O‘ZBEK TILI, ADABIYOTI VA FOLKLORI INSTITUTI HUZURIDAGI ILMIY DARAJALAR BERUVCHI DSc.02/30.12.2019.Fil.46.03 RAQAMLI ILMIY KENGASH ASOSIDAGI BIR MARTALIK ILMIY KENGASH

ALISHER NAVOIY NOMIDAGI TOSHKENT DAVLAT O‘ZBEK TILI VA ADABIYOTI UNIVERSITETI

ABDURAXMONOVA NILUFAR ZAYNOBIDDIN QIZI

O‘ZBEK TILI ELEKTRON KORPUSINING KOMPYUTER MODELLARI

10.00.11 –Til nazariyasi. Amaliy va kompyuter lingvistikasi

FILOLOGIYA FANLARI DOKTORI (DSc) DISSERTATSIYASI AVTOREFERATI

Toshkent-2021

UO‘K: 811.512.133-115

Doktorlik (DSc) dissertatsiyasi avtoreferati mundarijasi

Оглавнение автореферата докторской (DSc) диссертация

Contents of abstract of doctoral (DSc) dissertation

O‘ZBEK TILI, ADABIYOTI VA FOLKLORI INSTITUTI HUZURIDAGI ILMIY DARAJALAR BERUVCHI DSc.02/30.12.2019.Fil.46.03 RAQAMLI ILMIY KENGASH ASOSIDAGI BIR MARTALIK ILMIY KENGASH

ALISHER NAVOIY NOMIDAGI TOSHKENT DAVLAT O‘ZBEK TILI VA ADABIYOTI UNIVERSITETI

ABDURAXMONOVA NILUFAR ZAYNOBIDDIN QIZI

O‘ZBEK TILI ELEKTRON KORPUSINING KOMPYUTER MODELLARI

10.00.11 –Til nazariyasi. Amaliy va kompyuter lingvistikasi

FILOLOGIYA FANLARI DOKTORI (DSc) DISSERTATSIYASI AVTOREFERATI

Toshkent-2021

Fan doktori (DSc) dissertatsiyasi mavzusi O'zbekiston Respublikasi Vazirlar Mahkamasi huzuridagi Oliy attestatsiya komissiyasida B2019.2.DSc/Fil1174 raqam bilan ro'yxatga olingan.

Dissertatsiya Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universitetida bajarilgan.

Dissertatsiya avtoreferati uch tilda (o'zbek, ingliz, rus (rezyume) Ilmiy kengashning veb-sahifasi (www.uztafi@academy.uz) va «Ziyonet» (www.ziyonet.uz) Axborot ta'lim portalida joylashtirilgan.

Ilmiy maslahatchi:	**Dadaboyev Hamidulla Aripovich** filologiya fanlari doktori, professor
Rasmiy opponentlar:	**Muhamedova Saodat Xudayberdiyevna** filologiya fanlari doktori, professor
	Urinbayeva Dilbar Bazarovna filologiya fanlari doktori, dotsent
	Nazirova Elmira Shodmanovna texnika fanlari doktori, dotsent
Yetakchi tashkilot:	**Farg'ona davlat universiteti**

Dissertatsiya himoyasi O'zbek tili, adabiyoti va folklori instituti huzuridagi Ilmiy darajalar beruvchi DSc.02/30.12.2019.Fil.46.03 raqamli ilmiy kengashi huzuridagi bir martalik Ilmiy kengashida 2021-yil «24» 12 soat 11 dagi majlisida bo'lib o'tadi. (Manzil: 100060, Toshkent shahri, Mirobod tumani, Shahrisabz tor ko'chasi, 5-uy. Tel.: (871) 233-71-44; faks: (871) 233-71-44; e-mail:uztafi@academy.uz.).

Dissertatsiya bilan O'zbekiston Respublikasi Fanlar akademiyasining Asosiy kutubxonasida tanishish mumkin (52 raqami bilan ro'yxatga olingan). (Manzil: Toshkent shahri, Ziyolilar ko'chasi, 13-uy. Tel.: (99871) 262-74-58

Dissertatsiya avtoreferati 2021- yil «13» 12 da tarqatildi.
(2021-yil «5» 13 12 dagi raqamli reyestr bayonnomasi).

N.M.Mahmudov
Ilmiy darajalar beruvchi
ilmiy kengash raisi, f.f.d., professor

G'.M.Ismoilov
Ilmiy darajalar beruvchi ilmiy
kengash ilmiy kotibi,
f.f.n., katta ilmiy xodim

D. S. Xudoyberganova
Ilmiy darajalar beruvchi
ilmiy kengash qoshidagi Ilmiy
seminar raisi, f.f.d., prof.

KIRISH (doktorlik (DSc) dissertatsiyasi annotatsiyasi)

Tadqiqot mavzusining dolzarbligi va zarurati. Jahon amaliy tilshunosligida korpusshunoslik, korpus texnologiyasi va korpus lingvistikasi muammolarini o‘rganish XX asrning ikkinchi yarmida rivojlanish bosqichiga ko‘tarildi. XX asrda dastlabki ingliz tili korpusining paydo bo‘lishi boshqa dunyo tillarida ham katta hajmdagi elektron korpuslarning yaratilishida muhim rol o‘ynadi. Endilikda tabiiy tilning lisoniy modellari va nutqiy imkoniyatlarini kompyuter tiliga o‘tkazish, til bilan bog‘liq masalalarni axborot texnologiyalari hamda metodlari yordamida tabiiy tilni o‘rganishda korpus obyekt vazifasini bajarmoqda. Tabiiy tillarning rivojlanish barqarorligi, ularning milliy sofligini saqlab qolishga qaratilayotgan bir davrda tillarning elektron korpuslarini yanada takomillashtirish va yangi texnologiyalarni yaratish axborot asrining dolzarb masalalardan biri hisoblanadi.

Dunyo kompyuter lingvistikasi integrallashgan soha sifatida shiddat bilan rivojlanib borayotgan davrda qator usullar va metodlar orqali til muammolarini hal qilishda muhim vosita bo‘lib xizmat qilmoqda. Kompyuter texnologiyalarining til rivojiga va aksincha, til texnologiyasining kompyuter texnologiyalariga ijobiy ta’siri orqali qator ilmiy yutuqlarga erishildi. Jahonda kompyuter lingvistikasining turli yo‘nalishlari bo‘yicha ko‘plab olimlar tomonidan ilmiy izlanishlar olib borildi[1]. Buning natijasida fanning mashina tarjimasi, korpus lingvistikasi, kompyuter leksikografiyasi, tahrirlovchi dastur, nutqiy sintezator, lingvostatistik tahlil qiluvchi dastur, matnlarni referatlash va tasniflash kabi qator yo‘nalishlari vujudga keldi.

Bugungi kunda o‘zbek kompyuter lingvistikasi, tabiiy tilni qayta ishlash (NLP), mashinali ta’lim (machine learning), ma’lumot qidiruvi (Data Mining) kabi yangi sohalar kesimida rivojlanmoqda hamda ushbu sohalarda erishilayotgan ilmiy-amaliy natijalar bir-birini to‘ldirib bormoqda. O‘zbekiston Respublikasi Prezidentining 2019-yil 21-oktabrdagi “O‘zbek tilining davlat tili sifatidagi nufuzi va mavqeyini tubdan oshirish chora-tadbirlari to‘g‘risida”gi PF-5850-son Farmoniga muvofiq davlat tilining zamonaviy axborot texnologiyalari va kommunikatsiyalariga integratsiyallashuvini ta’minlashda qator ustuvor vazifalar ko‘rsatib o‘tildi[2]. O‘zbek tilining mavqeini mustahkamlash va uni til siyosati darajasida nufuzini oshirishda tilning raqamli resursini yaratishga doir qator dasturlar ishlab chiqilmoqda.

Ushbu tadqiqot O‘zbekiston Respublikasi Prezidentining 2016-yil 13-maydagi “Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universitetini

[1] Jurafskiy D., Martin J. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2007. – P. 12-13; Karin A., Beng A. Advances corpus-based contrastive linguistics. USA: John Benjamins, 2013 – P. 25-54; Koehn P., Och F.J., Marcu D. Statistical phrase based translation // Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT/NAACL). Proceedings of the Joint Conference 2003; Mitkov R. The Oxford handbook of Computational linguistics. Oxford university press, 2003; Kurdi M.Z. Natural Language Processing and Computational Linguistics: Speech, Morphology and Syntax. – Great Britain: Wiley-ISTE, 2016. – 300 p.; Чардин И.С. Лингвистические корпусы с разметкой на основе грамматики зависимостей и их применение при автоматическом синтактическом анализе: Автореф.дисс. …д-ра филол. наук. – Москва, 2004. – 24 с.

[2] O‘zbekiston Respublikasi Prezidenti Shavkat Mirziyoyevning 2020-yil 20-oktabrdagi «Mamlakatimizda o‘zbek tilini yanada rivojlantirish va til siyosatini takomillashtirish chora-tadbirlari to‘g‘risida»gi PF-6084-son farmoni // https://lex.uz/docs/5058351.

tashkil etish to‘g‘risida”gi PF-4797-sonli, 2017-yil 7-fevraldagi “O‘zbekiston Respublikasini yanada rivojlantirish bo‘yicha Harakatlar strategiyasi to‘g‘risida”gi PF-4947-sonli Farmonlari, 2019-yil 4-oktabrdagi PQ-4479-sonli “O‘zbekiston Respublikasining “Davlat tili haqida”gi Qonuni qabul qilinganligining o‘ttiz yilligini keng nishonlash to‘g‘risida”gi Qarori, O‘zbekiston Respublikasi Prezidentining 2019-yil 21-oktabrdagi PF-5850-sonli “O‘zbek tilining davlat tili sifatidagi nufuzi va mavqeini tubdan oshirish chora-tadbirlari to‘g‘risida”gi Farmoni, 2013-yil 27-iyundagi “O‘zbekiston Respublikasi Milliy axborot-kommunikatsiya tizimini yanada rivojlantirish chora-tadbirlari haqida”gi PQ-1989-sonli, 2017-yil 17-fevraldagi “Fanlar akademiyasi faoliyati, ilmiy tadqiqot ishlarini tashkil etish, boshqarish va moliyalashtirishni yanada takomillashtirish chora-tadbirlari to‘g‘risida”gi PQ-2789-sonli qarorlari hamda mazkur faoliyatga tegishli me’yoriy-huquqiy hujjatlarda belgilangan vazifalarni amalga oshirishda yuqorida belgilangan tilni rivojlantirishga doir ustuvor vazifalarni muayyan darajada hal qilishga xizmat qiladi.

Tadqiqotning respublika fan va texnologiyalari rivojlanishining ustuvor yo‘nalishlariga mosligi. Dissertatsiya respublika fan va texnologiyalari rivojlanishining I. «Axborotlashgan jamiyat va demokratik davlatni ijtimoiy, huquqiy, iqtisodiy, madaniy, ma’naviy-ma’rifiy rivojlantirish, innovatsion iqtisodiyotni rivojlantirish» ustuvor yo‘nalishi doirasida bajarilgan.

Dissertatsiyaning mavzu bo‘yicha xorijiy ilmiy tadqiqotlar sharhi[3].

Jahon kompyuter va korpus tilshunosligida korpus texnologiyasi, korpusshunoslik, korpus lingvistikasi va korpusga asoslangan yondashuv kabi yo‘nalishlarga doir izlanishlarda quyidagi ilmiy natijalarga erishilgan: zamonaviy ingliz tili, qadimgi ingliz tili, zamonaviy xorijiy tillar, kam sonli tillar, xavf ostida qolgan tillarni korpus yordamida o‘rganish va NLP va korpus metodlari ishlab chiqilgan, korpusning konkordans, tokenayzer, stemmizator, lemmatizator yaratilgan (Lankaster universiteti, Buyuk Britaniya); korpusga asoslangan diskurs tahlilga doir tadqiqotlar bajarilgan (Birmingem universiteti, Buyuk Britaniya); korpus menejerlari interfeyslari joriy etilgan (Vaseda universiteti,Yaponiya[4], Iova universiteti, AQSH); sintaktik annotatsiyalangan korpus (Penn Treebank) interfeysining formal-funksional modellari yaratilgan (Pensilvaniya universiteti, AQSH); dastlabki annotatsiyalangan Braun korpusining dastlabki varianti ishlab chiqilgan (Braun universiteti, AQSH); morfologik annotatsiyalangan korpus va morfoanalizator masalalariga doir nazariy va amaliy izlanishlar olib borilgan (Leypsig universiteti, Germaniya; Kolumbiya universiteti, AQSH; Abu Dabi Nyu York universiteti, Birlashgan Arab Amirligi); semantik lug‘at asosida morfologik omonimiyadan xoli korpus tahlillari olib borilgan, (Sankt-Peterburg universiteti, Rossiya); tobelanish nazariyasiga asoslangan grammatika va avtomatik usullar yordamida sintaktik razmetkalangan lingvistik korpus platformalari yaratilgan (Rossiya fanlar akademiyasi Axborotlarni uzatish Instituti qoshidagi Kompyuter

[3] Dissertatsiyasiga mavzusiga aloqador xorijiy tadqiqotlar sharhi http://ucrel.lancs.ac.uk/, google.scholar.com; https://www.researchgate.net, https://www.aclweb.org/anthology/L16-1207.pdf, http://qazcorpus.kz/indexen/, http://www.openslr.org/resources.php, http://ddi.itu.edu.tr/en/toolsandresources

[4] Laurens Antoni tomonidan korpus menedjerlari, jumladan AntConc, ParaConc, SegmentAnt kabilar yaratilgan.

lingvistikasi laboratoriyasi); sintaktik analizator yaratishning algoritmik tavsifi ishlab chiqilgan (M.V.Lomonosov nomidagi Moskva davlat universiteti, Rossiya); ta'limiy va lingvistik annotasiyalangan korpus tamoyillari va mezonlari o'rganilgan (Kaliforniya univeristeti, Kembrij universiteti, AQSH); vebga asoslangan Crawler instrumentariysi qo'llangan (Berlin universiteti, Germaniya); turkiy tillar bo'yicha lingvistik annotatsiyalangan korpusning morfologik, sintaktik va semantik teglash tizimi ishlab chiqilgan (Amaliy Semiotika instituti, Tatariston); ikki bosqichli morfologik analizatorni korpus tahliliga moslashtirish, gaplarni segmentlash, tokenayzer, sintaktik annotatsiyalangan korpusning universal tobelik nazariyasi, ko'p komponentli so'zlarning standartlarini ishlab chiqish kabi masalalar tadqiq qilingan (Istanbul texnika universiteti, Turkiya); nutq uslublari bo'yicha qozoq tilining millliy korpus menejeri ishlab chiqilgan (Ahmad Boytursunov nomidagi til instituti, Sun'iy intellekt instituti, Qozog'iston); korpusga asoslangan til ta'limi, tobelanish nazariyasiga asoslangan parsing, korpus morfologik tahlilida FST texnologiyasi, mualliflik korpusi, milly korpus yaratishning dasturiy va lingvistik ta'minoti, korpusning morfologik va semantik analizatori, mashina tarjimasining parallel korpuslar asosida neyro texnologiyalarni yaratish, o'zbek tilining ta'limiy va web-korpusini[5] shakllantirishga doir nazariy va amaliy tadqiqotlar olib borilmoqda (Toshkent davlat o'zbek tili va adabiyoti universiteti, O'zbekiston Milliy universiteti, Toshkent axborot-texnologiyalari universiteti, Samarqand davlat universiteti TATU filiali).

Muammoning o'rganilganlik darajasi. Dunyo tajribasida korpus yaratishning lingvistik, matematik va dasturiy tomonlari olimlarining tadqiqotlarida o'z ifodasini topgan[6]. Chunonchi, rus va ingliz tillari bo'yicha korpus lingvistikasi turli sohalar kesimida V.Zaxarov, A.Sedov, A.Baranov, R.Potapova, V.Rikov, U.Frensis, N.Leontyeva, V.Martin, S.Kubler, A.Laurens, E.Etwell, S.Hunston, L.Boizou, McKenneri, J.Grafmiller, J.Grieve, N.Grum, S. Hansson, K.McAuliff, M.Malberg, P.Milin, A.Murakami, R.Peych, A.Schembri, P.Tompson, B.Vinter, G.Lich kabi xorijiy olimlar[7] tomonidan hamda turkologiyada turk tili korpusi bo'yicha Aksan, Deniz Zeyrek, Kemal Oflazer, Umut Özge Bular; uyg'ur tili bo'yicha Yusup Aibaidulla, Kim-Teng Lua; boshqird tili bo'yicha L.A.Buskunbaeva, Z.Sirazitdinov; hakas tili bo'yicha Sheymovich, tatar tili

[5] http://uzschoolcorpara.uz; http://uzcorpus.uz/

[6] Kubler S., Zinsmeister H. Corpus linguistics and linguistically annotated corpora. – New York: Bloomsbury, 2015. –P. 321.; Martin W. Developing Linguistic Corpora: A Guide to Good Practice, OxfordBooks. 2005. http://www.ahds.ac.uk/creating/guides/linguistic-corpora/; Heike Z., Hinrichs E., Kübler S., Witt A. Linguistically annotated corpora: Quality assurance, reusability and sustainability / Corpus Linguistics: An International Handbook A. Lüdeling and M. Kytö (eds), Vol. 1, – Berlin: Mouton de Gruyter, –P. 759–76; Копотев М. Введение в корпусную лингвистику (учебное пособие). – Прага, 2014. – 264 с.; Atkins B., Zampolli A. Computational approach to the lexicon. – Oxford, 1994. –P. 494.

[7] Седов А.В. Математические модели, методы и алгоритмы построения размеченных корпусов текстов: Автореф. дис... канд. тех. наук. – Петрозаводск, 2013. – 22 с.; Anthony L. AntConc: Design and Development of a Freeware Corpus Analysis Toolkit for the Technical Writing Classroom, 2005. // IEEE International Professional Communication Conference Proceedings, – P.729-737; Atwell E. Development of tagsets for part-of-speech tagging / An international handbook. Corpus Linguistics: Mouton de Gruyter. 2008; Баранов А.Н., Михайлов М.Н., Сидоров Г.О. Динамический корпус текстов как новая технология прикладной лингвистики // Труды международного семинара Диалог'98 по компьютерной лингвистике и ее приложениям. – Т. 1998; Hunston S. Corpora in Applied Linguistics. Cambridge University Press, 2002. – 234 p.

bo‘yicha J.Suleymanov, A.Gatiatullin, O.Nevzorova, R.Gilmullin, B.Hakimov; qirimtatar tili bo‘yicha L.Kubedinova hamda tuva tili bo‘yicha A.Salchak kabi olimlarning ishlari diqqatga sazovor.

O‘zbekistonda kompyuter lingvistikasining shakllanishida dastlab lingvostatistikaga oid N.Yoqubova, M.Ayimbetov, S.Rizayev, S.Muhamedov kabi olimlar tomonidan izlanishlar olib borilgan[8]. S.Muhamedovning P.P.Piotrovskiy bilan hammualliflikda yozgan «Инженерная лингвистика и опыт системно – статистического исследования узбекских текстов» nomli kitobida lingvistik modellar, modellashtirish va uning tamoyillari, o‘zbekcha matnlarning kvantativ tahlillari o‘rganilgan. Shuningdek, so‘nggi o‘n yillikda kompyuter lingvistikasi sohasida A.Po‘latov, S.Muhamedova, A.Rahimov, Z.Xolmanova, N.Abduraxmonova kabi olimlarning sohaga doir o‘quv darslik va qo‘llanmalari nashrdan chiqdi[9]. Tilni kompyuter modellari, lingvistik ta’minotini yaratish, shuningdek, kompyuter metodlari yordamida lingvistik masalalarini yechishga yo‘naltirilgan monografik tadqiqotlar olib borildi[10].

Korpus lingvistikasida erishilgan natijalar kompyuter lingvistikasi va tabiiy tilni qayta ishlash sohalari uchun ham til texnologiyasi borasida tub burilishlar yasadi. Bunda tabiiy tilning mashina tilini yaratish, til bo‘yicha statistik ma’lumotga ega bo‘lish, sun’iy intellektning lisoniy va nutq modellarini shakllantirish, tilning leksikografik mashina fondini yaratish kabi qator amaliy ishlar uchun o‘rganish obyekti bo‘lib xizmat qiladi.

Korpus lingvistikasi turli fanlar uchun o‘rganish obyekti va vositasi sifatida foydalanilmoqda. Mazkur sohada Respublikamizning bir qator oliy ta’lim muassasalari, shuningdek, ilmiy tadqiqot institutlarida mazkur yo‘nalishlar bo‘yicha ilmiy izlanishlar olib borilmoqda. O‘zbek korpus lingvistikasida B.Mengliyev, Sh.Shahobiddinova, Z.Xolmanova, S.Karimov, L.Raupova, Sh.Hamroyeva, N.Abduraxmonova, G.Toirova, J.Djumabayeva, G.Ergasheva,

---

[8] Ризаев С. Ўзбек тилининг лингвостатистик тадқиқи: Филол. фан. д-ри. ... дисс. автореф. – Тошкент, 2008. – 50 б.; Мухамедов С.А. Статистический анализ лексико-морфологической структуры узбекских газетных текстов: Автореф. дисс. …канд. филол. наук. – Ташкент, 1980. –25 с.; Бектаев К.Б., Пиотровский Р.Г. Математическая лингвистика. – М.: Высшая школа, 1997. – 420 с.; Айымбетов М.К. Проблемы и методы квантитативно-типологического измерения близости тюркских языков (на материалах каракалпакского, казахского и узбекского языков): Автореф. дисс. …д-ра филол. наук. – Ташкент, 1997. – 47 с.

[9] Пўлатов А. Компьютер лингвистикаси. – Т.: Akademnashr, 2011. – 175 б.; Норов А. Компьютер лингвистикаси асослари. – Қарши, 2017. – 136 б.; Муҳаммедова С. Ҳаракат феъллари асосида компьютер дастурлари учун лингвистик таъмин яратиш. Методик қўлланма. – Тошкент, 2006.; Холманова З. Компьютер лингвистикаси (ўқув қўлланма) –Тошкент, 2019.; Abduraxmonova N.Z. Kompyuter lingvistikasi (darslik). – Toshkent: Nodirabegim, 2021. –398 b.

[10] Абдурахмонова Н.З. Инглизча матнларни ўзбек тилига таржима қилиш дастурининг лингвистик таъминоти (содда гаплар мисолида). филол.фан.бўйича фалсафа доктори (PhD)…дисс. – Тошкент, 2018. – 165 б.; Абжалова М. Ўзбек тилидаги матнларни таҳрир ва таҳлил қилувчи дастурнинг лингвистик модуллари (Расмий ва илмий услубдаги матнлар таҳрири дастури учун): филол.фан.бўйича фалсафа док. (PhD)…дисс. – Фарғона, 2019. – 164 б.; Хамроева Ш. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари: филол. фан. бўйича фалсафа док. (PhD)…дисс. – Қарши, 2018. – 250 б.; Эшмўминов А.А. Ўзбек тили миллий корпусининг синоним сўзлар базаси: филол. фан. бўйича фалсафа док. (PhD)…дисс. – Қарши, 2019. – 140 б.; Тоирова Г. Ўзбек тили миллий корпусини яратишнинг назарий ва амалий масалалари. – Германия: Globeedit, 2020. –168 б.

A.Eshmo‘minov, Sh.Gulyamova, M.Abjalovalarning tadqiqotlarini qayd etish o‘rinli[11].

Tadqiqotning dissertatsiya bajarilgan oliy ta’lim muassasasining ilmiy-tadqiqot rejalari bilan bog‘liqligi. Dissertatsiya Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universitetining “Development of the interdisciplinary master program on Computational Linguistics at Central Asian universities 585845-EPP-1-2017-1-ES-EPPKA2- CBHE-JP” (2018-2021) mavzusidagi xalqaro Erasmus+ loyihasi doirasida bajarilgan.

Tadqiqotning maqsadi o‘zbek tili elektron korpusini yaratishning konseptologik modellarini tuzish hamda lingvistik annotatsiyalangan til korpusi uchun kompyuter texnologiyalari metodlaridan foydalanishning samarali usullarini ishlab chiqishdan iborat.

Tadqiqot vazifalari:

o‘zbek tili elektron korpusini konseptologik va strukturaviy loyihalashda xorijiy tajriba amaliyotini o‘rganish;

tilning lingvistik korpusini yaratishda morfologik va sintaktik teglash va tahlil qilishning FST va UdPipe kabi avtomatik usullarini o‘zbek tiliga tatbiq qilish orqali lingvistik algoritmni tuzish hamda lisoniy modellarini mashina tiliga o‘tkazish;

matn fragmentining reprezentativligi va qidiriv birliklari (lemma va token)ni tahlil qilish uchun matn korpusining lingvistik va dasturiy ta’minotini tuzish;

o‘zbek tili uchun korpus yaratish texnologiyalari va metodlarini lingvistik instrumentariylar yordamida amalga oshirish;

korpus menejerining formal-funksional modellari asosida korpus interfeysini shakllantirish.

Tadqiqotning obyekti sifatida o‘zbek tilidagi rasmiy, ilmiy, badiiy va publitsistik uslubdagi turli janrlarga tegishli yozma matnlar hamda parallel korpus uchun ingliz va rus tilidagi rasmiy va ilmiy materiallar tanlangan.

Tadqiqotning predmetini elektron korpus yaratishning texnologiyalari va usullari, korpusni grammatik teglash va tahlil qilish modellari, o‘zbek tili korpus menejerining lingvistik va dasturiy ta’minoti tashkil qiladi.

Tadqiqot usullari. Tadqiqot mavzusini yoritishda tavsiflash, chog‘ishtirish, komponent tahlil, statistik, modellashtirish kabi metodlardan foydalanildi.

Tadqiqotning ilmiy yangiligi quyidagilardan iborat:

o‘zbek tilida ilk bor *FST* (finite state trunsducer) – chekli avtomat transyuteri vositasida korpusni morfologik teglashning avtomatik usuli yordamida lemmatizatsiya va tokenizatsiya jarayonlarining lingvistik tahlil bosqichlari ishlab chiqilgan hamda morfologik ma’lumotlar bazasi, morfotaktik qoidalar tizimi shakllantirilgan;

[11] Mengliyev B., Hamrayeva Sh. Korpus lingvistikasi: korpus tuzish va undan foydalanish. – T.: Globeedit. 2020. – 50 b.; Хамроева Ш. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари: филол. фанлари бўйича фалсафа докт. дисс. – Бухоро, 2018. – 165 б.; Mengliev B., Shahabitdinova Sh., Khamroeva Sh., Gulyamova Sh., Botirova A. The morphological analysis and synthesis of word forms in the linguistic analyzer // Linguistica Antverpiensia (1), 2021. – P. 703-712; Тоирова Г. Миллий корпус яратишнинг технологик жараёни хусусида. // Ўзбекистонда хорижий тиллар. Электрон илмий-методик журнал. – Тошкент, 2020. – № 2 (31) – Б.57-64. https://journal.fledu.uz/uz/ 2-31-2020.

korpusni sintaktik teglash va annotatsiyalashda universal tobelik nazariyasining UdPipe metodi orqali *CONLL-U* format namunasida o‘zbek tilining sintaktik modellari aniqlangan;

olib borilgan tadqiqotlarning amaliy ifodasi sifatida http://uzbekcorpus.uz sayti yaratilgan bo‘lib, o‘zbek tili korpus menejeri (qidiruv tizimi)ning lemma, token bo‘yicha hamda *n-gram* modeli asosida konkordanslarning qidiruv tizimi ishlab chiqilgan;

tarjima qilish algoritmi va o‘zbek tili parallel korpusining tarjima xotirasini yaratishda WordFast texnologiyasi foydalanilgan hamda o‘zbek tilining ingliz va rus tillarida tarjima birliklari muqobillarining ma’lumotlar bazasi tuzilgan;

korpusni grammatik teglash bosqichida Protege dasturi orqali ontologik modellashtirish usulining ilmiy asoslari isbotlangan;

korpusda qidirilayotgan so‘zning imloviy xatolarini aniqlash va unga yaqin bo‘lgan so‘zlar koeffisentini belgilashda qidiruv interfeysida ilova qilishda *Djaro Winkler* algoritmini qo‘llashning maqsadga muvofiq ekanligi asoslangan.

