QO‘QON DAVLAT PEDAGOGIKA INSTITUTI
HUZURIDAGI ILMIY DARAJALAR BERUVCHI
PhD.03/04.06.2021.FIL.132.01 RAQAMLI ILMIY KENGASH

SAMARQAND DAVLAT UNIVERSITETI

TURSUNOV MUHAMMADSOLIH SA’DIN O‘G‘LI

O‘ZBEK LINGVISTIK KORPUSINI YARATISH TAJRIBASIDAN (KONKORDANS, TOKENAYZER, LEMMATAYZER, RAZMETKALASH DASTURLARI ASOSIDA)

10.00.11 – Til nazariyasi. Amaliy va kompyuter lingvistikasi

FILOLOGIYA fanlari bo‘yicha falsafa doktori (PhD) dissertatsiyasi
AVTOREFERATI

Qo‘qon – 2024

UO‘K: 811.512.133'1:004.45(=512.133)

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

Оглавление автореферата диссертации доктора философии (PhD) по филологическим наукам

Content of Dissertation Abstract of Doctor of Philosophy (PhD) on Philology sciences

**QO‘QON DAVLAT PEDAGOGIKA INSTITUTI
HUZURIDAGI ILMIY DARAJALAR BERUVCHI
PhD.03/04.06.2021.FIL.132.01 RAQAMLI ILMIY KENGASH**

SAMARQAND DAVLAT UNIVERSITETI

TURSUNOV MUHAMMADSOLIH SA’DIN O‘G‘LI

O‘ZBEK LINGVISTIK KORPUSINI YARATISH TAJRIBASIDAN (KONKORDANS, TOKENAYZER, LEMMATAYZER, RAZMETKALASH DASTURLARI ASOSIDA)

10.00.11 – Til nazariyasi. Amaliy va kompyuter lingvistikasi

**FILOLOGIYA fanlari bo‘yicha falsafa doktori (PhD) dissertatsiyasi
AVTOREFERATI**

Qo‘qon – 2024

Falsafa doktori (PhD) dissertatsiyasi mavzusi O‘zbekiston Respublikasi Oliy ta’lim, fan va innovatsiyalar vazirligi huzuridagi Oliy attestasiya komissiyasida B2023.3.PhD/Fil3991 raqam bilan ro‘yxatga olingan.

Dissertatsiya Samarqand davlat universitetida bajarilgan.

Dissertatsiya avtoreferati uch tilda (o‘zbek, rus, ingliz (rezyume)) Qo‘qon davlat pedagogika institutining veb-sahifasida (www.kspi.uz) va «ZiyoNet» axborot-ta’lim portali www.ziyonet.uz manzillariga joylashtirilgan.

Ilmiy rahbar:	**Karimov Suyun Amirovich** filologiya fanlari doktori, professor
Rasmiy opponentlar:	**Abduraxmonova Nilufar Zaynobiddin qizi** filologiya fanlari doktori, professor
	Hamroyeva Shahlo Mirdjonovna filologiya fanlari doktori, dotsent
Yetakchi tashkilot:	**Jizzax davlat pedagogika universiteti**

Dissertatsiya himoyasi Qo‘qon davlat pedagogika instituti huzuridagi ilmiy daraja beruvchi PhD.03/04.06.2021.Fil.132.01 raqamli Ilmiy kengashning 2024-yil “_3_” _avgust_ soat _9^00_ dagi majlisida bo‘lib o‘tadi. (Manzil: 150700, Qo‘qon shahri, Turon ko‘chasi, 23-uy. Tel: (99873) 542-38-38; faks: (99873) 542-11-43; e-mail: quqondpi@umail.uz)

Dissertatsiya bilan Qo‘qon davlat pedagogika instituti Axborot-resurs markazida tanishish mumkin (_28_ - raqam bilan ro‘yxatga olingan). (Manzil: 150700, Qo‘qon shahri, Turon ko‘chasi, 23-uy. Tel: (99890) 508-64-42; e-mail: qdpi_arm@umail.uz)

Dissertatsiya avtoreferati 2024-yil “_9_” _iyul_ kuni tarqatildi.
(2024-yil “_9_” _iyul_ dagi _21_ - raqamli reyestr bayonnomasi)

M.X. Hakimov
Ilmiy daraja beruvchi ilmiy kengash raisi, filol.f.d., professor

A.To‘raxojayeva
Ilmiy daraja beruvchi ilmiy kengash kotibi, filol. f.f.n., dotsent

D.Jamoliddinova
Ilmiy daraja beruvchi ilmiy kengash qoshidagi Ilmiy seminar raisi o‘rinbosari, filol.f.d., professor

KIRISH (falsafa doktori (PhD) dissertatsiyasi annotatsiyasi)

Dissertatsiya mavzusining dolzarbligi va zarurati. Jahon tilshunosligi taraqqiyotida har bir xalq oʻz tilining milliy mavqeyini oshirish, uning asl holatini saqlash, kelajak avlodga butunligicha yetkazish maqsadida milliy til korpuslarini yaratmoqdalar. Zamonaviy tilshunoslikning asosiy yoʻnalishlaridan boʻlgan kompyuter lingvistikasi va korpus lingvistikasi rivojlanib, bu soha doirasida olib borilgan tadqiqotlarda milliy til va uning sofligini saqlashga keng e’tibor qaratilmoqda. Shuningdek, milliy til korpusining dasturiy ta’minotini yaratish, uni internet tarmogʻida ishlashini ta’minlash, milliy tilni kompyuter yordamida lingvistik aspektda tadqiq etish muhim masala hisoblanadi.

Dunyo tilshunosligida tilning axborot texnologiyalari bilan integratsiyasi, uni qayta ishlash, raqamlashtirish, elektron shaklda saqlash va turli tadqiqotlar olib borishda milliy til korpuslaridan samarali foydalanilmoqda. Chunonchi, til korpuslari nafaqat elektron jamlanma, balki til oʻrganuvchilarga, turli lingvistik tadqiqotlarni olib boruvchilarga qoʻllanma hamda leksikografik manba hisoblanadi.

Mamlakatimizda keyingi yillarda olib borilayotgan tilga doir islohatlarda ham oʻzbek tilining sofligini saqlash, uni dunyoga tanitish masalasi koʻtarilmoqda. Binobarin, Oʻzbekiston Respublikasi Prezidentining “Oʻzbek tilining davlat tili sifatidagi nufuzi va mavqeyini tubdan oshirish chora-tadbirlari toʻgʻrisida”gi PF-5850-son Farmonida davlat tilining axborot va kommunikasiya texnologiyalari, xususan, internet jahon axborot tarmogʻida munosib oʻrin egallashini ta’minlash, oʻzbek tilining kompyuter dasturlarini yaratish muhimligi ta’kidlandi[1]. Shuningdek, Oʻzbekiston Respublikasi Prezidentining 2020-yil 20-oktabrdagi “Mamlakatimizda oʻzbek tilini yanada rivojlantirish va til siyosatini takomillashtirish chora-tadbirlari toʻgʻrisida”gi PF-6084-sonli farmoni[2] bilan tasdiqlangan “2020-2030-yillarda oʻzbek tilini rivojlantirish va til siyosatini takomillashtirish konsepsiyasi”da oʻzbek tiliga oid barcha ilmiy, nazariy va amaliy ma’lumotlarni oʻzida jamlagan elektron koʻrinishdagi oʻzbek tili milliy korpusini yaratish vazifasi koʻrsatib oʻtilgan. Zamonaviy axborot texnologiyalari vositalaridan foydalanib, oʻzbek tili milliy korpusini yaratish va uning dasturiy ta’minotini ishlab chiqish ustuvor va dolzarb vazifadir.

Oʻzbekiston Respublikasi Prezidentining 2016-yil 13-maydagi PF-4797-son “Alisher Navoiy nomidagi Toshkent davlat oʻzbek tili va adabiyoti universiteti”ni tashkil etish toʻgʻrisida”, 2017-yil 7-fevraldagi PF-4947-son “Oʻzbekiston Respublikasini yanada rivojlantirish boʻyicha Harakatlar strategiyasi toʻgʻrisida”, 2019-yil 21-oktabrdagi PF-5850-son “Oʻzbek tilining davlat tili sifatidagi nufuzi va mavqeyini tubdan oshirish chora-tadbirlari toʻgʻrisida”, 2020-yil 20-oktabrdagi PF-6084-son “Mamlakatimizda oʻzbek tilini yanada rivojlantirish va til siyosatini takomillashtirish chora-tadbirlari toʻgʻrisida”, 2020-yil 29-oktabrdagi PF-6097-son

[1] Oʻzbekiston Respublikasi Prezidentining “Oʻzbek tilining davlat tili sifatidagi nufuzi va mavqeini tubdan oshirish chora-tadbirlari toʻgʻrisida”gi PF-5850-son Farmoni. // Xalq soʻzi, 2019 yil 22 oktyabr. №218 (7448).

[2] Oʻzbekiston Respublikasi Prezidentining 2020 yil 20 oktyabrdagi “Mamlakatimizda oʻzbek tilini yanada rivojlantirish va til siyosatini takomillashtirish chora-tadbirlari toʻgʻrisida”gi PF-6084-sonli farmoni (https://lex.uz/docs/5058351).

“Ilm-fanni 2030-yilgacha rivojlantirish konsepsiyasini tasdiqlash toʻgʻrisida”, 2022-yil 28-yanvardagi PF-60-son “2022–2026-yillarga moʻljallangan Yangi Oʻzbekistonning taraqqiyot strategiyasi toʻgʻrisida”gi Farmonlari hamda mazkur faoliyatga tegishli boshqa me’yoriy-huquqiy hujjatlarda belgilangan vazifalarni amalga oshirish mexanizmlarini takomillashtirishga ushbu dissertatsiya muayyan darajada xizmat qiladi.

Tadqiqotning respublika fan va texnologiyalari rivojlanishining ustuvor yoʻnalishlariga mosligi. Dissertatsiya respublika fan va texnologiyalari rivojlanishining I. “Axborotlashgan jamiyat va demokratik davlatni ijtimoiy, huquqiy, iqtisodiy, madaniy, ma’naviy-ma’rifiy rivojlantirishda innovatsion gʻoyalar tizimini shakllantirish va ularni amalga oshirish yoʻllari” ustuvor yoʻnalishiga muvofiq bajarilgan.

Muammoning oʻrganilganlik darajasi. Zamonaviy korpuslar – bu ma’lum bir tilda, elektron shakldagi matnlar toʻplamiga asoslangan axborot-ma’lumot tizimidir[3]. Har bir korpus filologik yondashuv bilan matnlar ustida ishlash uchun, albatta, lingvistik apparat va dasturiy ta’minot bilan ta’minlanishi lozim.

Birinchi korpus 1960-yillarda AQSHda Braun universitetida Genrix Kuchera va Nelson Frensis tomonidan yaratilgan[4] – u hozirgi Amerika inglizlarining Braun universiteti standart korpusi deb ataladi. Dastlabki korpuslarni quyidagicha baholash mumkin: boshlangʻich holatdagi hajmi 1 million soʻzdan iborat; unda lemmatizatsiya va razmetkalash yoʻq. Bunday korpuslar na hajmi boʻyicha, na tilni qamrab olish darajasi, na ma’lumotlar bilan ta’minlangan matnlar tarkibi boʻyicha talablarga javob bermaydi.

Bugungi kunda korpuslar lugʻatlar va grammatika kabi tilshunoslikning ajralmas qismiga aylandi. Korpuslarning paydo boʻlishi tilshunoslikning yangi yoʻnalishlaridan biri korpus lingvistikasi rivojini sifat bosqichiga koʻtardi. Dunyo miqyosida tan olingan lingvistik korpuslarga Rus milliy korpusi *(https://ruscorpora.ru/new/)*, Britaniya milliy korpusi *(http://www.natcorp.ox.ac.uk/, https://www.english-corpora.org/bnc/)*, Turk milliy korpusi *(https://www.tnc.org.tr/)*, Amerika milliy korpusi *(http://www.anc.org/)* kabilarni kiritish mumkin.

Matnlarni qayta ishlash, tokenlash va lemmalash, yarim avtomat morfologik razmetkalash, algoritmlar yordamida yaratilgan dasturiy ta’minotlar koʻplab xorijiy olimlar tomonidan tadqiq etilgan. Jumladan, Lourens Entoni, A.E.Polyakov, C.A.Chapelle, A.A.Abroskin, D.V.Sichinova, L.Yu.Sipisina, Oliver Meyson, Kagri Kolekin, V.P.Zaxarov, Yeshim Aksan, A.I.Izotov, Andreas Mengel, R.Garabik, Djavdet Suleymanov, Alfiya Galiyeva[5] va boshqalarning ishlarida korpus lingvistikasi masalalari oʻrganilgan.

[3] Николаев И.С., Митренина О.В., Ландо Т.М., Прикладная и компьютерная лингвистика. Книга – Санкт Петербург: СпбГУ. 2016. – С. 320. – ISBN 978-5-9710-3472-8.

[4] Kholkovskaia O., Role of the Brown Corpus in the History of Corpus Linguistics. Poster–Prague, 2017. – 1-4 p.

[5] Laurence A., AntConc: Design and Development of a Freeware Corpus Analysis Toolkit for the Technical Writing Classroom, IEEE International Professional Communication Conference Proceedings, 2005.– 29-737 p.; Поляков А.Э., Технология подготовки информации в национальном корпусе русского языка. http://www.ruscorpora.ru/ new/corpora-biblio.html.; Chapelle C.A. Computer applications in second language acquisition: Foundations for teaching, testing, and research. Cambridge, England: Cambridge University Press, 2001.; Аброскин А.А., Поиск

O‘zbek tilshunosligida ham ko‘plab olimlar va tadqiqotchilar milliy til korpusi va uning dasturiy taminotlari xususida ilmiy izlanishlar olib borganlar. Tabiiy tilni qayta ishlash, statistik tahlillar hamda korpus lingvistikasi bo‘yicha A.Po‘latov, S.A.Karimov, A.Qarshiyev, S.Matlatipov, U Tukeyev, M.Aripov, B.Mengliyev, N.Abdurahmonova, Sh.Hamroyeva[6] va boshqa ko‘plab tadqiqotchilarimiz[7] o‘zbek tilidagi muammolarni hal etish va ish samarasini oshirish maqsadida tilshunoslikda axborot texnologiyalaridan foydalanishning nazariy va amaliy asoslarini yaratdilar.

по корпусу: проблемы и методы их решения. Национальный корпус русского языка. – Нестор-История, 2009. – С.277–282.; Сичинава Д.В. Параллельные тексты в составе национального корпуса Русского языка: новые языки и новые задачи, национальный корпус русского языка: Исследования и разработки. – Москва, 2019. – С.41-60.; Щипицина Л.Ю. Информационные технологии в лингвистике. Учебное пособие. – Москва: Издательство «ФлИнта», 2013.; Mason O., Developing Software for Corpus Research, International Journal of English Studies, IJES, vol. 8 (1), 2008, pp. 141-156.; Çöltekin Ç., A Freely Available Morphological Analyzer for Turkish, Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010) – 820–827 p.; Victor Zakharov, Corpora of the Russian Language, Conference: 16th International Conference, DOI:10.1007/978-3-642-40585-3, TSD 2013.; Aksan Y., Aksan M. Building a National Corpus of Turkish: Design and Implementation, 2009. – 299-310 p.; Изотов И. Чешский национальный корпус и аналитический императив: опыт корпусного анализа малоупотребительных и маргинальных языковых единиц. Вестник огу №2/февраль, 2007. – С.4-11.; Mengel A., Lezius W., An XML-based representation format for syntactically annotated corpora, In Proceedings of the Second International Conference on Language Resources and Evaluation (LREC’00), Athens, Greece. European Language Resources Association (ELRA), 2000.; Гарабик Р., Словацкий национальный корпус. Труды международной конференции «Корпусная лингвистика–2004». Санкт-Петербург: Издательство С.-Петербургского университета, 2004. – С. 99-121.; Suleymanov D., Nevzorova O., Gatiatullin A., Gilmullin R., Khakimov B., National Corpus of the Tatar Language “Tugan Tel”: Grammatical Annotation and Implementation, Procedia - Social and Behavioral Sciences 95 (2013) 68 – 74 p.; Galieva A., Khakimov P., Gatiatullin A., On the Way to the Relevant Grammatical Tagset for the Tatar National Corpus, EPiC Series in Language and Linguistics, Volume 1, 2016, Pages 121-129.

