-1-ES-EPPKA2-CBHE-JP titled "University Cooperation Framework for Knowledge Transfer in Central Asia and China (UNICAC)." Specifically, the dissertation contributed to the expanding cross-lingual semantic search systems and databases; Developing language learning tools that emphasize syntactic features of different languages through data-driven analysis; Partially standardizing linguistic resources for use in international projects.

Approval of research results. The important ideas and practical results of this research have been approved at 3 international (including 1 Scopus) conferences and 6 republican scientific and practical conferences.

Publication of research results. A total of 14 works on the topic of the dissertation were published, including 5 articles in scientific publications recommended for publication by the Higher Attestation Commission of the Republic of Uzbekistan on the main scientific results of doctoral dissertations, including 4 in republican and 1 in foreign publications.

The structure and scope of the dissertation. Dissertation consists of introduction, three chapters, conclusion, used literature. The volume of the dissertation consists of 133 pages.

THE MAIN CONTENT OF THE DISSERTATION

Introduction establishes the relevance of the dissertation topic, provides a review of foreign scientific research on the topic of the dissertation and the level of

study of the problem, identifies the goals and objectives of the research, as well as its object and subject indicates the correspondence of the research work to the priority areas of development of science and technology, and provides information on the scientific novelty of the research, the reliability of the results, the theoretical and practical significance, the implementation of the results into practice, their publication, and the structure of the work.

The first chapter of the study, entitled **"Basic concepts of dependency parsing treebank analysis"** provides a comprehensive overview of the dependency parsing treebank and its syntactic analysis, context-free grammar, and provides a detailed explanation of Uzbek language technology and methods for creating a dependency parsing treebank, annotation schemes, and electronic tools used, with examples.

With the increasing demand for multilingual natural language processing, interest in Uzbek is also growing. Especially for the morphologically rich Uzbek language, identifying its dependency relationships and semantics is of great importance. Free word order and agglutinative morphology are widespread in Uzbek, and identifying relationships is essential for creating accurate and context-sensitive natural language processing tools. This research explores the importance, methodology, challenges, and areas of application of dependency relationship detection in Uzbek and provides its transformative role in the development of the field of natural language processing for Uzbek speakers.

Treebank dependency is the process of analyzing the grammatical structure of a sentence, in which the relationships between words are determined. In dependency analysis (parsing), each word is associated with a "head" word or adverb, which establishes a dependency relationship that indicates the syntactic role of each word in the sentence. For example, in the sentence "The student is reading a book", the word "book" (book) is associated [26] as the object of the verb "reading" (reading), and "Student" (student) serves as the subject.

Determining dependencies is important for many natural language processing applications, including machine translation, sentiment analysis, and question-answering systems. Determining hierarchical dependencies allows for the accurate interpretation of sentences by providing detailed syntactic structures, and producing logical and contextually appropriate results.

Dependency affix analysis in Uzbek is unique. Uzbek, a member of the Turkic language family, has several features that make affix analysis both difficult and useful.

Agglutinative morphology: Uzbek forms words by adding multiple suffixes to the root. For example, the word "o‘qiyman" (I read) contains the stem "o‘qi", the suffixes -iya, -p, and -man, which indicate tense and subject. Accurate parsing of these structures is essential for understanding the syntactic relationships between words. [27]

Free word order: Uzbek usually follows the Subject-Object-Verb (SOV) order, but this structure can vary significantly between written and spoken forms. This

[26]Shukurov M. O‘zbek tili va grammatikasi. – Toshkent. O‘zbekiston, 2010, 86-87-b.
[27] D.Nurmonova. Xozirgi o‘zbek adabiy tili: sintaksis. Ma’ruzlar matni - Toshkent. 2016. B-287.

freedom requires that affixers be flexible to different sentence structures while maintaining the clarity of syntactic relationships.

Case markers and morphological-syntactic correspondence: Uzbek uses case markers to indicate grammatical roles, such as subject, object, or possessor. These markers should be taken into account when determining relationships in Uzbek, as they directly affect the relationships between words.

Key components of determining relationships in Uzbek: Determining hierarchical relationships in Uzbek requires consideration of the specific structures of the language, detailed annotations, and linguistic definitions.

The following are the key components:

Word class identification: Word classes identify the syntactic category of each word (e.g., noun, verb, adjective). In Uzbek, identifying word classes is the first step, as it provides a basic layer of information for establishing relationships.

Headword selection and dependency relations in sentences: In Uzbek, words in a sentence are dependent on a "head" word, usually the main verb, which serves as the root of the dependency structure. In world practice, there are universal dependency relations, and the application of these rules is very important for Uzbek computer linguistics.

We define the following dependency relations for the Uzbek language:

• nsubj: Identifies the subject of the verb.

• obj: Indicates the object of the verb.

• obl: Identifies objective arguments, usually preceded by words or case modifiers.

• det: Denotes possessive tenses, for example, possessive tenses.

• amod: Adjective modifiers that modify nouns

An example of dependency analysis in Uzbek.

Let`s look at the Uzbek sentence: *My new friend went to school. (Mening yangi do'stim maktabga bordi.*

Table 1.

Example of hierarchical dependency analytics in Uzbek.

So'z	Lemma	POS	Bosh so'z	Iyerarxik bog'lanish
Mening	men	PRON	do'stim	nmod
yangi	yangi	ADJ	do'stim	amod
do'stim	do'st	NOUN	bordi	nsubj
maktabga	maktab	NOUN	bordi	obl
bordi	bormoq	VERB	ILDIZ (ROOT)	ILDIZ (ROOT)

In this example:

• *Mening* (my) is the possessive modifier (nmod) of the word *do'stim* (friend).

• *Yangi* (new) is the adjective modifier (amod) of the word *do'stim* (friend).

• *Do'stim* (friend) is the subject (nsubj) of the main verb *bordi* (went).

• *Maktabga* (school) is the objective modifier (obl) indicating the address.

This example also shows that the following types of problems are observed in dependency parsing relationships in Uzbek:

data scarcity: compared to widely used languages such as developed languages, large and annotated corpuses are almost non-existent in Uzbek, which makes it difficult to train hierarchical parsers with sufficient data. Limited resources often lead to low parsing accuracy in complex sentence structures;

dialect differences: there are several dialects of Uzbek, each with its own lexical and syntactic differences. A parser trained on standard Uzbek may have difficulty processing dialect differences, which requires dialect-specific training data or adaptive models;

morphological complexity: the agglutinative nature of Uzbek allows for extensive word formation through affixes. Each form can significantly change the role of a word in a sentence, making it important for parsers to recognize and process these forms correctly[28];

flexible syntax: flexible word orders are difficult to parse because parsers often rely on stable patterns to determine relationships. In Uzbek, sentences can be grammatically correct in different word orders, which complicates syntactic analysis.

The development of dependency parsing for the Uzbek language requires improved resources and sophisticated methods that focus on the specific features of the language. The dissertation presents the expected future directions of hierarchical parsing in the literature.

Chapter II of the dissertation is entitled "**Methods for creating a dependency parsing treebank of the Uzbek language and its practical and technological aspects**". Three special pre-processing stages were carried out on the corpus:

adding metadata - this is assigning an ID-identification symbol to each example sentence for tracking, clarity and uniqueness, translations, as well as sentences taken from social networks and web pages are identified. The hierarchical tree corpus data, metadata are linked with the numbering symbol (#);

error correction - this includes correcting errors, checking for gaps and filling in missing ones, formatting sentences with punctuation marks. Messages may also contain errors and awkward translations in sentences, which are corrected;

orthographic standardization is a necessary process for the Uzbek language, as the lack of clear spelling rules complicates natural language processing [29].

However, spelling standardization should not be overly rigid; instead, it is important to maintain balance. This process aims to preserve forms that are natural for native Uzbek speakers while avoiding unnecessary complexity in the spelling system. In the development of an Uzbek dependency treebank, changes in lemmas within the tree structure also play a crucial role in maintaining this balance.

[28] Bernd Bohnet. "Very Fast and Accurate Dependency Parsing". University of Stuttgart, Germany, 2010

[29] Bernd Bohnet. Very high accuracy and fast dependency parsing is not a contradiction.-University of Stuttgart, Germany, 2010.P93-94.