Tadqiqotning amaliy natijalari quyidagilardan iborat:

korpusning lingvistik bazasi uchun asos bo‘luvchi turli til sathlariga doir leksik birliklarning ma’lumotlar bazasi yaratilib, mualliflik guvohnomalari[12] olingan;

tadqiqotda erishilgan natijalar http://uzbekcorpus.uz saytida o‘z ifodasini topgan;

elektron korpusning kompyuter modellaridan biri sifatida olingan ilmiy xulosalar “Mahalla va oila” ilmiy-tadqiqot institutida amalga oshirilgan qisqa muddatli JHBL- 20-sonli “Oila, mahalla va gender tengligi mavzusidagi badiiy asarlarning elektron korpusini yaratish” (2020-2021) nomli loyihada foydalanilgan;

korpusning morfologik annotatsiyasi uchun qo‘llangan FST texnologiyasi vositasida *nominal (otli) va verbal (fe’lli*) guruhlarning morfotaktik qoidalari asosida lingvistik ma’lumotlar bazasi yaratilgan;

elektron korpusni grammatik teglash va annotatsiyalash jarayonida erishilgan natijalar turkiy tillarning grammatik tezaurusini yaratishda foydalanilgan[13];

tadqiqot natijalaridan korpusning morfologik va sintaktik analizatorini yaratishning kompyuter modellari orqali “Development of the interdisciplinary master program on Computational Linguistics at Central Asian universities 585845-EPP-1-2017-1-ES-EPPKA2- CBHE-JP” (2018-2021) mavzusidagi xalqaro Erasmus+ loyihasida qo‘llanilgan;

o‘zbek tili korpusi uchun lemmatizatsiya va tokenizatsiya jarayonlarining lingvistik tahlil bosqichlari ishlab chiqilgan hamda morfologik ma’lumotlar bazasi, morfotaktik qoidalar tizimini yaratish asosida “Lingvistik protsessor va elektron korpuslar uchun o‘zbek tilining morfologik bazasi”ni shakllantirishda qo‘llanilgan (O‘zbekiston Respublikasi Adliya vazirligi huzuridagi Intellektual

[12] Abduraxmonova N., Tuliyev U. Oila, mahalla va gender tengligi mavzusidagi badiiy asarlar elektron korpusining dasturiy ta’minoti. Intellektual mulk agentligi guvohnomasi 30.03.2021 DGU 10653; Abduraxmonova N., Tuliyev U., Maharov Q. “Oila ma’naviyati” konseptining elektron tezaurusining dasturiy ta’minoti. Intellektual mulk agentligi guvohnomasi 30.03.2021DGU 10655.

[13] ИРН: AP05132249 «Разработка электронных тезаурусов тюркских языков для создания систем многоязычного поиска и извлечения знаний» по договору № 132 от «12» марта 2018 г. http://alphabet.kz:9191/

mulk agentligi tomonidan 2021.19.02.da berilgan BGU 00394-raqamli ilmiy guvohnoma);

o‘zbek tili uchun ilk bor *FST* (finite state trunsducer) – chekli avtomat transyuteri vositasida korpusni morfologik teglashning avtomatik usuli *Uzmorphoanalyzer* – o‘zbekcha so‘zlarni morfologik tahlil qiluvchi dasturni yaratishda foydalanilgan (O‘zbekiston Respublikasi Adliya vazirligi huzuridagi Intellektual mulk agentligi tomonidan 2021.17.06.da berilgan DGU 11432 raqamli guvohnoma);

Tadqiqot natijalarning ishonchliligi korpus va kompyuter lingvistikasida metodologik jihatdan asoslangan nazariy qarashlar, umummetodologik asoslarga ega materiallarga tayanib, ilmiy xulosalarga kelingani, qolaversa, dissertatsiyada qo‘yilgan vazifalarni hal qilishning aprobatsiyasi, tavsiyalarning tajribada sinalganligi, olingan natijalarning vakolatli tuzilmalar tomonidan tasdiqlangani bilan izohlanadi.

Tadqiqot natijalarining ilmiy va amaliy ahamiyati. Tadqiqot natijalarining ilmiy ahamiyati tavsiya qilingan va ishlab chiqilgan tizimlashtirilgan matn tuzilmalari reprezentativligi hamda matnning turli lingvistik bosqichlarda tahlil qilish bo‘yicha kiritilgan takliflar kelgusida o‘zbekcha matnlarni qayta ishlash texnologiyasi, parallel matnlar korpusi, tor soha vakillari uchun o‘rganish obyekti til va nutq hodisalariga yo‘naltirilgan ilmiy tadqiqotlarni olib borishda yordam beradi.

Tadqiqot natijalarining amaliy ahamiyati tarjima amaliyotida muhim manba bo‘lishi, olingan ilmiy-nazariy qarashlar oliy o‘quv yurtlari uchun «Kompyuter lingvistikasi», «Korpus lingvistikasi», «Mashina tarjimasi» kabi fanlardan ma’ruzalar o‘qish, darslik va qo‘llanmalar yaratish, elektron lug‘atlar (tarjima, tezaurus) tuzish jarayonida foydalanish mumkinligi bilan izohlanadi.

Tadqiqot natijalarining joriy qilinishi. O‘zbek tili elektron korpusining kompyuter modellari bo‘yicha olingan ilmiy va amaliy natijalar sifatida lingvistik annotatsiyalash tamoyillari, lisoniy modellari, avtomatik usullari (FST texnologiyasi va UdPipe usuli), korpus menejeri, qidiruv tizimi uchun tahrirlovchi dastur kabi tadqiqot xulosalaridan:

tadqiqotda amaliy natija sifatida yaratilgan http://uzbekcorpus.uz sayti web resurs sifatida Qozon Federal universitetining Hisoblash matematikasi va axborot texnologiyalari institutida Kompyuter lingvistikasi faniga doir korpus lingvistikasi hamda morfologik analizator kabi mavzularni o‘qitishda foydalanilmoqda (Qozon federal universiteti Hisoblash matematikasi va Axborot texnologiyalari instituti ma’lumotnomasi, 2021). Natijada o‘zbek tilining elektron korpusi (http://uzbekcorpus.uz) http://www.turklang.net saytida turkiy til korpuslar ro‘yxatida ilova qilinib, turkiy tillarning struktur-funksional xususiyatlarini qiyoslash hamda lingvistik resurs sifatida tabiiy til texnologiyalardan foydalanish imkoniyatini yaratgan;

ilmiy natijalarning amaliy ifodasi sifatida korpus menejeri (qidiruv tizimi)ning lemma, token hamda *n-gram* modeli asosida konkordanslarning qidiruv tizimi orqali korpusda jamlangan ilmiy uslubdagi materiallar OT-F1-029. raqamli “O‘zbek davlatchiligi tarixida sug‘diy til va yozuvining tutgan o‘rni (mil.av. II –

milodiy XII asrlar)” (2017-2020) mavzusidagi fundamental grant loyihasini bajarishda maqolalar va monografiyalarni tayyorlashda ilmiy elektron manba sifatida hamda platformada yaratilgan parallel korpuslar va tarjima lug‘atlardan loyihaning amaliy natijalarini xorijiy jurnallarda chop etishda foydalanilgan. Natijada loyiha bo‘yicha olib tadqiqotlarning mazmuniy kontenti boyigan, ilmiy darajasi oshgan;

2017-2019-yillarda amalga oshirilgan OT-F1-78-“Hozirgi globallashuv davrida o‘zbek tili, uning tarixiy taraqqiyoti va istiqbollari (vazifaviy uslublar tahlili asosida)”nomli fundamental loyihasida o‘zbek tili lotin yozuvidagi ko‘p jildli “O‘zbek tilining izohli lug‘ati”ni tayyorlashda dissertatsiya natijasi sifatida yaratilgan o‘zbek tilining elektron korpus menejeri (qidiruv tizimi)ning lemma, token, konkordans hamda so‘z birikmasi bo‘yicha qidirish kabi funksional imkoniyatlaridan foydalanilgan. Shuningdek, elektron korpusdagi besh turdagi uslubda jamlangan yozma matn fragmentlari misolida “O‘zbek tili uslubshunosligi” o‘quv qo‘llanmasida foydalanilgan. (O‘zbekiston Respublikasi Fanlar akademiyasining 2021-yil 16-sentabrdagi №3/1255-2651sonli ma’lumotnomasi). Natijada loyiha bo‘yicha olib borilgan tadqiqotlarning mazmuniy kontenti kengaygan, maqolalarning turli kontekstda semantik maydonlarda voqelanishi korpus asosida aniqlangan;

o‘zbek tili korpus menejerining lemma, token hamda n-gram modeli asosida konkordanslarning qidiruv tizimiga doir amaliy natijalaridan “Mahalla va oila” ilmiy tadqiqot institutining JHBL-18 raqamli “Turli avlod vakillarining oilaviy qadriyatlarini tadqiq etish” (2020-2021) loyihasida foydalanilgan (“Mahalla va oila” ilmiy tadqiqot instituti 2021-yil 25-iyundagi 13-son ma’lumotnomasi). Natijada tadqiqotchilar uchun tilda aks etgan qadriyatlarning gender, yosh, va jinsga doir, qolaversa, milliy xususiyatlarini o‘rganish va ilmiy jihatdan statistikasini aniqlashda asos vazifasini bajargan;

o‘zbek tili elektron korpusini yaratishning amaliy natijalaridan JHBL-18 raqamli “O‘zbekistonda “Baxtiyor oila” web portalini yaratish” (2020-2021) loyihasida foydalanilgan. Elektron korpusda turli janrga oid oila mavzusidagi matnlarning xronologiyasi va ilmiy tavsifini tahlil qilish milliy qadriyatlar va ijtimoiy munosabatlarni o‘rganishda oilaning institut sifatida maqomini belgilashga xizmat qilgan (“Mahalla va oila” ilmiy tadqiqot instituti 2021-yil 25-iyundagi 12-son ma’lumotnomasi). Natijada loyiha mavzusiga doir tushunchalar yangi ilmiy dalillar bilan boyitilgan;

tadqiqotda amaliy natija sifatida yaratilgan o‘zbek tili korpus menejerining lemma, token bo‘yicha hamda n-gram modelli konkordans qidiruv tizimidan I-OT-2019-42 raqamli “O‘zbek va ingliz tillarining elektron (Inson qiyofasi, fe’l atvori, tabiat va milliy timsollar tasviri) poetik lug‘atini yaratish” (2019-2021) mavzusidagi amaliy grant loyihasida elektron lug‘at kontentini materiallari bilan boyitishda foydalanilgan (Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universitetining 01/4-1638 raqamli 2021-yil 18-sentabrdagi ma’lumotnomasi). Natijada o‘zbek tili elektron korpusining ichki korpus – parallel korpus tarjima xotirasi va tarjima segment birliklaridan foydalanib, lug‘atda aks etgan o‘zbek va ingliz tillarida inson qiyofasi, fe’l-atvori, tabiat va milliy

timsollarni ifodalovchi leksik birliklarni to‘plash va ularni ingliz va o‘zbek tillaridagi muqobillarini aniqlashga erishilgan;

O‘zbek tili korpusi uchun lemmatizatsiya va tokenizatsiya jarayonlarining lingvistik tahlil bosqichlari ishlab chiqilgan hamda morfologik ma’lumotlar bazasi, morfotaktik qoidalar tizimi hamda o‘zbek tili uchun ilk bor *FST* (finite state trunsducer) – chekli avtomat transyuteri vositasida korpusni morfologik teglashning avtomatik usullar, korpusni sintaktik teglash va annotatsiyalashda UdPipe metodi orqali o‘zbek tilining sintaktik modellari bazasi asosida o‘zbek tili uchun universal tobelik nazariyasining *CONLL-U* modeli 5111200 – O‘zbek tili va adabiyoti, 5120100 – Filologiya va tillarni o‘qitish (o‘zbek tili), 5120900 – O‘zbek-ingliz tarjima nazariyasi va amaliyoti yo‘nalishlari bakalavr talabalari, magistrlar, doktorantlar, tadqiqotchilar va soha mutaxassislari, shuningdek, malaka oshirish va pedagogik kadrlarni qayta tayyorlash kursi tinglovchilariga mo‘ljallangan Kompyuter lingvistikasida darsligida o‘z ifodasini topgan. (Oliy va o‘rta maxsus ta’lim vazirligining 2019-yil 20-iyuldagi 654-sonli buyrug‘i). Natijada darslikning g‘oyaviy asoslari nazariy va amaliy jihatdan asoslangan hamda ilmiy ma’lumotlar bilan boyitilgan.

Tadqiqot natijalarining aprobatsiyasi. Mazkur tadqiqot natijalari 21 ta xalqaro, 21 ta respublika ilmiy-amaliy anjumanda e’lon qilingan. Muallifning quyidagi https://scholar.google.com/citations?user=kEkD0kgAAAAJ&hl=en, https://www.researchgate.net/profile/Nilufar-Abdurakhmonova, https://orcid.org/0000-0001-9195-5723 kabi platforma profillarida ilmiy ishlari muhokama qilingan.

Tadqiqot natijalarning e’lon qilinganligi. Dissertatsiya mavzusi bo‘yicha 65 ta ilmiy ish, jumladan, O‘zbekiston Respublikasi Oliy attestatsiyasi komissiyasining doktorlik dissertatsiyalari asosiy ilmiy natijalarini chop etish tavsiya etilgan ilmiy nashrlarda 12 ta maqola (shundan ikkitasi xorijiy jurnalda), 1 ta monografiya, 3 ta mualliflik guvohnoma, 1ta Web of science va 3 ta Skopus bazasiga indekslangan nufuzli xalqaro hamda Respublika konferensiyalarda 46 ta ilmiy maqola va tezis e’lon qilingan.

Dissertasiyaning tuzilishi va hajmi. Dissertatsiya kirish, to‘rt bob, xulosa, foydalanilgan adabiyotlar ro‘yxatidan iborat bo‘lib, hajmi 220 sahifani tashkil qiladi.

DISSERTATSIYANING ASOSIY MAZMUNI

Kirish qismida tanlangan mavzuning dolzarbligi va zarurati asoslab berilgan, muammoning o‘rganilganlik darajasi, mavzu bo‘yicha xorijiy ilmiy tadqiqotlar sharhi, tadqiqotning maqsad va vazifalari, obyekti va predmeti, fan va texnologiyalar taraqqiyotining ustuvor yo‘nalishlariga mosligi ko‘rsatilgan, ilmiy yangiligi va amaliy natijalari bayon qilingan. Olingan natijalarning ishonchliligi asoslangan holda nazariy va amaliy ahamiyati ochib berilgan. Tadqiqot natijalarining amaliyotga joriy etilishi, ishning aprobatsiyasi, e’lon qilingan ishlar va dissertatsiyaning tuzilishi bo‘yicha ma’lumotlar keltirilgan.

Dissertatsiyaning birinchi bobi "*Korpus – til texnologiyasining o'rganish obyekti"* deb nomlangan. Bobning "*Zamonaviy korpus lingvistikasining analitik tahlili"* mavzusidagi birinchi faslida korpus lingvistikasi va korpus texnologiyalariga bildirilgan ilmiy qarashlar tahlil qilinib, korpusning model yoki metodologiya sifatida berilgan ilmiy tavsifi, kompyuter va korpus lingvistikasida korpus texnologiyasi va korpusga asoslangan tahlil kabi yo'nalishlarda jahon va turkologiyada Graeme Kennedi, Friedrich Wilhelm Kaeding, Martin Weisser, Viktor Zaxarov, Charlotte Taylor[14], Charlez Meer[15], Noam Chomskiy, Dougles Biber[16], Sandra Kübler[17] kabi olimlarning tadqiqotlari misolida tahlilga tortilgan va ularga munosabat bildirilgan. Dastlabki elektron yozma korpuslar sifatida Braun (1964), LOB (1978), FROWN (1999), FLOB (1998), Kolhapur (1978), ACE (1986), Lund korpusi; og'zaki korpuslar sifatida SEU, LLC, SEC (ingliz tilining o'g'zaki matn korpusi), "Map task" korpusi hamda HKCSE (inglizcha og'zaki matnlarning gong gonk korpusi) kabilar solishtirib, ularning kontenti va formal-funksional modellari tahlil qilingan.

"*Korpuslar taksonomiyasi*" deb nomlangan ikkinchi fasl korpuslarning turlari muayyan parametrlarga ko'ra tasniflashi masalasiga bag'ishlangan. Unga ko'ra korpuslar 1) ma'lumotlar bazasining turlari: a) og'zaki; b) yozma; d) aralash; 2) matnni muayyan tilda berilishi: a) ingliz; b) rus; d) nemis; 3) matn tarjimalarining paralleligiga ko'ra: a) ikki tillik; b) uch tillik; d) ko'p tillik; 4) uslubiga ko'ra: a) so'zlashuv; b) publitsistik; d) badiiy; e) rasmiy; f) ilmiy; 5) bazadan foydalanish imkoniyatiga ko'ra: a) ochiq; b) yopiq; 6) geografik holatiga ko'ra: muyyan geografik joy yoki bir nechta davlatga tegishli ekanligi; 7) korpus kontentiga ko'ra: a) umumiy; b) maxsus turga bo'linishi qayd etilgan.

M.Z.Kurdi og'zaki matn korpuslarini ma'lumotlar bazasiga kiritishda og'zaki nutqdan yozib olish, gapiruvchi buyruq operatorlari, inson-mashina dialogi, mashina yordamidagi inson-inson dialogi hamda ko'p modelli dialog tizimi orqali amalga oshirish mumkinligini qayd etadi. Muvozanatlashgan korpus, piramidali va imkoniyat darajasi cheklangan korpus turlari tematik ko'lamiga ko'ra farqlanadi. Lingvografik korpusning manbalar moduli, lingvografik tavsif, leksik birliklar kabi uch qismdan iborat ekanligi ilmiy asarlarda qayd etilgan fikrlar asosida dalillangan[18]. Ushbu faslda yuqorida keltirilgan tasniflarga xos turli tillarda yaratilgan korpuslarning tavsifi va funksional imkoniyatlari haqida ma'lumotlarga munosabat bildirilgan.

Birinchi bobning "*Kompyuter lingvistikasida korpusshunoslik masalalari"* deb nomlangan uchinchi faslida matnlarni yig'ish, tasniflash, annotatsiyalash, mashina uchun tabiiy tilni tushunish va mashina tarjimasi jarayonini lingvistik

[14] Taylor Ch. What is corpus linguistics? What the data says. // ICAME Journal, 2008. – № 32. – P. 179-200.
[15] Charlez M. English corpus linguistics: An introduction. – Cambridge: Cambridge University Press, 2004. – 168 p.
[16] Bern H., Heiko N. The Oxford Handbook of Linguistic Analysis. / Dougles Biber Corpus-based and Corpus-driven analysis of language variation and use. –UK: Oxford University, 2015. – P. 193.
[17] Sandra K., Heike Z. Corpus linguistics and linguistically annotated corpora. – New York: Bloomsbury Academic. 2015. – P. 4.
[18] Каримуллина Р. Н., Каримуллина Г. Н. О сводном лингвографическом корпусе татарского языка / Слово и словарь vocabulum et vocabularium // Сборник научных материалов МИНСК ИЗДАТЕЛСТВО «ЧЕТЫРЕ ЧЕТВЕРТИ» 2017. – С. 18.

jihatdan modellashtirish (NLP-natural language processing) kabi masalalarni yechishda korpusning amaliy jihatlari tahlilga tortilgan.

Ilmiy manbalarda korpus lingvistikasi tabiiy tilni qayta ishlash sohasining (NLP) bir yo‘nalishi sifatida qayd etiladi[19]. Ayrim ma’lumotlarda esa kompyuter lingvistikasi yoki amaliy lingvistikaning muayyan yo‘nalishi sifatida qaraladi. Jurafskiy, Martin (2008), Manning, Schütze (1999), Roark, Sproat (2007) kabi olimlarning nuqtayi nazaricha, matnlar korpusini annotatsiyalash va uni lingvistik tahlil qilishda kompyuter lingvistikasi o‘zining optimal usullariga ega bo‘lib, murakkab masalalarni yechishda muhim ahamiyat kasb etadi[20]. Graeme Kennedining fikricha[21], garchi korpus lingvistikasi kompyuter texnologiyalarning rivojlanish bosqichida mavjud bo‘lmasa-da, kompyuter xotirasi imkoniyatlarining kengayishi natijasida lingvistik tahlil muammolarini matn asosida o‘rganish sohada ayrim muammolarga yechim bo‘la oldi.

Turli maqsadlardan kelib chiqib korpus dizayni, hajmi hamda har bir korpusning individual xarakteri shakllantiriladi. Korpus kompyuter lingvistikasining mashina tarjimasi, nutq sintezatori, matn tahlili, sentiment tahlil va boshqa yana bir qator yo‘nalishlar uchun tadqiqot obyekti vazifasini bajaradi. Chunonchi, statistik mashina tarjimasi texnologiyasi katta hajmdagi korpuslardan olingan ma’lumotlarga asoslanadi. Shu bois kompyuter lingvistikasida til va nutq bilan bog‘liq hodisalar korpus texnologiyasi yordamida o‘rganiladi.

Dissertatsiya ishining “*Elektron korpus yaratish texnologiyalari va usullari*” deb nomlangan ikkinchi bobning birinchi fasli “Korpus yaratish texnologiyalari: lingvistik instrumentlar” deb nomlanadi. Ushbu faslda korpus yaratish uchun tayyor holda foydalanish mumkin bo‘lgan instrumentariy va platformalarning tadqiqiga doir masalalar yoritilgan. CLARIN platformasi umumiy til resurslari va texnologiyalari infrastrukturasiga doir turli lingvistik instrumentlarni qo‘llab-quvvatlaydi. Dunyo tillari bo‘yicha korpus texnologiyasiga mo‘ljallangan instrumentlar mazkur saytda https://corpus-analysis.com joylashgan bo‘lib, mazkur faslda AntGram, AntConc, AntWordProfiler, AntCorGen, BFSU Collocator, BFSU Sentence Collector, BNCWeb, LiC, FrameNet, WordSmith kabi korpus interfeyslari tayyor instrumentariylar sifatida korpus yaratish hamda matn bilan bog‘liq tahlillarni amalga oshirishning funksional imkoniyatlari ochib berilgan.

Bobning ikkinchi fasli “*Maxsus korpuslar yaratishning dasturiy vositalari*” deb nomlanib, foydalanuvchilarning muayyan maqsadlari yoki tadqiqot yo‘nalishiga ko‘ra uncha katta bo‘lmagan matn janrlari va hajmiga ko‘ra teng taqsimlangan o‘zbek tili uchun maxsus korpus yaratishning *AntConc va BootCat* ilovalaridan foydalanish imkoniyatlari ochib berilgan.

Laurens Antoni tomonidan yaratilgan *AntConc* dasturi so‘z va kalit so‘zlar generatori, konkordanseri, *n-gram* hamda so‘z birikmasi kabi korpus tahliliga

19 Mohamed Z.K. Natural Language Processing and Computational Linguistics: Speech, Morphology and Syntax. – Great Britain, USA: Wiley-ISTE 2016. –P. 12.

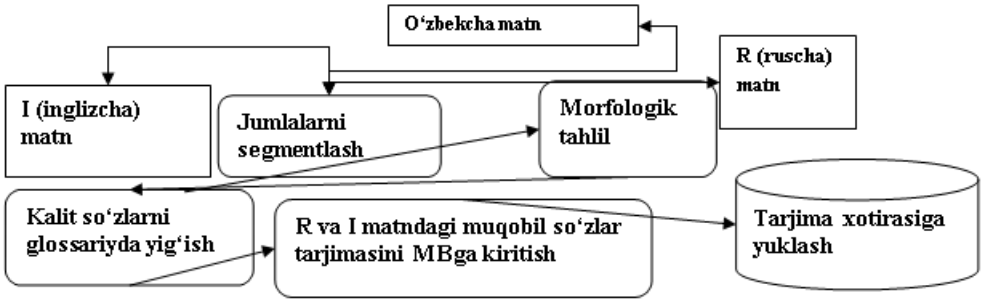
[20] Jurafsky D., Martin J. H. Speech and language processing, 2nd ed. Upper Saddle River: Prentice Hall, 2008; Manning, C., Schütze H. Foundations of statistical natural language processing. –Cambridge: The MIT Press, 1999; Roark, B., Sproat R. Computational approaches to morphology and syntax. – Oxford: Oxford University Press, 2009. – 316 p.

[21] Graeme K. An introduction to corpus linguistics. – London: Longman, 1998. – P. 2.

mo‘ljallangan. AntConc ilovasi uch asosiy parametrdan iborat: korpus fayllari, tablar, qidiruv interfeysi va uning moslamasi. Korpus fayllari ikki xil imkoniyatga ega: fayllarni alohida ochish (*open files*) yoki bir vaqtning o‘zida bir nechta papkalarga joylangan fayllarga murojaat qilish (*open directories*). Tablar panelida quyidagi instrumentlar joylashgan: qidiralayotgan so‘zning matn tarkibida kelish grafigi (plot), konkordanser (concordance-KWIC kines), so‘zlar chastotasini (frequincy) aniqlovchi parametr, kalit so‘zlar (key words). Menyu bo‘limidan foydalanilayotgan tilning yunikodini aniqlash, *n-gram* bo‘yicha so‘zlarning rangini tanlash, teglarning metama’lumotlarini berish tartibini o‘zgartirish yoki umuman yashirish vazifasini bajaradi; tokenlar uchun maxsus funksiya kiritilgan bo‘lib, harflar, raqamlar, tinish belgilar, simvollar, oraliq masofa kabi tokenlar guruhini belgilash imkoniyati mavjud. Shuningdek, ishimizda *BootCat* instrumenti internetdagi ma’lumotlarni qisqa muddatda to‘plash imkoniyatiga egaligi hamda uning yordamida muayyan termin orqali sohaviy terminologik bazani yaratish imkoniyati mavjudligi namunalar misolida dalillangan.

Bobning uchinchi fasli “*Web korpusni tuzishda lingvistik instrumentariylar va platformalar*” deb nomlanib, unda internet sahifalaridan olingan matnlar asosida korpus yaratishning Sketch Engine platformasi va unda ilk bor o‘zbek tili uchun yaratilgan *Turkic web-uzbek* korpusi haqida ma’lumot keltirilgan. Bundan tashqari WordSmith, BlackLab kabi dasturiy ilovalar yordamida korpus yaratish hamda matn korpusini avtomatik annotatsiyalash funksiyalariga ega WebLicht va CLARIN-D platformalari haqida ma’lumotlar keltirilgan.