[6] Polatov A., Muhamedova S., kompyuter lingvistikasi. O‘quv qo‘llanma. – Toshkent, 2009. – 104-s.; Karimov S., Qarshiyev A., Isroilova G. Abdulla Qahhor asarlari tilining lug‘ati. Alfavitli lug‘at. Chastotali lug‘at. Ters lug‘at. – T., 2007; O‘zbek tili korpusining dasturiy ta’minotini yaratishning dastlabki natijalari, Tursunov M.S., Qarshiyev A.B., Karimov S.A., Muhammad al-Xorazmiy avlodlari ilmiy-amaliy va axborot-tahliliy jurnal, 1(15), mart, 2021; Development of a modern corpus of computational linguistics, Tursunov M.S., Qarshiyev A.B., Karimov S.A., Conference: 2020 International Conference on Information Science and Communications Technologies (ICISCT),, DOI: 10.1109/ICISCT50599.2020.9351376, noyabr 2020.; Matlatipov S., Tukeyev U. and Aripov M. Towards the Uzbek Language Endings as a Language Resource. International Conference on Computational Collective Intelligence. Springer, Cham, 2020. pp. 729-740.; Mengliyev B. va b. O‘zbek tilining milliy korpusi // Ma‘rifat. – 26/04/2018; Менглиев Б., Ш. Ҳамроева. Лингвистик таъминот воситаларининг умумий тавсифи ҳамда тил корпусини яратишдаги аҳамияти. O‘zbek tili taraqqiyoti va xalqaro hamkorlik masalalari. –Toshkent, 2019. – P.17-20.; Abduraxmonova N., Elektron korpuslarning kompyuter modellari: Filol. fan. dok. diss. avtoref. – Toshkent, 2021. – 74 b.; Ҳамроева Ш. Ўзбек тили муаллифлик корпуси тузишнинг лингвистик асослари. Филол. фан...(PhD) дисс. –Қарши, 2018. –41 б.

[7] Eshmo‘minov A. O‘zbek tili milliy korpusining sinonim so‘zlar bazasi: Filol. fan. bo‘yicha fal. dok. (PhD) diss. – Qarshi, 2019. – 140 b.; Axmedova D. Atov birliklarini o‘zbek tili korpuslari uchun leksik-semantik teglashning lingvistik asos va modellari: Filol. fan. bo‘yicha fal. dok. (PhD) diss. – Buxoro, 2020. – 156 b.; Xoliyorov O‘. O‘zbek tili ta’limiy korpusini tuzishning lingvistik asoslari. Filol. fan. bo‘yicha fal. dok. (PhD) diss. avtoref. – Termiz, 2021. – 53 b.; Abduraxmonova N. Elektron korpuslarning kompyuter modellari: Filol. fan. dok. diss. avtoref. – Toshkent, 2021. – 74 b.; Abdullayeva O. O‘zbek tilining internet axborot matnlari korpusini shakllantirishning nazariy va amaliy asoslari. Filol. fan. bo‘yicha fal. dok. (PhD) diss. aftoref. – Toshkent, 2022. – 54 b.; Raxmanova A. O‘zbek tili Milliy korpusini yaratishda kompyuter usullari. Filol. fan. bo‘yicha fal. dok. (PhD) diss. – Farg‘ona, 2022. – 52 b.; Abjalova M. Korpus lingvistikasi. [Matn]: uslubiy qo‘llanma / M.A. Abjalova. – Toshkent: Bookmany print, 2022. – 103 b.; Elova D. O‘zbek tili korpusi birliklarining uslubiy teglarini yaratish tamoyillari va lingvistik ta’minot. Filol. fan. bo‘yicha fal. dok. (PhD) diss. avtoref. – Toshkent, 2022. – 65 b.; Xolova M. O‘zbek shevalari korpusini tuzishning lingvistik asoslari (Boysun tumani “j”lovchi shevalari misolida). Filol. fan. bo‘yicha fal. dok. (PhD) diss. avtoref. – Tyermez, 2022; Begmatova G. O‘zbek milliy kopusida idiomalar bazasini yaratish. Filol. fan. bo‘yicha fal. dok. (PhD) diss. avtoref. – Tyermez, 2021; Toirova G. O‘zbek tili milliy korpusini yaratishning nazariy va amaliy masalalari. Filol. fan. dok. (DSc) diss. avtoref. – Buxoro, 2021; Xidirov O. Milliy korpus uchun Parsing dasturi yaratishning lingvistik asoslari. – Jizzax, 2021.

Dissertatsiya tadqiqotining dissertatsiya bajarilgan oliy ta’lim muassasasining ilmiy-tadqiqot ishlari rejalari bilan bog‘liqligi. Tadqiqot Samarqand davlat universitetining ilmiy-tadqiqot ishlari rejalariga muvofiq “O‘zbek tilining milliy korpusini loyihalash va dasturiy majmua ishlab chiqish” (2021-2023) mavzusidagi amaliy loyiha doirasida bajarilgan.

Tadqiqotning maqsadi o‘zbek tili milliy korpusini tuzish va undan foydalanishning veb sahifaga asoslangan (konkordanser, tokenayzer, lemmatayzer, razmetkalash) dasturiy ta’minot ishlab chiqishdan iborat.

Tadqiqotning vazifalari:

mavzu doirasidagi ilmiy-nazariy manbalar asosida jahon, turkiy va o‘zbek tilshunosligida mavjud korpuslarni tahlil qilish, adabiyotlarni sharhlash;

o‘zbek tili milliy korpusining dasturiy ta’minotini ishlab chiqish;

o‘zbek tili matnlarini qayta ishlash, matnlarni tokenlash, lemmalash, razmetkalash modellari va algoritmlarini ishlab chiqish;

korpus dasturiy ta’minotlari orqali matnlarni O‘zbek tili milliy korpusiga kiritish va lingvistik aspektda tahrirlash;

o‘zbek tili milliy korpusini tuzish va undan foydalanishning uchun dasturiy ta’minot yaratish.

Tadqiqotning obyekti sifatida “Alpomish” dostonining Fozil Yo‘ldosh o‘g‘li varianti (Toshkent: Sharq, 2010); Abdunazar Poyonov varianti (Toshkent: Akademnashr, 2018); Mardonaqul Avliyoqul o‘g‘li variantlari (Toshkent: Fan, 2018) tanlangan.

Tadqiqotning predmetini sifatida o‘zbek tili milliy korpusining modellari, algoritmlari va dasturiy ta’minoti tashkil etadi.

Tadqiqotning usullari. Tadqiqot natijalarini ko‘rsatib o‘tishda modellar, algoritmlar, ma’lumotlar bazasi loyihasi, tahlil qilish metodlari va usullaridan foydalanildi.

Tadqiqotning ilmiy yangiligi quyidagilardan iborat:

o‘zbek tili milliy korpusi dasturiy ta’minotini yaratish orqali milliy til korpusining internet tarmog‘ida foydalanuvchiga taqdim etilishiga ko’ra Milliy korpusni yaratish, uni internet tarmog‘ida ishlashini ta’minlash, matnlarni lingvistik jihatdan tasnif etish hamda shu asosda har bir xalqning milliy tili va uning o‘ziga xos xususiyatlari sofligini saqlab qolish mumkinligi dalillangan;

o‘zbek tili milliy korpusini yaratish va boshqarishga mo‘ljallangan *uzbekcorpora.uz* dasturiy ta’minoti yaratilib, u so‘z va iboralarni korpus bo‘ylab qidirish (konkordans), razmetkasini aniqlash (so‘z morfologiyasi), lemmalash, tokenlash va chastotali lug‘atlarni yaratish kabi tarkibiy qismlardan iborat ekanligi asoslangan;

korpusni yaratishga mo‘ljallangan dasturlar va korpusdan foydalanishga xizmat qiladigan dasturlar matnlar bazasini shakllantirish, korpus lug‘atini tuzish va tahrirlash, matnlarni razmetkalash kabi vazifalarni bajarishi lingvistik aspektda dalillangan.

til korpusining dasturiy ta’minotida tanlangan matndagi so‘zlar yoki gaplarning grammatik va stilistik jihatdan xususiyatlarini samarador aniqlash va o‘zgartirish imkoniyati mavjud ekanligi isbotlangan.

Tadqiqotning amaliy natijalari quyidagilardan iborat:

o‘zbek tili milliy korpusini yaratish, mualliflik korpusini yaratish, korpusdan foydalangan holda tadqiqotlar bajarish, ushbu maqsadlarga erishish uchun loyihalashtirish, algoritmlar ishlab chiqish va dasturiy ta’minot yaratishdan iborat.

o‘zbek tili matnlarini qayta ishlash, matnlarni tokenlash, lemmalash, razmetkalash modellari va algoritmlari ishlab chiqilgan;

o‘zbek tili milliy korpusining dasturiy ta’minoti ishlab chiqilib, konkordansda mavjud kontekstlar berilgan shartlar asosida tayyorlanishi aniqlangan;

so‘zlarni morfologik (masalan, so‘z turkumlari va b.), uslubiy (masalan, publisistika, ilmiy, badiiy va b.) va davrlar (masalan, 1990-2000, 2000-2021) bo‘yicha qidirish imkoniyati mavjudligi asoslangan;

korpusning lingvistik natijalari sifatida “Alpomish” dostoni matni, uning til birliklari hamda grammatik tavsiflari dasturiy ta’minotga kiritilgan. “Alpomish” dostoni asosida korpusda lug‘at yaratilgan.

Tadqiqot natijalarining ishonchliligi qo‘yilgan muammoni hal qilishda yaratilgan dasturiy ta’minot va uning algoritmlari to‘liq va aniq ishlab chiqilganligi, ma’lumotlar bazasi loyihasi puxta loyihalanganligi, modellarni aniq ifodalanganligi hamda nazariy va amaliy bilimlarning bir-biriga mosligi bilan izohlanadi.

Tadqiqot natijalarining ilmiy va amaliy ahamiyati. Tadqiqot natijalarining ilmiy ahamiyati shundaki, o‘zbek tili milliy korpusini yaratish va loyihalashtirish mexanizmlari, matnlarni qayta ishlash algoritmlari va milliy korpus dasturiy ta’minotining nazariy asoslaridan o‘zbek tilshunosligining zamonaviy yo‘nalishlari hisoblangan kompyuter lingvistikasi va korpus lingvistikasi bo‘yicha mavjud ilmiy nazariyalarni boyitish va to‘ldirishda foydalanish mumkinligi bilan belgilanadi.

Tadqiqot natijalarining amaliy ahamiyati o‘zbek tili milliy korpusini yaratish va boshqarishga mo‘ljallangan *uzbekcorpora.uz* dasturiy ta’minotidan so‘z va iboralarni korpus bo‘ylab qidirish (konkordans), razmetkasini aniqlash (so‘z morfologiyasi), lemmalash, tokenlash va chastotali lug‘atlarni yaratishda foydalanish mumkinligi bilan izohlanadi. Ishlab chiqilgan dastur onlayn rejimda ishlaydigan bepul platforma hisoblanib, tilni tadqiq qilishda o‘rganuvchiga istalgan joyda, istalgan kompyuterda ishlash imkonini beradi.

Tadqiqot natijalarining joriy qilinishi. O‘zbek lingvistik korpusini yaratish tajribasidan (konkordans, tokenayzer, lemmatayzer, razmetkalash dasturlari asosida) olingan ilmiy natijalar asosida:

o‘zbek tili milliy korpusi dasturiy ta’minotini yaratish orqali milliy til korpusining internet tarmog‘ida foydalanuvchiga taqdim etilishiga ko’ra Milliy korpusni yaratish, uni internet tarmog‘ida ishlashini ta’minlash, matnlarni lingvistik jihatdan tasnif etish hamda shu asosda har bir xalqning milliy tili va uning o‘ziga xos xususiyatlari sofligini saqlab qolish mumkinligi kabi xulosalardan Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universitetida bajarilgan OT-F1-030 raqamli “O‘zbek adabiyoti tarixi” ko‘p jildlik monografiyani (7 jild) chop etish” mavzusidagi fundamental grant loyihasida foydalanilgan (Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universitetining

2023-yil 9-sentyabrdagi 01/4-1716-son ma’lumotnomasi). Natijada matnlarni raqamlashtirish, korpusga kiritish va tadqiqotlarni takomillashtirishga erishilgan;

o‘zbek tili milliy korpusini yaratish va boshqarishga mo‘ljallangan uzbekcorpora.uz dasturiy ta’minoti yaratilib, u so‘z va iboralarni korpus bo‘ylab qidirish (konkordans), razmetkasini aniqlash (so‘z morfologiyasi), lemmalash, tokenlash va chastotali lug‘atlarni yaratish kabi tarkibiy qismlardan iborat ekanligi hamda korpusni yaratishga mo‘ljallangan dasturlar va korpusdan foydalanishga xizmat qiladigan dasturlar matnlar bazasini shakllantirish, korpus lug‘atini tuzish va tahrirlash, matnlarni razmetkalash kabi vazifalarni bajarishi haqidagi ilmiy takliflar “Mahalla va oila” ilmiy-tadqiqot institutining JHBL-20-sonli “Oila, mahalla va gender tengligi mavzusida badiiy asarlarning elektron korpusini yaratish” mavzusidagi amaliy loyihasini bajarishda foydalanilgan (“Mahalla va oila” ilmiy-tadqiqot institutining 2023-yil 4-sentabrdagi 01-09/1608-son ma’lumotnomasi). Natijada uzbekcorpora.uz yordamida badiiy asarlardagi so‘z va iboralarni qidirish, ularning kontekstida qo‘llanilishini aniqlash, konkordanslar tuzish orqali kengaytirilgan kontekstlar hosil qilish, shuningdek, so‘zlarning morfologik razmetkasini aniqlash, lemmalash va tokenlash jarayonlari orqali matnlarning tuzilishi va ularning lingvistik xususiyatlarini o‘rganish, mahalla va gender tengligi mavzusidagi asarlarning lingvistik xususiyatlarini chuqurroq tahlil qilish va ulardan samarali foydalanish imkonini bergan;

til korpusining dasturiy ta’minotida tanlangan matndagi so‘zlar yoki gaplarning grammatik va stilistik jihatdan xususiyatlarini samarador aniqlash va o‘zgartirish imkoniyati mavjud ekanligi isbotlanganligi to’g’risida xulosalardan “Samarqand” viloyat telekanalida efirga uzatilgan “Diyor yangiliklari” kabi ko‘rsatuvining ssenariylarini tayyorlashda foydalanilgan (Samarqand viloyat teleradiokompaniyasining 2023-yil 20-apreldagi 01-07/138-son ma’lumotnomasi). Natijada keltirilgan ma’lumotlar ilmiy dalillar bilan boyitilib, ko‘rsatuvlar mazmuni ma’rifiy va amaliy jihatdan mukammallashuvi ta’minlangan.

Tadqiqot natijalarining aprobatsiyasi. Tadqiqot natijalari bo‘yicha 13 ta xalqaro va 19 ta respublika ilmiy-amaliy anjumanlarida ma’ruzalar qilingan.

Tadqiqot natijalarining e’lon qilinganligi. Dissertatsiya mavzusi bo‘yicha jami 11 ta ilmiy ish chop etilgan, jumladan, 1 ta Scopus bazasida mavjud jurnalda, 5 ta respublika hamda 5 ta xorijiy jurnallarda e’lon qilingan. 5 ta EHM uchun yaratilgan dasturiy vositalar uchun guvohnomalar olingan.