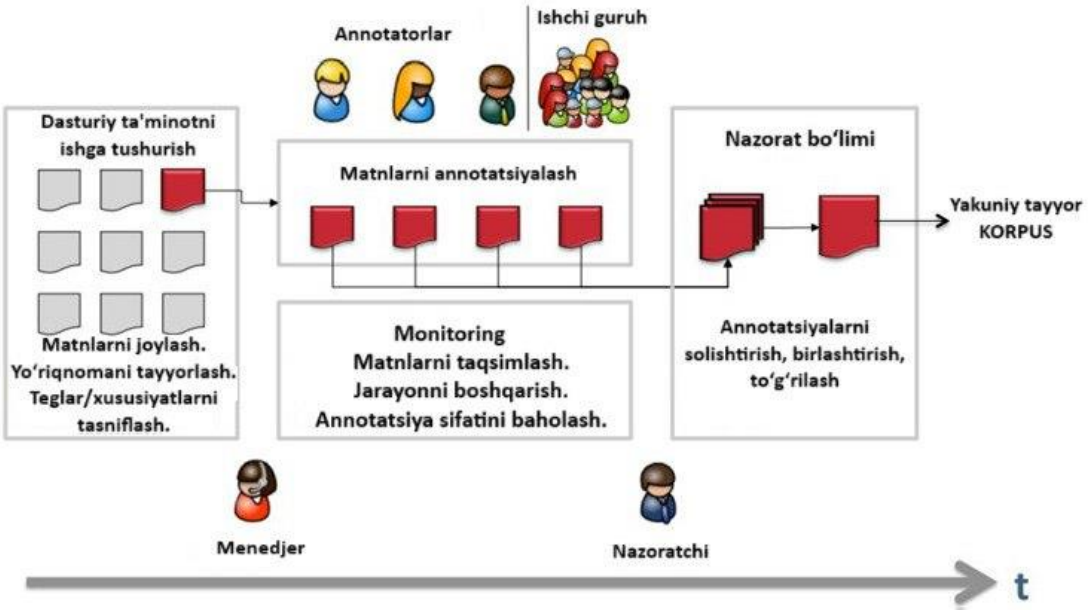


Figure 1. Dependency parsing treebank development process (sequence).

In the next stage, the sentence is divided into individual words or tokens (tokenization), for example, the sentence "Cats are running fast" is tokenized into ["Cats", "running", "fast"]. The tokenization process can handle punctuation marks, abbreviations, and various intervals between words. Segmentation of complex words and abbreviations also plays a special role (for example, "FA", "New York", "etc.").

In the Basis (Lemming) stage, the lexical form (lemma) of each word, i.e. the main form found in dictionaries, is determined. Since Uzbek is an agglutinative language, a pure lexical form that does not receive any grammatical adjuncts is usually chosen as the lemma.

The next step is to reduce each word in the sentence to its base or dictionary form (sentence lemma), which is useful in many natural language processing tasks. In Uzbek, lemma involves reducing a word to its base (dictionary form/base) through morphological analysis. Uzbek is an agglutinative language, which often uses suffixes to indicate tense, mood, person, and case. Therefore, lemma in Uzbek requires the removal of suffixes to obtain the stem. In lemma formation for the purpose of textual corpus formation, we can find the following examples: Sentence: I am reading a book and my brother is listening to music. (Lemmas: I, book, read, and, brother, music, listen); Sentence: He is preparing a lesson because there is an exam tomorrow. (Lemmas: He, lesson, prepare, because, tomorrow, exam, there); Sentence: If the weather is good, we will go for a walk. (Lemmas: If, weather, good, be, we, walk, go).

Here we can give the following examples from the process of establishing sentences:

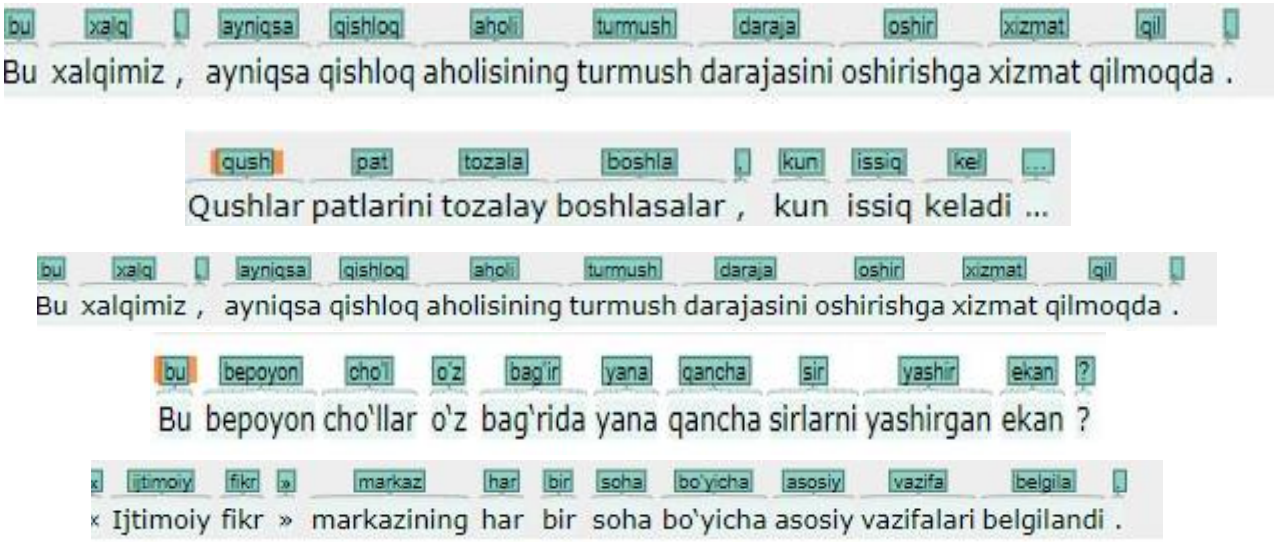


Figure 2. Identifying the roots (lemmas) of words in a complete sentence.

In words, the main root words were obtained in this form and the system automatically performed lemma. Each word and punctuation mark in the sentence was taken as a separate lemma.

Part-of-speech identification/tagging (POS). Part-of-speech is determined for each token, because ensuring accurate results in the lemma process depends on the definition of part-of-speech. In the process of part-of-speech identification, nouns, verbs, adjectives, and other grammatical features are assigned to each word. Libraries such as NLTK (Natural Language Toolkit) or spaCy in the Python programming language provide POS tags in the process of part-of-speech identification.

POS (Part-of-Speech) tags can be divided into the following three large groups:

-tokenization (word segmentation). Words are usually separated from each other by a space, but there are the following exceptions:

- compound or repeated words separated by a hyphen (“-”) (for example, kuta-kuta, bora-bora);

- the occurrence of case markers such as "bo‘lib" and "holda" added to the word "bo‘lib" is lexemed as a multi-word lexeme.

Table 2.

Part-of-speech Classification.

Open class words *Words in this class can be constantly replenished with new words. Open class words convey the main meaning of a sentence and include the following.* *For example: Nouns (NOUN, PROPN): book, student,*	**Closed class words (Closed class words)** *The words in this class are limited in quantity; they are almost not enriched with new words and often perform grammatical functions.* *For example: Pronouns (PRON): me, you, them.*	**Others (other)** *This category includes other categories that do not fall into the above two groups in terms of grammatical function, for example:* *Punctuation marks (PUNCT) period (.),*

Uzbekistan. Verbs (VERB): write, read, spoke.	*Conjunctions (CCONJ, SCONJ): and, or, but, because.*	*comma (,), question mark (?)Symbols (SYM): $, %, @, №.*
ADJ: adjective	ADP: adposition	PUNCT: punctuation
ADV: adverb	AUX: auxiliary verb	SYM: symbols
INTJ: interjection	CCONJ: conjunction	X: others
NOUN: noun	DET: determiner/modifier	
PROPN: proper noun	NUM: number	
VERB: verb	PRON: pronoun	

The Uzbek language belongs to the Turkic language family, and its syntactic structure has many unique features. The main features of Uzbek syntax are as follows:

In Uzbek, sentences are usually constructed in the order “Subject – Object – Verb” (ETF) (for example, in the sentence “Men kitobni o‘qiyman”-Men-I, kitobni-the book, read-o‘qiyman). In Uzbek, words are often modified by suffixes, which determine the meaning and syntactic role of the word. In Uzbek, connections between words are also often made through suffixes, so the word order is relatively free. When comparing Uzbek syntax with other languages, a number of interesting similarities and differences can be seen. English also has an ETF structure, but unlike Uzbek, the place and form of the verb are more rigid. In English, connections between words are mainly made through auxiliary verbs and conjunctions.