Bobning “*Parallel korpus yaratishning kompyuter usullari*” deb nomlangan to‘rtinchi fasli ikki tilli o‘zbekcha-inglizcha, inglizcha-o‘zbekcha, o‘zbekcha-ruscha, ruscha-o‘zbekcha parallel matn yaratishning gap segmentlariga ajratish, tarjima birliklari bo‘yicha ularning muqobil variantlarini aniqlash, tarjima xotirasi va glossariysini yaratishda *WordFast* texnologiyasining afzalliklari va funksional imkoniyatlari ko‘rsatib o‘tilgan. Shuningdek, olib borilgan izlanishlar natijasi sifatida parallel matn yaratishning algoritmi ishlab chiqilgan:



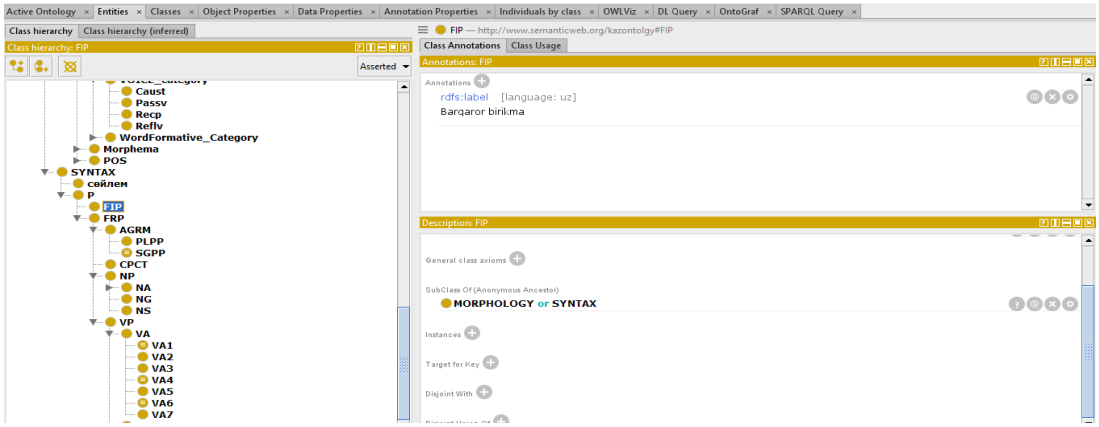
Dissertatsiyaning uchinchi bobi “*Korpusni grammatik teglash va tahlil qilish modellari”* deb nomlangan. Uning “*Lingvistik annotatsiyalash (razmetka) turlari*” deb nomlangan birinchi faslida matnni annotatsiyalash quyidagicha tasniflangan[22]: ekstralingvistik (metarazmetka) – matn (muallifi, nashr yili, nashriyot nomi, yili, janri, mavzusi) va uning muallifi haqida ma’lumot; 2) matn tuzilishi – sarlavha, satrboshi, gap, so‘z shakli; 3) lingvistik sathiga ko‘ra: a) morfologik (POS-tagging)

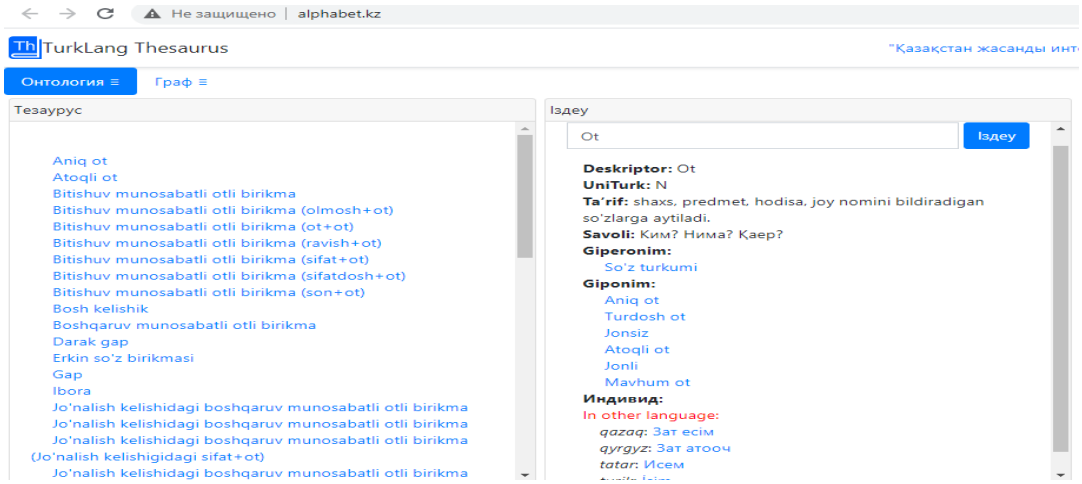
[22] Боярский К. К. Введение в компьютерную лингвистику. – Санкт-Петербург, 2013. – С. 28.

razmetka; b) sintaktik razmetka – bunda so‘zlar o‘rtasidagi sintaktik aloqalar va gap turlariga ko‘ra ma’lumotlar keltiriladi; d) semantik razmetka – matnda ifodalangan tushunchalarning predmet va predmet bo‘lmagan nomlari, faoliyat turlari hamda semantik munosabatlariga ko‘ra farqlanadi; e) anaforik razmetka – matnning biror elementi boshqa matnda anglashilgan mazmun bilan aloqadorlikda aniqlanadi va lingvistik jihatdan tahlil qilinadi; f) prosodik razmetka – urg‘u, ritm, mantiqiy urg‘u va boshqalariga ko‘ra matnning lingvistik jihatdan ifodalanishi qayd etilgan. Shuningdek, Lich[23] qoidasiga asoslangan annotatsiyalash[24] mezonlari sifatida quyidagilarga amal qilish maqsadga muvofiqligi qayd etilgan**:** 1) korpusdan annotatsiya qismi olib tashlanganda, u oddiy korpusga aylana olishi; 2) matndan olingan annotatsiyalar o‘z-o‘zini kengaytirish imkoniyatiga ega bo‘lishi hamda matn qolipiga mos bo‘lgan boshqa matndan annotatsiyalarni qabul qilishi; 3) annotatsiya modeli so‘nggi foydalanuvchi uchun qo‘llanish yo‘riqnomasiga asoslanishi; 4) annotatsiya kim tomonidan va qanday yaratilganligi aniq ko‘rsatilishi; 5) so‘nggi foydalanuvchi korpus annotatsiyasi bexato emasligini, shunchaki foydali instrument ekanligidan xabardor bo‘lishi; 6) annotatsiya strukturasi umum tomonidan qabul qilingan me’yorlarga asoslanishi; 7) hech bir annotatsiya o‘zidan oldingi annotatsiyalarni standart deb qabul qilmasligi va standart me’yorlar tajriba davomida yuzaga chiqishi mumkin. Annotatsiyani formatlashning uch xil shakli mavjud[25]: 1) chiziqli (gorizontal ko‘rinishdagi) format– har bir so‘zdan keyin uning grammatik ma’lumoti lingvistik metama’lumot o‘rnida keltiriladi; 2) kalit so‘zlar bilan ifodalanuvchi chiziqli format – korpus ma’lumoti bilan tashqi ma’lumot bilan bog‘lanib, lingvistik annotatsiyani ifodalaydi; 3) vertikal format: har bir qatorda tokenning muayyan morfosintaktik yoki belgilar guruhi bo‘yicha jadvalda ma’lumotlar beriladi.

“*Grammatik sathda ontologik modellashtirishning Protégé texnologiyasi tatbiqi”* deb nomlangan bobning ikkinchi faslida lingvistik razmetkalashning grammatik asoslari ontologik modellashtirish usuli orqali tahlilga tortilgan. Protégé dasturi yordamida ma’lumotlar bazasi uchun freymvork va ontologiya yaratishning maxsus tahrirlovchi dasturi sifatida Protégé-freym va Protégé-OWL instrumentariylarining funksional imkoniyatlari o‘rganilgan. Ushbu dastur orqali o‘zbek tilining grammatik modellari va ularning toksonomik munosabatlari tadqiq etilgan. Mazkur texnologiya muayyan sohada bilimlar bazasini o‘zaro bog‘lash va ularning formal modellarini yaratishda qator afzalliklarga ega[26]. Mazkur dasturiy muhit yordamida tur va tasnif munosabatlarni iyerarxik jihatdan lingvistik modellar asosida semantik tahlil qilish mumkin. Ontologik modellashtirish usulida o‘zbek tilining morfologik va sintaktik kategoriyalar tezaurusi yaratilgan[27].

[23] Geofrey L. – matn annotatsiya bo‘yicha qoidalarni rivojlantirgan ilk korpus tilshunosi.
[24] www.lancaster.ac.uk/fss/course/ling/corpus/corpus2/2MAXIMS>HTM
[25] Захаров В.П., Азарова И.В. и др. Моделирование в корпусной лингвистике: Специализированные корпусы русского языка. – Санкт-Петербургского университета, 2019. – С. 49.
[26] https://ru.wikipedia.org/wiki/Prot%C3%A9g%C3%A9
[27] http://alphabet.kz:9191/





Bobning uchinchi fasli "*O'zbekcha matnlarni morfologik teglash va tahlil qilishda FST texnologiyasi tadqiqi"* deb nomlanadi. Ushbu faslda lingvistik annotatsiyalangan korpus uchun FST texnologiyasi yordamida o'zbek tilini mashina tiliga o'tkazish va uning modellarini universal tobelanish platformasining teglash tizimiga moslab, morfologik tahlil bosqichlari ishlab chiqilgan.

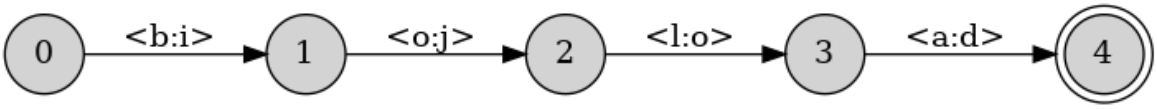
O'tgan asrning 90-yillarida morfologik tahlil qilish *generativ model*, *paradigmatik model*, *ikki tarkibli morfologik model*[28] kabi metodlarga asoslanilgan bo'lib, bunda ehtimollik darajasiga ko'ra lingvistik resursdan olingan natijalar statistik tahlil qilinadi. Avtomatik usulda morfologik tahlil qilish jarayoni quyidagi turlarga ko'ra tasniflanadi:

1. Mantiqiy ko'paytirish metodi orqali morfologik tahlil.
2. Lug'atsiz jadval orqali morfologik tahlil.
3. So'zlar asosiga asoslangan morfologik tahlil (ushbu tahlil turi Yevropa tillari uchun moslashtirilgan. Unda qo'shimcha yordamchi jadvallardan foydalaniladi).
4. So'zshakllarga (morfotaktik qoidalarga ko'ra) asoslangan morfologik tahlil.

Matnni normallashtirish jarayoni morfologik tahlil bilan bog'liq bo'lib, quyidagi bosqichlarga ega bo'ladi: tokenizatsiya (matndagi so'z shakllarini aniqlash)->lemmatizatsiya (so'zlarning lug'atdagi ifodasini aniqlash)->stemmizatsiya (derivativ so'zlarning o'zagini aniqlash).

[28] Cerstin M., Michael P. (eds.) Computational Morphology in the Framework of the SLIM Theory of Language / State of the Art in Computational Morphology. –Zurich, 2009. –P. 15.

FST texnologiyasi dastlab fin tili uchun Koskenniemi, ingliz tili morfologiyasi uchun Antvort tomonidan qo‘llangan[29]. Keyinchalik ikki bosqichli morfologik analizator yapon[30], koreys[31], turk[32], arab[33] va mongol[34] tillari doirasida foydalanilgan. Tadqiqot doirasida chekli morfologik avtomat uchun quyidagi fayllar yaratilgan: 1) “Qoidalar-Rules” – alifbo, fonologik qoidalar va fonetik hodisaga uchraydigan maxsus belgilar (bunda o‘zbek tili uchun kirill va lotin grafema asos qilib olindi; 2) Lug‘at–leksikon (bunda barcha so‘z turkumlarining o‘zak va sodda yasama shakli kiritilgan). FST texnologiyadan foydalanib, mavjud morfologik leksikon transyuterda, qoidalar ikki bosqichli jarayonda kompilyatsiya qilinadi. So‘ng mazkur leksikon va qoidalar birlashtiriladi. Buning uchun *compose* buyrug‘idan foydalaniladi. FST– chekli avtomat bo‘lib, ikkita *tasma* orqali harakatlanadigan model hisoblanadi. Masalan, *bola* so‘zi dasturiy ta’minotga kiritilganda, *ijod* so‘zi hosil bo‘lishi mumkin.



O‘zbek tili korpusining morfologik analizatorida lug‘at transyuteri va ikki bosqichli qoidalar transyutiri o‘zaro birlashtiriladi. Quyida FST yordamida o‘zbek tilining morfotaktik qoidalarini tahlil qilamiz. FST texnologiyasi deklarativ til bo‘lgani bois kompyuter modeliga o‘qitiladigan qoidalarni model holatiga keltiramiz.

LEXICON NumC
+SG: Poss1;
+PL:lar Poss2;

LEXICON NumC (consonant) – undosh harf bilan tugagan so‘zning son kategoriyasida kelishi. LEXICON NumC+SG=Poss1 bu birlik son kategoriyasining undosh bilan tugagan so‘zning 1-holatini ko‘rsatadi. LEXICON NumC+ *PL:lar* =Poss2 ushbu model undosh bilan tugagan so‘zga ko‘plik qo‘shimchasining 2-holatini bildiradi.

LEXICON NumV
+SG: Poss2 ;
+PL:lar Poss2;

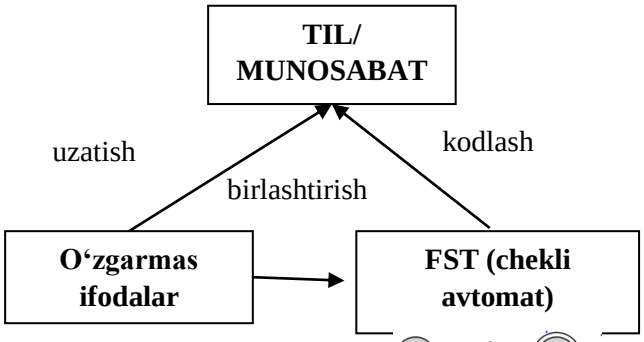
LEXICON NumV +SG= Poss2 unli bilan tugagan so‘zning birlik shaklini ifodalaydi. Demak, deklarativ tilda *LEXICON NumV* so‘zning modeli sifatida bundan keyin keladigan so‘zlarning pozitsiyalarini belgilashga xizmat qiladi.

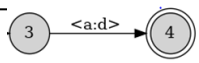
LEXICON Poss1

[29] Antworth E.L. PC-KIMMO: A Two-level Processor of Morphological Analysis, Summer Instıtute of Linguistics, Dallas, TX. 1990.
[30] Alam Y.S. Two-level Morphological Analysis of Japanese // Texas Linguistics Forum 22, 1983. –P. 229-252.
[31] Kim D. B., Lee S. J., Choi K.S., Kim G.C. A two-level morphological analysis of Korean // In Proceedings of the 15th conference on Computational linguistics - Volume 1 (COLING '94), 1994. –P. 535-539.
[32] Oflazer K. Two-level description of Turkish morphology // Literary and Linguistic Computing, Literary and Linguistic Computing Volume 9, Issue 2. 1994. –P. 137-148.
[33]Kenneth R. B. Arabic Finite State Morphological Analysis and Generation. // COLING-96. –Cophenagen, 1996. –P. 89-94.
[34] Jaimai P., Zundui T., Chagnaa A., Ock C.Y., PC-KIMMO-based Description of Mongolian Morphology // International Journal of Information Processing Systems Vol.1, No.1, 2005. –P. 41-48.

+PP1+PSG:m Case ;
+PP2+PSG:ng Case ;
+PP3+PSG:si Case ;
+PP1+PPL:miz Case ;
+PP2+PPL:ngiz Case ;
+PP3+PPL:i Case ;
0:0 Case ;

Yuqoridagi model so‘zga [LEXICON Poss1] qo‘shilishi mumkin bo‘lgan morfotaktik pozitsiyalari FST qoidalariga ko‘ra alohida kiritiladi. Har bir kiritiladigan model barcha so‘zlarning morfologik tahlili uchun yagona va universal bo‘lishi lozim. Morfologik teglangan har bir korpusdagi segment birlik so‘z va so‘z hosilasi (grammatik ko‘rsatkichi) bo‘yicha morfologik teglangan birliklar bilan til modeli bo‘yicha bazaga kiritiladi. Yuqoridagi model tahlil qilinsa, ixtiyoriy so‘zning grammatik ko‘rsatkichi o‘sha so‘zning nol ko‘rsatkichiga qarab: nominal turga kiruvchi so‘z birlik yoki ko‘plik shaklda keluvchi son kategoriyasi, so‘ng egalik qo‘shimchalar va undan keyin keluvchi kelishik kategoriyasiga ajraladi. Morfologik tegger va uning tahlilini amalga oshirish uchun nominal (otli) va verbal (fe’lli) guruhlarga ajratib olingan so‘zlar deklarativ til modeliga moslashtiriladi. FST texnologiyasining asosiy konsepti quyidagi uchlikka asoslangan[35]:



Korpusda qo‘llaniladigan 

si ham lotin va kirill grafikasida bo‘lgani bois ikki alifboni deklarativ tilda % belgi bilan birlashtiramiz. So‘ng orfografik qoidalarda uchraydigan har bir qoidaning ikki grafemaga moslashgan morfofonologik qoidasi yoziladi:

a. %##:0
%^Y:0 %^Y:y %^Y:й
%^G:g %^G:k %^G:q %^G:ғ %^G:г %G:к
%^K:k %^K:g %^K:г %^K:к
%^Q:q %^Q:g %^Q:ғ %Q:қ
%^A:a %^A:0 %^A:a
%^U:u %^U:0 %^U:y
%^I:i %^I:0 %^I:и

1.Istisnoli qoidalar to‘plami. Ushbu qoidalar to‘plamida fonetik o‘zgarishga uchraydigan barcha morfem qo‘shimchalar modeli yoziladi:

[35] Kenneth R.B., Lauri K. Finite State Morphology. –US: CSLI publication, 2003. – P.44.

"Y is dropped after consonant"
%^Y:0 <=> [:LatCons | :LatCons] [:‘]* [:%-]* _ ;
"Y remains after Latvowels"
%^Y:y <=> :LatVowel [:‘]* [:%-]* _ ;
"Y remains after Cyrvowels"
%^Y:й <=> :CyrVowel [:‘]* [:%-]* _ ;
"G remains g"
%^G:g <=> [:LatConsMinus |:LatVowel] _;
"G becomes k after k"
%^G:k <=> :k _;

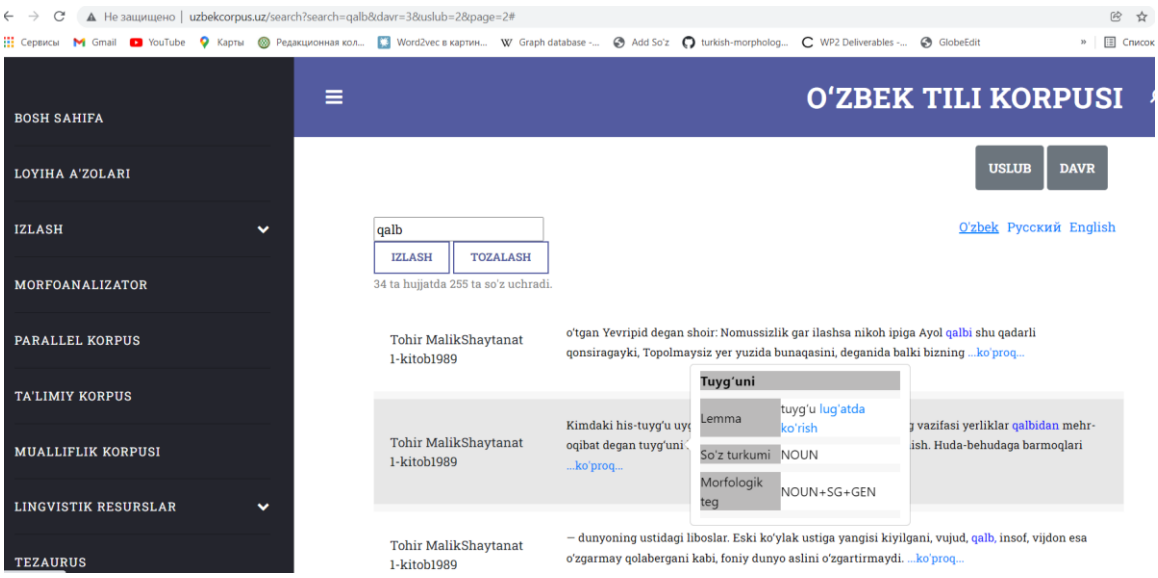
2. Keyingi bosqich ollomorflar uchun maxsus belgilar kiritib olish jarayoni.

Ollomorflarning FST uchun kiritiladigan har bir tegi (Multichar_Symbols) o‘zgaruvchi fonemalar va morfotaktik qoidalari uchun mashina tilida o‘qitib olinadi. Masalan, ^Y belgi –u va –yu yuklamasining o‘zgarish pozitsiyasini ko‘rsatadi.

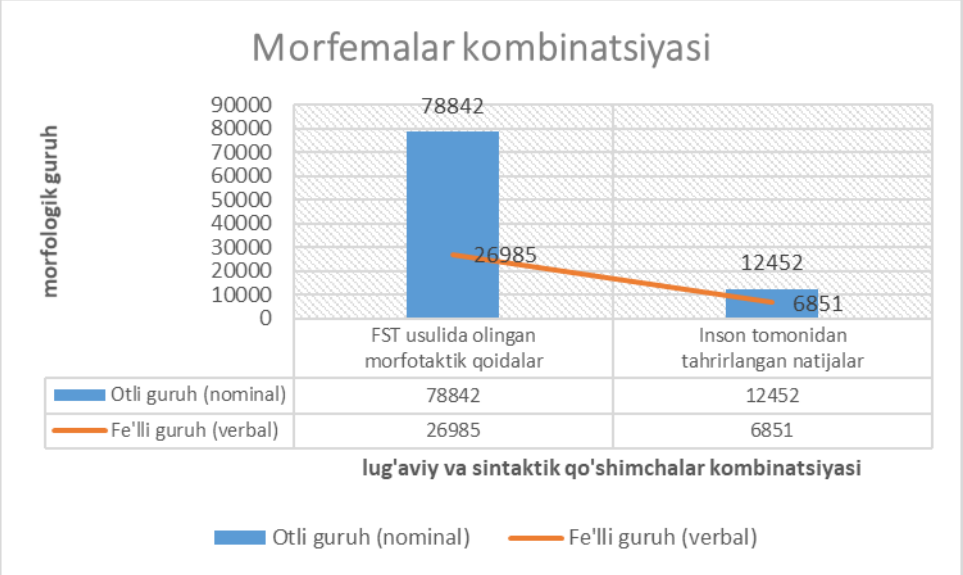
^Y
^K
^Q
^G
^A
^U
^I
^V
^W
^P
+VERB +CONV
+PAST +PRES +FUT
+PP1 +PP2 +PP3 +PSG +PPL
+P1 +P2 +P3 +SG +PL
+NOM +ACC +GEN +DAT +LOC +ABL
+PART
+NOUN
##

Natija namunasi:
shahar+NOUN+PL+PP3+PSG+ELIPT+GEN:shaharlarinikini
shahar+NOUN+PL+PP3+PSG+ACC:shaharlarining

Natijada o‘zbek tili elektron korpusning morfologik teglash qismi uchun FSTda erishilgan natijalardan foydalanildi:



Ishimizning morfologik bazasida atoqli otlar va iboralar ham alohida (*subdata*) baza sifatida kiritilgan. Bundan tashqari o‘zbek tilidagi ko‘makchi fe’lli so‘z qo‘shilmalari va ushbu fe’llarning har biriga qo‘shilish mumkin bo‘lgan orttirma nisbat qo‘shimchalarining morfotaktik belgilari ma’lumotlar bazasiga kiritildi. Shuningdek, mazkur korpus muvozanatlashgan korpus sanalgani bois matnlar kontenti sport, madaniyat, ma’rifat, din, ta’lim, fan, siyosat, qonun va adabiyot kabi turli janrdagi matnlarning umumiy soni teng miqdorda taqsimlandi.



Korpus menejerida matnda uchragan so‘zning asosi va so‘z turkumi ma’lumotlar bazasida mavjud bo‘lmasa, quyidagi algoritm orqali masalani yechish usulidan foydalanildi. Berilgan *w*-so‘z, $w = l \oplus \sum_{i=1}^{n} m_i$, *l*-lemma va $m_i \in \{m_1, \ldots, m_n\}$ (o‘ngdan chapga morfemalarni ajratish orqali aniqlangan) morfemalar ketma-ketligidan iborat bo‘lsa, u holda *l* lemma *V* – verbal guruh yoki *N*- nominal guruhga tegishliligini aniqlovchi *p* argumentning qiymati quyidagicha aniqlanadi:

$$p = \begin{cases} 0, agar\ n \geq 2\ va\ m_1 \in M_v, \\ 1, agar\ n \geq 2\ va\ m_1 \in M_n. \end{cases}$$

Bunda M_v-verbal so‘zlarga qo‘shiluvchi morfemalar to‘plami, M_n-nominal so‘zlarga qo‘shiluvchi morfemalar to‘plami. Agar *p*=0 bo‘lsa, u holda *l* lemma *V* guruhga, aks holda *N* guruhga tegishli bo‘ladi, ya’ni *l* haqida qaror qabul qilish unga qo‘shilgan morfemalar sonining kamida ikkita bo‘lishi va qo‘shilgan birinchi morfemaning qaysi guruhga tegishliligiga bog‘liq bo‘ladi.

“*O‘zbekcha matnlarni sintaktik teglash va tahlil qilishda UdPipe texnologiyasi*” deb nomlangan to‘rtinchi faslda o‘zbek tili korpusining kompyuter modeli sifatida sintaktik annotatsiyalash va teglash jarayoni tahlil qilindi. Ma’lumki, dunyo bo‘yicha sintaktik annotatsiyalangan korpuslarning dastlabki namunalardan biri sifatida ingliz tilining Lankaster-Leeds Treebankini ko‘rsatish mumkin. Tadqiqotimizda sintaktik teglash va tahlil qilish bo‘yicha chex tilining Praga sintaktik annotatsiyalangan korpusi (PDT), arab tilining arab tili sintaktik annotatsiyalangan korpusi (PADT), slavyan tilining slavyan (SDT), shved tilining Talbanken, nemis tilining Negra / TIGER hamda turk tilining Metu Sabanci kabi korpuslari orqali Sabine Buchholz[36] tomonidan olib borilgan ishlariga e’tibor qaratildi.

Sintaktik annotatsiyalash lingvistik korpusning eng muhim tahlil qilish bosqichlardan biri sanaladi. Sintaktik tahlil NLP va til texnologiyasining qator yo‘nalishlarida olib boriladigan muayyan darajadagi masalalarni yechishda muhim ahamiyat kasb etadi. Ayniqsa, matn tahlilida lisoniy modellar va gap bo‘laklarining sintaktik munosabatlarini kompyuter yordamida tahlil qilish anchagina murakkab masaladir. Sintaktik tahlil (parsing) matndagi sintaktik munosabatlar va strukturalarni avtomatik tarzda sintaktik strukturalarga ajratadi. Parsing kompyuter lingvistikasi va NLPning deyarli barcha yo‘nalishlari (mashina tarjimasi, savol-javob, axborot qidiruv tizimi, sentiment tahlil)da foydalaniladi. Mazkur sohada parsing uchun sintaktik teglashning turli formatlari va usullari ishlab chiqilgan. Ularga *tobelik grammatikasi, bevosita tarkibiy qismli grammatika, aralashgan nazariya (sintaktik guruh nazariyasi*), *erkin kontekstli grammatika* kabilar kiradi.

Universal tobelik nazariyasi www.universaldependency.org platformasida o‘rin topgan. 2018-yil hisobiga ko‘ra 132 ta sintaktik tahlil qiluvchi korpusning (treebank) 74 tasi http:// universaldependencies.org/ platformasidan o‘rin olgan. Ularning ba’zilari *tobelik nazariyasi asosida, ayrimlari tarkib toptiruvchilar nazariyasiga* asoslangan[37].