Dissertatsiyaning tuzilishi va hajmi. Dissertatsiya kirish, uch asosiy bob, umumiy xulosa, foydalanilgan adabiyotlar ro‘yxati va ilovalardan tashkil topgan bo‘lib, uning umumiy hajmi 120 sahifadan iborat.

DISSERTATSIYANING ASOSIY MAZMUNI

Dissertatsiyaning **Kirish** qismida o‘tkazilgan tadqiqotning dolzarbligi va zarurati asoslangan, tadqiqotning maqsadi va vazifalari, obyekti va predmeti tavsiflangan, respublika fan va texnologiyalar rivojlanishining ustuvor yo‘nalishlariga mosligi ko‘rsatilgan, tadqiqotning ilmiy yangiligi va amaliy natijalari bayon qilingan, olingan natijalarning ilmiy va amaliy ahamiyati ochib berilgan, nashr etilgan ishlar va dissertatsiya tuzilishi bo‘yicha ma’lumotlar keltirilgan.

Dissertatsiyaning **“O‘zbek tili milliy korpusini loyihalash”** deb nomlangan birinchi bobida korpus tushunchasi va uning turlari yoritilgan. Slavyan tillar va turkiy tillar korpuslari o‘rganilgan. O‘zbek tili milliy korpusining loyihalash, razmetkalangan korpus modellari hamda korpus yaratish bosqichlari bayon qilingan.

Birinchi bobning birinchi fasli **“Korpus tushunchasi va uning turlari”** deb nomlanib, unda korpus lingvistikasi, korpus turlari va o‘ziga xos xususiyatlari o‘rganilgan.

Lingvistik korpus – bu ma’lum bir tamoyillarga asosan jamlangan va umumiy standartlashtirilgan razmetkaga ega bo‘lgan matnlar majmuidir. Dastlab korpus (“Birinchi darajadagi korpus”) deganda biror bir umumiy xususiyatni (til, janr, muallif, davr) birlashtiradigan matnlar to‘plami tushunilgan. Zamonaviy kompyuter texnologiyalari katta hajmli matnlarga lingvistik ishlov berish jarayonlarini bir necha bor soddalashtirish va tezlatishga imkoniyat yaratdi hamda korpus tushunchasining mazmunini o‘zgartirdi. Zamonaviy korpus – bu muayyan bir tilning elektron shakldagi matnlar to‘plamiga asoslangan elektron axborot-ma’lumot tizimidir. Har bir korpus matnlar ustida filologik yondashuv bilan ishlashga xizmat qiladigan lingvistik va dasturiy ta’minotga ega.

Lingvistik korpuslarga namuna sifatida quyidagi eng taniqli va tan olingan korpuslarni keltirish mumkin: Rus tili milliy korpusi (*https://ruscorpora.ru/*), Britaniya milliy korpusi (*http://www.natcorp.ox.ac.uk/*, *https://www.english-corpora.org/bnc/*), Turk tili milliy korpusi (*https://www.tnc.org.tr/*), Amerika milliy korpusi (*http://www.anc.org/)* va boshqalar.

Ma’lumotlarni qayta ishlash va natijalarni tushunarli tarzda taqdim etishga xizmat qiluvchi dasturiy vositalarsiz til korpuslarini yaratib bo‘lmaydi. Ma’lum bir maqsadlarga mo‘ljallangan til korpuslari mavjud (*WordSmith Tools, MonoConc Pro, WordPilot*). Ular tilni keng qamrovda o‘rganishga imkon bermaydi. Milliy korpuslar tilning morfologik, semantik, sintaktik jihatlarini o‘rganishga, til bo‘yicha turli tahlillarni olishga va tilni keng tadqiq qilishga yordam beradi. Korpus funksionalligi tushunarli, sodda va qulay bo‘lgan interfeysga ega bo‘lishi lozim.

Matnlar korpusini tilshunos qo‘lidagi kuchli ish quroli deb aytish mumkin. Lingvistik korpus yordamida tilni tahlil qilish va o‘rganish orqali uning qonuniyatlarini tadqiq etish bilan shug‘ullanadigan fan – bu *korpus lingvistikasidir*. Korpus lingvistikasi tilshunoslikning bir bo‘lagi bo‘lib, kompyuter lingvistikasi va amaliy lingvistikaning yangi bir yo‘nalishini tashkil qiladi. Korpus matnlarning oddiy to‘plamlaridan (“kutubxona”, kolleksiya) farqli ravishda ular *razmetka*langan bo‘lishini talab etadi.

Korpus lingvistikasi ikkita ko‘rinishni o‘z ichiga oladi:

- matnlar korpuslarini yaratish va razmetkalash hamda ular uchun qidiruv vositalarini ishlab chiqish;
- korpuslar asosida tilshunoslik tadqiqotlarini olib borish.

Birinchi bobning ikkinchi fasli **“Mavjud til korpuslari tahlili”** deb nomlanib, unda slavyan va turkiy tillariga oid til korpuslarining dasturiy ta’minotlari, yaratish usullari tahlilga tortilgan.

Slavyan tillarga tegishli korpuslarni loyihalash nuqtayi nazaridan quyidagi til korpuslari oʻrganilgan va tahlil qilingan: **Rus tili milliy korpusi** – bu rus tilidagi matnlarni qidirish mumkin boʻlgan elektron onlayn korpus. U 2004-yil 27-aprelda tashkil etilgan[8]. Milliy korpusda cherkov slavyan, qadimgi rus (XI-XIV asrlar) va markaziy rus (XV-XVIII asr boshlari) matnlarining tarixiy korpusi ham mavjud. **Belarus tili milliy korpusi** – Belarus tilining umumiy kompyuter ma’lumotlar bazasini yaratish zarurati 2001-yilda Milliy akademiyasi Yakub Kolas nomidagi Tilshunoslik institutida boshlangan *“Belarus tilining lingvistik reprezentativligi muammosi va Belarus tilining korpusini qurish tamoyillari”* dasturi davomida paydo boʻldi. **Bolgar tili milliy korpusi –** Bolgar tili institutida prof. L.Andreichin boshchiligida “Kompyuter tilshunosligi va Bolgariya leksikologiyasi va leksikografiyasi” boʻlimi tadqiqotchilari tomonidan Bolgariya milliy korpusi tashkil etilgan. **Slovakiya tili milliy korpusi** asosan turli uslublar, janrlar, mavzular boʻyicha 1955-yillardagi slovakcha matnlarni oʻz ichiga olgan elektron ma’lumotlar bazasidir. Chex tili milliy korpusi Pragadagi Karl universiteti tomonidan yuritiladigan, chex tilidagi elektron shakldagi yozma matnlarning ochiq qidiriladigan ma’lumotlar bazasidir. Sayt chex va ingliz tillarida mavjud. **Polyak tili milliy korpusi** (*http://nkjp.pl/index.php?page=0&lang=1)* toʻrtta muassasaning umumiy tashabbusidir: Polsha Fanlar akademiyasining kompyuter fanlari instituti (koordinator), Polsha Fanlar akademiyasining Polsha tili instituti, Polsha ilmiy nashriyoti va Lodz universitetining Hisoblash va korpus lingvistikasi kafedrasi. Milliy korpus Fan va oliy ta’lim vazirligining ilmiy-tadqiqot loyihasi sifatida amalga oshirilgan.

Turkiy tillar hozirgi va qadimgi turkiy xalqlar va elatlarning tili boʻlib, Sibirdan Bolqon yarim oroligacha bir chiziq boʻylab choʻzilgan ulkan geografik hududda tarqalgan oʻzbek, uygʻur, qozoq, qirgʻiz, qoraqalpoq, saxa (yoqut), tuva, xakas, oltoy, karagas, shor, turkman, ozarbayjon, turk, qoraogʻuz, tatar, boshqird, chuvash, qoʻmiq, noʻgʻay, qorachoy bolqor, tofalar, chuvash kabi 25 dan ortiq til tushuniladi. **Turkiy tillarga** tegishli korpuslarni loyihalash nuqtayi nazaridan quyidagi til korpuslari oʻrganilgan va tahlil qilingan: **Turk tili milliy korpusi** (*https://www.tnc.org.tr/*) zamonaviy turk tili uchun muvozanatli, katta hajmdagi (50 million soʻz) va umumiy maqsadli korpus boʻlishi uchun moʻljallangan. Turk Milliy Korpusi (*TMK*) loyihasi Mersin universiteti tilshunoslari jamoasi tomonidan ishlab chiqilgan va Turkiya Ilmiy va texnologik tadqiqot kengashi tomonidan uch yil muddatga (2008-2011) moliyalashtiriladi[9]. **“Tugan Tel” tatar tili milliy korpusi** zamonaviy adabiy tatar tilining lingvistik manbasidir. Loyiha “2014-2020-yillarda Tatariston Respublikasining davlat tillarini va Tatariston Respublikasida boshqa tillarni saqlash, oʻrganish va rivojlantirish” Davlat dasturi doirasida amalga oshirilmoqda. **Qozoq tili milliy korpusi** (*http://qzcorpus.kz)* ning

[8] https://ruscorpora.ru/new/archive.html.

[9] Yeúim Aksan & Mustafa Aksan, Building a National Corpus of Turkish: Design and Implementation, p:299-309.

hajmi 30 million tokenni tashkil etadi, jumladan, 14 million token hajmli matn materiali, meta-teglar bilan ta’minlangan.

O‘zbek tilshunosligida ham korpus lingvistikasi sohasi jadal rivojlanmoqda. Jumladan Samarqand davlat universiteti va Toshkent axborot texnologiyalari universiteti Samarqand filiali hamkorligida “O‘zbek tili milliy korpusi” loyihasi (http://uzbekcorpora.uz/)[10], Toshkent davlat o‘zbek tili va adabiyoti universitetida “O‘zbek tilining ta’limiy korpusi” loyihasi (https://uzschoolcorpara.uz/)[11], O‘zbekiston milliy universitetida “O‘zbek tili korpusi” loyihalari (https://uzbekcorpus.uz/newIndex)[12] bajarilmoqda.

Mavjud korpuslarni tahlil qilish natijasida ko‘rish mumkinki, korpuslarni qurishda barcha tillar uchun yagona bo‘lgan metodologiya mavjud emas. Buning sababi shundaki, turli tillarda turli xil qoidalar, texnologik jarayonlar amal qiladi. Morfologik tavsiflar to‘plamlari turlicha bo‘lishiga qaramasdan, yuqorida qarab chiqilgan korpuslarning barchasida morfologik (yoki grammatik) razmetka ta’minlangan. Sintaksis razmetka esa ba’zi korpuslarda bor, lekin turlicha yondoshuvlar asosida amalga oshirilgan bo‘lsa, ba’zilarida umuman yo‘q. Korpuslarni tilni qamrab olish darajasi bo‘yicha solishtirganda, rus tili milliy korpusini ham xronologiya, ham janrlararo eng muvozanatlashgan va matnlar xilma-xilligi ta’minlangan korpus deb hisoblash mumkin. Lekin rus tili milliy korpusining bu yutug‘i uning asosiy kamchiligini ham yuzaga keltirgan. Har qanday yirik korpusda to‘liq va aniq razmetkalashning imkoniyati cheklangan bo‘ladi. Noaniqliklar, asosan, avtomatik razmetkalashda omonim so‘zlar tufayli paydo bo‘ladi.

Birinchi bobning uchinchi fasli **“O‘zbek tili milliy korpusini loyihalash: razmetkalangan matn korpusi modeli va bosqichlari”** deb nomlangan bo‘lib, unda korpus yaratish dasturiy ta’minotini loyihalash, korpus tarkibida matnlar bilan ishlash modellarini ishlab chiqish va uni yaratish bosqichlari yoritilgan.

Korpus yaratishda turli xil razmetkaning bo‘lish yoki bo‘lmasligini va agar bo‘lsa, uning aniqlik darajasini hisobga olish maqsadga muvofiq. Filolog mutaxassis har bir matnning razmetkalanishini so‘zma-so‘z sinchiklab tekshirishi, avtomatik razmetka xatolarini tuzatishi va omonim holatlarni bartaraf qilishi kerak. Ishning asosiy qismi qo‘l mehnati orqali bajarilishi mumkin. Shuning uchun filolog mutaxassisning kompyuterdagi ishiga mo‘ljallangan maxsus razmetkalash dasturlari yaratilishi zarur.

Korpus modelining sxematik shakli 1-ilovada tasvirlangan[13].

Korpusni T_i matnlarning jamlanmasi deb qaraymiz. Matn esa p_{ik} tinish belgilari, probel yoki satr oxiri belgilari bilan o‘zaro ajratilgan w_{ij} so‘zlar ketma-ketligidan iborat[14]:

$$Corpus = \{T_1, T_2, \ldots, T_n\},$$

[10] http://uzbekcorpora.uz/, murojaat sanasi: 26.01.2024

[11] https://uzschoolcorpara.uz/, murojaat sanasi: 26.01.2024

[12] https://uzbekcorpus.uz/newIndex, murojaat sanasi: 26.01.2024

[13] Седов А.В., Математические модели, методы и алгоритмы построения размеченных корпусов текстов, Петрозаводск, 2013. –С. 23.

[14] Седов А.В., Математические модели, методы и алгоритмы построения размеченных корпусов текстов, Петрозаводск, 2013. –С. 23.

$$T_i=\{w_{ij}\}U\{p_{ik}\},$$

yerda j- w_{ij} so‘zning i-matndagi o‘rnini, k- p_{ik} tinish belgining o‘rnini bildiruvchi qiymat. Shuningdek, har bir so‘zga uning matndagi joylashuv o‘rni mos qo‘yiladi:

$$Dispostion{:}w \rightarrow positions \in Positions,\ Positions= Pos1,\ Pos2,\ \ldots,\ Posq,$$

bu yerda w co‘z uchun uning matndagi koordinatasi q aniqlanadi. Matndagi boblar, paragraflar, abzaslar, gaplar, gap bo‘laklari – uning qo‘shimcha tuzilmaviy birliklaridir. Model matnning eng kichik tuzilmaviy birliklaridan (so‘zlardan) so‘z birikmalari hosil bo‘lishini inobatga oladi. Gapni so‘zlar to‘plami deb qaralsa, har bir so‘z birikmani qismiy to‘plam deb hisoblash mumkin:

$$SSj= wi1,\ wi2,\ \ldots,\ wik \subset S, \text{ bunda } \bigcup jSSj=S.$$

Har bir so‘z faqat bitta so‘z birikmasi tarkibiga kirishi mumkin, shuning uchun ikkita birikmaning kesishmasi bo‘sh to‘plamdan iborat bo‘ladi:

$$SSk \cap SSj=\emptyset, \text{ bunda } k \neq 1.$$

Har bir so‘z uchun matndagi o‘rnidan tashqari uning morfologik parametrlari to‘plami ham aniqlanishi kerak. Ot so‘z turkumiga tegishli so‘z uchun bu parametrlar – kelishik, son (birlik/ko‘plik), jonli/jonsiz.

Korpusni yaratish jarayoni bir necha bosqichlardan iborat:

1) ma’lumotlarni to‘plash: korpusga kiritiladigan matnlar ro‘yxatini tuzish, ularning manbasini aniqlash;

2) ma’lumotlarni kompyuterlashtirish: tanlangan ma’lumotlarni (elektron matnlar, skanerlangan matnlar, klaviaturadan terilgan matn va h.k.) yuklash;

3) metama’lumotlarni kodlashtirish: tanlangan matn haqidagi metama’lumotlarni yozish;

4) ma’lumotlarni razmetkalash: matnning hamma birliklarini teglash;

5) korpusda qidiruv interfeysini ishlab chiqish: korpusdan ma’lumotlar olishni qulaylashtiradigan va yengillashtiradigan grafik veb-interfeys yaratish;

6) korpusni ishga tushirish: dastlab sinov tariqasidagi versiyani, so‘ngra milliy va xalqaro darajada foydalanishga mo‘ljallangan birinchi versiya ishini yo‘lga qo‘yish.