Formal Corpus Format (CONLL-U)

Sentences consist of one or more-word strings, and word strings contain the following fields:

1. ID: Word index, an integer starting at 1 for each new sentence; may be an interval for multi-word units; may be a fractional number for empty nodes (fractional numbers may be less than 1, but must be greater than 0).
2. FORM: Word form or punctuation mark.
3. LEMMA: Lemma or stem of the word form.
4. UPOS: Part of speech mark.
5. XPOS: Optional language-specific (or any hierarchical tree model-specific) part of speech / morphological mark; if not present, an underscore is used.
6. FEATS: A list of morphological features from the feature list or a specified language-specific extension; an underscore is used if none exists.
7. HEAD: The head of the current word, which can be an ID value or zero (0).
8. DEPREL: The dependency relation (root if HEAD = 0) or its specified language-specific subtype relative to HEAD.
9. DEPS: An extended dependency graph in the form of a list of head-dependency pairs.

10.MISC: Any other annotation.

```
# text = Hozirda bunday kishilarga 85 ta faoliyat turi bilan shug‘ullanishga ruxsat berilgan.
1       Hozirda hozirda ADV     _       _       11      advmod  _       _
2       bunday  bunday  DET     _       _       3       det     _       _
3       kishilarga      kishi   NOUN    _       Case=Dat|Number=Plur    11      obj     _       _
4       85      85      NUM     _       NumType=Card    7       nummod  _       _
5       ta      ta      X       _       _       4       compound        _       _
6       faoliyat        faoliyat        NOUN    _       Case=Nom|Number=Sing    7       nmod    _       _
7       turi    tur     NOUN    _       Case=Nom|Number=Sing    9       obl     _       _
8       bilan   bilan   ADP     _       _       7       case    _       _
9       shug‘ullanishga shug‘ullan      VERB    _       Case=Dat|VerbForm=Vnoun 11      xcomp   _       _
10      ruxsat  ruxsat  NOUN    _       Case=Nom|Number=Sing    11      compound:lvc    _       _
11      berilgan        ber     VERB    _       Tense=Past|VerbForm=Fin|Voice=Pass      0       root    _       SpaceAfter=No
12                      PUNCT           11      punct
```

Figure 3. Appearance of a single sentence in the formal CONLL-U format when analyzing.

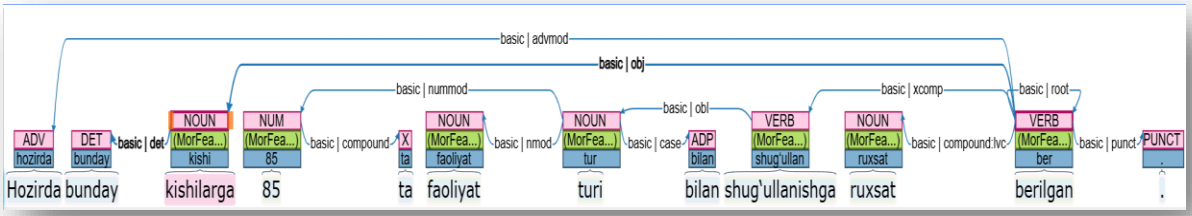


Figure 4. A completed view of the sentence in the dependency treebank.

The process of identifying word classes and entering them into the system was carried out. Part-of-speech (POS) tagging in the Uzbek corpus is the process of assigning grammatical categories or parts of speech to each word in the text. Each word in a sentence is tagged with a tag indicating whether it is a noun, verb, adjective, adverb, etc. POS tags[30] are an important preprocessing step in natural language processing that provides insight into the structure and meaning of sentences.

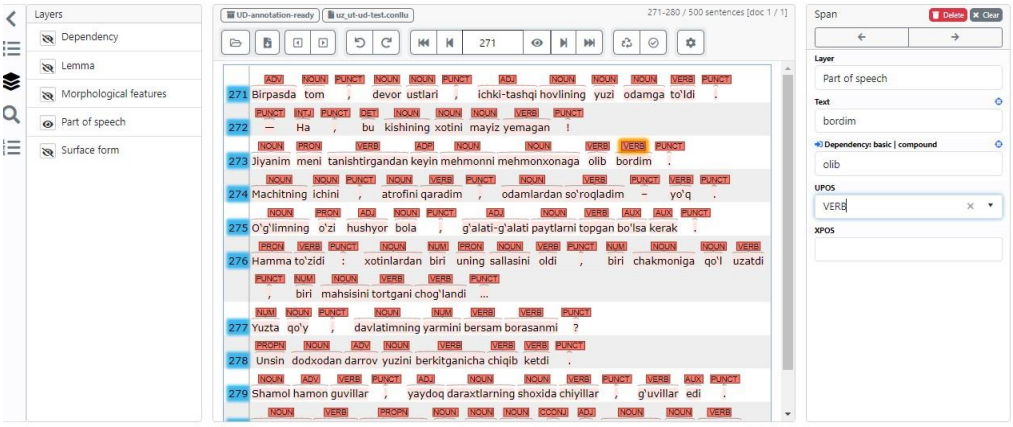


Figure 5. Identification of word classes (POS) in a sentence in the dependency treebank and their general appearance in the corpus.

[30] Zeman, D. Reusable Tagset Conversion Using Tagset Drivers. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC ’08). In Marrakech, Morocco.2008. pp 213 – 218.

Assigning POS tags using word class taggers: POS tags are special characters that are trained to identify parts of speech based on word morphology, structure, and syntax. Some common POS tags include: Noun, Verb, Adjective, Pronoun, Preposition, etc. POS taggers often use linguistic rules, existing models, or machine learning to make predictions.

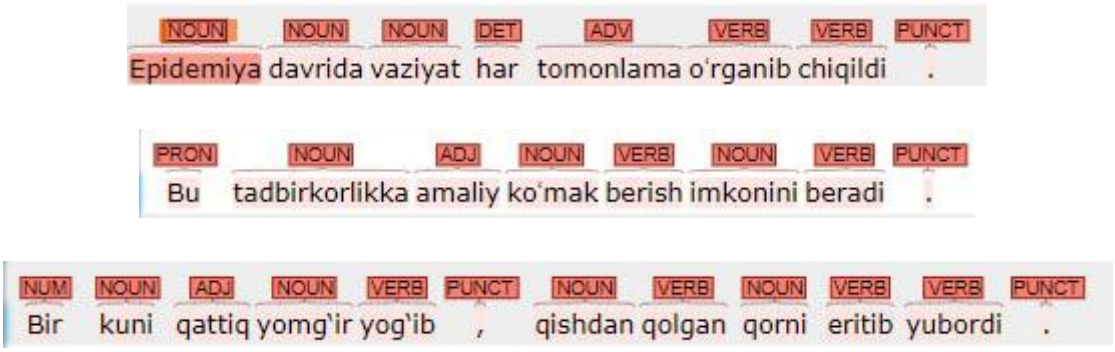


Figure 6. Separating groups of words in a sentence with tags (POS tags).

Methods for tagging word classes.

Rule-based labeling: How can context-sensitive labels be manually generated? Stox. Hidden Mar statistical) labeling: From probabilistic methods such as Hidden Margaritas (HMMs) or Conditional Random Fields (CRF), the Plig of a word is mapped to its context by assigning a possible state to it (HMMs);

Machine learning / deep learning efficient labeling: Neural networks, especially recurrent systems and transformers (e.g. BERT) are trained on large annotated corpora to predict POS labels. The three models learn to identify complex objects in the language and adequately remove[31]

POS tag examples with NLP libraries.