Ishimizda UdPipe texnologiyasi yordamida matnni sintaktik jihatdan teglash jarayoni amalga oshirilgan. Ushbu jarayon CoNLL-U formatda o‘z ifodasini topgan. Gaplarning sintaktik jihatdan tavsifini berishda bir nechta tarkibli so‘zlar (ibora, qo‘shma so‘z) alohida baza sifatida olingan. Odatda gaplarni segmenter yordamida segment birliklarga ajratishda gaplar orasidagi qator yoki satr boshiga

[36] Sabine B., Erwin M. CoNLL-X shared task on Multilingual Dependency Parsing https://www.aclweb.org/ anthology/ W06-2920.pdf. – P. 152.

[37] Przepiórkowski A., Patejuk A. From Lexical Functional Grammar to enhanced Universal Dependencies.// The UD-LFG treebank of Polish Lang Resources & Evaluation https://doi.org/10.1007/s10579-018-9433.

qaraladi. Bu borada Milan Straka gaplarni sintaktik jihatdan tahlil qilishning muayyan model orqali amalga oshirishning har bir algoritmik tavsifiga alohida to‘xtalib o‘tadi[38]. Unga ko‘ra UdPipe tayyor model sifatida *gaplarni segmentlash, tokenizatsiyalash, so‘z turkumlarini teglash, lemmatizatsiyalash va tobelanishga asoslangan sintaktik tahlil*ni kompleks ravishda amalga oshiradi.

Tadqiqotimizda morfologik jihatdan tahlil qilingan FST yordamida aniqlangan annotatsiyalash tizimidan foydalanilgan. Matnlarni sintaktik jihatdan teglashda har bir so‘zning uch belgisi *UPOS, XPOS, FEATS* alohida CoNLL-U format orqali tahlil qilindi. Morfologik teglash (Part-of-speech tagging) matnga kiritilgan har bir so‘zning so‘z turkumi bo‘yicha markerini belgilash jarayonidir. Teglash algoritmiga kiritiladigan ma’lumot – teglar majmui va so‘zlarning ketma-ketligidan iborat bo‘ladi. Bu yerda:

- **ID**: so‘z indeksi, har bir qator biror jumla boshlanganda yangi qatordan yoziladi.
- **FORM**: so‘z shakli yoki punktuatsion belgilar.
- **LEMMA**: lemma yoki so‘zning o‘zagi.
- **UPOS**: universal so‘z turkumi tegi (POS tag)
- **XPOS**: tilning maxsus so‘z turkumi tegi; agar bo‘lmasa, tag osti belgisi _ qo‘yiladi.
- **FEATS**: agar mavjud bo‘lsa, morfologik kategoriyalar tartibi
- **HEAD**: so‘zning ID raqami (so‘zning gapda nechinchi tartibda turgan indeksi) yoki nol (0) tartibi.
- **DEPREL**: Bosh so‘zga nisbatan tobelik munosabati. Sintaktik annotatsiyalangan korpuslarda ushbu tobelik munosabatlari to‘plamlari o‘rin olgan.
- **DEPS**: kengaytirilgan tobelik grafigi (ixtiyoriy), agar yozilmasa _ qo‘yiladi.
- **MISC**: boshqa annotatsiya kodi.

Ushbu UdPipe metodi fin va turk tili uchun qo‘llangan bois aynan agglyutinativlik tabiatiga xos tillar uchun mos tahlil usuli hisoblanadi. Tadqiqotimizda shu tartibda Nabi Jaloliddinning “Umar Hayyom” romanidan olingan ikki ming matn segment birlik sifatida CONLL-U formatda avtomatik usulda sintaktik tahlil qilingan modellar o‘rganildi. Ishimizda tanlangan formatning o‘ziga xosligi shundaki, CONLL-U orqali ko‘p tilli tizimlar uchun mos keladi hamda so‘zlarni o‘zaro solishtirish imkoniyati mavjud.

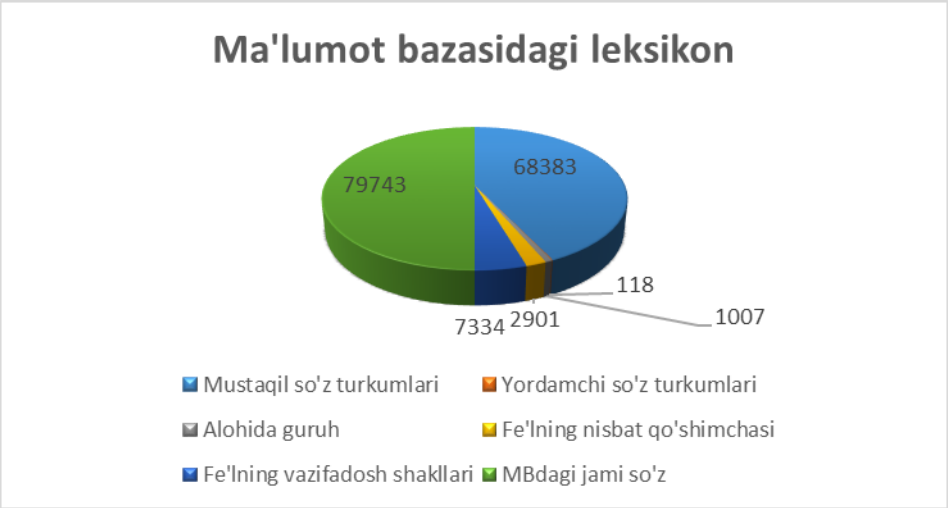
Teglashning eng oxirgi bosqichlari *UPOS, XPOS, FEATS* sifatida tahlil qilinadi.

[38]Straka M., Strakova J. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe https://ufal.mff.cuni.cz/~straka/papers/2017-conll_udpipe.pdf

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS		MISC
	Uchrashuvda	Uchrashuv	NOUN	OT	Case=Loc|Number=Singular|Number[psor]=Singular	5		_		_
	Innovatsiya	innovatsiya	NOUN	OT	Case=Loc|Number=Singular|Number[psor]=Singular	6		_		_
	Sohasida	soha	NOUN	OT	Case=Loc|Number=Singular|Number[psor]=Singular	6		_		_
	Xalqaro	xalqaro	ADJ	SF	Degree=Pos	4		_		_
	Hamkorlik	hamkorlik	NOUN	OT	Case=Nom|Number=Singular|Number[psor]=Singular	3		_		_
	,	,	PUNCT	TB	_	_	_	_		_
	Xalqaro	xalqaro	ADJ	SF	Degree=Pos	4		_		_
	Loyihalarni	loyiha	NOUN	OT	Case=Loc|Number=Singular|Number[psor]=Singular	3		_		_
	Ishlab	ishla	VERB	F	Verb=V|ADV	3		_		_
	Chiqish	chiq	VERB	F	Verb=V|INF	3		_		_
	Va	va	CONJ	B	_	_	_	_		_
	Gumanitar	gumanitar	ADJ	SF	Degree=Pos	3		_		_
	Hamkorlik	hamkorlik	NOUN	OT	Case=Nom|Number=Singular|Number[psor]=Singular	3		_		_
	Masalalari	masala	NOUN	OT	Case=Nom|Number=Singular|Number[psor]=Plural	2		_		_
	Muhokama	muhokama	NOUN	OT	Case=Loc|Number=Singular|Number[psor]=Singular	1		_		_
	Qilindi	qil	VERB	F	Number=Plur|Person=3|Tense=PresTense=Past|Voice=Pass	0		_		_
	.	.	PUNCT	TB	_	_	_	_		_

"O*'zbek tili korpusi menejerining lingvistik va dasturiy ta'minoti*" deb nomlangan dissertatsiyaning to'rtinchi bobning birinchi fasli "*Korpus loyihasini yaratishning konseptologik asoslari*" deb nomlanadi. Korpus menejeri (korpus brauzeri yoki korpus qidiruv tizimi) korpusda ma'lumotni samarali qidirish imkonini beruvchi tizim hisoblanadi. Korpus menejeri murakkab model bo'lib, tilning turli nutqiy shakllarini va so'zlarning birikuvchanlik xossalarini kontekst tarkibida statistik ko'rsatishga mo'ljallangan. Ushbu tizim foydalanuvchiga yaxlit gap strukturasini yoki muayyan vaziyatga xos atributlari lemma, teg bo'yicha yoki so'zlarning konkordansi va chastotasi kabi ma'lumotni berishga xizmat qiladi. Serverning korpus qidiruv tizimi korpus menejeri bo'lsa, foydalanuvchi uchun mo'ljallangan qismi korpus interfeysi deb ataladi. Korpus menejerlari turlicha bo'lishi mumkin. Masalan, Britaniya milliy korpusi webga asoslangan interfeys, Manatee (orqa font qismi) va Bonito (web interfeys) birgalikda bepul ochiq

resursli korpus boshqaruv tizimiga asoslangan Nosketch Engine, lingvistlar uchun dasturiy ta’minot paketi WordSmith Toolsni misol sifatida keltirish mumkin.

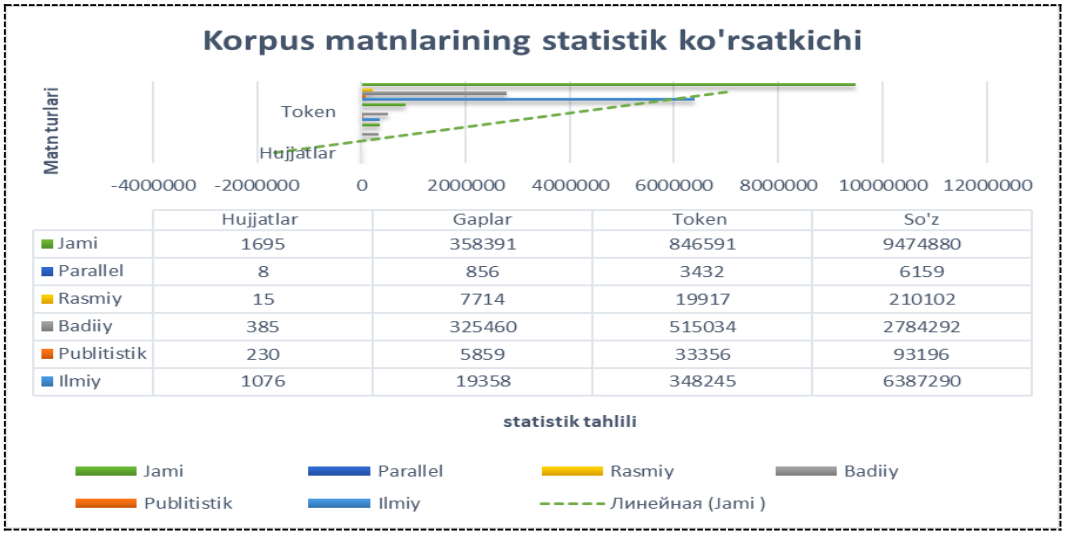


Bobning “*Matnlar reprezentativligi masalasi”* deb nomlangan ikkinchi faslida korpusning reprezentativligi matnlarning miqdor jihatdan yetarli ekanligi va janr jihatdan xilma-xilligi bilan belgilanishi ko‘rsatib o‘tilgan. V.P.Zaxarov va S.Y.Bogdanovaning fikricha[39], korpusning janriy-mavzuviy tuzilishi korpus matni uchun qanday manbalar tanlashiga e’tibor qaratish lozim. Unga ko‘ra, gazetalardagi kichik reklama matni alohida matn sifatida qaraladimi yoki ularni bir matnga birlashtirish to‘g‘rimi? Gazeta maqolasi matn sanaladimi yoki gazetaning bitta sonini yaxlit matn sifatida olish zarurmi? Har bir she’r bitta matnmi yoki she’riy to‘plamni umumiy holatda kiritish kerakmi? Bir-biriga javob tarzida yozilgan, mohiyatan bir mavzu muhokamasiga bag‘ishlangan nashr etilgan maktublar bitta matnmi yoki alohida korpus birligi sifatida olinadimi? Korpus tuzuvchi muallif ushbu savollarga korpusning turi va keyinchalik bajaradigan vazifasidan kelib chiqib javob berishi lozim. Milliy korpus yoki maxsus korpus ekanligiga qarab korpus birligi belgilanadi. Olimlar korpusni loyihalashtirish jarayonining muhim jihati sifatida xronologiyani ham nazarda tutadilar. Masalan, tilning zamonaviy korpusi deganda tilning qaysi leksik qatlami tushuniladi? Ushbu holatda turli janrlarda korpusning xronologik chegarasi turlicha bo‘lishi tabiiy. Korpusdan keng omma foydalanishi hamda xilma-xil vazifalarni bajarishga mo‘ljallanishi mumkin.

“O*‘zbek tili korpusining lingvistik va dasturiy ta’minoti*” deb nomlangan uchinchi faslida matn korpusining lingvistik ma’lumotlarini boshqarish tizimini amaliy jihatdan yaratish va uning auditoriya talabiga ko‘ra shakllantirish masalasi yoritilgan. Korpus menejeri statistik ma’lumot olish va foydalanuvchiga qulay tarzda natijalar bilan ta’minlashga mo‘ljallangan maxsus qidiruv tizimi hisoblanib, ko‘pincha bunday tizimlar tayyor yechimlarga asoslangan bo‘ladi. Bu qidiruv so‘rovlarini (ma’lumotlar namunalarini) bajarish tezligi, tizimning

[39] Захаров В.П., Азарова И.В. и др. Моделирование в корпусной лингвистике: Специализированные корпусы русского языка. – Издателство Санкт-Петербургского университета, 2019. – С. 16.

moslashuvchanligi, kengaytirilishi va kengayishi bilan bog‘liq muammolarni yechishga xizmat qiladi.



	Hujjatlar	Gaplar	Token	So'z
Jami	1695	358391	846591	9474880
Parallel	8	856	3432	6159
Rasmiy	15	7714	19917	210102
Badiiy	385	325460	515034	2784292
Publitistik	230	5859	33356	93196
Ilmiy	1076	19358	348245	6387290

O‘zbek tilining agglyutinativligi va qo‘shimchalar ketma-ketligi bo‘yicha grammatik shakllarning hosil qilishi so‘zlarning morfotaktik modelini yaratish zaruriyatini ko‘rsatdi. Shu bois korpus menejerida matndagi so‘zlarning asosini aniqlashda (stemmizator orqali) o‘zbek tilining so‘z turkumlari bazasi uchun 90 ming leksik birlikning so‘z turkumlari ichki guruhlar bo‘yicha tasniflanib, fe’lning vazifadosh (sifatdosh, harakat nomi va ravishdosh) va nisbat (orttirma, majhul, birgalik) qo‘shimchalarini qabul qiluvchi fe’l guruhlari to‘plami korpusning lingvistik ma’lumotlar bazasiga kiritildi. Ushbu guruhlar umumiy morfotaktik modellarni yaratishda asos vazifasini bajaradi hamda so‘zlarni qidirishda ushbu to‘plamlardan qidirilayotgan so‘zning lingvistik ma’lumotlari olinadi.

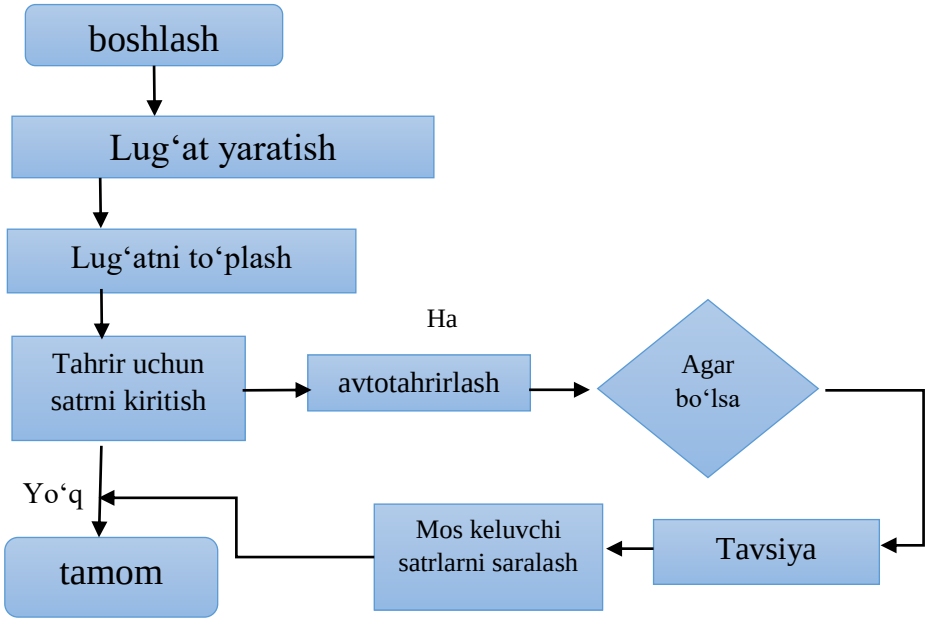
Korpus tarkibida parallel korpuslar uchun quyidagi ma’lumotlar bazasi shakllantirildi:

- ikki tilli tarjima lug‘ati: o‘zbekcha-inglizcha, inglizcha-o‘zbekcha, ruscha-o‘zbekcha, o‘zbekcha-ruscha;
- leksik ko‘p ma’noli so‘zlar bazasi;
- sintaktik jihatdan tasniflangan modellar tizimi;
- WordFast texnologiyasi yordamida parallel matnlar tarjima xotirasi;
- tarjima muqobillarining konkordanslar bazasi;
- terminologik birikmalar, frazeologizmlar va idiomalarning tarjimada berilishi.

O‘zbek tilining elektron korpusining ichki korpusi parallel korpus ham bo‘lib, ushbu korpusga o‘zbekcha-inglizcha va inglizcha-o‘zbekcha parallel matnlarning tarjimon xotirada segmentlangan birliklar bo‘yicha topish imkoniyati mavjud. Matn kontentiga ko‘ra ilmiy va rasmiy uslubdagi yozma matnlar bazadan o‘rin oldi.

“*Korpus qidiruv tizimida matn tahrirlovchi dasturi interfeysi*” deb nomlangan to‘rtinchi faslda so‘zni qidirish jarayonida foydalanuvchilar tomonidan yo‘l qo‘yilgan imlo xatolarni bartaraf etishning algoritmik yechimi ochib berilgan.

Tahrirlovchi dastur algoritmining blok-sxemasini quyidagicha ifodalash mumkin[40]:



Matnlarni tahrirlovchi dastur turk, tatar, qozoq, qirg‘iz tillari uchun yaratilgan. O‘zbek tili uchun tahrirlovchi dasturning lingvistik ta’minotini yaratish bo‘yicha ham bir nechta tadqiqotlar olib borilmoqda. Tahrirlovchi dasturning ta’minotini yaratish bo‘yicha quyidagi texnologiyalar mavjud:

1. Djaro Vinkler algoritmi orqali ehtimollik nazariyaga asoslangan texnologiya;
2. Sof lingvistik bazaga asoslangan texnologiya;
3. FST (finite state trancducer) algoritmiga asoslangan texnologiya.
4. Levenshteynning oraliq masofa nazariyasi

Tadqiqotimizda so‘zlarning xatolarini aniqlashda *Djaro Vinkler algoritmi*ni qo‘llashning afzalliklarini isbotlashga harakat qildik. Ushbu usul ikki o‘xshash satrni o‘zaro solishtirish uchun qo‘llaniladi, mazkur algoritmda qo‘shish, tushirib qoldirish va o‘rniga qo‘yish amallari bajariladi.

Berilgan so‘rovga ko‘ra natija topilmasa, korpusda qidirilayotgan so‘zning imlosi qayta tekshiriladi. Shunda noma’lum satrni (so‘zni) boshqa bir satr bilan solishtirganda satr uzunligi (elementlar soni)ni quyidagicha tekshiramiz:

$$d_j = \begin{cases} 0 & m\ bo'lganda = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) & qolgan\ holatlarda \end{cases}$$

qayerda:

$|s_i|$ - s_i *satr uzunligi*;

m – o‘xshash simvollar miqdori;

t – transpozitsiyalar miqdorining yarmi.

[40] Vibhakti V. Bhaire, Ashiki A. Jadhav, Pradnya A. Pashte, Mr. Magdum P.G SPELL CHECKER // International Journal of Scientific and Research Publications, Volume 5, Issue 4, April, 2015. 1 ISSN 2250-3153.

Agar o‘xshash bo‘lsa, va $\left[\frac{\max(|s_1|,|s_1|)}{2}\right] - 1$ dan oshib ketmasa, s_1 va s_2 satrlardan ikkita element olinadi. Satrdagi har bir c_1 dagi simvol c_2 dagi simvollar bilan solishtiriladi. Turli holatdagi o‘zaro o‘xshash raqamlar 2 ga bo‘linadi, bu transpozitsiya raqami hisoblanadi. Masalan, **ODAMLAR va OMADLAR** so‘zini solishtiradigan bo‘lsak, har ikki satrning uzunligi m=7 ga teng. Biroq undagi M va D simvollarning o‘rni turlicha holatda kelgan. Bundan O va A bir xil holatda joylashgan. Chunki D va M ning holati 1 dan kattaroq. Shuning uchun bularni o‘xshash simvollar, deb ayta olmaymiz.

Djaro Vinkler masofasida *p* o‘lchov koeffitsienti bo‘lib, u satr boshidan prefiks deb nomlangan *l* uzunligacha bo‘lgan bir-biriga o‘xshash satrlar uchun aniqlangan o‘lchov birligi hisoblanadi. Djaro Vinklerda kamida ikkita satr solishtiriladi:

$d_w = d_j + \left(lp(1 - d_j)\right)$

$d_j - s_1$ va s_2 satrlar uchun Djaro masofasi

l – satr boshidan 4-simvolgacha bo‘lgan *umumiy prefiks*

p – o‘lchov koeffitsienti. U 0,25 dan katta bo‘lmasligi kerak. Unga teskari holatlarda orasidagi masofa 1 dan katta bo‘lishi mumkin. Lekin u 0 va 1 orasida bo‘lishi kerak. Agar masofasi 0 ga teng kelsa, ikki satr o‘rtasida umuman o‘xshashlik yo‘qligini, agar 1 ga teng kelsa, ikki satr bir xil ekanligini bildiradi. Vinklerning ishlarida 0 va 1 koeffitsient birligi hisoblanadi. **TAROQ** va **TAYOQ** so‘zlarni solishtirilsa, quyidagi jadval hosil bo‘ladi.

z	T	A	R	O	Q
T	1	0	0	0	0
A	0	1	0	0	0
Y	0	0	0	0	0
O	0	0	0	1	0
Q	0	0	0	0	1

Yuqorida aks etgan TARAMOQ va QARAMOQ so‘zlari solishtiriladi: m=6, t=0, $|s_1|=|s_2|=7$:

$d_j = \frac{1}{3}\left(\frac{6}{7} + \frac{6}{7} + \frac{6}{6}\right) = 0.9047$

l=0, p=0.1:

$d_w = 0.9047 + \left(2 \cdot 0.1 \cdot (1 - 0.9047)\right) = 0.92376$

Piton dasturlash tilida ushbu natija sinab ko‘rildi, undagi natijasi quyidagicha bo‘ladi:

BAHOR	0.8666666666666667	0.9066666666666667
NAHOR	0.7333333333333334	0.7333333333333334
HABAR	0.7	0.7
KITOB	0.8666666666666667	0.9066666666666667
KILOB	0.7333333333333334	0.7866666666666667

Ushbu algoritm orqali muayyan so‘zdagi harflar tushirilib qoldirilishi, o‘rniga qo‘yishi, almashtirilishi mumkin. Shunga ko‘ra u so‘zlardagi harflar lingvistik bilimlar bazasi orqali tekshiriladi.

XULOSA

O‘zbek tili elektron korpusining kompyuter modellari bo‘yicha olib borilgan ilmiy tadqiqotlar asosida quyidagi xulosalarga kelindi:

1. Korpus lingvistikasida nutqning barcha ifodalarini kuzatish, tahlil qilish, o‘rganishning imkoniyati mavjud. Korpus matnlar majmui sifatida o‘rganilayotgan obyekt va predmetning tizimli majmuasidir. Korpus NLP va kompyuter lingvistikasining asosiy raqamli resursi sifatida til texnologiyalariga doir masalalarni echishga asosiy manba bo‘lib xizmat qiladi.

2. Modellashtirishning informatsion, kompyuter, matematik, biologik, raqamli, mantiqiy, statistik, struktur, grafik kabi qator turlari mavjud. O‘zbek tili elektron korpusining kompyuter modellari ikki xil yondashuvga asoslanadi: 1) korpus uchun yaratilgan tayyor lingvistik instrumentariylar va platformalar; 2) korpus taksonomiyasidan kelib chiqib, turli maqsadlarga mo‘ljallangan kompyuter lingvistikasi metodlari orqali yaratilgan formal-funksional modellar.

3. Elektron korpus tilning davriy tadriji mahsuli sifatida jamiyat laboratoriyasining o‘rganish obyekti vazifasini bajaradi. Shu bois korpusni uzluksiz ravishda boyitib, yangilab borishda korpus konseptologiyasini modellashtirishning universallik, standart, reprezentativlik kabi mezonlariga asoslaniladi.

4. Elektron korpuslar lingvistik ma’lumotlarni tahlil qilishda zamonaviy kompyuter metodlari va lingvistik tadqiqotlarning umumiy natijasidir. Ayniqsa, o‘zbek tilining maxsus (parallel, ta’limiy, mualliflik) korpuslari tilni turli aspektda o‘rganish va statistik jihatdan tadqiqot olib borish imkonini beradi.

5. Korpus texnologiyasi kompyuter lingvistikasining asosiy obyekti va predmeti sifatida matnlarni qayta ishlashda tabiiy tilning muayyan vaziyatdagi modeli vazifasini bajaradi. Tizimlashtirilgan va ma’lum me’yorlarga asoslangan tasnifiy matnlardan iborat elektron korpus lisoniy qoliplar, grammatik qoidalar, lug‘atlarni yanada takomillashtirishda muhim ahamiyat kasb etadi. Tilning pragmatik va kognitiv tasviri aks etgan og‘zaki va yozma matnlar majmuasi (korpus) ontologik bilimlar bazasi, semantik va neyro to‘rlar, sun’iy intellekt texnologiyasi hamda lingvoprotsessor uchun tilning lisoniy modellari yaratishda hamda nutqiy aktlarni o‘rganishda katta rol o‘ynaydi.

6. Filologiya sohasida ilmiy tadqiqotlar olib borishda lingvistik instrumentlaridan foydalanib maxsus korpuslarni yaratish statistik natijalarga erishishning kvantativ usuli bo‘lib xizmat qiladi.

7. Matnlarni tahlil qilishda muayyan turdagi instrumentariylar korpus foydalanuvchilarning maqsad va vazifalaridan kelib chiqib individual foydalanuvchilar interfeysi va korpus menejeriga ega bo‘ladi.

8. Parallel matnlar uchun segmentlash va lingvistik tahlil jarayonida Wordfast kabi instrumentlar samarali texnologik vosita sanaladi. Parallel matnlardagi konkordanslarni aniqlashda so‘z, so‘z birikmasi yoki barqaror birikmalarning u yoki bu tildagi muqobil ekvavalentligini aniqlashda tayyor lingvistik instrumentariylarga tayanish muhim. Parallel matnlardagi kalit so‘zlar uchun kontekstda tez-tez qo‘llanadigan birliklarni tarjimon xotiraga yuklash orqali parallel matnlarning qidiruv tizimini yaratishda foydalaniladi.