Dissertatsiyaning ikkinchi bobi **“O‘zbek tili milliy korpusining dasturiy ta’minoti”** deb nomlanadi. Ushbu bobda o‘zbek tili milliy korpusining dasturiy ta’minoti strukturasi va vazifalari, matnlar bazasini shakllantirish, korpus lug‘atini shakllantirish va tahrirlash, korpusga matnlar kiritish va matnlarni razmetkalash dasturlarining ishlash algoritmlari ko‘rib chiqilgan va tavsiflangan. Shuningdek, dasturiy ta’minot uchun internetdan foydalanish hamda dasturiy ta’minotning foydalanuvchi va boshqaruv interfeyslari tavsiflangan.

Ikkinchi bobning **“Dasturiy ta’minot strukturasi va vazifalari”** deb nomlangan birinchi faslida O‘zbek tili milliy korpusi dasturiy ta’minotining strukturasi va vazifalari, matnlar bazasini shakllantirish, korpus lug‘atini shakllantirish va tahrirlash, korpusga matnlar kiritish va matnlarni razmetkalash dasturlari yoritilgan.

O‘zbek tili milliy korpusi modeli va uni yaratish uchun bajariladigan ishlar bosqichlariga bog‘liq holda ikki qismdan iborat dasturiy ta’minot strukturasi ishlab chiqildi (*2-ilova*):

1) korpusni yaratishga mo‘ljallangan dasturlar;

2) korpusdan foydalanishga xizmat qiladigan dasturlar.

Korpusni yaratishga moʻljallangan dasturlar matnlar bazasini shakllantirish, korpus lugʻatini shakllantirish va tahrirlash, matnlarni razmetkalash kabi vazifalarni bajara olishi kerak (*3-ilova*).

Matnlar bazasini shakllantirishda quyidagi ishlar bajarilishi zarur:

- matnlarni raqamlashtirish, ularni tahrirlash;
- matn boʻyicha ma’lumotlarni faylga yozish;
- matnni bazaga kiritish.

Matnlarni raqamlashtirish vositalarini ularning dastlabki holati qaysi manbada (qogʻoz, **.pdf* formatdagi fayl) ekanligiga bogʻliq holda tanlaymiz.

Metarazmetka ma’lumotlari matn boʻyicha umumiy ma’lumotlardan iborat boʻlib, quyidagilarni oʻz ichiga oladi:

- matn nomi;
- matn muallifiga oid ma’lumotlar: ismi va familiyasi, jinsi, tugʻilgan sana va yili va h.k.;
- matn yozilgan vaqt;
- tematika va matn turi;
- janr;
- matn hajmi (soʻzlarda).

Dastur ushbu ma’lumotlarni foydalanuvchi tomonidan kiritilishini ta’minlaydi va ularni *MeteRazm* faylga yozib qoʻyadi. Soʻngra kiritilgan metarazmetka ma’lumotlar asosida maxsus yagona ism shakllantiriladi va matn saqlanib turgan **.docx* fayl shu ism bilan *Matn_Baza* papkaga koʻchirib yoziladi.

Matnni tokenlash. Matnni avtomatik qayta ishlashda, birinchi navbatda, undagi soʻzlarni ajratib olish, yoki boshqacha aytganda, matnni birliklarga boʻlaklash masalasi yuzaga keladi. Buning uchun ajratuvchi belgilarni (probel, tinish belgilari va h.k.) oʻz ichiga olmagan hamma qismiy satrlar matndan ajratilib olinishi lozim. Bu esa tokenlar toʻplami boʻladi[15]. Matnlarni avtomatik qayta ishlashning fundamental algoritmlaridan biri, berilgan matnni tokenlarga boʻlib tashlashdan iborat. Algoritm kirishiga matn berilib, chiqishida matndagi tokenlar roʻyxati olinadi. Bu algoritmni amalga oshiruvchi dasturni tokenayzer deb atashadi. Odatda, tokenlar soʻz shakllari bilan bir xil ma’noni beradi. Lekin leksik birliklarni ifodalash uchun “soʻz” emas, balki “token” termini ishlatiladi. Bunga sabab, ba’zi hollarda token sifatida soʻzdan kichikroq birliklar (alohida morfema) yoki soʻzdan kattaroq birliklar (soʻz birikmalari) ishlatilishi mumkin.

Matndagi soʻzlarning turli roʻyxatlarini tuzish. Bu yerda maqsad – matnning lugʻatga kiritilishi lozim boʻlgan birliklarini (soʻzlar, soʻz shakllari) ajratib olish, ularning roʻyxatini qilish va roʻyxatni foydalanuvchi talabi bilan turli shakllarda (alfavit-chastotali tartibdagi, chastota-alfavit tartibdagi va ters roʻyxat) taqdim etish. *Gram_Lugat* dasturning bu vazifani bajarishga mos moduli *Suz_Rhati* deb nomlanib, uning kirishida matndan ajratilgan tokenlar roʻyxati

[15] Николаев И.С., Митренина О.В., Ландо Т.М. (ред.) Прикладная и компьютерная лингвистика, Ленанд, 2016. - 320 с. - ISBN 978-5-9710-3472-8.

bo‘lsa, chiqishida so‘z va so‘z shakllarining tartiblangan ro‘yxati bo‘ladi. Bu ro‘yxatni shakllantirish uchun dastur grammatik lug‘at fayli *LUGAT*dan foydalanadi. Kirishda berilgan har bir token mavjud *LUGAT* fayldan izlab ko‘riladi. Agar token lug‘atdan topilsa, demak, uni lug‘atga kiritish zaruriyati yo‘q, tashlab yuborish mumkin. Aks holda, ya’ni token lug‘atdan topilmasa, u so‘z yoki so‘z shakli ekanligi tekshiriladi. So‘z yoki so‘z shakli bo‘lgan tokenlar alohida ro‘yxatga kiritiladi. Bu ro‘yxat foydalanuvchi talabiga mos holda, yo alfavit-chastotali tartibda, yo chastota-alfavit tartibda, yo ters ro‘yxat shaklida saralanadi va ekranga chiqariladi.

So‘zlarni grammatik lug‘atga kiritish. Bu vazifani bajaruvchi dasturiy modul *Lug_Kirit* deb nomlangan. Lug‘atga yangi birlik kiritish foydalanuvchi tomonidan interaktiv rejimda amalga oshiriladi. Bu jarayonni *Lug_Kirit* dasturi ta’minlaydi. Kompyuter ekranida *Suz_Rhati* dasturi tomonidan chiqarilgan ro‘yxat mavjud va unda bitta so‘z boshqa rang bilan ajratib ko‘rsatilgan. *Lug_Kirit* dasturi yordamida foydalanuvchi ro‘yxat bo‘yicha harakatlanib, joriy so‘zni tanlaydi va “*Enter*” tugmasini bosadi. Natijada ekranda kiritish oynasi paydo bo‘ladi. Foydalanuvchi bu oynada so‘ralgan ma’lumotlarni klaviaturadan terib kiritadi va “*Bazaga kiritish*” tugmasini bosadi. Shunda tanlangan so‘z va unga mos kiritilgan grammatik xarakteristikalar jamlanib, elektron lug‘at fayli *LUGAT*ga va boshqa yordamchi fayllarga yozib qo‘yiladi. Shundan so‘ng foydalanuvchi ro‘yxatdan navbatdagi so‘zni tanlaydi va bazaga kiritadi. Shu tariqa, berilgan ro‘yxatdagi barcha so‘zlar *LUGAT* faylga kiritilib, gramatik lug‘at qadamma-qadam boyitib boriladi.

Grammatik razmetkalash dasturi *Nuxt JS, Python* va *PostGreSql* ma’lumotlar bazasini boshqarish tizimida (MBBT) tuzilgan. Matnlarni korpusga kiritish jarayonini soddalashtirish uchun grammatik razmetkalash ikki bosqichda bajariladi: matnni dastlabki formatlash bosqichi va razmetkalash bosqichi. Dastlabki formatlash bosqichida matnning tuzilmaviy komponentlari aniqlanadi, ya’ni matnda so‘zlar, gaplar, abzaslar va paragraflar ajratiladi. Shuningdek, so‘zlarning matndagi o‘rni aniqlanadi. Ikkinchi bosqichda grammatik razmetkalash amalga oshiriladi. Razmetkalash natijasida matndagi har bir so‘z shaklga o‘zining morfologik parametrlari va satrli atributlari to‘plami biriktirib chiqiladi. Quyida bu ikkala bosqichning grammatik razmetkalash dasturi orqali bajarilishi batafsil bayon qilinadi.

Dastlabki formatlash bosqichi. Bu bosqichda korpusga kiritiladigan fayl dastlab qanday shaklda bo‘lishidan qat’iy nazar, u *MS Word* dasturi formatiga o‘tkazilishi kerak. *MS Word*ning 2007 va undan yuqori versiyalarini ishlatish zarur, chunki ularda matn **.docx* formatda bo‘lib, standart tizim asosida avtomatik ravishda teglangan holda bo‘ladi. Matnni tuzilmaviy komponentlarga (so‘zlar, gaplar, abzaslar, pargraflar) bo‘lish *MS Word* dasturi tomonidan avtomatik tarzda bajariladi. Matnda so‘zlarning o‘rnini bildiruvchi ma’lumotlar alohida faylga yoziladi.

Razmetkalash bosqichi. Razmetkalash deb matnga va uning komponentlariga maxsus teglarni biriktirib chiqishni tushunish kerak. Maxsus teglar ikki xil bo‘ladi: lingvistik teglar va ekstralingvistik (tashqi) teglar. Lingvistik teglar matn elementlarining leksik, grammatik va shunga o‘xshash boshqa xususiyatlarini tavsiflaydigan ma’lumotlardan iborat. Ekstralingvistik teglar esa muallif haqidagi va matn haqidagi ma’lumotlarni (muallif, nomlanishi, nashr yili va joyi, janr, tematika va hokazo) tavsiflaydi[16].

Dasturiy ta’minot foydalanuvchiga korpus tarkibidagi matnlar ustida tilshunoslikka oid tadqiqotlar olib borish va shu asosda xulosalar chiqarishga imkoniyat yaratishi kerak. Xususan, korpus foydalanuvchisi tomonidan dasturiy ta’minotga quyidagi vazifalar qo‘yiladi:

- konkordans tuzish;
- kontekstlarni nafaqat so‘zlar bo‘yicha, balki so‘z birikmalari bo‘yicha ham izlash;
- ro‘yxatlarni foydalanuvchi tomonidan tanlangan bir necha me’yorlar bo‘yicha tartiblash;
- topilgan so‘z shakllarini kengaytirilgan kontekstda tasvirlash imkoniyatini berish;
- korpusning alohida elementlari bo‘yicha statistik ma’lumotlar berish;
- natijalarni saqlash va chop qilish;
- nafaqat alohida fayllar bilan, balki o‘lchovi chegaralanmagan korpuslar bilan ham ishlay olish;
- so‘rovlarga tez javob qaytarish va natijalarni tez chiqarish.

Umumiy holda, bu vazifalarning bajarilishi tadqiqot uchun zaruriy axborotlarni topish, ularni to‘plash va kerakli tarzda foydalanuvchiga taqdim etish kabi jarayonlardan iborat bo‘ladi. Bu jarayonlarni amalga oshirish uchun korpus bo‘ylab axborot qidiriuv masalasi hal etilishi kerak.

Ikkinchi bobning ikkinchi fasli **“Dasturiy ta’minot interfeysi”** deb nomlanib, unda til korpusi uchun internet tarmog‘ining afzalliklari, ishlab chiqilgan dasturiy ta’minotning yaratilgan korpusdan foydalanish hamda korpus yaratish kabi funksiyalari bayon etilgan.

Internetda ilmiy tadqiqotlar natijalarini taqdim etish foydasiga bir nechta dalillar keltirilishi mumkin[17]. *Birinchidan,* bugungi kunda hech bo‘lmaganda mamlakatimizda ilm-fanning pragmatik darajaga o‘tishi, ya’ni ilmiy faoliyat natijasi jamiyat uchun zarur bo‘lgan haqiqiy ilovalar bo‘lishi kerak. Bunday holatda, internet orqali ilmiy tadqiqotlar taqdimoti ilmiy fikrning mahsulotini har tomonlama namoyish etishga yordam beradi. Turli xil *World Wide Web* xizmatlari orqali fikrlarni potensial mijozlar va manfaatdor tomonlarga yetkazish imkoniyati yaratiladi. Binobarin, internet tadqiqotchining marketing agenti bo‘lib xizmat qiladi hamda shakllantiradi. *Ikkinchidan,* internetda ilmiy ish natijalarini taqdim etish, masalan, veb-sayt shaklida tadqiqotni tizimlashtirish vazifasini bajaradi. Sayt tuzilishini o‘ylab, tadqiqotchi bilmasdan to‘plangan ilmiy materiallarni tasniflaydi

[16] Николаев И.С., Митренина О.В., Ландо Т.М. (ред.) Прикладная и компьютерная лингвистика, Ленанд, 2016. – 320 с. – ISBN 978-5-9710-3472-8.

[17] Green, L. The Internet: An Introduction to New Media/ L.Green. – New York: Berg, 2010; Silver, D. Critical Cyber culture Studies / D.Silver, A.Massanary. – New York: New York University Press, 2006; Браславский, П.И. Методы повышения эффективности поиска научной информации (на материале Internet).: автореф. дис... .канд.тех.наук. 05.13.16 / П. И. Браславский. — Екатеринбург, 2000. — 16 с.

va tizimlashtiradi, uni obyektiv qo‘shimcha ishlov berishga olib keladi. Shu nuqtayi nazardan internet jurnallarda nashrlar, konferensiyalarda ma’ruzalar olimning faoliyatini baholash mezonlari bilan teng huquqli bo‘ladi. Shuningdek, internet vositalari amaliy forumlarni tashkil qilish imkonini beradi. Unda munozaralar natijasida tadqiqotning istiqbolli yo‘nalishlarini ishlab chiqish, uning ta’sirini baholash, ya’ni ilmiy ishni har tomonlama tahlil qilish mumkin bo‘ladi.

O‘zbek tili milliy korpusidan foydalanish uchun internet tarmog‘iga ulanib, brauzer ochiladi hamda http://uzbekcorpora.uz/ yoki http://www.uzbekcorpora.uz/ havolasi orqali dasturga o‘tiladi (*4-ilova*).

Dasturning asosiy oynasi menyu panelidan, dasturning muhim qismiy dasturlariga o‘tish, foydali saytlarga o‘tish hamda o‘zbek tili bo‘yicha yangiliklardan iborat. Dasturiy ta’minotdan foydalanish bo‘yicha kengaytirilgan ma’lumotlar dissertatsiya ishida o‘z aksini topgan.

“Matnlar korpusini yaratish va tadqiq etish” deb nomlangan uchinchi bobda korpus uchun matnlar to‘plamini yig‘ish, raqamlashtirish va formatlash usullari keltirilgan. Korpus yordamida til bo‘yicha tadqiqot bajarish uchun konkordans tuzish va uni tahlil qilish dasturining algoritmlari sharhlangan, optimal metodlar foydalanilgan va dasturiy ta’minotda natijalar olish tavsiflangan. Tadqiqot obyekti bo‘lgan “Alpomish” dostoni matnlarining korpusga kiritish, alfavit va chastotali lug‘atlari tuzilgan hamda statistik tahlillari olingan.

Uchinchi bobning birinchi fasli **“Matnlar to‘plamini yig‘ish, raqamlashtirish va formatlash”** deb nomlanib, ushbu faslda matnlarni to‘plash, raqamlashtirish va uni dastur formatiga tushurish kabi ma’lumotlar berilgan. Korpus yaratish jarayonida matnlarning tanlangan to‘plami qanday shaklda ekanligi muhim ahamiyatga ega. Matnni korpus tarkibiga kiritishda u bir necha bor shakl o‘zgartishlarga duch keladi. Korpus uchun matn tayyorlash quyidagi bosqichlardan iborat:

1) matnni dastlabki razmetkalash – uni minimal *HTML* formatga o‘tkazish;
2) morfologik razmetkalash;
3) matnning metama’lumotlarini kiritish;
4) matnni korpus tarkibiga kiritish.