Table 3.

Sample of using NLTK library in Python.

```
import nltk

nltk.download('averaged_perceptron_tagger')

sentence = "The nimble, brown fox jumps over the lazy dog"

tokens = nltk.word_tokenize(sentence)

pos_tags = nltk.pos_tag(tokens) print(pos_tags)

# Output: ('nimble', 'JJ'), ('brown, 'JJ'), ('fox', 'NN'), (jumps', 'VBZ'), ('over', 'IN'), ('lazy', 'JJ'), ('dog', 'NN')]
```

[31] D. Zimbra, A. Abbasi and D. Zeng, The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation, ACM Transactions on Management Information Systems, Vols. xx, No. x, 9 2018

Chapter III of the study is entitled **"Effectiveness of the Uzbek dependency parsing treebank".** This section evaluates the final performance, quality, and scope of the corpus that embodies and demonstrates the hierarchical analysis of Uzbek sentences. The dependency-parsing treebank developed for the Uzbek language includes 1,200 syntactically annotated simple, compound, and complex sentences. The corpus consists of data from online news websites and samples from literary texts, comprising 700 and 500 sentences respectively. Table 4 presents the main size indicators of the corpus.

Table 4.

The total number of sentences in the created corpus and their analysis indicators.

Indicator	Quantity
Total number of sentences	1200 (news: 700; literary: 500)
Total number of tokens	18 000
Average sentence length	15 tokens
Types of UD syntactic connections	33 (full coverage)
UD parts-of-speech (POS)	17 (all included)

Transition-based dependency treebank parsing method is a family of artificial intelligence models that determine syntactic relationships between words in text analysis through sequential operations (shift, reduce, arc-left, arc-right) (Figure 8). It works in the following process:

The initial step is performed through the stack, buffer, and transition operations. Syntactic relationships are determined sequentially at each step.

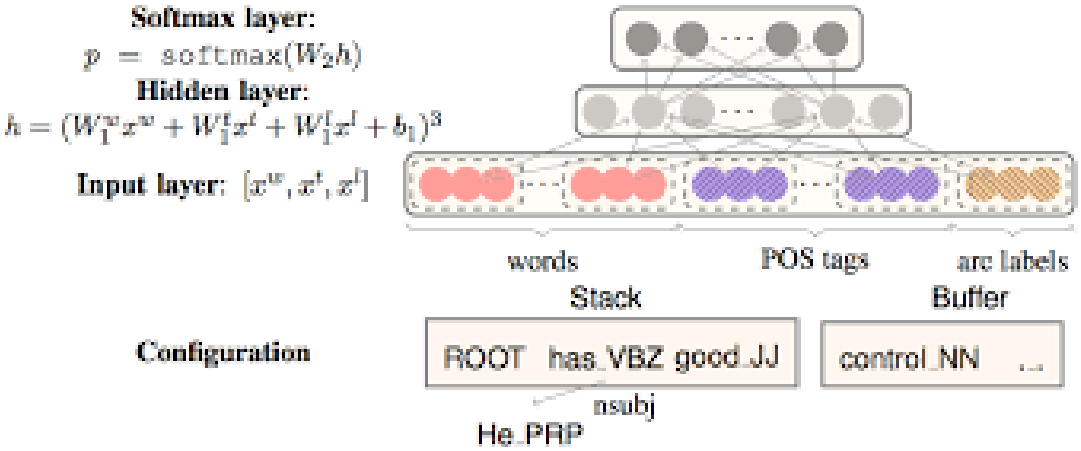


Figure 7. Transition-based dependency parsing artificial intelligence model

The main evaluation criteria are:

• **UAS (Unlabeled Attachment Score)** – The percentage of correctly identified attachments (without labels).

• **LAS (Labeled Attachment Score)** – The percentage of correctly identified attachments with a label.

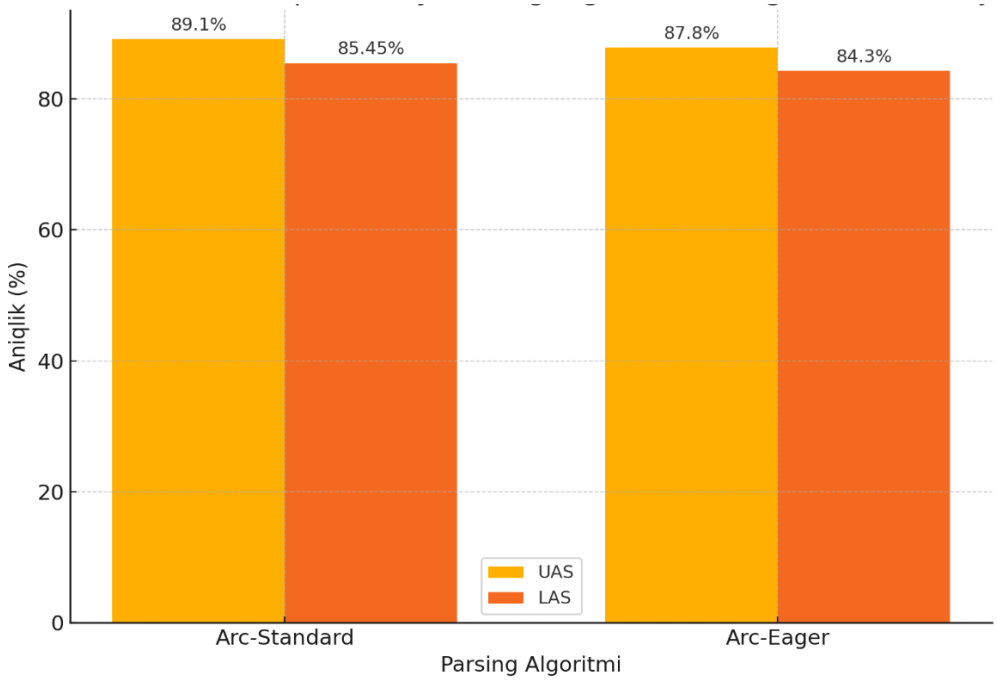


Figure 8. UAS/LAS evaluation results of the Arc Standard and Arc Eager artificial intelligence models.

In the above figure, the efficiency of Transition-based dependency parsing algorithms Arc-Standard and Arc-Eager is compared using UAS and LAS criteria.

UAS (Unlabeled Attachment Score). It determines whether a dependency is true or false, without considering its type. In this indicator, the Arc-Standard algorithm (89.10%) shows a higher result than the Arc-Eager algorithm (87.80%).

LAS (Labeled Attachment Score). Along with determining the dependency, it also determines its type. In this indicator, the Arc-Standard algorithm also gave a better result (85.45%) than the Arc-Eager algorithm (84.30%).

*Arc-Standard algorithm. Arc-Standard is a stack-based algorithm and uses transition operations such as SHIFT, LEFT-ARC, RIGHT-ARC. Each operation is performed only on the top elements of the stack. Arc-Standard usually produces accurate and high-quality parsing results gives.

*Arc-Eager algorithm. The Arc-Eager algorithm is also a transition-based parsing algorithm. Unlike Arc-Standard, the Arc-Eager algorithm tries to detect dependencies as early as possible and performs LEFT-ARC and RIGHT-ARC operations differently. Arc-Eager is generally faster, but in some cases its accuracy is slightly lower than Arc-Standard.