9. Korpusni lingvistik jihatdan annotatsiyalash (razmetka) undan qay tartibda foydalanishning bosh tamoyili hisoblanadi. Matnlarning razmetkasi uning metama'lumotida (yozma matnlar uchun uning struktur birliklari, sarlavhasi, muallifi, janri, nashri, yili; og‘zaki matnlarning obyekti sifatida olingan so‘zlovchilarning yoshi, jinsi, kasbi, millati) aks etadi. Kiritilgan ma'lumotlarni mashina o‘qiy oladigan formatda matnni kodlash yo‘riqnomasi (Text Encoding Initiative (T.E.I.)) orqali standart tilda beriladi.

10. Lingvistik annotatsiyalash leksik, morfologik, sintaktik, semantik, prosodik (diskurs), anaforik, temporal turlarga tegishli bo‘ladi. Annotatsiya jarayonida muayyan xulosani chiqarish uchun annotatsiya modellaridan (annotation scheme) foydalaniladi. Ular annotatsiya qo‘llanmalarida qayd etiladi. Matnni annotatsiyalash qo‘l mehnati yoxud avtomat va yarim avtomat usulda amalga oshiriladi.

11. Annotatsiyaning mazmuniy jihatlari yuqorida qayd etilgan lingvistik annotatsiyalash turlaridan tashqari annotatsiya formatini ham belgilash muhim mezonlardan biridir. O‘zbek tilining sintaktik tahlili uchun CONLL-U format asos qilib olindi.

12. O‘zbek tilining morfologik va sintaktik teglash tizimi Protégé texnologiyasi orqali ontologik modellashtirishda iyerarxik munosabatga asoslanildi. Grammatik teglash tizimi UniTurk doirasida turk, tatar, qozoq va qirg‘iz tillari uchun ham assotsiativ shaklda munosabatlar tizimiga bog‘langan, natijada korpus tahlili uchun turkiy tillarning teglash tizimi yaratilgan.

13. Chekli avtomat metodi (FST) yordamida o‘zbek tilining morfologik tahlil qiluvchi dasturning mashina fondi yaratildi: 1) “Qoidalar-Rules” bunda alifbo, fonologik qoidalar va fonetik hodisaga uchraydigan maxsus fonemalar (bunda o‘zbek tili uchun har ikki grafema asos qilib olindi: kirill va lotin; 2) Lug‘at-leksikon (bunda barcha so‘z turkumlarining o‘zak va sodda yasama shakli kiritiladi).

14. Korpusning morfologik tahlili va teglash jarayonida FST texnologiyasiga asoslanilishi uning ehtimollik variatsiyalarini natija sifatida olishda yordam beradi. FST orqali ulkan hajmdagi matn korpusi va u yerdan olingan holatlar tizimidan foydalanib, ehtimollik modellar yaratiladi, so‘zning imlosini nafaqat tekshirishda, balki uni to‘g‘irlashda (foydalanuvchiga bir nechta variant taklif qilgan holda) ko‘maklashadi.

15. Matn etarli miqdorda hajmga ega bo‘lishi, turli tipdagi ma'lumotlarni o‘zining tabiiy kontekstdagi shaklida korpusda ifodalanishi, tayyorlangan va yaratilgan ma'lumotlardan ko‘p marotaba foydalanish imkoniyatining mavjudligi, korpus menejerining foydalanuvchilar uchun olingan natijalarni mos shaklda taqdim qilish, talabdan kelib chiqqan holda korpus funksiyalari imkoniyatalarini yaratish hamda beriladigan so‘rovlar natijasida erishiladigan statistik ma'lumotlarni aks ettirish, saralangan va jamlangan matnlarni boshqarish va lingvistik ma'lumotlarni muayyan maqsadlar uchun foydalanishga yo‘naltirish korpus yaxlitligini ta'minlaydi.

16. Lingvistik annotatsiyalangan korpuslar kompyuter lingstikasi yo‘nalishlari uchun (mashina tarjimasi, nutq sintezatori, semtiment tahlil, spell-cheker kabilar) obyekt vazifasini bajaradi. Parallel korpuslar yordamida shakllantirilgan tarjima xotirasi matnlarni avtomatik tarjima qilishda lingvistik resurs vazifasini bajaradi.

17. Korpusning qidiruv tizimi (menejeri) foydalanuvchilar hamda kompyuter lingvistikasi mutaxassislari uchun amalda qo‘llash imkoniyatlari bir-biridan farqlanadi. Har ikki subyekt uchun matn xatolarini bartaraf etishda Jaro Vinkler algoritmi va FST texnologiyasidan foydalanish inson resursi sarflaydigan vaqt va mehnatni tejash hamda o‘zbek tilining grammatik fondini yaratishda vosita bo‘lib xizmat qiladi.

ONE-TIME SCIENTIFIC COUNCIL BASED ON SCIENTIFIC COUNCIL WITH THE NUMBER DSC.02 / 30.12.2019.FIL.46.03 AWARDING SCIENTIFIC DEGREES UNDER THE INSTITUTE OF UZBEK LANGUAGE, LITERATURE AND FOLKLORE

TASHKENT STATE UNIVERSITY OF UZBEK LANGUAGE AND LITERATURE NAMED AFTER ALISHER NAVOI

ABDURAKHMONOVA NILUFAR ZAYNOBIDDIN QIZI

COMPUTATIONAL MODELS OF ELECTRONIC UZBEK CORPUS

10.00.11 – Language theory. Applied and computational linguistics

ABSTRACT OF DOCTOR OF SCIENCE (DSC) ON PHILOLOGICAL SCIENCES

Tashkent-2021

The theme of doctor science (DSc) thesis was registered by the Supreme Attestation Commission at the Cabinet of Ministries of the Republic of Uzbekistan under B2019.2. DSc/Fil1174.

The doctor of science thesis was implemented at Tashkent State university of Uzbek language and literature named after Alisher Navoi.

The abstract of the thesis in three languages (Uzbek, English, Russian (summary)) is logged on the web-sites of the Scientific Council (www.uztafi@academy.uz) and the information-educational portal "ZiyoNet" (www.ziyonet.uz).

Scientific advisor:	**Khamudilla Dadaboev Aripovich** Doctor of philological sciences, professor
Rasmiy opponentlar:	**Muhamedova Saodat Xudayberdiyevna** Doctor of philological sciences, professor
	Urinbayeva Dilbar Bazarovna Doctor of philological sciences, associate professor
	Nazirova Elmira Shodmanovna Doctor of technical sciences, associate professor
Leading organization:	**Fergana State university**

Defense of the Dissertation will take place on «24» 12 2021, at «11» a.m. at a meeting Scientific Council DSc.02/30.12.2019. Fil.46.03 under Scientific Degree at Institute of Uzbek language, literature and folklore of the Uzbekistan Academy Sciences (address: 100060, Tashkent, str. Mirobod district, Shakhrisabz street, 5-house. Tel.: (99871) 233-36-50; fax: (99871) 233-71-44; e-mail: uztafi@academy.uz).

Dissertation could be introduced at the Main Library of the Uzbekistan Academy Sciences (registered under No. 54). (address: 100100, Tashkent, str. Ziyolilar, 13. Tel: (99871) 262-74-58).

The abstract of the dissertation was distributed on «13» 12 2021.
(Registry report number 5 on «13» 12 2021)

N.M.Mahmudov
Chairman of the Scientific Council awarding scientific degrees, Doctor of Philological sciences, Professor

G.M.Ismailov
Secretary of Scientific Council awarding scientific degrees, Candidate of Philological sciences, Senior Scientific Researcher

D. S. Xudoyberganova
Chairman of Scientific Seminar at the Scientific Council awarding scientific degrees, Doctor of Philological sciences, Professor

INTRODUCTION (the abstract of the doctoral (DSc) dissertation)

Actuality and necessity of the research theme. The studies on corpus, corpus technology and corpus linguistics in the world Applied linguistics reached a development stage in the second half of the twentieth century. The first English corpus played an important role in the creation of huge seized electronic corpora for other world languages in the XX century as well. Corpus serves as an investigation object to transform lingual models and speech capabilities of natural language into computer language, to solve language issues using computer technology and methods. One of the global issues in the information age is to keep the sustainable development of natural languages, the further improvement of the electronic corpus and new technologies.

At rapid developing world, corpus technology in computational linguistics serves as an important tool to tackle linguistic problems. A number of scientific progress was made through the positive impact of computer technology on language development and vice versa. Many scholars around the world conducted researches in various areas of computational linguistics[41]. As a result, the number of areas of science have emerged, such as machine translation, corpus linguistics, computer lexicography, spellchecker, speech synthesizer, linguastatistical analysis software, annotation and text classification.

Currently Uzbek computational linguistics is developing in the crossfields such as natural language processing (NLP), machine learning, data mining. Overall practical attainments in aforementioned fields are beneficial each other. It is important to keep mind that to provide an integration with the State language through modern information technologies, pointed out in a number of priorities in accordance with the Decree of the President of the Republic of Uzbekistan on October 21, 2019 № PF-5850 "On measures to radically increase the prestige and status of Uzbek language as the state language"[42]. Several programs have been developed to strengthen the status of Uzbek language and increase its prestige at the level of language policy that focused on creation digital resources of the language.

This investigation implements some functions in official priorities that comprising the Decree of the President of the Republic of Uzbekistan on May 13, 2016 "On the establishment of the Tashkent State University of Uzbek Language and Literature named after Alisher Navoi" No. PF-4797, on February 7, 2017.

[41] Jurafskiy D., Martin J. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2007. – P. 12-13; Karin A., Bengt A. Advances corpus-based contrastive linguistics. USA: John Benjamins, 2013 – P. 25-54; Koehn P., Och F.J., Marcu D. Statistical phrase based translation // Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT/NAACL). Proceedings of the Joint Conference 2003; Mitkov R. The Oxford handbook of Computational linguistics. Oxford university press, 2003; Kurdi M.Z. Natural Language Processing and Computational Linguistics: Speech, Morphology and Syntax. – Great Britain, USA: Wiley-ISTE 2016. – 300 p.; Чардин И.С. Лингвистические корпусы с разметкой на основе грамматики зависимостей и их применение при автоматическом синтактическом анализе: Автореф.дисс. …д-ра филол. наук. – Москва, 2004 – 24 с.

[42] "Mamlakatimizda o'zbek tilini yanada rivojlantirish va til siyosatini takomillashtirish chora-tadbirlari to'g'risida" O'zbekiston Respublikasi Prezidentining Farmoni https://lex.uz/docs/-5058351

Decree No. PF-4947 “On the Strategy of Actions for the Further Development of the Republic of Uzbekistan”, the Law of the Republic of Uzbekistan “On the State Language” No. PQ-4479 on October 4, 2019 Resolution of the President of the Republic of Uzbekistan on October 21, 2019 No PF-5850 “On measures to radically increase the prestige and status of the Uzbek language as the state language” Decree No. PP-1989 of June 27, 2013 “On measures to further develop the national information and communication system of the Republic of Uzbekistan”, on February 17, 2017 “On the activities of the Academy of Sciences, the organization of research work , further management and financing Resolution No. PQ-2789 “On measures to improve the language” .

The relevance of the research to the priorities of the development of science and technology in the country. The dissertation was conducted within the framework of the first direction of the development of science and technology of the Republic I. “Social, legal, economic, cultural, spiritual and educational development of an information society and a democratic state, the development of innovative economy”.

Review of foreign research on the topic of the dissertation[43].

The following scientific results have been achieved in areas such as world corpus technology and corpus-based analysis: modern English, Old English, modern foreign languages, a small number of languages, endangered languages studied by the corpus; in the field of NLP corpus methods have been developed and corpus concordance, tokenizer, stemmizer, lemmatizer were created in this field (University of Lancaster, UK); studies on corpus-based discourse analysis were implemented (University of Birmingham, UK); corpus manager interfaces have been developed (Waseda University, Japan, Iowa University, USA); formal-functional models of the syntactically annotated corpus (Penn Treebank) interface were created (University of Pennsylvania, USA); the first version of the first annotated Brown corpus was built (Brown University, USA); implemented theoretical and practical researches on annotated corpus and morphoanalyzer (University of Leipzig, Germany; Columbia University, USA; Abu Dhabi New York University, United Arab Emirates); corpus analysis developed without morphological homonymy based on semantic dictionary (St. Petersburg University, Russia); syntactically annotated linguistic corpus platforms created via grammar based on theory of dependency and automatic methods of parsers (Laboratory of Computational linguistics at the Institute of Information Transmission of the Russian Academy of Sciences); developed algorithm of the creation of parser (Moscow State University named after M. Lomonosov, Russia); the principles and criteria of learner corpus and linguistically annotated corpus were studied (University of California, University of Cambridge, USA); implemented web-based Crawler tool (University of Berlin, Germany); a system of morphological, syntactic and semantic tagging of the linguistically annotated corpus in Turkic languages was developed (Institute of Applied Semiotics,

[43] Dissertatsiyasiga mavzusiga aloqador xorijiy tadqiqotlar sharhi http://ucrel.lancs.ac.uk/, google.scholar.com; https://www.researchgate.net , https://www.aclweb.org/anthology/L16-1207.pdf, http://qazcorpus.kz/indexen/ , http://www.openslr.org/resources.php, http://ddi.itu.edu.tr/en/toolsandresources

Tatarstan); issues such as adaptation of two-stage morphological analyzer to corpus analysis, sentence segmentation, tokenizer, universal dependence theory of syntactically annotated corpus, development of standards for multi-component words (Istanbul Technical University, Turkey); created a national corpus manager for the Kazakh language (Ahmad Boytursunov Language Institute, Institute of Artificial Intelligence, Kazakhstan); corpus-based language education, parsing based on the theory of dependency, FST technology in corpus morphological analysis, authorship corpus, software and linguistics of national corpus creation, corpus morphological and semantic analyzer, creation of statistical based machine translation of Uzbek language, theoretical and practical research on the formation of the learner and web corpus[44] (Tashkent State University of the Uzbek Language and Literature, National University of Uzbekistan, Tashkent University of Information Technologies, TUIT branch of Samarkand State University).

Degree of study of the problem. In the world experience, the linguistic, mathematical and programmatic aspects of building a corpus have been reflected in researches by scientists[45]. Particulary, V. Zakharov, A. Sedov, A. Baranov, R. Potapova, V. Rikov, U. Francis, N. Leontyeva, V. Martin, S. Kubler, A.Laurens, E.Etwell, S.Hunston, L.Boizou, McKenneri, J.Grafmiller, J.Grieve, N.Grum, S. Hansson, K.McAuliff, M.Malberg, P.Milin, A.Murakami, R. Patch, A. Schembri, P. Thompson, B. Winter, G. Leach conducted researches in different fields of corpus linguistics[46]. In Turkology Aksan, Deniz Zeyrek, Kemal Oflazer, Umut Özge on the Turkish language corpus; Yusuf Aibaidulla, Kim-Teng Lua on Uyghur corpus; L.A. Buskunbaeva, Z. Sirazitdinov on Bashkir corpus; Sheymovich on Khakas corpus, J.Suleymanov, A.Gatiatullin, O.Nevzorova, R.Gilmullin, B.Khakimov on Tatar corpus; L.Kubedinova on Crimean Tatar corpus and A.Salchak on Tuvan corpus are noteworthy as contributors to develop Turkic corpora.

It can be said that initially formation of computational linguistics was in our country by N.Yakubova, M. Ayimbetov, S. Rizayev, S. Mukhamedov's reseaches on investigating linguostatistics[47]. Linguistic models, modeling and its principles,

---

[44] http://uzschoolcorpara.uz; http://uzcorpus.uz/

[45] Kubler S., Zinsmeister H. Corpus linguistics and linguistically annotated corpora. – New York: Bloomsbury, 2015. –P. 321.; Martin W. Developing Linguistic Corpora: A Guide to Good Practice, OxfordBooks. 2005. http://www.ahds.ac.uk/creating/guides/linguistic-corpora/; Heike Z., Hinrichs E., Kübler S., Witt A. Linguistically annotated corpora: Quality assurance, reusability and sustainability / Corpus Linguistics: An International Handbook A. Lüdeling and M. Kytö (eds), Vol. 1, – Berlin: Mouton de Gruyter, –P. 759–76; Копотев М. Введение в корпусную лингвистику (учебное пособие). – Прага, 2014. – 264 с.; Atkins B., Zampolli A. Computational approach to the lexicon. – Oxford, 1994. –P. 494.

[46] Седов А.В. Математические модели, методы и алгоритмы построения размеченных корпусов текстов: Автореф. дис... канд. тех. наук. – Петрозаводск, 2013. – 22 с.; Anthony L. AntConc: Design and Development of a Freeware Corpus Analysis Toolkit for the Technical Writing Classroom, 2005. // IEEE International Professional Communication Conference Proceedings, – P.729-737; Atwell E. Development of tagsets for part-of-speech tagging / An international handbook. Corpus Linguistics: Mouton de Gruyter. 2008; Баранов А.Н., Михайлов М.Н., Сидоров Г.О. Динамический корпус текстов как новая технология прикладной лингвистики // Труды международного семинара Диалог'98 по компьютерной лингвистике и ее приложениям. – Т. 1998; Hunston S. Corpora in Applied Linguistics. – Cambridge: Cambridge University Press, 2002. – 234 p.

[47] Ризаев С. Ўзбек тилининг лингвостатистик тадқиқи: Филол. фан. д-ри. ... дисс. автореф. – Тошкент, 2008. – 50 б.; Мухамедов С.А. Статистический анализ лексико- морфологической структуры узбекских газетных текстов: Автореф. дисс. …канд. филол. наук. – Ташкент, 1980. –25 с.; Бектаев К.Б, Пиотровский Р.Г.

statistical analysis of Uzbek texts obsereved by scholars S.Muhamedov, P.Piotrovsky titled “Egineering linguistics and experimental system-statistical study of Uzbek texts”. Furthemore, in the last decade in the field of computational linguistics published sevaral textbooks and manuals by authorship A.Pulatov, S.Mukhamedova, A.Rahimov, Z.Kholmanova, N.Abdurakhmonova[48]. On creating computer language models, linguistic database, as well as the solution of linguistic problems were investigated by computer methods at the stage of monographic plan[49].

Overall results in corpus linguistics made radical shifts in language technology for both the fields of computational linguistics and natural language processing. It serves as a research object for a number of practical activities, such as the creation a machine language of natural language, obtaining statistical information on language, the formation the language and speech models of artificial intelligence, compiling a machine readable linguistic resources.

Scientific researches in this area are carried out in several higher educational institutions and research institutes in our Republic. It is important to keep mind that scientists who contribute to develop Uzbek corpus linguistics, namely B. Mengliyev, Sh. Shahobiddinova, Z. Kholmanova, S. Karimov, L. Raupova, N.Abdurakhmonova, Sh. Khamroyeva, G. Toirova, G. Ikromova, J. Djumabayeva, G. Ergasheva, A. Eshmuminov, Sh.Gulyamova, M.Abjalova.

Relationship of the dissertation with the research plans of the higher education institution where the dissertation was implemented. Dissertation was carried out within the framework of the international Erasmus + project “Development of the interdisciplinary master program on Computational Linguistics at Central Asian universities 585845-EPP-1-2017-1-ES-EPPKA2-CBHE-JP” inTashkent State University of Uzbek Language and Literature named after Alisher Navoi (2018-2021).

The purpose of the study: creation of conceptual models of electronic corpus of the Uzbek language and development of effective ways viaon computer technology methods for linguistically annotated language corpus.

Математическая лингвистика. – М.: Высшая школа, 1997. – 420 с.; Айымбетов М.К. Проблемы и методы квантитативно-типологического измерения близости тюркских языков (на материалах каракалпакского, казахского и узбекского языков): Автореф.дисс. …д-ра филол. наук. – Ташкент, 1997. – 47 с.

[48] Пўлатов А. Компьютер лингвистикаси. – Т.: Akademnashr, 2011. – 175 б.; Норов А. Компьютер лингвистикаси асослари. – Қарши, 2017. – 136 б.; Муҳаммедова С. Ҳаракат феъллари асосида компьютер дастурлари учун лингвистик таъмин яратиш. Методик қўлланма. – Тошкент, 2006.; Холманова З. Компьютер лингвистикаси (ўқув қўлланма) –Тошкент, 2019.; Abduraxmonova N.Z. Kompyuter lingvistikasi (darslik). – Toshkent: Nodirabegim, 2021. –398 b.

[49] Абдурахмонова Н.З. Инглизча матнларни ўзбек тилига таржима qилиш дастурининг лингвистик таъминоти (содда гаплар мисолида). фил.фан.бўйича фалсафа доктори (PhD)…дисс. – Тошкент, 2018, 165 б..; Абжалова М. Ўзбек тилидаги матнларни таҳрир ва таҳлил қилувчи дастурнинг лингвистик модуллари (Расмий ва илмий услубдаги матнлар таҳрири дастури учун): филол.фан.бўйича фалсафа док. (PhD)…дисс. – Фарғона, 2019. – 164 б.; Хамроева Ш. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари: филол. фан. бўйича фалсафа док. (PhD)…дисс. – Қарши, 2018. – 250 б.; Эшмўминов А.А. Ўзбек тили миллий корпусининг синоним сўзлар базаси: филол. фан. бўйича фалсафа док. (PhD)…дисс. – Қарши, 2019. – 140 б.; Тоирова Г. Ўзбек тили миллий корпусини яратишнинг назарий ва амалий масалалари. – Германия: Globeedit, 2020. – 168 б.

Research tasks:

to study foreign experience on projecting the conceptual and structural principles of Uzbek electronic corpus;

to develop a linguistic algorithm transfering language models into machine by applying automatic methods on morphological and syntactic tagging and analysis in Uzbek, such as FST and UdPipe for creating a linguistic corpus;

to create linguistic algorithm and software of the corpus for providing the representation of the text fragment in the interface and obtain the results of search units;

to form the corpus interface on formal-functional models basis of the corpus manager.

The object of research is official, scientific, artistic and journalistic texts in various genres and official and scientific materials in English and Russian for the parallel corpora.

The subject of the research is technologies and methods of creating an electronic corpus, models of grammatical tagging and analysis of the corpus, Uzbek language, linguistic and software of the corpus manager.

Research methods. Methods such as descriptive, comparative, component analysis, statistics, modeling were used to conduct on the research topic.

The scientific novelty of the research is as follows:

developed linguistic analysis of lemmatization and tokenization processes by appliying FST – finite state trunsducer for the first time for the Uzbek language as automatic morphological tagging stages for Uzbek corpus and created morphological database, and morphotactic rules system;

the CONLL-U model of the theory of universal dependency for the Uzbek language is formed on syntactic models basis of the Uzbek language by means of UdPipe method on syntactic tagging and annotation process;

as a practical reserach result http://uzbekcorpus.uz has created and developed corpus manager of n-gram model concordance, lemma, token and word combination;

WordFast technology was used to create the translation algorithm and the translation memory of the parallel corpus for the Uzbek language, and created linguistic database of alignments of the Uzbek translation units in English and Russian;

proved the scientific foundations of ontological modeling method through the Protege program at the stage of grammatical tagging of the corpus;

based on the expediency of using the Jaro Winkler algorithm in the application in the search interface to determine the spelling errors of the word in the corpus and to identify the coefficient of words close to variations.

The practical results of the research are as follows:

created database of the language various lexical units which based on linguictic database of the corpus and at the result obtained authorship certificates[50];

[50] Abduraxmonova N., Tuliyev U. Oila, mahalla va gender tengligi mavzusidagi badiiy asarlar elektron korpusining dasturiy ta’minoti. Intellektual mulk agentligi guvohnomasi 30.03.2021 DGU 10653; Abduraxmonova N., Tuliyev

the outcomes of the investigation are represented at http://uzbekcorpus.uz;

some frame of the computer models of the electronic corpus in dissertation was applyed in the short-term project JHBL-20 titled “Creation of an electronic corpus of literary works on the theme of *family, neighborhood and gender equality”* (2020-2021), implemented at the Research Institute “Makhalla and Family”;

created a linguistic database based on the morphotactic rules dividing by nominal and verbal groups of parts of speech by applying FST technology for morphological annotation of the corpus;

the results obtaining in the process of grammatical tagging and annotation of the electronic corpus were used in the creation of the grammatical thesaurus of Turkic languages[51];

implemented as the results of computer models for creating morphoanalyzer and parser of corpus for the investigation in the scope of the international project Erasmus+ “Development of the interdisciplinary master program on Computational Linguistics at Central Asian universities” (585845-EPP-1-2017-1-ES-EPPKA2-CBHE-JP, 2018-2021);

developed the linguistic analysis of lemmatization and tokenization processes for the Uzbek corpus and created “Morphological database of the Uzbek language for linguistic processors and electronic corpora” based on the creation of a morphological database, morphotactic rules system in the dissertation. (Authorshipr certificate given by the Intellectual Property Agency under the Ministry of Justice of the Republic of Uzbekistan 2021.19.02, No. BGU 00394);

the results of automated morphological tagging of Uzbek Corpus by means FST (finite state trunsducer) were used – *Uzmorphoanalyzer* web-program which directed morphological analysis of Uzbek words (2021.17.06. DGU 11432 certificate issued by the Intellectual Property Agency).

Reliability of research results It is confirmed that scientific conclusions based on materials with general methodological basis and methodologically grounded theoretical point of views in corpus and computational linguistics, moreover, approbation of the tasks set in the dissertation, practical testing of recommendations, obtained results are confirmed by the competent authorities.

Scientific and practical significance of research results. The scientific significance of the outcomes of the research and the representation of the structured text and suggestions for the analysis of the text at different linguistic stages help to conduct researches on Uzbek text processing technology, parallel text corpus language and speech phenomena, which are the object of study for the special sphere.

U., Maharov Q. “Oila ma’naviyati” konseptining elektron tezaurusining dasturiy ta’minoti. Intellektual mulk agentligi guvohnomasi 30.03.2021, DGU 10655.

[51] ИРН: AP05132249 «Разработка электронных тезаурусов тюркских языков для создания систем многоязычного поиска и извлечения знаний» по договору № 132 от «12» марта 2018 г. http://alphabet.kz:9191/

The scientific importance of dissertation the results is the representativeness of structured text structures and proposals for the analysis of the text at different linguistic stages assists in conducting targeted research.

The practical significance of the research results is that it is basic source in the translation studies. This is explained by the fact that it can be used in the process of providing lectures, creating coursebook and mannuals, and electronic dictionaries (translation, thesaurus) in the subjects of Corpus linguistics, Machine translation, Computational linguistics at universities.