Bu bosqichlardan o‘tib borish davomida dastlabki matnga biriktirilayotgan ma’lumotlar tobora ortib boradi. Birinchi bosqichda matnga uning tuzilishi haqidagi ma’lumot kiritiladi. Matnning turli xil elementlari (so‘z, abzas, paragraf va h.k) ajratilib, tegishli belgilar qo‘yiladi. Ikkinchi bosqichda esa lingvistik ma’lumotlar kiritiladi. Dissertatsiyaning ikkinchi bobida tavsiflangan dasturiy ta’minot yordamida morfologik razmetkalash amalga oshiriladi. Uchinchi bosqichda matnning “pasporti” hosil qilinadi. U matn bo‘yicha metama’lumotlardan iborat bo‘lib, matndan alohida tayyorlanadi. Oxirgi bosqichda matn korpus tarkibiga kiritiladi va shundan so‘ng uning ustida tegishli dasturlar yordamida tilshunoslik tadqiqotlari o‘tkazilishi mumkin bo‘ladi.

Raqamli va raqamli bo‘lmagan matnlar korpus uchun matn manbalari sifatida ishlatilishi mumkin. Tabiiyki, ikkinchi holatda, qandaydir tarzda matnni kompyuterga kiritish kerak bo‘ladi: uni qayta yozish yoki skanerlash kerak. Misol uchun, “Alpomish” dostonining Fozil Yo‘ldosh o‘g‘li variantining *DOCX* formati bizda mavjud yemas. Shuning uchun bu kitobni raqamlashtirishimiz kerak bo‘ladi. Bunda qo‘lyozma manbani yoki kitobni skanerlash jarayonida *PDF* formatga o‘tkaziladi va undagi yozuvlarni tanib oluvchi konvertor (masalan, *Fine Reader*) dasturlar yordamida Microsoft Office Word dasturida ochadigan *DOCX* formatga o‘tkaziladi. Bunda, albatta, konvertor skaner qilingan elektron kitobni *DOCX* formatga o‘tkazishda kitobning asl holati bo‘yicha o‘tkazishi juda qiyin. Shuning uchun qo‘l mehnati asosida *DOCX* formatdagi matn, asl holatdagi matn bilan bir xillikka keltiriladi. Bunda imloviy buzilishlar, matndagi belgilarni noto‘g‘ri tanib olish kabi xatoliklar tuzatiladi.

Matnni formatlash. Matnlar turli xil PDF, rasmli, dokumentli va boshqa formatlarda bo‘ladi. Korpusga matnlarni kiritishdan avval mavjud matnli fayllarni Microsoft Officening 2010-yil va undan yuqori bo‘lgan versiyasidagi **.docx* formatiga o‘tkazish kerak bo‘ladi. Boshqa formatdagi matnlarni **.docx* formatiga maxsus dasturlar yordamida o‘tkaziladi va **.docx* formatiga o‘tkazish jarayonida matnning asl holati buzilishi mumkin. Bunda matndagi imloviy xatolar qo‘l mehnati yordamida matnning asl holati bilan bir xillikka keltiriladi. Undan so‘ng matnni korpusga yuklash mumkin bo‘ladi. Ushbu tadqiqotda korpusda matnlarni saqlash uchun *JSON* formatdan foydalanilgan (5-ilova).

Uchinchi bobning ikkinchi fasli **“Konkordanslar tuzish va tahlil qilish”** deb yuritilib, unda konkordans to‘g‘risida tushunchalar va ishlab chiqilgan dasturiy ta’minotda konkordansdan foydalanish va tahlillar olish keltirib o‘tilgan. Boshqa korpuslar singari *uzbekcorpora.uz* tizimining asosiy vazifalaridan biri bo‘lgan korpus bo‘ylab qidiruv tizimi amalga oshiriladi. Bunda qidirilayotgan so‘zning xususiyatlari to‘g‘risida ma’lumotlar beradi.

Konkordans – bu matnni o‘rganishning an’anaviy, uzoq vaqtdan beri ma’lum bo‘lgan, ammo hali ham matnni o‘rganishning yetarlicha o‘rganilmagan usuli. U bevosita va kengaytirilgan kontekstdagi so‘zlarning to‘liq indeksini beradi[18].

O‘zbek tili milliy korpusida qidiruv natijalariga kirishni ta’minlash alohida vazifa bo‘lib, sayt yetarlicha ma’lumotga ega bo‘lishi kerak, ammo bir vaqtning o‘zida keraksiz ma’lumotlar ortiqcha yuklanmasligi lozim. Ushbu funksiya tadqiqot uchun murojaat qilgan mutaxassis filologlar uchun mo‘ljallangan. Foydalanuvchiga axborot berishning umumiy tamoyillari[19] o‘rganildi va ular asosida dastlabki qidiruv tizimi loyihasi va dasturiy ta’minoti tayyorlandi. Qidiruv tizimidan foydalanish tushunarli bo‘lishi uchun sahifalar iloji boricha sodda ko‘rinishga keltirilgan. Shu sababli qidiruv sahifasining asosiy qismida ortiqcha ma’lumotlar qo‘yilmagan. Bu ish maydonini vizual ravishda kengaytirish imkonini berdi. Filologlar tomonidan ilgari surilgan talablardan kelib chiqqan holda, yaratilgan axborot resursi foydalanuvchiga ma’lumot olishning bir nechta variantlarini taqdim etishi kerak. Qidiruv natijasida so‘zning kontekstini olish uchun ikkita varaintni taqdim etadi: 1. Qidirilayotgan so‘z ishtirok etgan gapni kontekst sifatida olish. 2. Qidirilayotgan so‘zni old tomonidan va orqa tomonidan

[18] Glushakov S.V., Программирование Wеб-страниц / S.V.Glushakov, I.A.Jakin, T.S.Xachirov. – Xarkov: “Folio”, 2005. – 390 s.

[19] Fedorchuk A., Как создаются wеб-сайты. Краткий курс / A.Fedorchuk – Sankt-Peterburg, 2000. – 224 s.

olinishi kerak bo‘lgan so‘zlar miqdorini ko‘rsatgan holda kontekstni olish. Kontekst – bu tanlangan so‘zni atrofini o‘rab turgan so‘zlar bilan birgalikda olish.

Matndan so‘zni topish: Yuqorida tavsiflangan funksiyalardan birining harakatlari natijasida so‘rov yaratilgan va serverga yuborilgandan so‘ng, kerakli parametrlar orqali so‘zlarni qidirish sodir bo‘ladi. Algoritm natijasida olingan mantiqiy ifoda so‘rovga almashtiriladi va qayta ishlashga yuboriladi. Natijada, har birida so‘z bo‘lgan katakchalar imlolar orqali va matndagi so‘zning o‘rnini o‘z ichiga olgan qatorlar qidirila boshlanadi (matn raqami, bob raqami, paragraf raqami, jumla raqami va so‘z raqami). Berilgan koordinatalarga muvofiq konteksti qidiriladi va ko‘rsatiladi. Asl matnning manzili serverda aniqlanadi. Jadvallardan matn boshiga nisbatan gap boshining ofseti topiladi. Shundan so‘ng, fayl o‘qiladi va keyin ekranda ko‘rsatiladi.

Kontekstning chiqishi: Bizda so‘z uchun aniq identifikator mavjud bo‘lganligi sababli, biz izlayotgan so‘zni matndagi joylashuvida o‘rab turgan oldingi va keyingi so‘zlardan qanchasini olishni ko‘rsatishimiz yoki so‘z ishtirok etgan gapni olishimiz mumkin. Bunda korpusdagi matnli fayllardan topilgan kontekstlar ro‘yxatini foydalanuvchiga chiqarishda, har bir matndan ko‘pi bilan 10 tadan kontekst chiqaradi. Bitta matndagi to‘liq kontekstlar ro‘yxatini ko‘rish uchun ushbu matnning nomi ko‘rsatilgan maydonda kontekstlari soni ham ko‘rsatiladi. Ushbu kontekstlari sonini ustidan sichqonchani chap tugmasi bosilsa, alohida sahifada matnning kontekstlar ro‘yxati to‘liq taqdim etiladi.

So‘zning chiqishi va uning morfologik parametrlari: Yuqorida aytib o‘tganimizdek, biz har bir so‘zni matndan ma’lumotlar bazasidagi o‘xshashiga bog‘lash imkoniyatiga egamiz. Bu noyob identifikator yordamida amalga oshiriladi. Foydalanuvchi sichqonchaning chap tugmasi bilan ma’lum bir so‘zni bosadi, buning natijasida so‘zning noyob identifikatorini o‘qiydigan va so‘zning o‘zi hamda uning o‘ziga xos xususiyatlari ko‘rsatiladigan oynani ochadigan protsedura chaqiriladi. Buning uchun ma’lumotlarni yig‘ish bloki va parametrlarni dekodlash bloki ishga tushiriladi.

Kengaytirilgan kontekst chiqishi: Topilgan kontekstlarni ko‘rsatgandan so‘ng foydalanuvchi kengaytirilgan kontekstga o‘tishni taklif qiladi. Kengaytirilgan kontekst sifatida kontekst ishtirok etgan paragrafni chiqarib ko‘rsatadi. Kengaytirilgan kontekstda ham har bir so‘zning parametrlarini ko‘rish imkoniyati mavjud.

Uchinchi bobning uchinchi fasli **“Alpomish dostoni matnlarining alfavit va chastotali lug‘atlari hamda statistik tahlillari”** deb nomlanib, “Alpomish” dostonining matni lug‘atini yaratish uchun uch variant tanlab olindi hamda alfavit, chastotali lug‘atlari tuzilgan va statistik tahlillari olingan. Birinchisi doston variantlaridan eng mukammali bo‘lgan 2010-yilda “Sharq” nashriyoti tomonidan chop etilgan Fozil Yo‘ldosh o‘g‘li varianti, Abdunazar Poyonov varianti, 2018-yilda O‘zbekiston Respublikasi Fanlar Akademiyasi “Fan” nashriyoti tomonidan chop etilgan Mardonaqul Avliyoqul variantlari material vazifasini o‘tadi. Bu material adib ijodiy tafakkuri, badiiy mahorati va tildan foydalanishdagi o‘ziga xos uslubi haqida birmuncha to‘la va asosli xulosalar chiqarishga imkon beradi.

Ikkinchi masala so‘zlikni tuzish usulini belgilash hisoblanadi. Jahon leksikografiyasida lug‘at tuzishning tematik, kombinator, g‘oyaviy, alfavitli kabi usullari orasida alfavitli so‘zlik tuzish usuli ma’qullangan va yozuvchilik leksikografiyasida shu usuldan ko‘proq foydalanilmoqda. Bu o‘rinda ham ana shu an’anaga sodiq qolindi. Ayni paytda, u chastotali va ters lug‘atlar hisobiga boyitildi.

Lug‘atda so‘zlar matnda qanday holatda uchragan bo‘lsa, shu holatda keltirildi va kompyuter yordamida ularni statistik jihatdan ishlashning imkoni paydo bo‘ldi. “Alpomish” dostonining uchta variantining statistik tahlil natijalari olindi (1-jadval).

1. “Alpomish” dostonining Fozil Yo‘ldosh o‘g‘li variantining elektron lug‘ati olindi. Doston MS Word dasturida tayyorlangan 623 Kb hajmli *.*txt* formatdagi fayldan iborat. Dostonning ushbu varianti dastur yordamida qayta ishlangan so‘ng unda jami 14413 ta so‘z ishlatilgan va ularning qo‘llanilishi 82106 so‘zni tashkil qilgan. Elektron lug‘atda 82106 ta birlik hosil bo‘ldi.

2. “Alpomish” dostonining Abdunazar Poyonov variantining elektron lug‘ati olindi. Doston MS Word dasturida tayyorlangan 631 Kb hajmli *.*txt* formatdagi fayldan iborat. Dostonning ushbu varianti dastur yordamida qayta ishlangan so‘ng unda jami 19098 ta so‘z ishlatilgan va ularning qo‘llanilishi 78632 so‘zni tashkil qilgan. Elektron lug‘atda 78632 ta birlik hosil bo‘ldi.

3. “Alpomish” dostonining Mardonaqul Avliyoqul variantining elektron lug‘ati olindi. Doston MS Word dasturida tayyorlangan 488 Kb hajmli *.*txt* formatdagi fayldan iborat. Dostonning ushbu varianti dastur yordamida qayta ishlangandan so‘ng unda jami 13520 ta so‘z ishlatilgan va ularning qo‘llanilishi 62100 so‘zni tashkil qilgan. Elektron lug‘atda 62100 ta birlik hosil bo‘ldi.

“Alpomish” dostonining Fozil Yo‘ldosh o‘g‘li variantini asoslar va turkumlar bo‘yicha statistik tahlil natijalari olindi. Bunda lug‘atdagi so‘zlarning turkumlarini biriktirib chiqishda 12 turdagi so‘z turkumlaridan foydalanildi: fe’l, ot, sifat, son, olmosh, ravish, bog‘lovchi, ko‘makchi, yuklama, modal so‘z, undov so‘z va taqlid so‘z. Fozil Yo‘ldosh o‘g‘li variantida 14413 ta so‘z shakllari bir-birini takrorlamay kelgan. Ushbu lug‘atdagi so‘z shakllaridan asoslar lug‘ati tuzilganda 3126 ta asos so‘zdan iborat lug‘at hosil bo‘ldi (*1-ilova*).

Kompyuter yordamida ishlash jarayonida quyidagi kamchiliklar ko‘zga tashlandi:

– lug‘atda ayrim so‘zlar, masalan, obrazlarning nomlari qo‘shtirnoq ichida keltirilgan. Kompyuter qo‘shtirnoq ichidagi har bir so‘zni alohida o‘qimoqda. Qo‘shtirnoqdan keyin ayrim qo‘shimchalar bor bo‘lsa, masalan, *“qo‘l ushlatar”ini, “Qultoy”dan* yoki *“Qultoy bobo”ning* deb yozilgan bo‘lsa, uning tarkibidagi -*ini* va -*ning* ham alohida so‘z tarzida qayd etmoqda. *“Qultoy”dan* deb yozilgan so‘zdagi qo‘shtirnoqni hisobga olmasdan *Qultoydan* tarzida keltirmoqda. Matndan Qultoydan so‘zi izlanganda esa kompyuter uni topib bera olmayapti. Chunki qo‘shtirnoq bunga xalaqit bermoqda. Matnda *Qultoyni, Qultoyda, Qultoyning, Qultoyga, Qultoydan* kabi so‘zlarining barchasida *Qultoy* qo‘shtirnoqqa olingan. Ularning lug‘atda berilishida qo‘shtirnoqdan holi qoldirilgan;

– matnda ba’zi so‘zlarning oxirgi yoki o‘rtasida harflar yoki qo‘shimchalar to‘rtburchak qavsga olingan. Masalan, *rostdi[r], dash[t]i, Qo‘ng‘irotni[ng].* Bunda so‘zni lug‘atga kiritishda ikki xil variantini olinadi. Misol uchun, *rostdi[r]* so‘zi lug‘atga ham rostdi, ham rostdir tarzida kiritildi;

– ikki so‘z birikuvining qo‘shma so‘z ekanligi, faqat, inson fikri orqaligina anglashiladi. Shuning uchun hozircha qo‘shma so‘zlarning lug‘atda aks etishi muammoligicha qolmoqda;

– ayrim so‘zlar, masalan, *Ultontoz* so‘zi personaj tilida aytilishiga qarab Ulton, Ultonshoh shaklida berilgan. Ularni ham ikkita so‘z sifatida qayd etishga to‘g‘ri keladi. Shunday bo‘lmasa, lug‘atning «kompyuter prinsipi» buzilib ketadi;

– *Yig‘in-tudani, damba-dam, ort-sirtidan, qilmading-ku* kabilar matnda chiziqcha bilan berilgan va bu qo‘shilmalarni ham kompyuter lug‘atda juft so‘zlar sifatida qayd yetgan;

– matndagi qisqartma so‘zlar ham lug‘atga kiritilgan;

– ba’zi so‘zlar matnda ikki xil berilgan: *boryapti – borayapti, bo‘lmasmikan – bo‘lmasmikin, ko‘targuli – ko‘targulik* ko‘rinishida ikki xil yozilgan. Kompyuter ularni alohida-alohida lug‘at birliklari sifatida qayd etgan. Bu o‘rinda lug‘atga qo‘shimcha ishlov berishga yoki ularni shu holicha qoldirib, ikki holatni ikki so‘z sifatida e’tirof etishga to‘g‘ri keladi. Biz ularni o‘z holicha qoldirishni maqsadga muvofiq deb hisobladik;

– kompyuter matndagi raqamlarni qayd etmagan. Masalan, matnda *8 ta, 10 ta* yoki *8 tasi, 10 tasi* deb yozilgan bo‘lsa, *-ta* yoki *-tasi* ni alohida so‘z sifatida ko‘rsatgan;

– qandaydir bir harf boshqa shriftda yozilib qolgan bo‘lsa-yu, umumiy matndagi shriftlarga mos kelmasa ham kompyuter so‘zni o‘qimaydi yoki lug‘atda uni alohida qayd etadi.