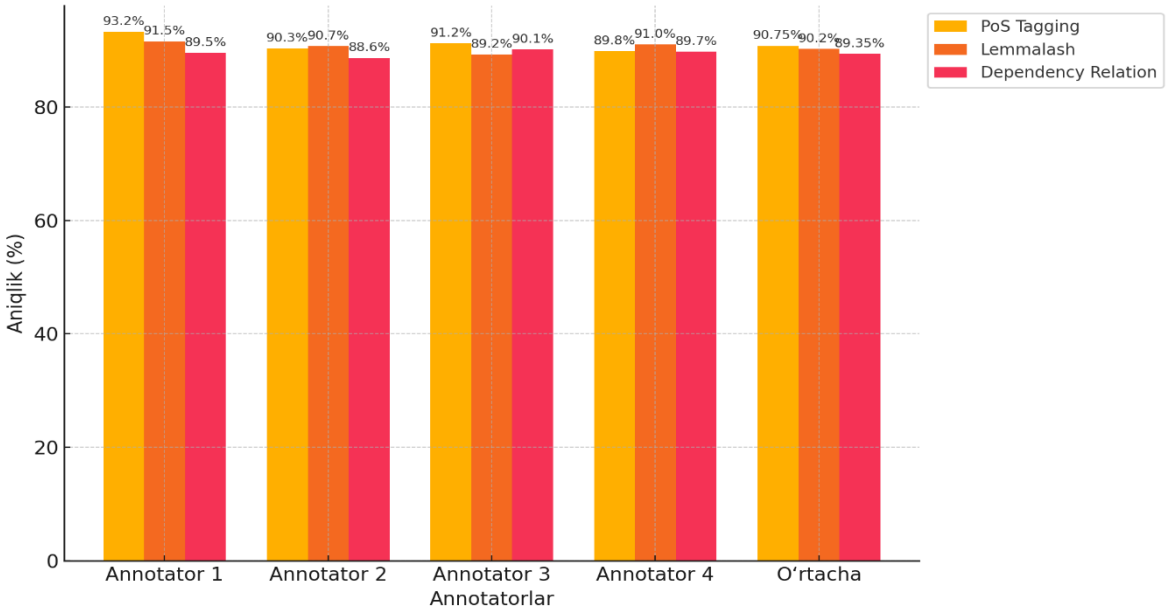


Figure 9. Evaluation of the work efficiency of annotators

To verify the quality of the annotation process, the performance of the annotators was evaluated using the following key indicators:

PoS Tagging (Determination of Word Classes) The overall average accuracy of the annotators in determining the classes of words was **90.75%**. This indicates that the annotators have a high level of agreement in determining the classes of words.

Lemming (Determination of the lexical form of words) The overall accuracy of the annotators in lemmatizing words was **90.20%**. This indicates that there is good agreement among the annotators in determining the base forms of words.

The average accuracy in determining the relationship relationships was **89.35%**. This result indicates that the annotators have the ability to accurately and high-quality annotation in determining syntactic connections.

These high results obtained through the software platform used in the annotation process confirm the quality of the annotation and the success of the annotation process. This means that the dependency treebank being created for the Uzbek language is ready for reliable use in the future.

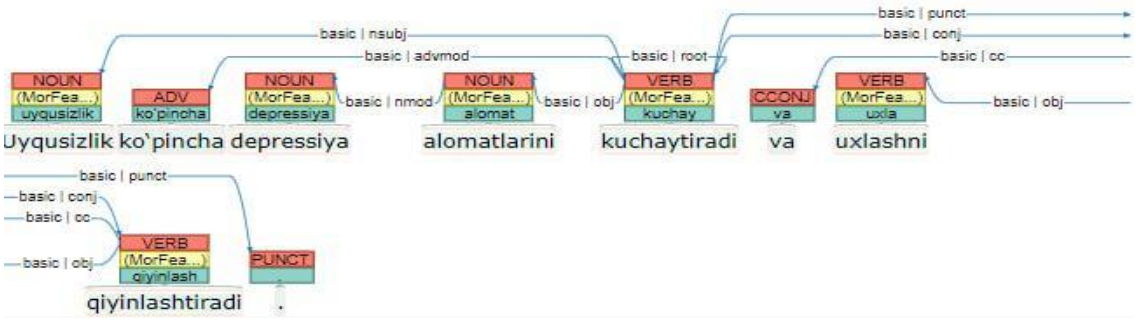


Figure 10. The result of dependency parsing (hierarchical analysis) of a sentence in the Uzbek language is depicted.

The verb “did” is chosen as the root (main point of connection). This is the main action of the sentence and all other parts are connected to this word.

The main connections are as follows:

- *did (root, VERB);*
- *trainer* (nsubj, NOUN) – the performer (subject) of the verb;
- *assistant* (nmod, NOUN) is a determiner connected to the word “*trainer*” (what kind of trainer? - *assistant trainer*);
- *debate* (obj, NOUN) – the object of the verb (what is the action directed at?);
- *previous* (amod, ADJ) is an adjective connected to the word “*discussion*” (which debate? - previous debate);
- *thoroughly* (advmod, ADV) – an adverb connected to the verb (how did he analyze? - *thoroughly*);
- *analysis* (compound:lvc, NOUN) Together with the verb, it forms a compound verb (*analyzed*);
- *said* (conj, VERB) The second action (did and said) that is equally connected to the main verb (*did*);
- *and* (cc, CCONJ) As a connecting word (did and said);
- *instructions* (obj, NOUN) are the object of the verb "did" (what did he say? - *instructions*);
- *own* (nmod, PRON) indicates belonging to the word "instructions" (whose instructions? - *his*);
- (punct, PUNCT) – a punctuation mark that indicates the end of a sentence.

CONCLUSION

1. Identification of morpho-syntactic features: Since the Uzbek language is morphologically rich, it was important to correctly identify various morphological and syntactic features, such as tense, person, and case of the verb, in the process of compiling the corpus. In the process of dependency parsing, the features of nouns, adjectives, verbs, and other main word classes were identified in detail and their place in the syntactic structure was determined.

2. Selection of the main universal word class tags used in the corpus: In the corpus of the Uzbek language dependency parsing, tags based on the Universal Dependencies (UD) standards were used. For example, tags representing general syntactic relations such as nsubj (nominal subject), obj (object), advmod (adverbial modifier) were used. In this, how certain structures specific to Turkic languages are expressed in the universal dependency system was tested and the results were documented.

3. Reflecting the free word order of the Uzbek language: Since the word order in the Uzbek language is relatively free, it was observed in the dependency analysis how changes in word order affect syntactic connections. As a result, it was achieved to correctly reflect the syntactic relationships between the root and its dependent words.

4. Difficulties encountered in the annotation process: When compiling the dependency parsing treebank for the Uzbek language, annotation was difficult in some complex structures (for example, parts of speech and case units, or independent speech units). In particular, different morphological forms have different semantic functions, and in some cases, additional developments were required to correctly define them.

5. Identification of lexical features as an adjunct: In order to correctly reflect some specific morphological and syntactic forms of the Uzbek language, lexical features, such as auxiliary verbs and prepositional words, were reanalyzed based on special rules. As a result, the functional role of some words in the corpus was more clearly reflected.

6. The importance of a high-quality dependency parsing treebank : Creating a high-quality hierachical analysis corpus for the Uzbek language will allow for a deeper study of the syntactic structure of the Uzbek language. This corpus may also be useful in studying other Turkic languages or in identifying common syntactic rules among Turkic languages.

7. Future improvements: Further expansion of the Uzbek dependency treebank and enrichment with texts from other fields or styles will further increase the scope of this corpus in the future. At the same time, there is also the possibility of introducing structures that are not available in the corpus or do not comply with the current analysis (annotation) rules.

8. There are 17 standart tags for tagging morphological features in the process of natural languages, of which 15 are mainly used. The tags include: ADJ, ADP, ADV, CCONJ, DET, INTJ, NOUN, NUM, PART, PRON, PROPN and SCONJ. At the same time, the symbols AUX (auxiliary), SYM (symbol), and X (other) can also be found in some places.

9. 16 Standart annotation features have been documented in the natural language process. This dependency parsing treebank uses 15 predefined morphological features: Aspect, Case, Clusivity, Degree, Deixis, Mood, Neutral, Number, PartType, Person, Polarity, PronType and Voice. In addition, two tags are included because they are specific to the Uzbek language, namely the determiner suffix and the neutral demonstrative suffix.

10. It is important to explore the possibilities of automating the annotation of the Uzbek language corpus, for example, using a rule-based algorithm. This rule-based algorithm may be more suitable for this Uzbek dependency parsing treebank, as it requires less data compared to statistical algorithms.