Implementation of research results. The scientific and practical results of computer models of the Uzbek electronic corpus include the principles of linguistic annotation, linguistic models, automated methods (FST technology and UdPipe method), corpus manager, search engine editing software implemented following aspects:

the site http://uzbekcorpus.uz, created as a practical result of the research, is used as a web resource in teaching subjects such as Corpus linguistics and morphological analyzer in Computer Linguistics at the Institute of Computational Mathematics and Information Technology in Kazan Federal University (Kazan Federal University Institute of Computational Mathematics and Information Technology) reference book, 2021). As a result, the electronic corpus of the Uzbek language (http://uzbekcorpus.uz) was added to the list of Turkic corpora at http://www.turklang.net, hence it hels to compare the structural and functional features of the Turkic languages and the use of natural language technologies as a linguistic resource. created an opportunity;

as a practical expression of scientific findings of dissertation, used in the preparation articles and monographs provided by a fundamental project trough 2017-2020, named OT-F1-029 “The place of the Sogdian language and script in the history of Uzbek statehood (II-XII centuries BC) concentrated in the corpus through the search system by the lemma, token and of concordance n-gram model of the corpus manager (search engine), parallel corpus and translation dictionaries in the platform were used effectively in publishing in world scientific journals. As a result, the content of the research carried out on the project has been enriched, the scientific quality has increased;

as a result of the dissertation of functional features of the electronic linguistic corpus manager, such as lemma, token, concordance and word combination search system applied in the preparation of a multi-volume “Explanatory Dictionary of the Uzbek language” in Latin in the fundamental project OT-F1-78 “The Uzbek language in current globalization epoch, its historical development and prospects (based on the analysis of functional styles)” done through 2017-2019. Moreover, it was used to provide fragments of text in the textbook “Sylistics of the Uzbek language” by taken by in five types of style in the electronic corpus (Reference by No. №3 / 1255-2651 of the Academy of Sciences of the Republic of Uzbekistan, September 16, 2021). As a result, the content of the research conducted on the project has expanded, the occurrence of articles in different contexts with the semantic fields was determined on the basis of the corpus;

the corpus search system by n-gram model concordans, lemma, token of corpus manager as findings in the dissertation used JHBL-18 “Study of family values of different generations” at of the Research Institute “Neighborhood and Family” (2020-2021) (Reference No. 13 on June 25, 2021 of the Research Institute “Mahalla and Family”). As a result, it has served as a basis for researchers to study national customs and characteristics of the language according to the gender, age, gender and to identify scientific statistics in the reserach.

practical results on creation Uzbek electronic corpus implemented for the the project JHBL-18 “Creation of a web portal titled “*Happy Family*” in Uzbekistan” (2020-2021). The systematization and chronological classification of texts on the subject of the family in various genres reflected in the corpus served as a source for the scientific analysis of values and social relations as an institution of the family (Reference No. 12 of June 25, 2021 of the Research Institute "Mahalla and Family"). As a result, the concepts of the project theme are enriched with new scientific facts;

applied search system by lemma, token, concordance of Uzbek corpus as one of practical achievments of the research implemented to expand materials of electronic dictionary of practical project I-OT-2019-42 “Creation of an electronic poetic dictionary of Uzbek and English languages (Appearance, character of people, image of nature and national symbols) (2019-2021)” (Tashkent State University of Uzbek Language and Literature named after Alisher Navoi 01 / Reference No. 4-1638 dated September 18, 2021). As a result, achieved to collect lexical units which represent appearance and characters of people, nature and national figures and identified their English and Uzbek translation alignments by using the sub (parallel) corpus translation memory and translation segment units;

linguistic analysis stages of lemmatization and tokenization processes have been developed for the Uzbek language corpus, and morphological database, morphotactic rules system and for the first time for the Uzbek language FST (finite state transducer) – automatic methods of morphological tagging of the corpus using a finite state transducer, CONLL-U model of universal dependency theory for Uzbek language on syntactic models the basis of Uzbek language by UdPipe method in syntactic tagging and annotation of Corpus are represented in the textbook on Computational linguistics for masters, doctoral students, researchers and specialists, as well as students of advanced training and retraining courses for teachers, and undergraduate students who are directed 5111200 - Uzbek language and literature, 5120100 - philology and language teaching (Uzbek language), 5120900 - Uzbek-English translation theory and practice are reflected. (Order of the Ministry of Higher and Secondary Special Education No. 654 of July 20, 2019). As a result, the ideological basis of the textbook is theoretically and practically based and enriched with scientific information.

Approbation of research results. The dissertation results have been published in 21 international and 21 national scientific conferences. Author’s scientific works have been discussed on the following https://scholar.google.com/citations?user=kEkD0kgAAAAJ&hl=en, https://www.researchgate.net/profile/

Nilufar-Abdurakhmonova, https://orcid.org/0000-0001-9195-5723 scientific platforms.

Publication of research results. Overall 65 scientific works were published on the topic of the dissertation, including 12 articles in scientific publications recommended for publication of the main scientific results of doctoral dissertations of the Higher Attestation Commission of the Republic of Uzbekistan (two of them in foreign journals), 1 monograph, 5 copyright certificates, 46 scientific articles and theses were published at the prestigious international conferences indexing Web of science and 3 articles indexed in Scopus, as well as at national and international conferences.

The structure and the size of the dissertation. The dissertation consists of an introduction, four chapters, a conclusion, a list of references and 220 pages.

MAIN CONTENT OF THE DISSERTATION

The introduction analyzed the relevance and necessity of the chosen topic, the level of study of the problem, the review of foreign research on the topic, the goals and objectives of the research, the object and subject, the relevance of science and technology, scientific novelty and practical results are described. Based on the reliability of obtained the results, the theoretical and practical significance of the topic is revealed. Given information on the implementation of research results in practice, approbation of the work, published works and the structure of the dissertation.

The first chapter of the dissertation is entitled "*Corpus as the object of study of language technology"*. The first paragraph of this chapter is entitled "*Analytical study of contemporary Corpus Linguistics*" analyzes scientific viewpoints by scholars on corpus linguistics and corpus technologies. Linguistic issues on corpus technology and corpus-based analysis explored on discussing corpus either model or methodology by scholars's works: Graeme Kennedy, Friedrich Wilhelm Kaeding, Martin Weisser, Victor Zakharov, Charlotte Taylor[52], Charles Meer[53], Noam Chomsky, Dougles Bieber[54], Sandra Kubler[55]. The first electronic written corpus were Brown (1964), LOB (1978), FROWN (1999), FLOB (1998), Kolhapur (1978), ACE (1986), Lund corpus; corpora such as SEU, LLC, SEC (English text corpus), Map task corpus, and HKCSE (Hong Konk corpus of English oral texts) were compared with oral corpora, and their content and formal-functional models were analyzed in this chapter.

The second paragraph entitled *"Taxonomy of Corpora"*, is devoted to the classification of corpora according to various parameters. In relation to this, corpora might be according to: 1) types of databases: a) verbal; b) written; d) mixed; 2) the text given in some language: a) English; b) Russian; d) German…; 3)

[52] Taylor Ch. What is corpus linguistics? What the data says // ICAME Journal, 2008. – P. 32. 179-200.
[53] Charlez M. English corpus linguistics: An introduction. Cambridge University Press, 2004. – 168 p.
[54] Bern H., Heiko N. The Oxford Handbook of Linguistic Analysis. / Dougles Biber Corpus-based and Corpus-driven analysis of language variation and use. – UK: Oxford university, 2015. – 193 p.
[55] Sandra K. Heike Z. Corpus Linguistics and Linguistically Annotated Corpora. –New York: Bloomsbury Academic, 2015. – P. 43.

by the parallel texts: a) bilingual; b) trilingual; d) multilingual; 4) by styles: a) spoken language; b) journalistic; d) artistic; e) official; f) scientific; 5) depending by availability access of the database: a) open; b) closed; 6) by geographical location: a) by particular geographical location; b) by several countries; 7) corpus content type: a) general; b) special.

Hence, M.Z. Kurdi classified oral corpus according to input texts in database: speech recording, speaking command operators, human-machine dialogue, machine-assisted human-human dialogue and multi-model dialogue system. Balanced, pyramid and limited corpus types are subdivided according to thematic levels. Module of references, lingua graphic description, lexical units represented as three components of corpus in some references[56]. This chapter analyzes the point of views on the description and functional capabilities of corpora created in different languages specific to the aforementioned classifications.

The third paragraph entitled "*Corpus studies issues in Computational Linguistics",* analyzes the issues such as compiling the texts, classification, annotation, machine language understanding, and linguistic modeling of machine translation (NLP-natural language processing).

In scientific sources, corpus linguistics is referred to as a branch of natural language processing (NLP) [57]. Some sources refer to it as computational linguistics or applied linguistics. According to scholars' the point of views such as Jurafsky, Martin (2008), Manning, Schütze (1999), Roark, Sproat (2007) Computational linguistics has special optimal methods of annotating the corpus of text and its linguistic analysis, which can solve complex problems. According to Graeme Kennedy[58], although Corpus linguistics did not exist at the same time as the development of computers, the problems of text-based linguistic analysis were solved as a result of the expansion of computer memory capabilities. The design, size and individual character of each corpus are formed for different purposes. The corpus is the object of research for computer translation, machine translation, speech synthesizer, text analysis, sentiment analysis, and a number of other areas. For example, statistical machine translation technology is based on data from large corpora. Therefore, in computational linguistics, language and speech phenomena are studied using corpus technology.

The second chapter of dissertation is entitled "*Technologies and methods of creating an electronic corpus"* and its the first paragrapgh is entitled *"Creation corpus technology: linguistic tools".* This paragrapgh deals with the study of tools and platforms that can be used to create a corpus. The CLARIN platform supports a variety of linguistic tools for common language resources and technology infrastructure. Instruments for corpus technology in world languages are available at https://corpus-analysis.com/, and in this section AntGram, AntConc, AntWordProfiler, AntCorGen, BFSU Collocator, BFSU Sentence Collector,

[56] Каримуллина Р. Н., Каримуллина Г. Н. О сводном лингвографическом корпусе татарского языка/ слово и словар vocabulum et vocabularium // Сборник научных материалов МИНСК ИЗДАТЕЛСТВО «ЧЕТЫРЕ ЧЕТВЕРТИ» 2017. – С. 18.

57 Mohamed Z.K. Natural Language Processing and Computational Linguistics: Speech, Morphology and Syntax. – Great Britain, USA: Wiley-ISTE, 2016. –P. 12.

[58] Graeme K. An introduction to corpus linguistics. – London: Longman, 1998. – P. 2.

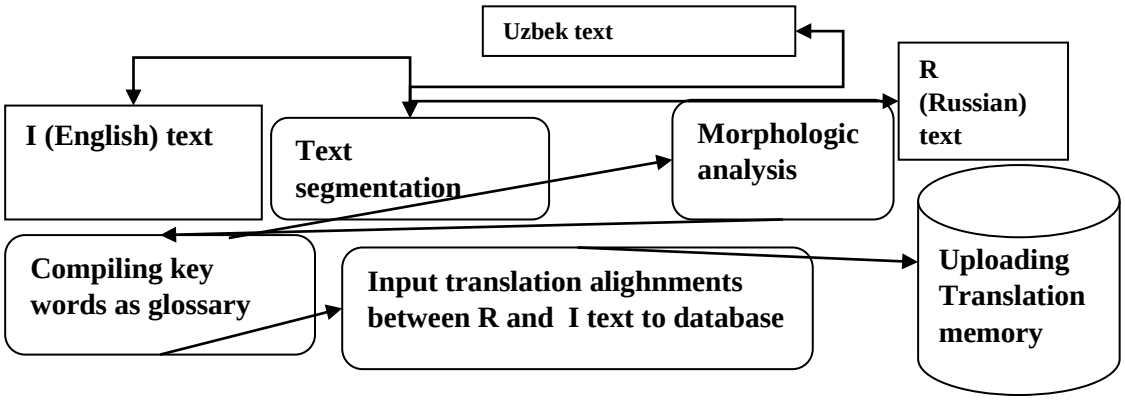
BNCWeb, LiC, FrameNet, Corpus interfaces such as WordSmith reveal the functionality of creating a corpus as a ready-made tool, as well as performing text-related analyzes.

The second paragraph is entitled *"Software for creating special corpora*", discusses as examples of the *AntConc and BootCat* applications for creating a special corpus for Uzbek language, which are distributed according to the genres and size of the text, which are not too large for specific purposes or areas of researches.

The *AntConc* program, developed by Laurens Anthony, is designed for corpus analysis such as word and keyword generator, concordance, *n-gram*, and phrase. The AntConc application consists of three main parameters: corpus files, tabs, the search interface, and its device. Corpus files have two options: *open files* or *open directories* at the same time. The tab panel contains the following tools: plot, concordance-KWIC, frequencies, keywords. The menu section defines the Unicode of using language, selects the color of the words on the *n-gram*, changes the order in which the metadata of the tags is presented or hides it altogether; there is a special function for tokens, which allows you to specify a group of tokens, such as letters, numbers, punctuation, symbols, spacing. The *BootCat* tool also has the function to gather information on the Internet in a short period of time, which can be used to create a sectoral terminological database using a specific term.

The third paragraph is entitled "*Linguistic tools and platforms for creating a web corpus*", which provides information about the Sketch Engine platform for creating a corpus based on texts from websites and the first *Turkic web-Uzbek* corpus for the Uzbek language. In addition, WebLicht and CLARIN-D platforms, which have the ability to create a corpus using software applications such as WordSmith, BlackLab, and automatically annotate the text corpus are presented as examples in this chapter.

The fourth section of the chapter entitled "*Computer methods of creating a parallel corpus*" shows the division into speech segments of bilingual Uzbek-English, English-Uzbek, Uzbek-Russian, Russian-Uzbek parallel text, translation units, the advantages and functionality of *WordFast* technology in identifying their



alingments, creating translation memory and lexicon. Moreover, an algorithm for creating parallel text was developed as a result of research for of Uzbek.

The third chapter of the dissertation is entitled *"Models of grammatical tagging and analysis of the corpus"*. In its first paragraph entitled *"Types of*

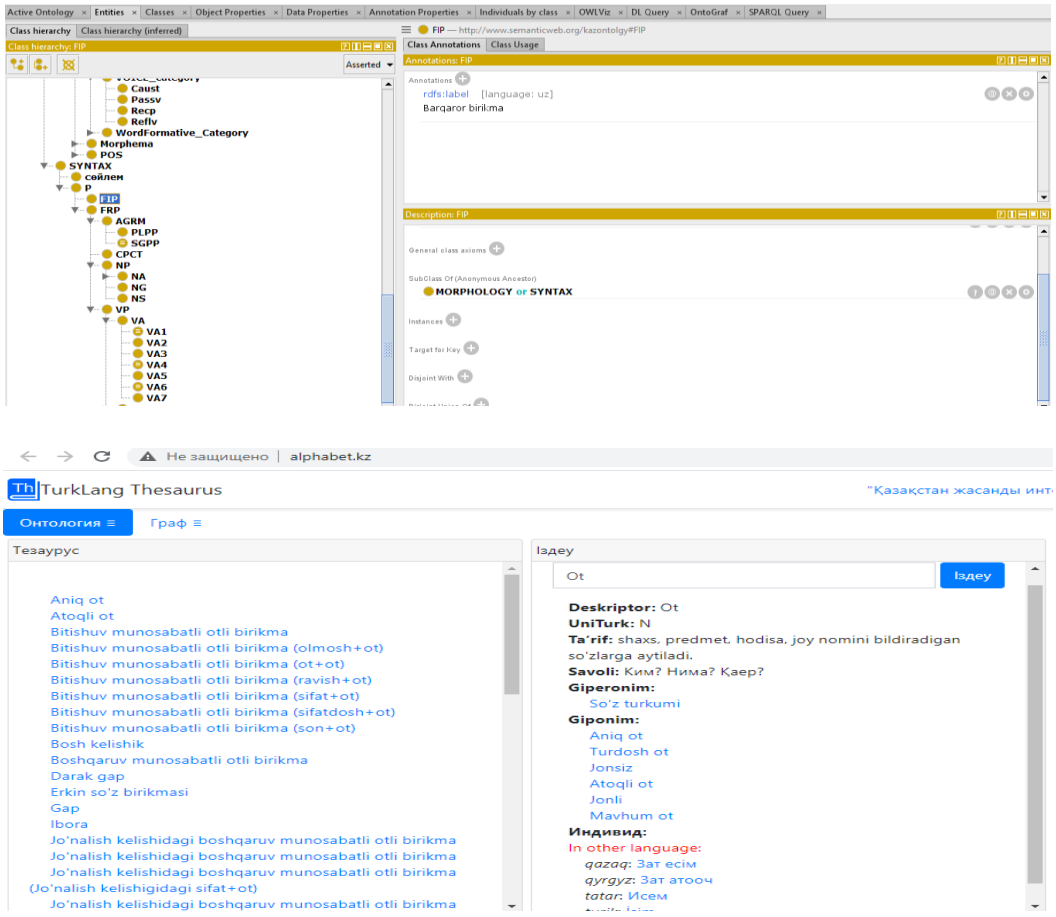
Linguistic Annotation", text annotation is classified as follows[59]: extra linguistic (metadata)- information about the text (author, year of publication, name of publishing house, year, genre, subject) and information about its author; 2) text structure - title, heading, sentence, word form; 3) according to the linguistic level: a) morphological (POS-tagging) markup; b) syntactic markup - this includes information on the syntactic relationships between words and the types of sentences; d) semantic markup - differs according to the subject and non-subject names, types of activities and semantic relations of the concepts expressed in the text; e) anaphoric markup - an element of the text is identified and linguistically analyzed in relation to the content understood in another text; f) prosodic markup - the linguistic expression of the text in terms of stress, rhythm, logical stress, etc. It is also advisable to use the following as annotation**[60]** criteria based on the Leech[61] rule: 1) when the annotation part is removed from the corpus, it can become a normal corpus to use; 2) annotations from the text have the ability to expand themselves and receive annotations from other texts that correspond to the text template; 3) the annotation model is based on the end-user instructions; 4) the creator of the annotation should indicate how to apply them; 5) the last user is aware that the corpus annotation is not inaccurate, but simply a useful tool; 6) the structure of the annotation is based on generally accepted norms; 7) none of annotation must be accepted to previous annotations as standard, and standard norms may arise during the experiment.

This section also provides information on the types of annotation format. According to this there are three different forms of annotation[62]: 1) linear (horizontal) format - after each word, its grammatical information is given instead of linguistic metadata; 2) linear format, represented by keywords - represents the linguistic annotation, externally connected with the corpus information; 3) vertical format: each row contains data for a specific morphosyntactic or character group of the token.

In the second paragraph entitled *"Implementation of Protégé technology of ontological modeling at the grammatical level"*, the grammatical basis of linguistic marking is analyzed by the method of ontological modeling. Protégé has protégé-frame and protégé-OWL editing tools as a free, open-source editing program for creating frameworks and ontologies for databases. The program explores grammatical models of the Uzbek language and their taxonomic relationships. This technology has a number of advantages in interconnecting knowledge bases in a particular field and creating their formal models[63]. Using this environment, mechanisms for hierarchical analysis of species and classification relationships (Figure 1-2) and semantic analysis based on the created models have been

[59] Боярский К. К. Введение в компютерную лингвистику. – Санкт-Петербург, 2013. – С. 28.
[60] www.lancaster.ac.uk/fss/course/ling/corpus/corpus2/2MAXIMS>HTM
[61] Geofrey Leech - matn annotatsiya bo'yicha qoidalarni rivojlantirgan ilk korpus tilshunosi.
[62] Захаров В.П., Азарова И.В. и др. Моделирование в корпусной лингвистике: Специализированные корпусы русского языка, ИЗДАТЕЛСТВО САНКТ-ПЕТЕРБУРГСКОГО университета, 2019. – С. 49.
[63] https://ru.wikipedia.org/wiki/Prot%C3%A9g%C3%A9

developed. The method of ontological modeling created a thesaurus of morphological and syntactic categories of the Uzbek language[64].



The third part of the paragraph is entitled *"Study of FST technology in morphological tagging and analysis of Uzbek texts"*. This subsection identifies methods to transfer natural language to machine language using FST technology for a linguistically annotated corpus and morphological tagging and analysis of its models using the tagging system of the universal subordination platform.

In the 1990s, morphological analysis was based on methods such as the *generative model, the paradigmatic model, and the two-component morphological model*[65], in which the results obtained from a linguistic resource are statistically analyzed according to the degree of probability. The process of automatic morphological analysis is classified into the following types:

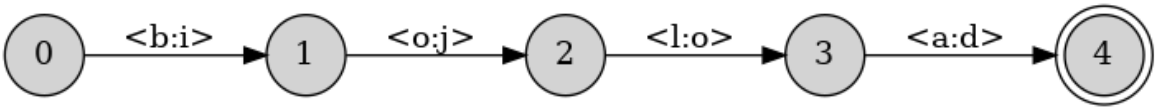
1. Morphological analysis by the method of logical multiplication.
2. Morphological analysis using a dictionary-free table.
3. Word-based morphological analysis (word-based morphological analysis is adapted for European languages. It uses additional auxiliary tables).
4. Morphological analysis according to word forms (according to morphotactic rules).

[64] http://alphabet.kz:9191/

[65] Cerstin M., Michael P. (eds.). JSLIM – Computational Morphology in the Framework of the SLIM Theory of Language / State of the Art in Computational Morphology. – Zurich, 2009. –P. 15.

The process of text normalization is associated with morphological analysis and has the following stages: tokenization (identification word forms in the text) -> lemmatization (determination of words in the dictionary) -> stemming (derivation of derivative words core detection).

FST technology was initially used by Koskenniemi for Finnish, and Antworth for English morphology[66]. Later, a two-stage morphological analyzer was used in Japanese[67], Korean[68], Turkish[69], Arabic[70], and Mongolian[71]. The following files were created for FST in our work: 1) "Rules" - alphabet, phonological rules and special phonemes that occur in a phonetic phenomenon (based on both graphemes for the Uzbek language: Cyrillic and Latin); 2) Lexicon (which includes the basic and simple constructions of all word groups). In an existing morphological lexicon transducer using FST technology, the rules are compiled in a two-stage process. These lexicons and rules are then combined. This is done using command *compose*. A finite automaton transducer is a finite automaton that goes through two *belts.* For example, when a *o'g'il* word is entered into software, the word *ijod* can be generated.



Dictionary transducer and a two-stage rule tranducer generated at the stage of the morphological analyzer for the Uzbek language corpus. Below we analyze the morphotactic rules of the Uzbek language using FST. Regarding technology, FST technology is a declarative language, and we transferred linguistic models of as computer models.

LEXICON NumC
+SG: Poss1;
+PL:lar Poss2;

LEXICON NumC (consonant) is a numeric category of a word ending with a consonant. LEXICON NumC + SG = Poss1 This unit represents the 1st position of a word category ending with a consonant. LEXICON NumC + PL: lar = Poss2 This model represents the 2nd position of the plural suffix in a word ending with a consonant.

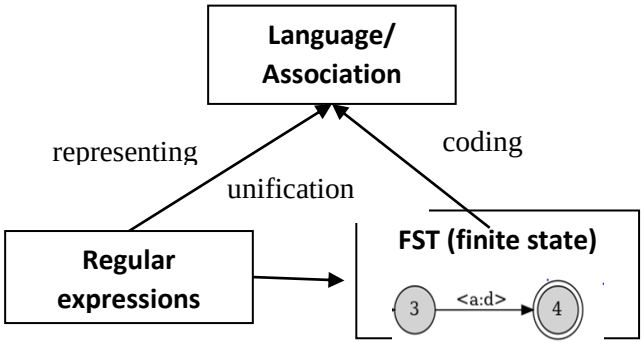
LEXICON NumV
+SG: Poss2 ;
+PL:lar Poss2;

[66] Antworth, E.L. PC-KIMMO: A Two-level Processor of Morphological Analysis. Summer Institute of Linguistics. Dallas, 1990.
[67] Alam Y.S. Two-level Morphological Analysis of Japanese // Linguistics Forum. –Texas, 1983. –P. 229-252
[68] Kim D.B., Lee S.J., Choi K.S., Kim G.C. A two-level morphological analysis of Korean. In Proceedings of the 15th conference on Computational linguistics - Volume 1 (COLING '94), 1994. – P. 535-539.
[69] Oflazer K. Two-level description of Turkish morphology // Literary and Linguistic Computing, Literary and Linguistic Computing Volume 9, Issue2. 1994. – P. 137-148.
[70] Beesley K.R. Arabic Finite State Morphological Analysis and Generation, In COLING-96, Cophenagen. – P. 89-94.
[71] Jaimai P., Zundui T., Chagnaa A., Ock C.Y., PC-KIMMO-based Description of Mongolian Morphology, International Journal of Information Processing Systems Vol.1, No.1, 2005. – P. 41-48.

LEXICON NumV + SG = Poss2 Represents the singular form of a word ending with a vowel. Thus, in the declarative language, *LEXICON NumV* serves as a model of the word to determine the positions of subsequent words.

```
LEXICON Poss1
  +PP1+PSG:m      Case  ;
  +PP2+PSG:ng     Case  ;
  +PP3+PSG:si     Case  ;
  +PP1+PPL:miz    Case  ;
  +PP2+PPL:ngiz   Case  ;
  +PP3+PPL:i      Case  ;
   0:0            Case  ;
```

The morphotactic positions that can be added to the above model word [LEXICON Poss1] are entered separately according to FST rules. Each input model must be unique and universal for the morphological analysis of all words. The segment in each morphologically tagged corpus is entered into the database according to the language model with morphologically tagged units by word unit and word derivation (grammatical index). If the above model is analyzed, the grammatical index of an arbitrary word depends on the zero index of that word: the nominal type is a category of numbers that come in singular or plural form, followed by the possessive suffixes, then the category of case. To derive the morphological tag and its analysis, words divided into nominal (Noun, Num., Pron., Adj.) and verbal (verb) groups are adapted to the declarative language model. The basic concept of FST technology is based on the following triangle[72]:



The corpus text considered for search system are both in Latin and Cyrillic graphics, we combined the two alphabets in the declarative language with the % sign. Then a morphophonological rule for each of the rules found in the spelling rules, corresponding to two graphemes:

```
a.        %##:0
  %^Y:0 %^Y:y %^Y:й
  %^G:g %^G:k %^G:q %^G:ғ %^G:г %G:к
  %^K:k %^K:g %^K:г %^K:к
  %^Q:q %^Q:g %^Q:ғ %Q:қ
  %^A:a %^A:0 %^A:a
```

[72] Kenneth R.B., Lauri K. Finite State Morphology. – US: CSLI publication, 2003. – P.44.

%^U:u %^U:0 %^U:y
%^I:i %^I:0 %^I:и

1. A set of exceptional rules. This set of rules describes the model of all morpheme suffixes that undergo phonetic change:

"Y is dropped after consonant"
%^Y:0 <=> [:LatCons | :LatCons] [:‘]* [:%-]* _ ;
"Y remains after Latvowels"
%^Y:y <=> :LatVowel [:‘]* [:%-]* _ ;
"Y remains after Cyrvowels"
%^Y:й <=> :CyrVowel [:‘]* [:%-]* _ ;
"G remains g"
%^G:g <=> [:LatConsMinus |:LatVowel] _;
"G becomes k after k"
%^G:k <=> :k _;

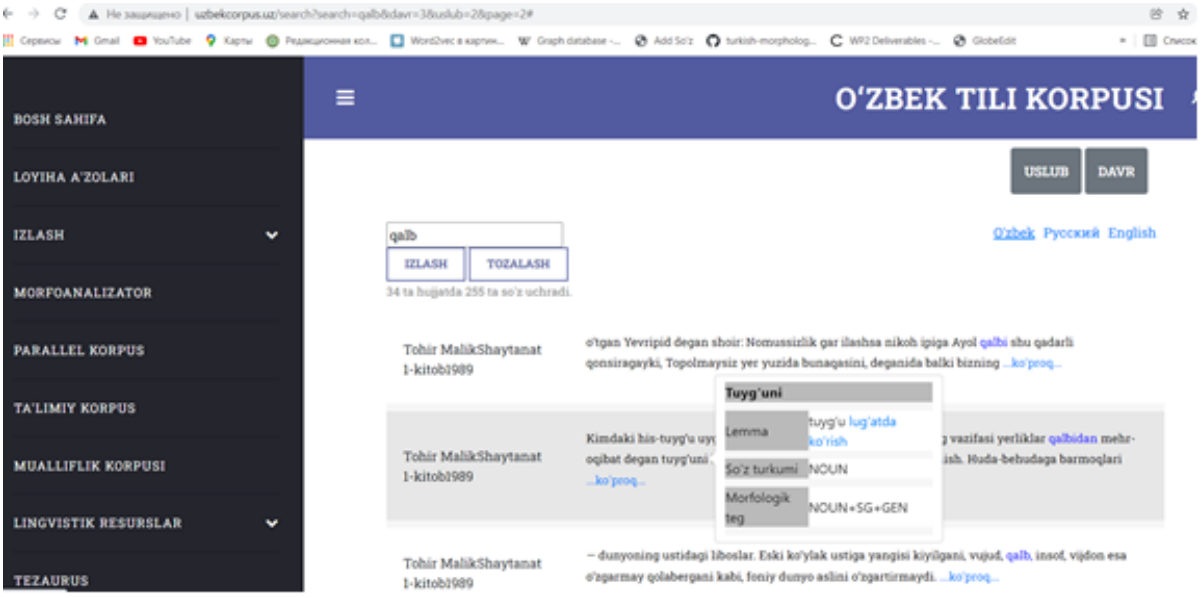
2. The next step is the process of entering special characters for allomorphs.