XULOSA

1. Jadal suratlar bilan rivojlanib borayotgan ijtimoiy hayotda zamonaviy tilshunoslikning matn lingvistikasi, kompyuter lingvistikasi va korpus lingvistikasi kabi yo‘nalishlari o‘zaro uyg‘unlikda ish ko‘rib, milliy til korpuslari yaratilmoqda. Milliy til korpusi dasturiy ta’minotini yaratish, uni Internet tarmog‘ida ishlashini ta’minlash, kompyuter yordamida matnlarni lingvistik tadqiq etish hamda shu asosida har bir xalqning milliy tili va uning o‘ziga xos jihatlarini, lingvistik xususiyatlarini o‘rganish, ularning asl holatini saqlash, kelajak avlodga sofligicha yetkazish masalalariga keng e’tibor qaratilmoqda.

2. Adabiyotlarni va mavjud korpuslarni o‘rganish asosida dasturiy ta’minotning ikki qismi – korpusni yaratishga mo‘ljallangan dasturlar va korpusdan foydalanishga xizmat qiladigan dasturlar orqali matnlar bazasini shakllantirish, korpus lug‘atini tuzish va tahrirlash, matnlarni razmetkalash mumkinligi belgilandi.

3. Mavjud korpuslar tahlili asosida quyidagi xulosalar olindi: 1. Korpuslarni yaratishda barcha tillar uchun yagona bo‘lgan metodologiya mavjud emas; 2. Morfologik xarakteristikalar to‘plamlari turlicha bo‘lishiga qaramay, milliy

korpuslarning barchasida morfologik (yoki grammatik) razmetka ta’minlangan; 3. Har qanday yirik korpusda to‘liq va aniq razmetkalashning imkoniyati cheklangan bo‘lib, noaniqliklar, asosan, avtomatik razmetkalashda omonim so‘zlar tufayli paydo bo‘ladi; 4. Milliy korpus tilning qonuniyat va xususiyatlarini iloji boricha ko‘p taqdim eta olishi va tarkibi bo‘yicha muvozanatlashgan bo‘lishi kerak; 5. Matnlarning bir turi tildagi jami matnlar ichida qanday hajmiy ulushga ega bo‘lsa, ana shunday miqdor milliy korpusda ham saqlanib qolishi kerak.

4. Korpus yaratishda turli xil razmetkaning bo‘lish yoki bo‘lmasligi va agar bo‘lsa, uning aniqlik darajasini hisobga olish maqsadga muvofiqdir. Korpus dasturiy ta’minoti tarkibida razmetkalash dasturlari bo‘lib, ular quyidagi sifatlarga ega bo‘ladi: 1. Foydalanuvchi so‘zlar yoki gaplarning grammatik yoki sintaksis xarakteristikalarini yetarli darajada to‘liq aniqlash va o‘zgartirish imkoniyatiga ega; 2. Tahlil maksimal darajada qulay va sodda bo‘lib, tez bajariladi.

5. Korpus bilan ishlashda, nafaqat, lokal tarzda foydalanish, balki internet orqali foydalanish imkonini beradigan veb-resurslar va *uzbekcorpora.uz* veb-dasturi ishlab chiqildi. Bu esa bir vaqtning o‘zida bir necha foydalanuvchining istalgan masofadan turib, korpusni boyitishi yoki korpusdan foydalanishi uchun imkoniyat yaratdi.

6. So‘zni fayllardan qidirish algoritmi ishlab chiqildi. Qidiruv tizimi yordamida so‘z qidirilganda, dastavval, so‘zni korpus lug‘atidan izlaydi, agar so‘z mavjud bo‘lsa, so‘zning ID si konteyner jadvaldan topiladi.

7. Matnlarni korpusga kiritish uchun, uni raqamlashtirish va korpus talabiga ko‘ra formatlashtirish lozim bo‘ladi. Raqamlashtirish jarayonida ko‘proq matnni lingvistik xatolarini tuzatishga e’tibor berish lozim. Bu esa foydalanuvchiga lingvistik ma’lumotlarni yuqori aniqlikda olish imkonini beradi.

8. Korpus tarkibidagi matnlarni va ularning lingvistik ma’lumotlarini saqlash uchun *XML,* formatga ko‘ra *JSON* formatdan foydalanish afzalroq ekanligi aniqlandi. *JSON* formatning *XML* formatdan afzallik tomonlari: 1. *JSON* yakuniy tegdan foydalanmaydi; 2. *JSON* qisqaroq; 3. *JSON* tezroq o‘qish va yozish; 4. *JSON* massivlardan foydalanishi mumkinligi bilan xarakterlanadi.

9. Korpus qidiruv tizimida konkordanslar tuzish qoidalari o‘rganildi. Qidirilayotgan so‘z bo‘yicha butun korpus yoki ma’lum matn bo‘yicha konkordanslar tuzish imkoniyati yaratildi. Qidirilayotgan so‘zning kontekstlarini foydalanuvchiga turli ko‘rinishlarda taqdim etish usullari ishlab chiqildi.

10. Matnlarning alfavit va chastotali lug‘atlari tuzildi hamda statistik tahlillari olindi. Matnning statistik lug‘atlarini yaratish dasturiy ta’minotida lug‘atni to‘liq va aniq tuzishda yo‘l qo‘yiladigan ba’zi kamchiliklar dalillar asosida ko‘rsatilib, tavsiyalar berildi.

11. Kompyuter yordamida matnning lingvistik jihatdan tahrirlash filologiya yo‘nalishidagi tadqiqotlar uchun amaliy ahamiyatga ega bo‘lib, u matnda ishlatilgan leksik birliklar miqdori va grammatik belgilarini hamda matn tili va uslubiga xos xususiyatlarni aniqlash imkonini yaratadi. Bu esa zamonaviy tilshunoslikning korpus lingvistkasi va kompyuter lingvistikasi sohalari rivojini ta’minlab, bu sohalarning sifat bosqichiga ko‘tarilishida amaliy ahamiyat kasb etadi.

SCIENTIFIC COUNCIL AWARDING
SCIENTIFIC DEGREE PhD.03/04.06.2021.FIL.132.01
AT KOKAND STATE PEDAGOGICAL INSTITUTE

SAMARKAND STATE UNIVERSITY

TURSUNOV MUKHAMMADSOLIKH SADIN OGLI

FROM THE EXPERIENCE OF CREATING THE UZBEK LINGUISTIC CORPUS (BASED ON THE PROGRAMS OF CONCORDANCE, TOKENIZER, LEMMATIZER, MARKUP)

10.00.11 – Language theory, applied and computer linguistics

ABSTRACT
of the doctor of philosophy (PhD) on philological sciences

Kokand – 2024

The theme of PhD dissertation is registered by Supreme Attestation Commission at the Cabinet of Ministry of the Republic of Uzbekistan under the number № B2023.3.PhD/Fil3991.

The dissertation has been prepared in Samarkand State University

The abstract of the PhD dissertation is posted in three (Uzbek, English and Russian (Resume)) languages on the website of the Scientific Council (www.kspi.uz) and on the website «ZiyoNet» Information-educational portal (www.ziyonet.uz)

Scientific supervisor:	**Karimov Suyun Amirovich** Doctor of Philological sciences, Professor
Official opponents:	**Abdurakhmanova Nilufar Zainabuddin kizi** Doctor of Philological sciences, Professor
	Khamroeva Shahlo Mirdjonovna Doctor of Philological sciences, associate professor
Leading organization:	**Jizzakh State Pedagogical University**

Defense of the Dissertation will take place on «3» August 2024, at 9^{00} p.m. at a meeting of Scientific Council PhD.03/04.06.2021.Fil.132.01 under Kokand State Pedagogical Institute. Address: 150700, Kokand city, Turon street, 23. Tel: (99873) 542-38-38; fax: (99873) 542-11-43; e-mail: quqondpi@umail.uz).

Dissertation could be reviewed at information-resource center of Kokand State Pedagogical Institute (registration number 28). Address: 150700, Kokand city, Turon str., 23. Tel: (99890) 508-64-42; e-mail: qdpi_arm@umail.uz)

Dissertation abstract sent out on «9» July 2024.
(Mailing report number 21 on «9» July 2024).

O'ZBEKISTON RESPUBLIKASI OLIY TA'LIM, FAN VA INNOVATSIYALAR VAZIRLIGI MUQIMIY NOMIDAGI QO'QON DAVLAT PEDAGOGIKA INSTITUTI

M.Kh. Khakimov
Chairman of Scientific Council Awarding scientific degree, Doctor of Philological sciences, Professor

A. Turakhojaeva
Secretary of Scientific Council Awarding scientific degree, doctor of philosophy, associate professor

D.Jamoliddinova
Deputy chairman of Scientific Seminar at the Scientific Council awarding scientific degree, Doctor of Philological sciences, Professor

INTRODUCTION (Doctor of Philosophy (PhD) Dissertation Annotatic

Relevance and necessity of the dissertation theme. In the developme the world linguistics, each nation has been creating national language corpora in order to increase the national status of its language, preserve its original state, and pass it on to the next generation in its entirety. Being the main directions of modern linguistics, computational linguistics and corpus linguistics are developing, and in the studies conducted in this field, much attention is paid to preserving the national language and its purity. Moreover, creating the software of the national language corpus, ensuring its operation on the Internet, researching the national language in the linguistic aspect with the help of a computer are of great importance.

In the world linguistics, national language corpora are effectively used in the integration of language with information technologies, its processing, digitization, electronic storage and conducting various investigations. As the language corpora are not only an electronic collection, but also a guide and a lexicographic resource for language learners and scholars conducting various linguistic studies.

The issue of preserving the purity of the Uzbek language and introducing it to the world is also being raised in the language reforms carried out in our country in recent years. Consequently, in the Decree PD-5850 of the President of the Republic of Uzbekistan «On measures to fundamentally increase the prestige and status of the Uzbek language as a state language», the importance of taking the prestigious place in the sphere of information and communication technologies, in particular, in Internet global information network, along with creating Uzbek language computer programs was emphasized[20]. Furthermore, approved by the Decree PD-6084[21] of the President of the Republic of Uzbekistan dated October 20, 2020 «On measures to further develop the Uzbek language and improve language policy in our country» the task of creating a national corpus of the Uzbek language, which includes all scientific, theoretical and practical information about the Uzbek language, is specified in the «concept of the development of the Uzbek language and improvement of the language policy in 2020-2030». Therefore, creating a national corpus of the Uzbek language using modern information technology tools is a priority and urgent task.

The dissertation serves to a certain extent in carrying out the tasks specified in regulatory legal documents related to this activity as PD-4797 of the President of the Republic of Uzbekistan, dated May 13, 2016 «On the establishment of the «Tashkent State University of Uzbek Language and Literature» named after Alisher Navoi», PD-4947 dated February 7, 2017 «On the strategy of actions for the further development of the Republic of Uzbekistan», PD-5850 dated October

[20] O‘zbekiston Respublikasi Prezidentining “O‘zbek tilining davlat tili sifatidagi nufuzi va mavqeini tubdan oshirish chora-tadbirlari to‘g‘risida”gi PF-5850-son Farmoni. //Xalq so‘zi, 2019 yil 22 oktyabr. №218 (7448).

[21] O‘zbekiston Respublikasi Prezidentining 2020 yil 20 oktyabrdagi “Mamlakatimizda o‘zbek tilini yanada rivojlantirish va til siyosatini takomillashtirish chora-tadbirlari to‘g‘risida”gi PF-6084-sonli farmoni (https://lex.uz/docs/5058351).

21, 2019 «Measures to fundamentally increase the prestige and status of the Uzbek language as a state language», PD-6084 dated October 20, 2020 «On measures to further develop the Uzbek language in our country and improve the language policy», PD-6097 dated October 29, 2020 «On approving the conception of development of science until 2030», PD-60 of January 28, 2022 «On the development strategy of New Uzbekistan for 2022-2026» and other acts, related to this activity.

The relevance of the research to the priority directions of the development of the science and technology of the Republic. The dissertation was carried out in accordance with the priority direction of the republican science and technology development «Formation of a system of innovative ideas in the social, legal, economic, cultural, spiritual and educational development of the information society and democratic state and ways of their implementation».

Scope of study of the problem. Modern corpora is an information system based on a collection of texts in a certain language in electronic form[22]. Each corpus must be equipped with linguistic hardware and software to work on texts through a linguistic approach.

The first corpus was created by Heinrich Kuchera and Nelson Francis at Brown University in the USA in the 1960s[23] - it is now called the Brown University Standard Corpus of American English. The initial corpus can be estimated as follows: the volume in the initial state consists of 1 mln. words; it lacks lemmatization and marking-up. Such corpora do not meet the requirements either in terms of size, language coverage, or content of informative texts.

Today, corpora have become an integral part of linguistics, like dictionaries and grammar. The emergence of corpora, one of the new directions of linguistics, raised the development of corpus linguistics to the qualitative stage. Globally recognized linguistic corpora include the Russian National Corpus (https://ruscorpora.ru/new/), the British National Corpus (http://www.natcorp.ox.ac.uk/, https://www.english-corpora .org/bnc/), Turkish National Corpus (https://www.tnc.org.tr/), American National Corpus (http://www.anc.org/) .

Text processing, tokenization and lemmatization, semi-automatic morphological classification, software created with the help of algorithms have been researched by many foreign scientists. In particular, Lawrence Anthony, A.E.Polyakov, C.A.Chapelle, A.A.Abroskin, D.V.Sichinova, L.Yu.Sipisina, Oliver Mason, Kagri Kolekin, V.P.Zakharov, Yeshim Aksan, A.I.Izotov, Andreas Mengel, R. Garabik, Djavdet Suleimanov, Alfiya Galiyeva[24] and some others have studied issues of corpus linguistics.

[22] Николаев И.С., Митренина О.В., Ландо Т.М., Прикладная и компьютерная лингвистика.Книга– Санкт Петербург: СпбГУ. 2016. – С. 320. –ISBN 978-5-9710-3472-8.

[23] Kholkovskaia O. Role of the Brown Corpus in the History of Corpus Linguistics. Poster–Prague, 2017. – 1-4 p.