НАУЧНЫЙ СОВЕТ DSc.03/25.08.2021.Fil.01.16 ПО ПРИСУЖДЕНИЮ УЧЕНЫХ СТЕПЕНЕЙ ПРИ НАЦИОНАЛЬНОМ УНИВЕРСИТЕТЕ УЗБЕКИСТАНА ИМЕНИ МИРЗО УЛУГБЕКА

НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ УЗБЕКИСТАНА

РАХИМБОЕВА ХУЛКАР ГАЙРАТОВНА

СОЗДАНИЕ КОРПУСА ИЕРАРХИЧЕСКОГО АНАЛИЗА ПРЕДЛОЖЕНИЙ УЗБЕКСКОГО ЯЗЫКА

10.00.11 – Теория языка. Прикладная и компьютерная лингвистика

АВТОРЕФЕРАТ
ДИССЕРТАЦИИ ДОКТОРА ФИЛОСОФИИ (PhD)
ПО ФИЛОЛОГИЧЕСКИМ НАУКАМ

Ташкент – 2025

Тема диссертации доктора философии (PhD) по филологическим наукам зарегистрирована в Высшей аттестационной комиссии при Министерстве высшего образования, науки и инноваций Республики Узбекистан за № B2022.2.PhD/Fil2702.

Диссертация выполнена в Ургенчском государственном университете.

Автореферат диссертации размещен на трех языках (узбекском, английском, русском (резюме)) размещён на сайте Ученого совета (www.nuu.uz) и на информационно-образовательном портале «ZiyoNet» (www.ziyonet.uz).

Научный руководитель:	**Садуллаева Нилуфар Азимовна** доктор филологических наук, профессор
Официальные оппоненты:	**Абдурахмонова Нилуфар Зайнобиддин кызы** доктор филологических наук, профессор
	Тоирова Гули Ибрагимовна доктор филологических наук, доцент
Ведущяя организация:	**Ургенчский государственный университет**

Защита диссертации состоится «___»_____2025 года в _____ часов на заседании Научного совета **DSc.03/25.08.2021.Fil.01.16** по присуждению ученых степеней при Национальном университете Узбекистана. (Адрес: 100174, Ташкент, улица Университет, дом 4. Факультет узбекской филологии, 1 этаж, каб. 108. Tel.: (9987) 246-02-24; факс: (9987) 246 02-24; e-mail: nauka@nuu.uz.

С диссертацией можно ознакомиться в Информационно-ресурсном центре Национальном университете Узбекистана (зарегистрирована за № _____). Адрес: 100174, Ташкент, улица Университет, дом 4. Административное здание УЗМУ, 2 этаж, каб. 4. Tel.: (9987) 246 02-24; факс: (9987) 246 02-24.

Автореферат диссертации разослан «_____» _____2025 года.

(протокол реестра рассылки за № _____ от «_____» _____ 2025 г.)

Н.А.Рахманов
Председатель разового научного совета по присуждению ученых степеней, докт. филол. наук, профессор

М.Б.Хужамкулова
Ученый секретарь разового ученого совета по присуждению ученых степеней, канд, филол, наук, доцент

А.Э.Маматов
Председатель научного семинара при разовом научном совете по присуждению ученых степеней, докт, филол. наук, профессор

ВВЕДЕНИЕ (аннотация диссертации доктора философии (PhD)

Актуальность и востребованность темы диссертации. Известно, что различные сектора в странах по всему миру напрямую подвержены влиянию информационного века, и значение коммуникации и информации чрезвычайно велико. В то же время компьютерные технологии проникли во все сферы жизни, играя ключевую роль в достижении целей нового времени. В условиях глобализации возникают определенные трудности из-за различий между машинным и человеческим языком. Более того, уникальные характеристики человеческого языка все более настоятельно требуют решения этих проблем. С этой точки зрения вычислительная лингвистика имеет большое значение.

В мировой лингвистике иерархические анализаторные корпусы имеют важное значение для проведения глубокого анализа синтаксической структуры текстов. Для продвижения языковых навыков и представления языка миру его теоретическое и научное применение недостаточно – также важно внедрять язык через современные и соответствующие ресурсы обработки естественного языка (NLP). Вычислительная лингвистика изучает язык с помощью моделей искусственного интеллекта и включает анализ больших коллекций текстов (корпусов).

Сегодня в нашей стране такие области, как корпусная лингвистика и искусственный интеллект в рамках компьютерных наук, все еще находятся на стадии развития, и многие аспекты еще не были исследованы. Ресурсы для лингвистической обработки естественного языка в этой области остаются значительно ограниченными. Создание структурированного корпуса с деревьями зависимостей (с синтаксическим и морфологическим анализом), охватывающего достаточное количество предложений, позволяет развивать лингвистические исследования узбекского языка и повысить его признание в международном научном сообществе.

Данная научно-исследовательская работа служит реализации задач, отмеченных в таких нормативно-правовых документах, связанных с данной отраслью, как Указ Президента Республики Узбекистан от 5 октября 2020 года № УП-6079 «Об утверждении Стратегии “Цифровой Узбекистан – 2030” и мерах по ее эффективной реализации», Постановление № ПП-4996 от 17 февраля 2021 года «О мерах по созданию условий для ускоренного внедрения технологий искусственного интеллекта», Указ № УП-5850 от 21 октября 2019 года «О мерах по кардинальному повышению роли и авторитета узбекского языка в качестве государственного языка», Указ № УП-6084 от 20 октября 2020 года «О мерах по дальнейшему развитию узбекского языка и совершенствованию языковой политики в стране». Как отметил Президент нашей страны, «Повышение эффективности научных исследований, связанных с уникальными особенностями, диалектами,

историческим развитием и перспективами узбекского языка»,[32] является одной из повседневных потребностей нашей страны.

Соответствие исследования приоритетным направлениям развития науки и технологий республики. Диссертация выполнена в рамках приоритетного направления развития науки и технологий республики: “Формирование системы инновационных идей и пути их реализации в социальном, правовом, экономическом, культурном, духовнопросветительском развитии информационного общества и демократического государства”.

Степень изученности проблемы. В мировой лингвистике тема Универсальной зависимости связа Юрна с именами Дж. Нивре, Д. Земана, де Марнеффа, Д. Мэннинга, Даниэля Джурафски[33]

В нашей республике ведущими деятелями узбекской компьютерной лингвистики являются А.К.Пулатов, М.М.Арипов, А.Э.Маматов, Н.З.Абдурахманова, М.М.Курбанова, М.М.Мусаев, Н.А.Садуллаева, Ш.А.Назиров, М.Х.Хакимов, О.Хамдамов, Н.А.Игнатьев, Г.Р.Матлатипов и др.[34] Научные исследования, проводимые под их руководством были направлены на создание формальной модели грамматики узбекского языка, формализации структур предложений в узбекском, турецком и каракалпакском языках, создании словаря для тюркских языков, моделировании процессов формирования речевых сигналов, создании синтезаторов речи, создании логико-лингвистических и математических моделей языковых объектов для многоязычного моделируемого машинного перевода.

Известный узбекский ученый, доктор физико-математических наук, профессор Абдумаджид Пулатов на протяжении многих лет проводил

---

[32] O‘zbekiston Respublikasi Prezidenti Sh.M.Mirziyoyevning 2019 yil 21 oktyabrdagi “Milliy o‘zligimiz va mustaqil davlatchiligimiz timsoli” mavzusida o‘zbek tiliga davlat tili maqomi berilganining o‘ttiz yilligiga bag‘ishlangan tantanali marosimdagi nutqi // Xalq so‘zi, 2019 yil, 22 oktyabr.

[33] De Marneffe, Christopher D. Manning, Joakim Nivre, Daniel Zeman. Universal Dependencies. Computational Linguistics: Volume 47, Issue 2, 2021. – P. 255. – URL: https://direct.mit.edu/coli/article/47/2/255/98516/ Universal-Dependencies. 59 Nivre J., Zeman D.;

Nivre J., Zeman D. Universal Dependencies: A Cross-Linguistic Perspective on Grammar and Lexicon // Proceedings of the Workshop on Grammar and Lexicon: Interactions and Interfaces (GramLex 2016). – 2016. – P. 8–17. – URL: https://aclanthology.org/W16-3806.pdf.