Each tag of the allomorphs entered for FST (Multichar_Symbols) is taught in machine language for variable phonemes and morphotactic rules. For example, the ^ Y symbol indicates the change position of the *-u* and *-yu*.

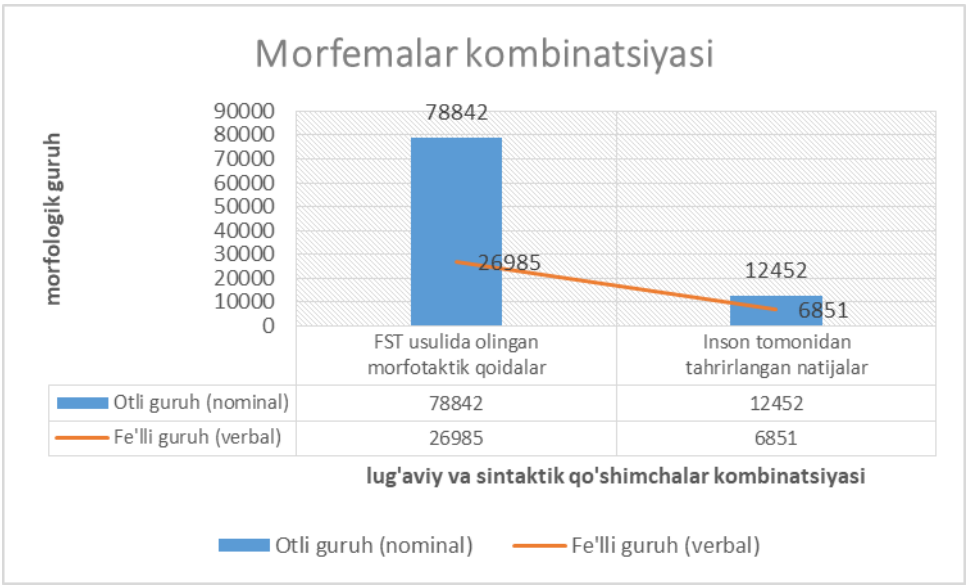
^Y
^K
^Q
^G
^A
^U
^I
^V
^W
^P
+VERB +CONV
+PAST +PRES +FUT
+PP1 +PP2 +PP3 +PSG +PPL
+P1 +P2 +P3 +SG +PL
+NOM +ACC +GEN +DAT +LOC +ABL
+PART
+NOUN
##

Consequently, the results for the morphological tagging of the Uzbek language electronic corpus are obtained:

Result fragment: **shahar+NOUN+PL+PP3+PSG+ELIPT+GEN:shaharlarinikini**
shahar+NOUN+PL+PP3+PSG+ACC: shaharlarining



In the morphological database of our work, proper names and phrases are also included as a separate (*subdata*) base. In addition, the database includes morphotactic features of Uzbek auxiliary verb conjugations and accusative pronouns that can be added to each of these verbs. Our corpus as balanced corpus the content of the texts consists of a generally equal proportion in *various genres, such as sports, culture, enlightenment, religion, education, science, politics, law, and literature.*



In search phase if the stem or phrase in the text does not exist in the database, the problem is solved using the following algorithm. Let *w* be a given word, $w = l \oplus \sum_{i=1}^{n} m_i$, *l*-lemma, and $m_i \in \{m_1, \ldots, m_n\}$ a sequence of morphemes (defined by separating morphemes from right to left). In this case, the value of the argument *p*, which determines whether the *l* lemma belongs to the *V*-verbal group or the *N*-nominal group, is determined as follows:

$$p = \begin{cases} 0, agar\ n \geq 2\ va\ m_1 \in M_v, \\ 1, agar\ n \geq 2\ va\ m_1 \in M_n. \end{cases}$$

Here is a set of morphemes added to M_v-verbal words, and a set of morphemes added to M_n-nominal words. If *p = 0*, then l lemma belongs to group *V*, otherwise to group *N*, i.e. the decision about *l* must be at least two of the number of morphemes added to it and which of the first morphemes is added depends on the group.

The fourth paragraph entitled *"UdPipe technology for syntactic tagging and analysis of Uzbek texts"*, analyzes the process of syntactic annotation and tagging as another model of the Uzbek language corpus. The Lancaster-Leeds Treebank in English is one of the first examples of syntactically annotated corporations around the world. In our study, we paid attention correspondingly the investigation of Sabine Buchholz[73] who analyzed comparatively on the syntactic types of linguistic tagging as syntactic structures through examining Prague Syntactic Annotated Corpus (PDT) for Czech language, Arabic Syntactic Annotated Corpus (PADT), Slavic (SDT) of the Slavic language, Talbanken for the Swedish language, Negra / TIGER for German, and Metu Sabanci corpus for Turkish. Syntactic annotation is one of the most important stages of the linguistic corpus analysis. Syntactic parsing plays an important role in solving certain levels of problems in a number of areas of NLP and language technology. Computer analysis of the basic linguistic models and parts of speech syntactic relationships is especially difficult in text analysis. Syntactic analysis (parsing) automatically analyzes syntactic relationships and structures in the text. Parsing is used in almost all areas in computational linguistics and NLP (machine translation, question and answer, information retrieval system, sentiment analysis). Various formats and methods of syntactic tagging for parsing have been developed in this field. These include subordinate grammar, *direct component grammar, mixed theory (syntactic group theory), and free context grammar*.

The theory of universal dependency is available at www.universaldependency.org. As according to information 2018, 74 of the 132 syntactic analysis corpora (treebank) are available on the platform http://universdependencies.org/. Some are based on *the theory of dependency and some are based on the theory of components*.

In our work, the process of syntactic tagging of text using UdPipe technology was carried out. This process is reflected in the CONLL-U format. In the syntactic description of sentences, words with several components (phrase, compound word) are taken as a separate base. Typically, when segmenting sentences into segment units using a segmenter, the line or line spacing between sentences is taken into account. In this regard, Milan Straka focuses on each algorithmic description of the implementation of a syntactic analysis of sentences by a specific model[74].

[73] Sabine B., Erwin M. CoNLL-X shared task on Multilingual Dependency Parsing https://www.aclweb.org/anthology/W06-2920.pdf – P. 152.

[74] Straka M., Strakova J. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe https://ufal.mff.cuni.cz/~straka/papers/2017-conll_udpipe.pdf

According to him, UdPipe as a ready-made model performs a complex *syntactic analysis based on segmentation, tokenization, word grouping, lemmatization and subordination*.

In our study, morphological analyzing and annotation was applied by FST-technology. In syntactical analysis, the three characters of each word, UPOS, XPOS, and FEATS, were analyzed separately in CONLL-U format. Part-of-speech tagging is the process of marking each word entered in a text by word group. The information entered into the tagging algorithm consists of a set of tags and a sequence of words. Here:

- **ID**: word index, each line is written on a new line at the beginning of a sentence.
- **FORM**: word form or punctuation.
- **LEMMA**: lemma or stem of a word.
- **UPOS**: universal phrase tag
- **XPOS**: language-specific word tag; if not, the underscore is _.
- **FEATS**: order of morphological categories, if any
- **HEAD: ID** number of the word (index of the word in the fourth order in the sentence) or zero (0) order).
- **DEPREL**: Dependent attitude to the keyword. Syntactically annotated corpora contain these sets of dependencies.
- **DEPS**: Extended dependency graph (optional), set to _ if not written.
- **MISC**: other annotation code

Since this UdPipe method is used for Finnish and Turkish, it is considered to be a suitable component analysis for agglutinative languages.

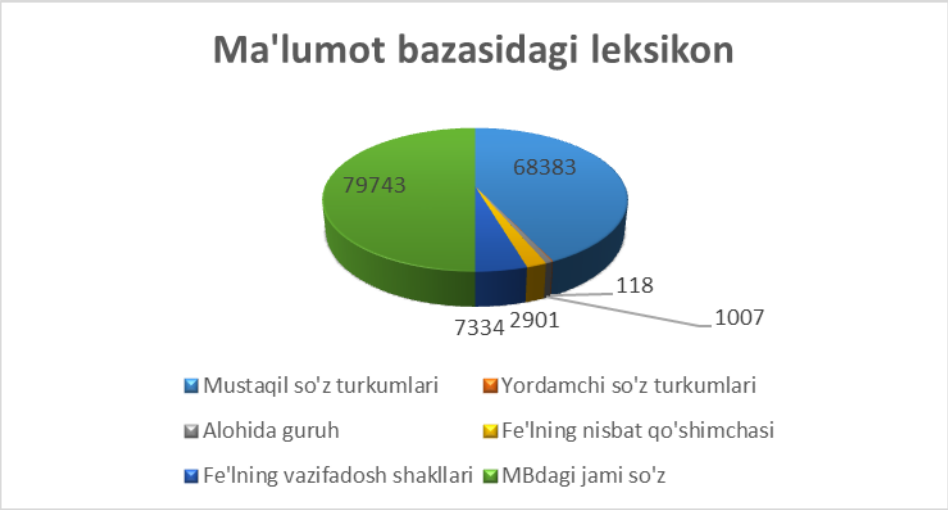
In our investigation for syntactic analysis in this order, two thousand texts taken from the novel “Umar Khayyam” by Nabi Jaloliddin were syntactically annotated and edited by hand in CONLL-U format as a segment unit. The uniqueness of the CONLL-U format chosen in our work is that it is suitable for multilingual systems and has the ability to compare words.

The final stages of tagging are analyzed as *UPOS, XPOS, FEATS.*

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS		MISC
	Uchrashuvda	Uchrashuv	NOUN	OT	Case=Loc|Number=Singular|Number[psor]=Singular	5		_		_
	Innovatsiya	Innovatsiya	NOUN	OT	Case=Loc|Number=Singular|Number[psor]=Singular	6		_		_
	Sohasida	Soha	NOUN	OT	Case=Loc|Number=Singular|Number[psor]=Singular	6		_		_
	Xalqaro	Xalqaro	ADJ	SF	Degree=Pos	4		_		_
	Hamkorlik	hamkorlik	NOUN	OT	Case=Nom|Number=Singular|Number[psor]=Singular	3		_		_

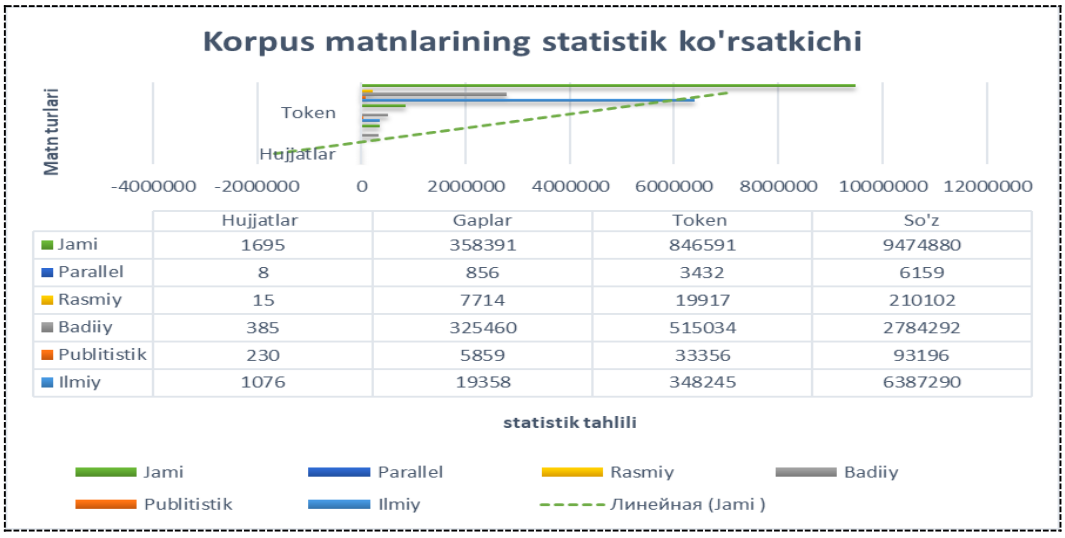
	,	,	PUNCT	TB	_	_	_	_		_
	Xalqaro	Xalqaro	ADJ	SF	Degree=Pos	4		_		_
	Loyihalarni	Loyiha	NOUN	OT	Case=Loc|Number=Singular|Number[psor]=Singular	3		_		_
	Ishlab	Ishla	VERB	F	Verb=V|ADV	3		_		_
	Chiqish	Chiq	VERB	F	Verb=V|INF	3		_		_
	Va	Va	CONJ	B	_	_	_	_		_
	Gumanitar	gumanitar	ADJ	SF	Degree=Pos	3		_		_
	Hamkorlik	hamkorlik	NOUN	OT	Case=Nom|Number=Singular|Number[psor]=Singular	3		_		_
	Masalalari	Masala	NOUN	OT	Case=Nom|Number=Singular|Number[psor]=Plural	2		_		_
	Muhokama	Muhokama	NOUN	OT	Case=Loc|Number=Singular|Number[psor]=Singular	1		_		_
	Qilindi	Qil	VERB	F	Number=Plur|Person=3|Tense=Pres Tense=Past|Voice=Pass	0		_		_
	.	.	PUNCT	TB	_	_	_	_		_

The fourth chapter of the dissertation is entitled *"Linguistic database and Software of Uzbek Corpus Manager". The first paragraph of this chapter is e*ntitled *"Conceptual Foundations for Creating Corpus Project"*. Corpus Manager (corpus browser or corpus search engine) is a system that allows to effectively search for information in the corpus. The corpus manager is a complex model designed to statistically represent different speech forms of a language and the cohesive properties of words in context. This system serves to provide the user with information such as the structure of a whole sentence or the attributes specific to a particular situation, such as lemma, tag, or the concordance and frequency of words. If the server's corpus search engine is the corpus manager, the user-friendly part is called the corpus interface. Corpus managers are different. For example, the



The British National Corpus has web-based interface, Nosketch Engine has combining Manatee (back font) and Bonito (web interface) in free open source corpus management system, and WordSmith Tools software package for specialized linguists.

The second paragraph of the work, entitled "*The Problem of Text Representation"*, states that the representativeness of the corpus is determined by the quantity of the texts and the diversity of the genres. According to V.P.Zakharov and S.Y.Bogdanova[75], it should be noted that to consider to what sources the genre-thematic structure for designing content of the text corpora. According to this, question is the text in small advertisements in newspapers considered as separate text or should it be combined into general text? Is a newspaper article a text or whole newspaper content is one text? Should each poem be a single text or it might be chosen as whole text of collection of poems? Is a one text published letters written in question-answering which discussed general topic or each of letter should be chosen as one text? The author of the corpus should answer these questions based on the corpus typology and linguistic functions. Depending on whether it is a national corpus or a special corpus, a corpus unit should be determined. Scientists also consider chronology as an important aspect of the corpus design process. For example, which lexical layer of language is understood as the modern language for corpus? In this case, it is natural that the chronological boundaries of the corpus differ in different genres. The tool can be used for the general and special linguistic issues.



	Hujjatlar	Gaplar	Token	So'z
Jami	1695	358391	846591	9474880
Parallel	8	856	3432	6159
Rasmiy	15	7714	19917	210102
Badiiy	385	325460	515034	2784292
Publitistik	230	5859	33356	93196
Ilmiy	1076	19358	348245	6387290

The third paragraph, entitled *"Linguistic and software of the Uzbek language corpus"*, deals with the practical creation of a system for managing the linguistic data of the text corpus and its formation according to the needs of the audience. Corpus Manager is a specialized search engine designed to retrieve statistics and provides user-friendly results, frequently based on ready-made solutions. This solves problems related to the speed of search queries (data samples), system flexibility, and scalability.

The formation of grammatical forms on the agglutinativeness of the Uzbek language and the sequence of suffixes makes it necessary to analyze the morpheme structure of words. Therefore, the corpus manager needs a database of Uzbek phrases to determine the basis of words. For corpus analysis, the functional forms

[75] Захаров В.П., Азарова И.В. и др. Моделирование в корпусной лингвистике: Специализированные корпусы русского языка. ИЗДАТЕЛСТВО САНКТ-ПЕТЕРБУРГСКОГО университета, 2019. – С. 16.

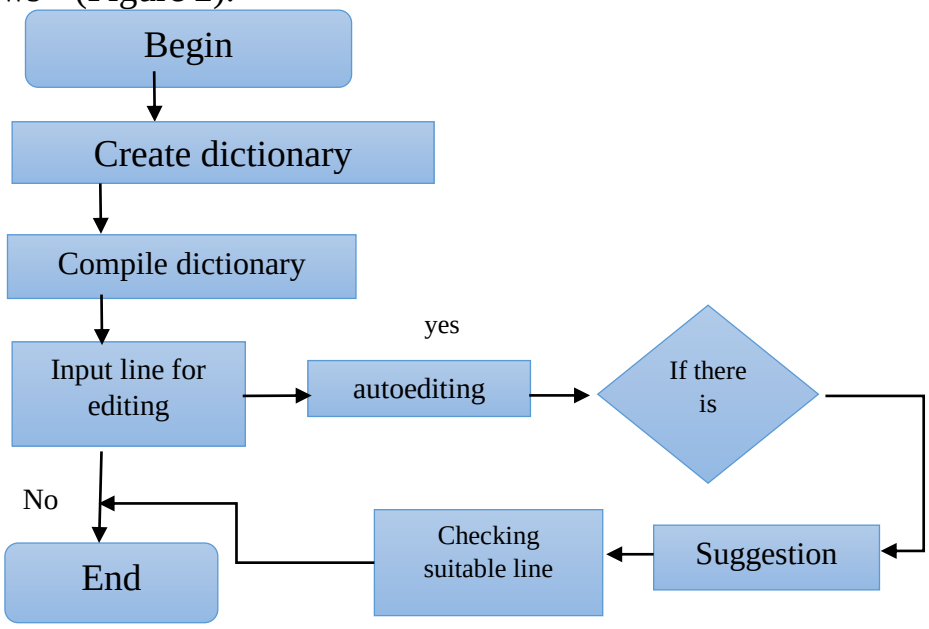
of the verb, in addition to the main word groups of 90,000 words, *are adjectives, nouns, and adverbs; ratios: the database is grouped* by a certain amount of coding, such as addition, subtraction, and sharing. These groups form the basis of general morphotactic models, and when searching for words, information is obtained about each word in the group included in these sets.

The following database for parallel sub-corpus has been created in electronic corpus:

- bilingual translation dictionary: Uzbek-English, English-Uzbek, Russian-Uzbek, Uzbek-Russian;
- database of lexical polysemous word;
- classified syntactically models system;
- translation memory of parallel texts by WordFast technology;
- database of translation concordance;
- translation of terminological combinations, phraseologies and idioms

The fourth paragraph is entitled *"Text Editor Software Interface in Corpus Search Engine,"* reveals an algorithmic solution for correcting spelling errors made by users during a word search.

The block diagram of the editing program algorithm can be expressed as follows[76] (Figure 2):



The text editing program has been made for Turkish, Tatar, Kazakh and Kyrgyz languages. There are some investigations on linguistic database to develop spellchecker for the Uzbek language. The following technologies are available for creating software for editing program:

1. Probability theory-based technology using the Jaro Winkler algorithm;
2. Technology based on pure linguistic database;
3. Technology based on FST (finite state trancducer) algorithm;
4. Levenstein's distance theory.

[76] Bhaire V., Jadhav A., Pashte P., Magdum P.G SPELL CHECKER // International Journal of Scientific and Research Publications, Volume 5, Issue 4, April, 2015. http://www.ijsrp.org/research-paper-0415.php?rp=P403950

In our study, we decided to use *the Jaro Winkler algorithm* to detect word errors. This metric is used to compare two similar lines, and the algorithm performs addition, subtraction, and substitution.

If no results are found for the query, the spelling of the word searched in the corpus will be double-checked. Then we check the length of the line (number of elements) by comparing the unknown line (word) with another line as follows:

$$d_j = \begin{cases} 0 & m\ bo'lganda = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) & qolgan\ holatlarda \end{cases}$$

where:

$|s_i|$ - S_i- lentgh of string;

m – the number of similar characters;

t – half the amount of transpositions.

If they are similar, and $\left[\frac{\max(|s_1|,|s_1|)}{2}\right] - 1$ do not exceed, s_1 and s_2 two elements are taken from the rows. The character in c_1 on each line is compared to the characters in c_2. Similar numbers in different states are divisible by 2, which is the transposition number. For example, if we compare the words **ODAM** and **OMAD**, the length of both lines is *m = 7*. However, the characters *M* and *D* are in different positions. From this *O* and *A* are in the same position. Because the position of *D* and *M* is greater than *1*. Therefore, we can not say that they are similar symbols.

Jaro Winkler's distance is a measure of *p*, which is the unit of measurement for similar lines from the beginning of a line to a length *l* called the prefix. At least two lines are compared in Jaro Winkler: $d_w = d_j + \left(lp(1 - d_j)\right)$ $d_j - s_1$ and s_2 Djaro distance for rows

z	T	A	R	O	Q
T	1	0	0	0	0
A	0	1	0	0	0
Y	0	0	0	0	0
O	0	0	0	1	0
Q	0	0	0	0	1

l –is *a common prefix* from *the beginning of a line* to 4 characters

p– is the measurement factor. It should not exceed 0.25. In the opposite case, the distance between them can be greater than 1. Nevertheless, it should be between *0* and *1*. If the distance is 0, it means that there is no similarity between the two lines, if it is 1, it means that the two lines are the same. In Winkler's work, the coefficients *0* and *1* are calculated as units. We also accepted this value in our work:

Comparing the words **TAROQ** and **TAYOQ** gives the following table.

Comparing the **TARAMOQ** and **QARAMOQ** shown above: m=6, t=0, $|s_1|=|s_2|=7$: $d_j = \frac{1}{3}\left(\frac{6}{7} + \frac{6}{7} + \frac{6}{6}\right) = 0.9047$ l=0, p=0.1:

When Python tests this result in a programming language, the result is as follows:

BAHOR	0.8666666666666667	0.9066666666666667
NAHOR	0.7333333333333334	0.7333333333333334
HABAR	0.7	0.7
KITOB	0.8666666666666667	0.9066666666666667
KILOB	0.7333333333333334	0.7866666666666667

With this algorithm, the letters in a particular word can be dropped, replaced, or input. Accordingly, the letters in those words are checked against a linguistic knowledge base.

CONCLUSION

Based on scientific research on computer models of the Uzbek language electronic corpus, the following conclusions were drawn:

1. Corpus linguistics has the capabilty to observe, analyze, and study almost all expressions of speech. A corpus is a systematic complex of objects and objects studied as a set of texts. The corpus serves as a major digital resource for NLP and computational linguistics, as well as a major resource for solving language technology issues.

2. There is a number of informational, computer, mathematical, biological, numerical, logical, statistical, structural, graphical modeling types. Computer models of the Uzbek language electronic corpus are based on two different approaches: 1) ready-made linguistic equipment and platforms designed for the corpus; 2) formal-functional models created using computer linguistic methods for different purposes based on corpus taxonomy.

3. The electronic corpus, as a product of the periodic evolution of language, serves as the object of study of the laboratory of society. Therefore, the continuous enrichment and renewal of the corpus is based on the criteria of modeling the corpus concept, such as universality, standard, representativeness.

4. Electronic corpora are the result of generalization of modern computer methods and linguistic research in the analysis of linguistic data. In particular, the special corpus of the Uzbek language allows to study the language from different aspect and statistically.

5. Corpus technology as the main object and subject of computational linguistics serves as a model of natural language in a particular situation in the processing of texts. The electronic corpus, which consists of structured and standardized classification texts, plays an important role in further improving language models, grammar rules, and dictionaries. The set of oral and written texts (corpus), which reflects the pragmatic and cognitive representation of language, crucial tool to study of ontological knowledge database, semantic and neural networks and artificial intelligence technology, as well as linguistic models of language and speech acts for the language processor.

6. The creation of special corpora using linguistic tools in scientific research in the field of philology is a quantity method of achieving statistical results.

7. In text analysis, certain types of equipment have an individual user interface and a corpus manager based on the goals and objectives of the corpus users.

8. Instruments such as Wordfast are an effective technological tool in the process of segmentation and linguistic analysis for parallel texts. In determining concordance in parallel texts, it is important to rely on ready-made linguistic equipment to determine the alternative equivalence of a word, phrase, or stable

combination in a given language. For keywords in parallel texts, it is used to create a search engine for parallel texts by loading frequently used units in the context into the interpreter's memory.

9. Linguistic annotation of the corpus (markup) is the main principle of its use. The annotation of the texts is reflected in metadata (for written texts, its structural units, title, author, genre, publication, year; age, gender, profession, nationality of the speakers taken as parameter for oral texts). The entered information is provided in a machine-readable format in a standard language using the Text Encoding Initiative (T.E.I.).

10. Linguistic annotation belongs to lexical, morphological, syntactic, semantic, prosodic (discourse), anaphoric, temporal types. Annotation schemes are used in the annotation process to draw certain conclusions. They are listed in the annotation manual. Text annotation can be done manually or automatically or semi-automatically.

11. It is basic criteria of semantic sides of the annotation besides linguistic annotation mentioned above, it is significant to identify annotation format.The CONLL-U format is chosen as the basis for the syntactic analysis of the Uzbek language.

12. The system of morphological and syntactic tagging of the Uzbek language is based on a hierarchical approach to ontological model by Protégé technology. The grammatical tagging system is also associatively linked to the Turkic, Tatar, Kazakh, and Kyrgyz languages within UniTurk, resulting in a Turkic language tagging system for corpus analysis.

13. The machine fund of the program of morphological analysis of the Uzbek language is created by means of the limited automatic method (FST): 1) "Rules" - alphabet, phonological rules and special phonemes that occur in a phonetic phenomenon (based on both graphemes for the Uzbek language: Cyrillic and Latin; 2) Lexicon (which includes the basic and simple constructions of all word groups).

14. The morphological analysis of the corpus and its reliance on FST technology in the tagging process help to obtain its probability variations of morphemes. Probability models are created using FST, using a large corpus of text and a system of corpora derived from each other, to assist not only spell checker of the words, but also correct them (by offering several options for the user).

15. The text should be of sufficient size, different types of information should be represented in the corpus in its natural context, there should be multiple access to the prepared and created data, the corpus manager should present the results to the users in appropriate form, creating on-demand corpus functionality and displaying statistical data obtained as a result of queries, managing selected and aggregated texts, and directing linguistic data to specific uses ensure the integrity of the corpus.

16. Linguistically annotated corpora serve as objects for computational linguistics (machine translation, speech synthesizer, semantic analysis, spell-checker, etc.). Translation memory, formed using parallel corpora, serves as a linguistic resource for automatic translation of texts.

17. The practical application of the corpus search system (manager) for users and computational linguistics specialists differs from each other. For both subjects, the use of Djaro Winkler algorithm and FST technology in the elimination of text errors serves as a tool to save time and efforts, as well as to create a grammatical fund of Uzbek language.