[24] Laurence A. AntConc: Design and Development of a Freeware Corpus Analysis Toolkit for the Technical Writing Classroom, IEEE International Professional Communication Conference Proceedings, 2005.– 29-737 p.; Поляков А.Э. Технология подготовки информации в национальном корпусе русского языка. http://www.ruscorpora.ru/ new/corpora-biblio.html.; Chapelle C.A. Computer applications in second language acquisition: Foundations for teaching, testing, and research. Cambridge, England: Cambridge University Press, 2001.; Аброскин А.А. Поиск по

In Uzbek linguistics, many scientists and researchers have conducted scientific research on the national language corpus and its software. A.Polatov, S.A.Karimov, A.Karshiyev, S.Matlatipov, U.Tukeyev, M.Aripov, B.Mengliyev, N.Abdurahmonova, Sh.Khamroyeva[25] have created the theoretical and practical foundations on natural language processing, statistical analysis and corpus linguistics, some others[26] on using information technologies in linguistics in order to solve problems in the Uzbek language and increase work efficiency.

корпусу: проблемы и методы их решения. Национальный корпус русского языка. – Нестор-История, 2009. – С.277–282.; Сичинава Д.В. Параллельные тексты в составе национального корпуса Русского языка: новые языки и новые задачи, национальный корпус русского языка: Исследования и разработки. – Москва, 2019. – С.41-60.; Щипицина Л.Ю. Информационные технологии в лингвистике. Учебное пособие. – Москва: Издательство «ФлИнта», 2013.; Mason O., Developing Software for Corpus Research, International Journal of English Studies, IJES, vol. 8 (1), 2008, pp. 141-156.; Çöltekin Ç., A Freely Available Morphological Analyzer for Turkish, Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010) – 820–827 p.; Victor Zakharov, Corpora of the Russian Language, Conference: 16th International Conference, DOI:10.1007/978-3-642-40585-3, TSD 2013.; Aksan Y., Aksan M. Building a National Corpus of Turkish: Design and Implementation, 2009. – 299-310 p.; Изотов И. Чешский национальный корпус и аналитический императив: опыт корпусного анализа малоупотребительных и маргинальных языковых единиц. Вестник огу №2/февраль, 2007. – С.4-11.; Mengel A., Lezius W., An XML-based representation format for syntactically annotated corpora, In Proceedings of the Second International Conference on Language Resources and Evaluation (LREC’00), Athens, Greece. European Language Resources Association (ELRA), 2000.; Гарабик Р., Словацкий национальный корпус. Труды международной конференции «Корпусная лингвистика–2004». Санкт-Петербург: Издательство С.-Петербургского университета, 2004. – С. 99-121.; Dzhavdet Suleymanov, Olga Nevzorova, Ayrat Gatiatullin, Rinat Gilmullin,Bulat Khakimov, National Corpus of the Tatar Language “Tugan Tel”: Grammatical Annotation and Implementation, Procedia - Social and Behavioral Sciences 95 (2013) 68 – 74 p.; Galieva A., Khakimov P., Gatiatullin A., On the Way to the Relevant Grammatical Tagset for the Tatar National Corpus, EPiC Series in Language and Linguistics, Volume 1, 2016, Pages 121-129.

[25] Polatov A., S.Muhamedova, kompyuter lingvistikasi. O‘quv qo‘llanma. – Toshkent, 2009. – 104-s.; Karimov S., Qarshiyev A., Isroilova G. Abdulla Qahhor asarlari tilining lug‘ati. Alfavitli lug‘at. Chastotali lug‘at. Ters lug‘at. – T., 2007; O‘zbek tili korpusining dasturiy ta’minotini yaratishning dastlabki natijalari, Tursunov M.S., Qarshiyev A.B., Karimov S.A., Muhammad al-Xorazmiy avlodlari ilmiy-amaliy va axborot-tahliliy jurnal, 1(15), mart, 2021; Development of a modern corpus of computational linguistics, Tursunov M.S., Qarshiyev A.B., Karimov S.A., Conference: 2020 International Conference on Information Science and Communications Technologies (ICISCT),, DOI: 10.1109/ICISCT50599.2020.9351376, noyabr 2020.; Matlatipov S., Tukeyev U. and Aripov M. Towards the Uzbek Language Endings as a Language Resource. International Conference on Computational Collective Intelligence. Springer, Cham, 2020. pp. 729-740.; Mengliyev B. va b. O‘zbek tilining milliy korpusi // Ma‘rifat. – 26/04/2018; Менглиев Б., Ш. Ҳамроева. Лингвистик таъминот воситаларининг умумий тавсифи ҳамда тил корпусини яратишдаги аҳамияти. O‘zbek tili taraqqiyoti va xalqaro hamkorlik masalalari. –Toshkent, 2019. – P.17-20.; Abduraxmonova N., Elektron korpuslarning kompyuter modellari: Filol. fan. dok. diss. avtoref. – Toshkent, 2021. – 74 b.; Ҳамроева Ш. Ўзбек тили муаллифлик корпуси тузишнинг лингвистик асослари. Филол. фан...(PhD) дисс. –Қарши, 2018. –41 б.

[26] Eshmo‘minov A. O‘zbek tili milliy korpusining sinonim so‘zlar bazasi: Filol. fan. bo‘yicha fal. dok. (PhD) diss. – Qarshi, 2019. – 140 b.; Axmedova D. Atov birliklarini o‘zbek tili korpuslari uchun leksik-semantik teglashning lingvistik asos va modellari: Filol. fan. bo‘yicha fal. dok. (PhD) diss. – Buxoro, 2020. – 156 b.; Xoliyorov O‘. O‘zbek tili ta’limiy korpusini tuzishning lingvistik asoslari. Filol. fan. bo‘yicha fal. dok. (PhD) diss. avtoref. – Termiz, 2021. – 53 b.; Abduraxmonova N. Elektron korpuslarning kompyuter modellari: Filol. fan. dok. diss. avtoref. – Toshkent, 2021. – 74 b.; Abdullayeva O. O‘zbek tilining internet axborot matnlari korpusini shakllantirishning nazariy va amaliy asoslari. Filol. fan. bo‘yicha fal. dok. (PhD) diss. aftoref. – Toshkent, 2022. – 54 b.; Raxmanova A. O‘zbek tili Milliy korpusini yaratishda kompyuter usullari. Filol. fan. bo‘yicha fal. dok. (PhD) diss. – Farg‘ona, 2022. – 52 b.; Abjalova M. Korpus lingvistikasi. [Matn]: uslubiy qo‘llanma / M.A. Abjalova. – Toshkent: Bookmany print, 2022. – 103 b.; Elova D. O‘zbek tili korpusi birliklarining uslubiy teglarini yaratish tamoyillari va lingvistik ta’minot. Filol. fan. bo‘yicha fal. dok. (PhD) diss. avtoref. – Toshkent, 2022. – 65 b.; Xolova M. O‘zbek shevalari korpusini tuzishning lingvistik asoslari (Boysun tumani “j”lovchi shevalari misolida). Filol. fan. bo‘yicha fal. dok. (PhD) diss. avtoref. – Tyermez, 2022; Begmatova G. O‘zbek milliy kopusida idiomalar bazasini yaratish. Filol. fan. bo‘yicha fal. dok. (PhD) diss. avtoref. – Tyermez, 2021; Toirova G. O‘zbek tili milliy korpusini yaratishning nazariy va amaliy masalalari. Filol. fan. dok. (DSc) diss. avtoref. – Buxoro, 2021; Xidirov O. Milliy korpus uchun Parsing dasturi yaratishning lingvistik asoslari. – Jizzax, 2021.

The connection of the research with the research plans of the higher educational institution where the dissertation was accomplished. The research was carried out at the Department of Uzbek Linguistics of Samarkand State University as part of the practical project of the 30th type of 2021 of the Ministry of Innovation of the Republic of Uzbekistan on the topic «Designing the national corpus of the Uzbek language and developing a software package» (2021-2023).

The aim of the research is to create a web-based (concordance, tokenizer, lemmatizer, marking-up) software for creating the national corpus of the Uzbek language.

to analyze the existing corpora in the world, Turkic and Uzbek linguistics, interpretation of literature on the basis of scientific and theoretical sources within the theme;

to develop the software of the national corpus of the Uzbek language;

to develop the Uzbek language text processing, text tokenization, lemming, marking-up models and algorithms;

to input texts into the national corpus of the Uzbek language through corpus software and edit them in a linguistic aspect.

The object of the research. As the object of the research, the variants of the epic «Alpomish» by Fazil Yoldosh ogli (Tashkent: Sharq, 2010); Abdunazar Poyonov (Tashkent: Akademnashr, 2018); Mardonaqul Avliyaqul ogli (Tashkent: Fan, 2018) are chosen.

The subject of the research. The models, algorithms and software of the national corpus of the Uzbek language were taken as the subject of the research.

The research methods. Models, algorithms, database development technologies, and web-oriented software tools, analysis methods were used to enlighten the dissertation work.

The scientific novelty of the research is as follows:

the possibilities of creating a national corpus through the software of the National Corpus of the Uzbek language, ensuring its operation on the Internet, linguistic classification of texts using a computer, studying on the basis of this national language of each people and its characteristics, and preserving its purity have been identified;

through the creation of the software uzbekcorpora.uz, designed for the creation and management of the national corpus of the Uzbek language, it has been substantiated that the search for words and phrases throughout the entire corpus (concordance), definition of markup (word morphology), lemmatization, tokenization and creation of frequency dictionaries are its integral components;

it has been proven from the linguistic point of view that the programs designed to create the corpus and the programs serve to use the corpus perform such tasks as forming a text base, creating and editing a corpus dictionary, and marking-up texts.

corpus-based software has been proven to have the ability to sufficiently identify and modify the grammatical and stylistic features of words or sentences in any text.

The practical results of the research are as follows:

the practical importance of the results of the dissertation is determined by the creation of the national corpus of the Uzbek language, the creation of the author's corpus, conducting research using the corpus, designing and creating algorithms to achieve these goals. This will serve as a program for improving the national corpus of the Uzbek language in the future.

Uzbek language text processing, text tokenization, lemmatizing, and marking-up models and algorithms have been developed;

the software of the national corpus of the Uzbek language has been developed and it has been determined that the existing contexts in the concordance will be prepared based on the given conditions;

it is based on the possibility of searching words by morphological features (for example, word groups, etc.), stylistic features (for example, journalistic, scientific, artistic, etc.) and according to the periods (for example, 1990-2000, 2000-2021);

as linguistic results of the corpus, the text of the epic «Alpomish», its language units and grammatical descriptions are included in the software. Based on the epic «Alpomish», a dictionary was created in the corpus.

The reliability of the research results is explained by the complete and precise development of the software and its algorithms created to solve the given problem, the careful design of the database project, the clear representation of the models, and the compatibility of theoretical and practical knowledge. It is explained that the supply is based on proven sources.

Scientific and practical significance of the research results. The scientific significance of the research results is that the mechanisms of creation and design of the national corpus of the Uzbek language, text processing algorithms and the theoretical foundations of the national corpus software are based on the existing scientific research on computer linguistics and corpus linguistics, which are considered modern directions of Uzbek linguistics, which is determined by the possibility of use in enriching and completing theories.

The practical significance of the results of the research is to search (concordance) of words and phrases in the corpus of the *uzbekcorpora.uz* software, which is intended for the creation and management of the national corpus of the Uzbek language, to determine the marking-up (morphology of words), lemmatizing, tokenization and is explained by the fact that it can be used to create frequency dictionaries. The *uzbekcorpora.uz* software for editing the corpus database is a free online platform that allows the learner to work anywhere, on any computer, in language research. The research work is determined by the fact that it is important in the scientific improvement of dissertations, monographs, textbooks and manuals created in such disciplines as text linguistics, computer linguistics and corpus linguistics.

Implementation of the research results. Based on the scientific results obtained from the experience of creating the Uzbek linguistic corpus (on the basis of concordance, tokenizer, lemmatizer, marking-up programs):

the inferences on the issues as creating the National Corpus by using the software of the National Corpus of the Uzbek language, linguistic classification of texts with the help of a computer have been used in the fundamental grant project OT-F1-030 on the topic of «Publishing a multivolume monograph (7 volumes) «History of Uzbek language»» (Reference № 01/4-1716 dated September 9, 2023, of Tashkent State University of Uzbek Language and Literature named after Alisher Navoi). As a result, digitization of texts, inclusion in the corpus and improvement of investigations were achieved.

the views on creating the national corpus of the Uzbek language, searching the words and phrases in corpus (concordance) using uzbekcorpora.uz software, determining the mark-up (morphology of words), lemming, tokenization and creating frequency dictionaries have been used in the practical project JHBL-20 «Creating an electronic corpus of artistic works on the topic of family, neighborhood and gender equality» of the research institute «Community and Family» (reference 01-09/1608 dated September 4, 2023, of the research institute «Community and Family»). As a result, it served as a scientific-practical resource in the formation of the electronic corpus of literary works.

materials on the tasks of creating the programs designed to create a corpus and serve to use the corpus, creating and editing a corpus dictionary, marking-up the texts, which it is possible to determine and change the grammatical and stylistic features of words or sentences in any text have been used in the preparation of scripts for programs such as «News of the Land» of «Samarkand» regional TV channel (reference № 01-07/138 dated April 20, 2023 of the Samarkand regional television and radio company). As a result, the information provided is enriched by scientific evidence, and the content of the shows is improved from an educational and practical point of view.

Approbation of the research results. Research results were presented at 13 international and 19 national scientific-practical conferences.

Publication of research results. 6 scientific works were published on the dissertation theme, including 1 published in a journal available in the Scopus database, 5 republican and 5 foreign journals. Certificates were obtained for 5 software tools created for EC.

The structure and volume. The dissertation consists of an introduction, three main chapters, a summary, a list of used literature and appendices, and its total volume is 120 pages.

THE MAIN CONTENT OF THE DISSERTATION

In the introductory part of the dissertation, the relevance and necessity of the conducted research is based, the aim and tasks, object and subject of the research are described, the concordance with the priority directions of the development of science and technology of the republic is shown, the scientific novelty and practical results of the research are described, the scientific and practical results of the obtained results are depicted, its importance is revealed, published works and information on the structure of the dissertation are presented.

The first chapter of the dissertation entitled «**Designing the National Corpus of the Uzbek Language»** describes the concept of the corpus and its types. The corpora of Slavic and Turkic languages were studied. The stages of designing the national corpus of the Uzbek language, marked corpus models and corpus creation are described.

The first part of the first chapter is called «**The notion of corpus and its types»,** which examines corpus linguistics, corpus types and their characteristics.

Linguistic corpus is a set of texts that are compiled according to certain principles and have a common standardized format. Initially, a corpus («First-level corpus») was understood as a set of texts that share a common feature (language, genre, author, and period). Modern computer technologies made it possible to simplify and speed up the processes of linguistic processing of large texts and changed the content of the notion of corpus. A modern corpus is an electronic information system based on a collection of texts in electronic form of a particular language. Each corpus has linguistic software that serves to work on texts through a linguistic approach.

Examples of linguistic corpora include the following most famous and recognized corpora: Russian National Corpus (https://ruscorpora.ru/), British National Corpus (http://www.natcorp.ox.ac.uk/, https: //www.english-corpora.org/bnc/), Turkish National Corpus (https://www.tnc.org.tr/), American National Corpus (http://www.anc.org/), etc.

It is impossible to create language corpora without software tools to process data and present results in a comprehensive way. There are language corpora designed for specific purposes (*WordSmith Tools, MonoConc Pro, WordPilot*). They do not allow extensively learning the language. National corpora help to study the morphological, semantic, syntactic aspects of the language, to obtain various analyses of the language and to conduct extensive language research. The functionality of the corpus should have clear, simple and convenient interface.

Text corpus can be said to be a powerful tool in the hand of a linguist. *Corpus linguistics* is a science that studies the laws of language by analyzing and studying it with the help of a linguistic corpus. Being a branch of linguistics, corpus linguistics forms a new branch of computational linguistics and applied linguistics. A corpus, unlike a simple collection of texts (a «library», a collection), requires that they be marked-up.

Corpus linguistics includes two aspects:

• creation and classification of text corpora and development of search tools for them;

• conducting linguistic research based on corpora.

The second part of the first chapter is called «**Analysis of existing language corpora»**, in which software and creation methods of language corpora of Slavic and Turkic languages are analyzed.