D. Jurafsky, JH. Martin. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition: Prentice Hall, 2000

[34] N.Z.Abdurakhmonova, K Urdishev., Corpus Based Teaching Uzbek As A Foreign Language.-Tashkent: Journal of Foreign Language Teaching and Applied Linguistics.2019.// Nilufar Abdurakhmonova, Ismailov Alisher, Rano Sayfulleyeva., MorphUz: Morphological analyzer for the Uzbek language.- 2022 7th International Conference on Computer Science and Engineering (UBMK). 2022/9/14.pp. 61-66.// N Abdurakhmonova., Dependency Parsing Based On Uzbek Corpus, Proceedings of the International Conference on Language technology for all (LT4all), 2019.// Nilufar Z Abdurakhmonova, Alisher S Ismailov, Davlatyor Mengliev., Developing NLP tool for linguistic analysis of Turkic languages, 2022 IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON).// Nilufar Abdurahmonova.,Kompyuter lingvistikasi. Nodirabegim. Toshkent, 2021.

Mirsaid Aripov, Bibigul Razzoqova Altinbay Sharibbek, Nilufar Abdurakhmonova., Ontology of grammar rules as example of Noun the Uzbek language and Kazakh languages, VI international scientific conference "Modern problems of the applied mathematics and information technology - Al Khorezmiy -2018". Pp37-38// Mirsaid Aripov, Baxodir Begalov, Uzoqboy Begimqulov, Mirsalim Mamarajabov., Axborot texnologiyalari, oshkent-2009.

A. Polatov, S. Muhammedova. Kompyuter lingvistikasi. -T.,2008; A. Po'latov. Kompyuter lingvistikasi. -T., 2011: A. Rahimov. Kompyuter lingvistikasi asoslari. –T.,2011.

исследования по рассматриваемой области, издавал такие книги и пособия по языкознанию, как «Английский язык», «Дунёвий ўзбек тили. Ўзбек тилида феъл шакллари ва уларнинг рус, инглиз тилларидаги кўринишлари», «Компьютерная лингвистика», «Английский язык для самостоятельного изучения» и другие[35]. Также У.Шарипов и И.Юлдошев разработали учебник «Тилшунослик асослари»,[36] А.Рагимова – учебник «Основы компьютерной лингвистики»[37], Г.Р.Матлатипов – учебник «Технологии машинного перевода»[38], Н.З. Абдурахмановой – учебник «Корпусная лингвистика».

В мире проведен ряд исследований по разработке компьютерно-ориентированных моделей естественных языков, в том числе работы Ноама Хомского (США), Кристофера Д. Мэннинга (США), Д. Юраффского (США), Карлоса Гомеса-Родригеса (Испания), Хоакима Нивре, Мигеля А. Алонсо (Испания), Марко Кульмана (Швеция), А. Джеймса, Г. Фанта, Р. Дельмонти [39] и других. В странах содружества независимых государств были изучены также научные и методические труды таких ученых, посвященные компьютерной лингвистике, как Е.И.Большаковой и А.Шарипбая.

Цель исследования – построение иерархически реляционного корпуса для обработки текстовых данных на узбекском языке, хранящихся в электронном виде, оценка с помощью модели искусственного интеллекта и создание программного обеспечения.

Задачи исследования. Для достижения указанной цели решаются следующие задачи:

подготовить руководство по аннотированию корпуса синтаксического анализа зависимостей узбекского языка, включая лемматизацию, морфологическую разметку, определение членов предложения и разработку правил зависимостей;

---

[35] A.Po'latov. Kompyuter lingvistikasi. -T., 2011:

[36] Sharipov O‘,. Yo‘ldoshev I. Tilshunoslik asoslari,-Toshkent: Nizomiy nomidagi Toshkent Davlat pedagogika universiteti, 2006.

[37] A.Rahimov. Kompyuter lingvistikasi asoslari. –T.,2011.

[38] G‘.R.Matlatipov, “Mashina tarjimasi texnologiyalari”. O‘quv qo‘llanma, 2024.

[39] Carlos Gómez-Rodríguez, Joakim Nivre., A transition-based parser for 2-planar dependency structures, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010. PP1492-1501.

Joakim Nivre, Laura Rimell, Ryan McDonald, Carlos Gomez-Rodriguez., Evaluation of dependency parsers on unbounded dependencies. Proceedings of the 23rd international conference on computational linguistics. 2010. PP833-841.

Joan D Gonzalez-Franco, Jorge E Preciado-Velasco, Jose E Lozano-Rizk, Raul Rivera-Rodriguez, Jorge Torres-Rodriguez, Miguel A Alonso-Arevalo., Comparison of Supervised Learning Algorithms on a 5G Dataset Reduced via Principal Component Analysis (PCA)- Future Internet. 2023. P335.

Ralph Debusmann, Denys Duchier, Alexander Koller, Marco Kuhlmann, Gert Smolka, Stefan Thater., A relational syntax-semantics interface based on dependency grammar, COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics. 2004.PP 176-182.

Dann Juraffskiy., James H.Martin. Speech and language processing. -New Jersey: Upper Saddle River, February 2008.

Ch. Manning., H. Schutze. Foundations of Statistical Natural language Processing, MIT Press. Cambridge, MA: May 1999.

Andrew Ng. Machine Learning yearning, technical strategy for AI Engineers in the Era of Deep Learning, August 2023.

собрать корпусную базу данных, состоящую из простых и сложных предложений на узбекском языке, и провести предварительную обработку (pre-processing);

определить лемму каждого слова в предложениях, включённых в узбекский корпус, проанализировать их морфологические характеристики и с помощью автоматического парсера получить результаты и представить полный корпус анализа зависимостей;

определить синтаксические свойства слов в предложениях узбекского корпуса, построить их зависимостные структуры и проанализировать подчинённость с использованием модели искусственного интеллекта.

Объект исследования – Тексты, отобранные с веб-сайтов Kun.uz и Daryo.uz

Предмет исследования – Создание корпуса иерархического (древовидного) анализа узбекского языка: содержание, принципы, методы, формы, грамматические связи, алгоритмы анализа и процессы аннотирования.

Методы исследования. Для решения исследовательской темы были применены методы классификации, описания, автоматического синтаксического анализа (parsing)[40], аннотирования с использованием инструкций, статистики, переноса и корпусного анализа.

Научная новизна исследования:

создан иерархический корпус зависимостей узбекского языка;

создано руководство по аннотации, лемматизации, морфологической характеристике, выделению частей речи и разработке правил связи иерархического корпуса узбекского языка;

правила синтаксических зависимостей и морфологические особенности предложений на узбекском языке были глубоко проанализированы и аннотированы в рамках корпуса;

построена модель искусственного интеллекта для обработки естественного языка с учётом специфики узбекского языка, а надёжность корпуса была оценена с помощью синтаксического анализа, выполненного данной моделью.

Практические результаты исследования.

служит основным языковым ресурсом для автоматического перевода, анализа текста, систем вопрос-ответ и голосовых помощников в области обработки естественного языка (NLP) и искусственного интеллекта на узбекском языке;

обеспечивает точную синтаксическую базу данных для лингвистических исследований, позволяя глубоко анализировать грамматическую систему узбекского языка;

[40] *Парсинг (*на узбекском: *tahlil qilish, ajratish, tarkibiy qismlarga ajratish*) – это процесс определения грамматической структуры данного предложения, то есть установление синтаксических связей и зависимостей между словами внутри него.

способствует разработке интерактивных инструментов и образовательных программ на основе лингвистических моделей в учебном процессе.

Достоверность результатов исследования.

Надежность результатов корпуса определялась с помощью методов математической оценки UAS (оценка немаркированной привязанности) и LAS (оценка маркированной привязанности).

Научная и практическая значимость результатов исследования.

Научная значимость определяется тем, что корпус иерархических отношений служит базой для исследований в области языкознания и компьютерной лингвистики, представляя синтаксическую структуру узбекского языка в цифровой форме, что позволяет проводить более глубокий анализ синтаксических явлений, а также сравнительные научные исследования с другими языками. Указанный корпус представляет синтаксическую структуру узбекского языка в числовой форме, что способствует аналитическим исследованиям и совершенствованию языковых моделей.