РАЗОВЫЙ НАУЧНЫЙ СОВЕТ ПРИ НАУЧНОМ СОВЕТЕ DSc.02 / 30.12.2019.FIL.46.03 ПО ПРИСУЖДЕНИЮ УЧЁНОЙ СТЕПЕНИ ПРИ ИНСТИТУТЕ УЗБЕКСКОГО ЯЗЫКА, ЛИТЕРАТУРЫ И ФОЛЬКЛОРА

ТАШКЕНТСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ УЗБЕКСКОГО ЯЗЫКА И ЛИТЕРАТУРЫ ИМЕНИ АЛИШЕРА НАВОИ

АБДУРАХМОНОВА НИЛУФАР ЗАЙНОБИДДИН КИЗИ

КОМПЬЮТЕРНЫЕ МОДЕЛИ ЭЛЕКТРОННОГО КОРПУСА УЗБЕКСКОГО ЯЗЫКА

10.00.11 - Теория языка. Прикладная и компьютерная лингвистика

АВТОРЕФЕРАТ ДИССЕРТАЦИИ ДОКТОРА ФИЛОЛОГИЧЕСКИХ НАУК (DSc)

Ташкент-2021

Тема диссертации доктора наук (DSc) зарегистрирована в Высшей аттестационной комиссии при Кабинете Министров Республики Узбекистан за № B2019.2. DSc/Fil1174.

Диссертация выполнена в Ташкентском государственном университете узбекского языка и литературы имени Алишера Навои.

Автореферат диссертации размещен на трех языках (узбекском, английском, русском (резюме)) на веб-странице Научного совета (www.uztafi@academy.uz) и в информационно-образовательном портале "ZiyoNet" (www.ziyonet.uz).

Научный руководитель: **Дадабоев Хамидилло Арипович**
доктор филологических наук, профессор

Официальные оппоненты: **Мухамедова Саодат Худайбердиевна**
доктор филологических наук, профессор

Уринбаева Дилбар Базаровна
доктор филологических наук, доцент

Элмирова Назира Шодмановна
доктор технических наук, доцент

Ведущая организация: **Ферганский государственный университет**

Защита диссертации состоится на заседании разового научного совета на основе научного совета DSC.02 / 30.12.2019. Fil.46.03 при Институте узбекского языка, литературы и фольклора 12 «24» 202_ г. в 11 часов. (Адрес: 100060, город Ташкент, улица Шахрисабзская, дом 5. Тел.: (871) 233-71-44; факс: (871) 233-71-44; e-mail: uztafi@academy.uz).

С диссертацией можно ознакомиться в Главной библиотеке Академии наук Узбекистана. (зарегистрирована за №51). (Адрес: 700100, город Ташкент, улица Зиёлилар, дом 13. Тел.: (99871) 262-74-58.)

Автореферат диссертации разослан 12 «13» 2021 года

(реестр протокола рассылки № 5 от 13 декабря 2021 г.)

Н.М.Махмудов
Председатель научного совета по присуждению ученых степеней, д.ф.н., профессор

Г.М.Исмоилов
Ученый секретарь научного совета по присуждению ученых степеней к.ф.н., старший научный сотрудник

Д.С.Худойберганова
Председатель научного семинара при научном совете по присуждению ученых степеней, д.ф.н., профессор

ВВЕДЕНИЕ (автореферат докторской диссертации)

Актуальность и необходимость темы исследования. Изучение корпуса, технологии корпуса и корпусной лингвистики в мировой прикладной лингвистике начало развиваться во второй половине двадцатого века. Появление первого корпуса английского языка в 21 веке сыграло важную роль в создании огромных электронных корпусов и для других мировых языков. Корпусная лингвистика – ведущая сфера для преобразования языковых моделей и речевых возможностей естественного языка в компьютерный язык, а также для решения языковых проблем с использованием информационных технологий и методов. Одна из актуальных проблем в информационный век, в то время как уделяется внимание на устойчивое развитие всех языков, на сохранении их национальных особенностей, это дальнейшее совершенствование электронного корпуса языков и создание новых технологий.

Цель исследования: создание концептуальных моделей создания электронного корпуса узбекского языка и разработка эффективных способов использования методов компьютерных технологий для лингвистически аннотированного языкового корпуса**.**

Объект исследования: письменные тексты в различных жанрах официального, научного, художественного и публицистического стиля, а также официальные и научные материалы на английском и русском языках для параллельных корпусов.

Научная новизна исследования заключается в следующем**:**

впервые в узбекском языке посредством *FST* (finite state trunsducer) – конечного автоматического преобразователя с помощью автоматического способа морфологической маркировки разработаны этапы лингвистического анализа процессов лемматизации и токенизации для корпуса узбекского языка, а также сформированы база морфологических сведений, система морфотактических правил;

при синтаксическом тегировании и аннотировании корпуса определены синтаксические модели узбекского языка на примере формата CONLL-U с помощью метода UdPipe универсальной теории подчинения;

в качестве практического результата исследования создан сайт http://uzbekcorpus.uz и по лемме, токен корпусного менеджера узбекского языка и на основе модели *n-gram* разработана поисковая система конкордансов;

при создании алгоритма перевода и памяти переводов параллельного корпуса узбекского языка использована технология WordFast, а также составлена лингвистическая база данных альтернативных единиц перевода узбекского языка на английский и русский языки;

доказаны научные основы метода онтологического моделирования посредством программы Protege на этапе грамматической разметки корпуса;

обосновано что для определения орфографических ошибок слова, которое ищется в корпусе и определения коэффициента близких к нему слов

при приложении в интерфейсе поиска целесообразно использование алгоритма Джаро Винклера.

Внедрение результатов исследований.

Научные и практические результаты по компьютерным моделям электронного корпуса узбекского языка включают принципы лингвистической аннотации, лингвистические модели, автоматизированные методы (технология FST и метод UdPipe), корпус менеджер, программное обеспечение для редактирования поисковых систем, реализованные в следующих аспектах:

сайт http://uzbekcorpus.uz, созданный как практический результат исследования, используется в качестве веб-ресурса при обучении студентов таким предметам, как корпусная лингвистика и морфологический анализатор в Институте вычислительной математики и информационных технологий Казанского федерального университета (Справка Института вычислительной математики и информационных технологий Казанского федерального университета от 2021 года). В результате электронный корпус узбекского языка (http://uzbekcorpus.uz) приложен в списке корпусов тюркских языков (http://www.turklang.net), создана возможность сопоставления структурных и функциональных свойств тюркских языков и использования естественных языковых технологий в качестве лингвистического ресурса;

научные материалы, собранные в корпусе посредством поисковых систем конкордансов на основе модели *n-gram* и леммы, токен корпусного менеджера, который выступает как практическое выражение научных результатов, использованы в качестве научного электронного источника при выполнении фундаментального проекта за номером ОТ-F1-029на тему “Место Согдийского языка и письменности в истории Узбекской государственности (II век до нашей эры – XII век нашей эры)” (2017-2020), при подготовке статьей, монографий, а также параллельные корпусы и переводческие словари, созданные в платформе использованы при публикации практических результатов проекта в зарубежных журналах. В результате, обогащен смысловой контент исследований по проекту, повышена их научная степень;

функциональные возможности электронного корпусного менеджера узбекского такие как: система поиска леммы, лексемы, конкорданса и словосочетания, применены при составлении многотомного «Толкового словаря узбекского языка» на латыни в фундаментальном проекте ОТ-F1-78 «Узбекский язык в эпоху глобализации, его историческое развитие и перспективы (на основе анализа функциональных стилей)», который осуществлен в 2017-2019 гг. Также, фрагменты письменных текстов, собранных в электронном корпусе по пяти стилям использованы как текстовые примеры в учебнике «Методика узбекского языка» (Справка № 3 / 1255-2651 от 16 сентября 2021 г. Академии наук Республики Узбекистан). В результате, расширился смысловой контент исследований, проводимых в рамках проекта, на основе корпуса определялась встречаемость статей в разных контекстах с семантическими полями;

практические результаты диссертации по системе поиска n-грамм модели, конкордансов, леммы, токена корпус-менеджера использованы в проекте за номером JHBL-18 «Изучение семейных ценностей разных поколений» в НИИ «Соседство и семья» (2020-2021 г.) (Справка № 13 НИИ «Махалля и Семья» от 25 июня 2021 г.). В результате он послужил основой для исследователей для изучения пола, возраста, гендерной и научной статистики, национальных особенностей языковых ценностей;

Практические результаты создания электронного корпуса узбекского языка были использованы в проекте JHBL-18 «Создание в Узбекистане веб-портала «Счастливая семья» (2020-2021 гг.). Анализ научного описания и хронологии текстов на тему семьи в различных жанрах, отраженных в корпусе, послужили изучению ценностей и социальных отношений, определении статуса семьи как института (Справка № 12 от 25 июня 2021 г. НИИ «Махалля и семья»). В результате понятия, которые относятся к теме проекта, обогащаются новыми научными фактами;

практические результаты диссертации по системе поиска n-грамм модели конкордансов, леммы, токена корпус-менеджера использованы при обогащении материалами словарный контент практического проекта за номером I-OT-2019-42 на тему «Создание электронного поэтического словаря узбекского и английского языков (Внешний вид, характер людей, образ природы и национальные символы) (2019-2021 гг.)» (Справка № 01 /4-1638 от 18 сентября 2021 года Ташкентского государственного университета узбекского языка и литературы имени Алишера Навои). В результате используя внутренний корпус электронного корпуса узбекского языка – переводная память и переводческие сегментные единицы параллельного корпуса удалось собрать лексические единицы, которые представляют внешний вид и характеры людей, природы и национальных фигур, и определит их альтернативы на английском и узбекском языках;

практические результаты (такие как: в узбекском языке посредством *FST* (finite state trunsducer) – конечного автоматического преобразователя с помощью автоматического способа морфологической маркировки разработаны этапы лингвистического анализа процессов лемматизации и токенизации для корпуса узбекского языка, а также сформированы база морфологических сведений, система морфотактических правил; при синтаксическом тегировании и аннотировании корпуса определены синтаксические модели узбекского языка на примере формата CONLL-U с помощью метода UdPipe универсальной теории подчинения) отражены в учебнике Компьютерная лингвистика для студентов, магистрантов, докторантов, научных сотрудников и специалистов, а также студентов курсов повышения квалификации и переподготовки преподавателей по направлениям 5111200 - узбекский язык и литература, 5120100 - преподавание филологии и языков (узбекский язык), 5120900 - Теория и практика узбекско-английского перевода. (Приказ Министерства высшего и среднего специального образования № 654 от 20 июля 2019 г.). В результате

теоретически и практически обоснована идеологическая основа учебника и обогащена информацией.

Структура и объем диссертации. Диссертация состоит из введения, четырех глав, заключения, списка использованной литературы и объём составляет 220 страниц.

E’LON QILINGAN ISHLAR RO‘YXATI
LIST OF PUBLISHED WORKS
СПИСОК ОПУБЛИКОВАННЫХ РАБОТ
I bo‘lim (I часть, part I)

1. Abduraxmonova N. O‘zbek tili elektron korpusining kompyuter modellari. Monografiya. Globedit, 2021. 200 b.
2. Abdurakhmonova N., Tuliyev U. Morphological analysis by finite state transducer (FST) for Uzbek-English machine translation // Horijiy filologiya. – Samarkand, 2018. –№3. –P. 64-71.
3. Абдурахмонова Н. Замонавий корпусларнинг компьютер моделлари // Ўзбекистонда хорижий тиллар. 2020. –№ 1(30). – Б. 50–58. https://doi.org/ 10.36078/
4. Abduraxmonova N., Haydarov M. O‘zbek tilida WordNet yaratish masalalariga doir // O‘zbekistonda xorijiy tillar (ilmiy metodik elektron jurnal), 2019. –№ 4 (27). –B. 19-28.
5. Abduraxmonova N. O‘zbek tili korpusini yaratishda lingvistik annotatsiyalash tamoyillari // So‘z san’ati, 2021. –№ 1 (4). –B.164-17.
6. Abdurakhmonova N. Linguistic Issues of Creating Parallel Corpora for Uzbek Multilingual Machine Translation System // BuxDU ilmiy axborotnomasi, 2020. –№ 6 (27). –P. 60-68.
7. Абдурахмонова Н., Тешабоев А. Машина таржимасида матн таҳлили ва уни моделлаштириш // Scientific Bulletin. Physical and Mathematical Research. 2019. –№ 1 (2). –P. 101-107.
8. Abduraxmonova N. Hayrullayeva G. Ot so‘z turkumini o‘qitishda korpusga asoslangan ta’lim texnologiyasi // Ilm sarchashmalari, 2019. –№ 4. – B. 112-116.
9. Abduraxmonova N. O‘zbek tili korpusini morfologik teglashda FST texnologiyasi tatbiqi // So‘z san’ati, 2021. –№ 3 (4). –B. 319-325.
10. Abduraxmonova N. O‘zbek tili elektron korpusi uchun matnlar reprezentativligi masalasi // FarDU axborotnomasi, 2021. –№ 4. – B. 169-173.
11. Abduraxmonova N. O‘zbek tili elektron korpusining lingvistik va dasturiy ta’minoti // O‘zMU xabarlari, 2021. –№ 1/5/2. – B. 158-163.
12. Abdurakhmonova N. Formal-Functional Models of The Uzbek Electron Corpus // International Journal of Anglisticum. Literature, Linguistics and Interdisciplinary Studies. – Tetovo, 2021. –№ 8 (10). – P. 59-62.
13. Matlatipov S., Aripov M., Abdurakhmonova N., Modeling WordNet type thesaurus for Uzbek language semantic dictionary // International Journal of Systems Engineering, 2018, –№ 1: 2628. http://www.sciencepublishinggroup.com/ j/ijse
14. Abdurakhmonova N. Syntactic models for parsing of Uzbek corpus // Вестник науки и образования, 2020. –№ 25 (103), – C. 15-19.
15. Abdurakhmonova N. Syntactic parsing based on Uzbek corpus / Proceedings of the Language Technologies for All (LT4All) – Paris: UNESCO Headquarters, 2019. 5-6 December. –P. 164–168.

16. Abdurakhmonova N. Ontological model of Uzbek language (as example morphology) / Computational Models in Language and Speech –TEL-2018 international conference. –Kazan, 2018. – P. 5-10.

17. Abdurakhmonova N. A two-level morphological analysis of the Uzbek corpus / МАТЕРИАЛЫ IV международного научного конгресса иностранная филология. социальная и нацональная вариативность языка и литературы. – Симферополь, 2019. – P. 425-431.

18. Abduraxmonova N. O‘zbek tili korpusi til texnologiyasining lingvistik resursi sifatida / “O‘zbek tilining raqamli lingvistik resursini yaratish istiqbollari” mavzusidagi respublika ilmiy-amaliy konferensiya, 2020. – B. 263-267.

II bo‘lim (II часть, II part)

19. Abduraxmonova N. Kompyuter lingvistikasi (darslik). –Toshkent: Nodirabegim, 2021. –398 b.

20. Sharipbay A. A., Mukanova A. S., Niyazova R.S, Razakhova B.SH., Omarbekova A.S., Abdurakhmonova N.Z. ONTOLOGIES, SEMANTIC TECHNOLOGIES (Electronic textbook) –Nursultan, 2021. – P. 88.

21. Abdurahmonova N. Mashina tarjimasining lingvistik ta’minoti (monografiya). –Toshkent: Muharrir, 2018. –165 b.

22. Abduraxmonova N. Lingvistik protsessor va elektron korpuslar uchun o‘zbek tilining morfologik bazasi mualliflik guvohnomasi 19.02.2021, BGU 00394.

23. Abdurakhmonova N., Menliyev D., Tuliyev U. Uzmorphoanalyzer – o‘zbekcha so‘zlarni morfologik analiz qiluvchi dastur mualliflik guvohnomasi 17.06.2021, DGU 11432

24. Abduraxmonova N., Tuliyev U. Oila, mahalla va gender tengligi mavzusidagi badiiy asarlar elektron korpusining dasturiy ta’minoti mualliflik guvohnomasi 30.03.2021, DGU 10653

25. Abduraxmonova N., Tuliyev U., Maxarov Q. “Oila ma’naviyati” konseptining elektron tezaurusining dasturiy ta’minoti 30.03.2021, DGU 10655.

26. Abdurakhmonova N., Turdiyev A. The use of morphological ontological models for Turkish-Uzbek machine translation // O‘zMU xabarlari, 2019. – № 1 (3), – P.128-132.

27. Abdurakhmonova N. Corpus Based Teaching Uzbek as A Foreign Language //Journal of Foreign Language Teaching and Applied Linguistics, 2019. – № 6. – P.131-137.

28. Alessandro A., Usmanov T., Khamdamov U., Abdurakhmonova N., Mamasaidov M. UZWORDNET: A Lexical-Semantic Database for the Uzbek Language/11th International Global Wordnet Conference (GWC2021) (*Indexing in Scopus*), 2021. – P. 8-19. https://www.aclweb.org/3anthology/2021.gwc-1.2.pdf

29. Khusainov A., Suleymanov D., Gilmullin R., Abdurakhmonova N. Recent Results of “TurkLang-7” Project / XVI международная конференция по компьютерной и когнитивной лингвистике TEL’2020 (12-13 ноября) –Казань, 2020. (*Indexing in Scopus*), Vol.2780, 2020. – P. 90-101 http://ceur-ws.org/Vol-2780/

30. Suleymanov D., Gatiatullin A., Prokopyev N., Abdurakhmonova N. Turkic Morpheme Web Portal as a Platform for Turkology Research / International Conference on Information Science and Communications Technologies ICISCT 2020 (*Indexing in Scopus*) Applications, Trends and Opportunities. –Tashkent, 2020 (4-6 November), https://ieeexplore.ieee.org/ document/9351500.

31. Abdurakhmonova N., Aripov M., Duarte A., Georgouli K., Gomez-Rodriguez C., Kandalina Y., Maia B., Sharipbay A., Tukeyev U., Urazboev G., Vetulani Z., Yusupov O. ERASMUS+ Project Experiences On The Computational Linguistics Master Degree In The Central Asia Universities / 14th annual International Technology, Education and Development Conference –INTED-2020. –Valencia, 2020. (*Indexing Web Of Science*), – P. 4803-4812.

32. Abduraxmonova N., Tuliyev U. O‘zbek tili korpus menedjerining funksional imkoniyatlari / “Ахборот тизимлари ва технологияларининг замонавий жамиятдаги ўрни” мавзусида Республика миқёсидаги илмий-амалий конференция материаллари тўплами.–Namangan, 2021 (30 март). – B. 212-214.

33. Abdurakhmonova N., Isroilov J. Personal names spell-checking – a study related to Uzbek // Iranian Journal of Social Sciences and Humanities Research. UCT. J. Soc. Human. Resear. (UJSSHR). –Iran, 2018. – № 2 (6). – P. 1-6.

34. Abdurakhmonova N. Isroilov J., Matlatipov S. New keyboard or Unicode for Uzbek Latin alphabet in process of the text via computational technology // O‘zbek tili va adabiyoti ta’limi, 2018. – № 6. – P. 6-8.

35. Abdurakhmonova N., Norov A., M. Aripov Composing modification of the Uzbek Phonetic alphabet based on international Phonetic alphabet // International Journal of Advanced Research in Science, Engineering and Technology, 2019. – № 6 (9) – P. 10731-10735.

36. Abdurakhmonova N. Formal-functional models of Uzbek electronic Corpus / “Kompyuter lingvistikasi: muammo va yechimlar” xalqaro ilmiy-amaliy onlayn konferensiya, –Toshkent, 2021. – B. 65-68.

37. Gatiatullin A., Abduraxmonova N. “Turkiy morfema” portali o‘zbek tili elektron korpusi uchun lingvistik annotatsiyalash tizimi sifatida / “O‘zbek milliy va ta’limiy korpuslarini yaratishning nazariy hamda amaliy masalalari” mavzusidagi xalqaro ilmiy-amaliy konferensiya. –Toshkent, 2021. – B.65-68.

38. Abduraxmonova N. O‘zbek tilini dunyoga olib chiqish zaruriyati / o‘zbek tilining raqamli lingvistik resursini yaratish istiqbollari mavzusidagi respublika ilmiy-amaliy konferensiyasi. – Toshkent, 2020. – B.145-148.

39. Abduraxmonova N. Inglizcha-o‘zbekcha tarjima lug‘ati uchun semantik ma’lumotlar bazasi / Translation, information, communication – political and social bridge” nomli xalqaro konferensiya. –Samarqand, 2018 (4-may). – B. 147-148.

40. Абдурахмонова Н. О проблемах создания онтологии узбекского языка / Азербайджанский язык: вчера и сегодня, посвященной 125—летнему юбилею Бекира Чобанзаде международная конференция. – Ozarbayjon, 2019. – C.15-19.

41. Abduraxmonova N. O‘zbek tilining ontologiyasini yaratish masalalari / Til va adabiyot ta’limida zamonaviy axborot va pedagogik texnologiyalar Respublika konferensiya. –Toshkent, 2018. – B.172-174.

42. Abduraxmonova N. Tezaurus yaratishda leksik munosabatlar tavsifi / Til va adabiyot ta’limida zamonaviy axborot va pedagogik texnologiyalar Respublika konferensiya / –Toshkent, 2018. – B.177-180.

43. Abdurakhmonova N. Database of mobile application of specific terms on computational linguistics / Papers presented to the 3th forum of social sciences “The Great Steppe”. –Astana, 2018. –P. 220-224.

44. Abduraxmonova N. Mashina tarjimasi lingvistik ta’minoti uchun o‘zbek tilidagi ko‘makchi fe’llarning semantik maydonlarini berish masalasi / “Oliy ta’limda darsdan tashqari mashg`ulotlar – talabalarning kasbiy komptentligini shakllantirish omili” mavzusidagi respublika ilmiy-amaliy konferensiya. –Navoiy, 2018. – B. 67-68.

45. Abdurakhmonova N., Tuliyev U. Spell checking analysis of Uzbek text using Djaro Winkler algorithm / Turklang.2018 international conference. – Tashkent, 2018. –P. 310-314.

46. Abdurakhmonova N. Ontology of grammar rules as example of noun of Uzbek and Kazakh languages / Turklang.2018 international conference. – Tashkent, 2018 (13-15 September). – P. 37-38.

47. Abdurakhmonova N. Necessity Uzbek corpus for machine translation / Modern linguistics, translatology, linguodidactics international conference. – Tashkent, 2018 (4-8 October) –P. 120-122.

48. Abduraxmonova N. Til texnologiyasini rivojlantirishda o‘zbek tili milliy korpusining innovatsion ahamiyati / “Yoshlar Innovatsion faolligini oshirishning dolzarb vazifalari” mavzusidagi Respublika ilmiy-amaliy konferensiya. –Toshkent, 2019. – B. 159-162.

49. Абдурахмонова Н., Собиров А. Корпус ёрдамида тезаурус яратишнинг концептуал аҳамияти / Translation, information, communication – political and social bridge Proceedings of the international scientific conference. – Самарқанд, 2019. – P. 36-39.

50. Abduraxmonova N., Aripov M., Norov A. Syntactic Structures for Ontological Models (As Example of Uzbek) / Turklang international conference. – UFA, 2020. – P. 94-105.

51. Абдурахмонова Н., Собиров А. Семантическое поле, связанное со словом *маънавият* и его лингвистические свойства на основе корпусного анализа / Papers Presented To The fourth Forum Of Social Sciences “The Great Steppe” By International Turkic Academy. –Nur-Sultan, 2019. –C. 719-727.

52. Abdurakhmonova N., Ahmedova H. Syntactic modelling machine translation between Russian and Uzbek / Turklang. 2019 international conference. –Simforopl, 2019.

53. Абдурахмонова Н. Разработка онтологической модели синтаксических правил узбекского языка / Actual problems of applied mathematics and information technologies (Abstracts on the international scientific conference). –Tashkent, 2019. (14-noyabr) – C. 281-283.

54. Абдурахмонова Н., Хайруллаева Г., Урдишев К. Использование корпуса для обучения синонимии на узбекском языке / Электронная письменность народов Российской Федерации: опыт, проблемы и перспективы Материалы II Международной научной конференции. –Уфа, 2019. –С. 99-104.

55. Abdurakhmonova N. Brief Outlook on Creation of Uzbek Thesaurus by Sketch Engine Technology / “Filologiyaning dolzarb masalalari” mavzusidagi Respublika ilmiy-uslubiy anjumani. – Qo‘qon, 2020. – B. 145-147.

56. Abduraxmonova N., Sadikova M. Korpus lingvistikasida matnlarni lingvistik annotatsiyalash tamoyillari / “Ilm-fan ta’limining rivojlanish istiqbollari” mavzusidagi ilmiy konferensiya. –Toshkent, 2020. – B. 332-336.

57. Abduraxmonova N., Hayrullayeva G. Sifat so‘z turkumini o‘qitishda korpusga asoslangan ta’lim texnologiyasi / Міжнародної науково-практичної інтернет-конференції «Тенденції та перспективи розвитку науки і освіти в умовах глобалізації». – Переяслав – 2020. – С. 26-3.

58. Abduraxmonova N., Babajanov U. Ochiq manbali terminologik ma’lumotlar bazasini yaratishning lingvistik aspekti / “Davlat tili o‘qitishning dolzarb masalalari: muammo va yechimlar” xalqaro ilmiy-amaliy anjuman materiallari to‘plami. –Farg‘ona, 2020. – B. 97-100.

59. Abdurakhmonova N. Zakharov V. Creating multilingual thesaurus as example of the concept of “Empire” in Russian, English and Uzbek based on corpus analysis / “Davlat tili o‘qitishning dolzarb masalalari: muammo va yechimlar” xalqaro ilmiy-amaliy anjuman materiallari to‘plami, –Farg‘ona, 2020. – B. 139-142.

60. Abduraxmonova N., Babajanov U. O‘zbek tili terminlar bazasi lingvistik portalining funksional imkoniyatlari xususida / “Davlat tili – ijtimoiy taraqqiyot va milliy yuksalish mezoni” (O‘zbek tiliga Davlat tili maqomi berilganligining 31 yilligiga bag‘ishlangan respublika ilmiy-amaliy konferensiya materiallari), – Buxoro, 2020. – B. 252-255.

61. Abduraxmonova N., Mengliyev D. O‘zbek tili tezaurusini yaratishda konsept va semantik munosabatlar tasnifi / Международной научно-теоретической Online конференции на тему «Вопросы внедрения в практику научно-инновационных технологий в повышении качества обучения и воспитания» – Нукус, 2021.– С. 320-322.

62. Abduraxmonova N., Xayrullayeva G. O‘zbek tilining ta’limiy korpusi kontentini yaratish xususida / “Ўзбек тилининг рақамли лингвистик ресурсини яратиш истиқболлари” мавзусидаги республика илмий-амалий конференция. – Toshkent, 2020. – B. 281-284.

63. Abduraxmonova N. Gatiatullin A. O‘zbek tilining lingvistik ma’lumotlar bazasini shakllantirishda “Turkiy morfema” portali instrument sifatida / O‘zbek tilini dunyo miqyosida keng targ‘ib qilish bo‘yicha hamkorlik istiqbollari xalqaro ilmiy-amaliy anjuman. –Toshkent, 2020. – B. 125-130.

64. Abdurakhmonova N. Creation Lexical Alignments for Uzbek-Russian Parallel Corpora // Интеллект. Язык. Компьютер. Вып.19. XVI

международная конференция по компьютерной и когнитивной лингвистике TEL'2020. –Казань, 2020. –P. 180-191.

65. Abdurakhmonova N., Aripov M. Синтаксические структуры для онтологических моделей (на примере узбекского языка) / VIII Международной конференции по компьютерной обработке тюркских языков TurkLang’2020. –Уфа, 2020. – P. 95-105.

Avtoreferat «O‘zMU xabarlari» jurnali tahririyatida tahrirdan o‘tkazildi.

Bosishga ruxsat etildi 6.12.2021.
Buyurtma № 66. Adadi 100 nusxa. Hajmi 4,5 b/t.
«Times New Roman» garniturasi. Bichimi 60x84 $^{1}/_{16}$
OOO “AKTIV PRINT” bosmaxonasida chop etildi.
Toshkent, Chilonzor 25, Lutfiy 1A.