Практическая значимость результатов исследования определяется тем, что, что корпус иерархических отношений служит необходимым ресурсом для разработки программ обработки естественного языка, включая автоматический перевод, чат-боты, анализ содержимого текста, автоматическое реферирование текста и сентиментальный анализ. Указанное, в свою очередь, повысит качество интеллектуальных и информационно-поисковых систем, работающих на базе узбекского языка, что расширит возможности использования цифровых технологий в повседневной жизни. Корпус также служит готовым ресурсом на узбекском языке для разработки цифровых сервисов и программного обеспечения, что позволяет местным и глобальным IT-компаниям создавать свои продукты более высокого качества и в соответствии с требованиями пользователей.

Внедрение результатов исследований. Научные результаты исследования по теме «Создание корпуса для иерархического анализа предложений на узбекском языке» были использованы в рамках гранта «UzUDT: Универсальный корпус деревьев связей для обработки естественного языка на узбекском языке и его семантического анализа» (REP-25112021/113), финансируемого проектом Всемирного банка «Модернизация национальной инновационной системы Узбекистана», созданного при Министерстве инновационного развития Республики Узбекистан. В частности, научные результаты диссертации были использованы в рамках проекта для проведения леммного анализа слов в узбекских текстах, выявления иерархии синтаксических связей, их разметки и оценки, аннотирования и оценки групп слов в тексте, а также разработки базы данных морфологических признаков.

Выводы и результаты исследования были использованы в практическом проекте А-ФА-2019-9 «Исследование древних редких рукописей и источников, создание их оцифрованной библиотеки», действующем в

Хорезмской академии Маъмуна. Результаты нашли применение при анализе творческих образцов представителей классической литературы, обнаруженных в исторических произведениях диссертационных материалов. В результате удалось правильно интерпретировать слова и фразы, встречающиеся в тексте исторических произведений; произвести лингвистический анализ языка литературных источников XIX века, сопоставление его с языковыми особенностями предшествующих и более поздних периодов.

Выводы по работе были использованы в рамках проекта Erasmus+ №598340-EPP-1-2018-1-ES-EPPKA2-CBHE-JP University Cooperation Framework for Knowledge Transfer in Central Asia and China (UNICAC)**.** В результате появилась возможность расширить кросс-языковые семантические поисковые системы и базы данных; использования аналитики на основе данных для изучения языков и создания образовательных инструментов, подчеркивающих синтаксические особенности языков; произвести стандартизацию языковых ресурсов для использования их в международных проектах.

Апробация результатов исследования. Результаты исследования были апробированы на 9 научно-практических конференциях, в частности на 3 международных (в том числе 1 Scopus), 6 республиканских конференциях.

Опубликованность результатов исследования. По теме диссертации опубликовано 14 научных работ, в том числе 5 статей в научных изданиях, рекомендованных Высшей аттестационной комиссией Республики Узбекистан для публикации основных результатов докторских диссертаций, из них 4 – в республиканских и 1 – в зарубежном научном журнале.

Структура и объем диссертации. Диссертационная работа состоит из введения, трех глав, заключения и рекомендаций, списка использованной литературы. Общий объём работы составляет 133 страниц.

E’LON QILINGAN ISHLAR RO‘YXATI
СПИСОК ОПУБЛИКОВАННЫХ РАБОТ
LIST OF PUBLISHED WORKS

I bo‘lim (I часть; PartI)

1. Raximboyeva X.G. Uzbek dependency treebank in universal dependencies // Journal of Advanced linguistic studies. Vol. 11. No. 2. ISSN 2251-4075. – Delhi, India, 2024. Jul-Dec. – P. 283-292 (10.00.00. № 13).
2. Raximboyeva X.G. Sentence annotation for Uzbek language using dependency relations // O‘zMU xabarlari. – Toshkent, 2023. – № 1/12. – B. 351-354 (10.00.00. № 15).
3. Raximboyeva X.G. O‘zbek tilida gaplarni universal ierarxik korpusda annotatsiya qilish masalalari // Xorazm Ma’mun akademiyasi axborotnomasi. ISSN 2091-573 X. – Xorazm, 2024. – № 11/4. – B. 213-215 (10.00.00. № 21).
4. Raximboyeva X.G. Morfological and syntactic features of sentences in the Uzbek language / International Conference on modern Science and scientific studies. Vol. 3. Issue 1. ISSN (E) 2845-3730/SJIF 2024:6809. – France, 2024. November 19. – P. 197-199.
5. Raximboyeva X.G. Developing dependency parsing datasets for Uzbek: Challenges and achievements / “ТІЛТАНУДАҒЫ ТҰЛҒАЛАР” академиктер Әбдуәли Қайдар, Рәбиға Сыздық, Шора Сарыбаевтың 100 жылдық мерейтойларына арналған халықаралық ғылыми-практикалық конференция. – Astana, 2024. – B. 198-200.
6. Raximboyeva X.G. Uzbek Treebanking: Dependency relations and formalisms / “Amaliy matematikaning zamonaviy muammolari va istiqbollari” nomli Qarshi Davlat Universiteti ilmiy-amaliy konferensiyasi. – Qarshi, 2024. 24-25-may. – B. 354-356.
7. Raximboyeva X.G‘. Iyerarxik tahlil korpusining asosiy tushunchalari / Vol. 17. No. 2. “Ta’limning zamonaviy transformatsiyasi” nomli Respublika onlayn ilmiy-amaliy konferensiya. – T., 2025. 17-to‘plam. 2-son. – B. 196-200.

II bo‘lim (II часть; PartII)

8. Raximboyeva X.G. Methods for creating dependency parcing treebank, schemes for annotation and used software tools // O‘zbekiston Milliy Universiteti xabarlari. – Toshkent, 2023. – № 1/5/1. ISSN 2181-7324. – B. 297-299 (10.00.00. № 15).
9. Raximboyeva X.G. Grammar Induction and the Formation of Combinatory Categorial Grammar through Uzbek Language Sentences // Ilm Sarchashmalari. Urganch, 2022. – № 7. Index 1072. – B. 180-184 (10.00.00. № 3).
10. Raximboyeva X.G. Requirement of Parsing Characteristics for semantic parsing / International scientific-practical Conference “Modern Philological Paradigms: Interactions of tradition and innovation III”. – Tashkent, 2023. – № 1297. May 8. – P. 570-572.
11. Raximboyeva X.G. Introduction to the guidelines for the syntactic annotation for the Uzbek language / Alisher Navoiy nomidagi Toshkent davlat

o‘zbek tili va adabiyoti universiteti “O‘zbek tili taraqqiyoti va hamkorlik masalalari” mavzusidagi konferensiya materiallari. – T., 2023. 19-oktyabr. – B. 127-131.

12. Raximboyeva X.G. Grammatik bog‘lanishlar va sintaktik-ierarxik tahlil jarayoni (parser) o‘zbek tili misolida / “Zamonaviy filologik paradigmalar: an’analar va innovatsion yondashuvlarning o‘zaro ta’siri II” mavzusidagi Xalqaro ilmiy konferensiya. – Toshkent, 2022. – № 144. Aprel 5. – B. 778-784.

13. Raximboyeva X.G. Uzbek Sentiment Analysis based on local Restaurant Reviews / The International Conference on Agglutinative Language Technologies as a challenge of Natural Language Processing (ALTNLP). – Koper, Slovenia, CEUR Workshop Proceedings, 3315, 2022. – P. 126-136. (3, Scopus, IF=0.202)

14. Raximboyeva X.G. O‘zbek tili ierarxik korpusi (Treebankida) teglash tushunchasi va uning NLPdagi roli / “O‘zbek tili tatqiqi va ta’limi muammolari” mavzusidagi Xalqaro ilmiy-amaliy konferensiya. – Toshkent, O‘zbekiston, 2024. Vol. 2. – № 18.10. – B. 136-139.

Avtoreferat “O‘zMU xabarlari” jurnalida tahrirdan o‘tkazildi.

Bosishga ruxsat etildi: 08.05.2025-yil.
Bichimi 60x84 1/16, “Times New Roman”
garniturada raqamli bosma usulida bosildi.
Shartli bosma tabog‘i: 3,5. Adadi: 100. Buyurtma №: 192.

“TRAINMAX” MChJ bosmaxonasida chop etildi.
100194, Toshkent shahri, Yunusobod-11, 62-uy.