The following language corpora have been studied and analyzed from the point of view of designing corpora for **Slavic languages**: The National Corpus of the **Russian Language** is a searchable electronic online corpus of texts in the Russian language. It was established on April 27, 2004[27]. The National Corpus also

[27]https://ruscorpora.ru/new/archive.html.

contains historical corpus of Church Slavonic, Old Russian (XI-XIV centuries) and Central Russian (XV-early XVIII centuries) texts. The need to create a national corpus of the **Belarusian language** – a general computer database of the Belarusian language appeared on implementing the program «The problem of linguistic representativeness of the Belarusian language and the principles of creating a corpus of the Belarusian language», which was launched in 2001 at the Institute of Linguistics named after Yakub Kolas of the National Academy. National Corpus of the **Bulgarian Language** – The Bulgarian National Corpus was created by the researchers of the Department of «Computer Linguistics and Bulgarian Lexicology and Lexicography» at the Bulgarian Language Institute under the leadership of Prof. L.Andreichin. The National Corpus of the **Slovak Language** is an electronic database of Slovak texts dating back to 1955, mainly in various styles, genres, and themes. The National Corpus of the **Czech Language** is an open searchable database of Czech written texts in electronic form maintained by Charles University in Prague. The site is available in Czech and English. The National Corpus of the **Polish Language** (http://nkjp.pl/index.php?page=0& lang=1) is a joint initiative of four institutions: Institute of Computer Science of the Polish Academy of Sciences (coordinator), Institute of the Polish Language of the Polish Academy of Sciences, Polish Scientific Publishing House and the Department of Computational and Corpus Linguistics, University of Lodz. The national corpus was implemented as a research project of the Ministry of Science and Higher Education.

Being a language of modern and ancient Turkic peoples, the **Turkic languages** include more than 25 languages of the peoples, spread over a huge geographical area stretching along a line from Siberia to the Balkan Peninsula as Uzbek, Uyghur, Kazakh, Kyrgyz, Karakalpak, Sakha (Yakut), Tuva, Khakas, Altai, Karagas, Shor, Turkmen, Azerbaijan, Turkish, Karaoguz, Tatar, Bashkir, Chuvash, Komik, Nogai, Karachoi Balgor, Tofalar, Chuvash. The following language corpora have been studied and analyzed from the point of view of designing corpora for **Turkish languages**: The National Corpus of the **Turkish Language** (https://www.tnc.org.tr/) is a balanced, large-scale (50 million words) for the modern Turkish language for general purposes. The Turkish National Corpus (TNC) project was developed by a team of linguists from Mersin University and is funded by the Scientific and Technological Research Council of Turkey for a period of three years (2008-2011)[28]. The national corpus of the **Tatar language «Tugan Tel»** is a linguistic source of the modern literary Tatar language. The project is implemented within the framework of the State program «Preservation, study and development of the state languages of the Republic of Tatarstan and other languages in the Republic of Tatarstan in 2014-2020». The volume of the national corpus of the **Kazakh language** (http://qzcorpus.kz) is 30 million tokens, including 14 million tokens of text material provided with meta-tags.

[28] Yeúim Aksan & Mustafa Aksan, Building a National Corpus of Turkish: Design and Implementation, p:299-309.

The field of corpus linguistics is also rapidly developing in Uzbek linguistics. In particular, the project “National corpus of the Uzbek language” (http://uzbekcorpora.uz/)[29] in cooperation with the Samarkand branch of the Samarkand State University and the Taoshkent University of Information Technologies, the “Educational corpus of the Uzbek language” at the Tashkent State University of the Uzbek Language and Literature “ project (https://uzschoolcorpara.uz/)[30], “Uzbek language corpus” projects (https://uzbekcorpus.uz/newIndex)[31] are being implemented at the National University of Uzbekistan.

As a result of the analysis of existing corpora, it can be seen that there is no single methodology for creating corpora for all languages. The reason for this is that different rules and technological processes apply in different languages. Although the sets of morphological descriptions are different, morphological (or grammatical) notation is provided in all the corpora discussed above. Syntax markup is present in some corpora, but implemented based on different approaches, and in some do not exist at all. When comparing the corpora by the level of language coverage, the national corpus of the Russian language can be considered the most balanced corpus, both chronologically and in terms of genre, and providing a variety of texts. But this achievement of the national corpus of the Russian language also brought to light its main shortcoming. In any large corpus, the possibility of complete and accurate markup is limited. Ambiguities are mainly caused by homonyms in automatic parsing.

The third part of the first chapter is called «**Designing the National Corpus of the Uzbek Language: Model and Stages of the Marked-up Text Corpus**», which describes the stages of designing corpus creation software, developing models for working with texts in the corpus, and creating it.

When creating a corpus, it is advisable to take into account the presence or absence of different markups and, if so, the degree of accuracy. A philologist should carefully check the mark-up of each text word by word, correct automatic mark-up errors and eliminate homonymous cases. Most of the work can be done by hand. Therefore, it is necessary to create special mark-up programs for the computer work of a philologist.

The schematic form of the corpus model is shown in App. 1[32].

We consider the corpus as a collection of T_i texts. The text consists of a sequence of words separated by punctuation marks p_{ik}, spaces or end-of-line marksw_{ij}[33]:

$$Corpus = \{T_1, T_2, ..., T_n\},$$
$$T_i = \{w_{ij}\} U \{p_{ik}\},$$

[29] http://uzbekcorpora.uz/

30 https://uzschoolcorpara.uz/

31 https://uzbekcorpus.uz/newIndex

[32] Седов А.В., Математические модели, методы и алгоритмы построения размеченных корпусов текстов, Петрозаводск, 2013. –С. 23.

[33]Седов А.В., Математические модели, методы и алгоритмы построения размеченных корпусов текстов, Петрозаводск, 2013.. –С. 23.

where j-w_{ij} is the position of the word in the i-text, and k-p_{ik} is the value indicating the position of the punctuation mark. Also, each word is assigned its place in the text:

$Dispostion{:}w \rightarrow positions \in Positions,\ Positions= Pos1,\ Pos2,\ \ldots,\ Posq,$

where, for w, its coordinate q in the text is determined. Chapters, paragraphs, sections, sentences, sentence fragments in the text are its additional structural units. The model takes into account the formation of phrases from the smallest structural units (words) of the text. If a sentence is considered a set of words, then each phrase can be considered a subset:

$SSj= wi1,\ wi2,\ \ldots,\ wik \subset S,$ bunda $\bigcup jSSj=S.$

Each word can be part of only one phrase, so the intersection of two phrases is an empty set:

$SSk \cap SSj=\varnothing,$ bunda $k \neq 1.$

For each word, in addition to its place in the text, a set of its morphological parameters should be determined. For a noun, these parameters are agreement,

The process of creating a corpus consists of several stages:

1) collecting data: compiling a list of texts to be included in the corpus, determining their source;

2) computerization of data: loading selected data (electronic texts, scanned texts, text typed from the keyboard, etc.);

3) encoding metadata: writing metadata about the selected text;

4) data marking-up: tagging all text units;

5) development of a search interface in the corpus: creating a graphic web interface that facilitates and make easy the retrieval of information from the corpus;

6) launching the corpus: first the trial version, then the first version intended for national and international use.

The second chapter of the dissertation is called **«Software of the national corpus of the Uzbek language»**. In this chapter, the structure and tasks of the software of the national corpus of the Uzbek language, the formation of the text database, the formation and editing of the corpus dictionary, the input of texts in the corpus and the algorithms of the text marking-up programs are considered and described. It also describes the use of the Internet for the software, as well as the user and management interfaces of the software.

In the first part of the second chapter, entitled «**Software structure and functions»,** the structure and functions of the software of the national corpus of the Uzbek language, the formation of the text database, the formation and editing of the corpus dictionary, the input of texts in the corpus and the programs for the marking-up of texts are covered.

Depending on the model of the national corpus of the Uzbek language and the stages of the work to be performed for its creation, a software structure consisting of two parts was developed (App. 2):

1) programs designed to create a corpus;

2) programs which serve to use the corpus.

Programs designed to create a corpus should be able to perform tasks such as creating a text database, creating and editing a corpus dictionary, and text formatting (App. 3).

When creating a text base, the following tasks must be performed:

- digitization of texts, their editing;
- writing text information into a file;
- including the text into the database.

We choose text digitization tools depending on their original source (paper, *.pdf format file).

Metamarking-up data consists of general information about the text, including:

- text name;
- information about the author of the text: first and last name, gender, date and year of birth, etc.;
- the time when the text was written;
- theme and type of text;
- genre;
- text size (in words).

The program ensures that these data are entered by the user and writes them in the MeteRazm file. Then, based on the input metadata, a special unique name is formed, and the *.docx file in which the text is saved is copied to the Text Base folder with this name.

Text tokenization. In automatic text processing, first of all, there is a problem of marking words out, or in other words, dividing the text into units. For this, all partial lines that do not contain markers (spaces, punctuation marks, etc.) should be marked out in the text. This will be a set of tokens[34]. One of the fundamental algorithms of automatic text processing consists in dividing the given text into tokens. The algorithm is given a text as input, and the output is a list of tokens in the text. The program that implements this algorithm is called a tokenizer. Generally, tokens have the same meaning as word forms. However, to represent lexical units, the term «token» is used, not «word». This is because, in some cases, units smaller than words (individual morphemes) or units larger than words (phrases) can be used as tokens.

Making different lists of words in the text. The aim here is to mark the text units (words, word forms) that should be included in the dictionary, make a list of them, and make the list in different forms (alphabetic-frequency order, frequency-alphabetic and reverse list) at the request of the user. *Gram_Dictionary* program module suitable for this task is called *Suz_Rhati*, and its input is a list of tokens marked out in the text, and its output is an ordered list of words and word forms. The program uses the grammar dictionary file DICTIONARY to generate this list. Each token given on input is looked up in the existing DICTIONARY file. If the token is found in the dictionary, then it does not need to be included in the

[34] Николаев И.С., Митренина О.В., Ландо Т.М. (ред.) Прикладная и компьютерная лингвистика, Ленанд, 2016. - 320 с. - ISBN 978-5-9710-3472-8.

dictionary, it can be discarded. Otherwise, i.e. if the token is not found in the dictionary, it is checked whether it is a word or word form. Tokens that are words or word forms are listed separately. This list is sorted and displayed according to the user's request, either in alphabetical-frequency order, or in frequency-alphabetical order, or in the form of a reverse list.

Adding words to the grammar dictionary. The software module that performs this task is called *Lug_Kirit*. Adding a new unit to the dictionary is done by the user in interactive mode. This process is provided by the *Lug_Kirit* program. On the computer screen, there is a list produced by the *Suz_Rhati* program, in which one word is highlighted with a different color. Using the *Lug_Kirit* program, the user moves through the list, selects the current word and presses the «Enter» button. As a result, an input window will appear on the screen. The user enters the information requested in this window from the keyboard and clicks the «Input to database» button. Then the selected word and the corresponding grammatical characteristics are summarized and recorded in the electronic dictionary file *LUCAT* and other auxiliary files. After that, the user chooses the next word from the list and enters it into the database. In this way, all the words in the given list are included in the *DICTIONARY* file, and the grammatical dictionary is gradually enriched.

A grammar parsing program built on *Nuxt JS, Python, and the PostGreSql* database management system (MBBT). To simplify the process of entering texts into the corpus, grammaticalization is performed in two stages: the stage of initial formatting of the text and the stage of parsing. At the initial formatting stage, the structural components of the text are determined, that is, words, sentences, paragraphs and sections are separated in the text. In addition to it, the place of words in the text is determined. In the second stage, grammatical classification is performed. As a result of parsing, each word in the text is attached to a form with its own set of morphological parameters and line attributes. Below is a detailed explanation of how both of these stages are done using a grammar parsing program.

Initial formatting step. At this stage, regardless of the original format of the file to be included in the corpus, it should be converted to *MS Word* format. It is necessary to use *MS Word* versions 2007 and higher, because in them the text is in **.docx* format and is automatically tagged based on the standard system. Dividing the text into structural components (words, sentences, paragraphs, sections) is performed automatically by the *MS Word* program. Information indicating the position of words in the text is written in a separate file.

Marking-up stage. Marking-up should be understood as attaching special tags to the text and its components. There are two types of special tags: linguistic tags and extralinguistic (external) tags. Linguistic tags consist of information that describes the lexical, grammatical, and other similar properties of text elements.

Extralinguistic tags describe information about the author and the text (author, title, year and place of publication, genre, subject area, etc.)[35].

The software should enable the user to conduct linguistic research on the texts contained in the corpus and draw conclusions based on this. In particular, the following tasks are assigned to the software by the corpus user:

- creating a concordance;
- searching for contexts not only by words, but also by word combinations;
- sorting lists according to several criteria selected by the user;
- providing an opportunity to describe the found word forms in an expanded context;
- providing statistical information on separate elements of the corpus;
- saving and printing results;
- ability to work not only with individual files, but also with unlimited size corpuses;
- immediate response to requests and release of results.

In general, the performance of these tasks consists of processes such as finding the necessary information for research, collecting it and presenting it to the user in the right way. In order to implement these processes, the issue of information search across the corpus should be resolved.

The second part of the second chapter is called «**Software interface**» and it describes the advantages of the Internet for the language corpus, the functions of the developed software, such as using the created corpus and creating the corpus.

Several arguments can be made in favor of presenting the results of scientific research on the Internet[36]. First, today, at least in our country, science should move to a pragmatic level, that is, the result of scientific activity should be real applications necessary for society. In this case, the presentation of scientific research through the Internet helps to comprehensively demonstrate the product of scientific thought. Through various *World Wide Web* services, it is possible to convey ideas to potential customers and interested parties. Consequently, the Internet serves as a marketing agent for the researcher. Secondly, presenting the results of scientific work on the Internet, for example, serves as a systematization of research in the form of a website. Thinking about the structure of the site, the researcher unknowingly classifies and systematizes the collected scientific materials, leads to objective additional processing. From this perspective, publications in online journals and lectures at conferences have equal rights with the criteria for evaluating the scientist's activity. In addition to it, Internet tools allow organizing practical forums. As a result of discussions, it will be possible to develop promising directions of research, to evaluate its impact, that is, to analyze the scientific work in all respects.

[35] Николаев И.С., Митренина О.В., Ландо Т.М. (ред.) Прикладная и компьютерная лингвистика, Ленанд, 2016. - 320 с. - ISBN 978-5-9710-3472-8.

[36] Green, L. The Internet: An Introduction to New Media/ L.Green. – New York: Berg, 2010; Silver, D. Critical Cyber culture Studies / D.Silver, A.Massanary. – New York: New York University Press, 2006; Браславский, П.И. Методы повышения эффективности поиска научной информации (на материале Internet).: автореф. дис... .канд.тех.наук. 05.13.16 / П. И. Браславский. — Екатеринбург, 2000. — 16 с.

To use the **national corpus of the Uzbek language**, connect to the Internet, open a browser and go to the program via the link http://uzbekcorpora.uz/ or http://www.uzbekcorpora.uz/ (App. 4).

The main window of the program consists of a menu bar, switching to important programs, switching to useful sites and news on the Uzbek language. Extended information on the use of software is taken place in the dissertation work.

The third chapter, titled «**Creating and Researching a Corpus of Texts»,** presents methods for collecting, digitizing, and formatting a collection of texts for a corpus. Algorithms of concordance creation programs and analysis of it for conducting language research with the help of the corpus are explained, optimal methods are used, and the results obtained in the software are described. The text of the *Alpomish* epic, which is the object of the research, was included in the corpus, alphabetic and frequency dictionaries were compiled, and statistical analyzes were obtained.

The first part of the third chapter is called «**Collection, Digitization and Formatting of the Text Collection»**, and in this part information is given about the collection of texts, digitization and its transfer into the program format. The forms of the selected set of the texts are important in the process of creating the corpus. When entering the text into the corpus, it undergoes several changes of form. The preparation of the text for the corpus consists of the following steps:

1) initial text formatting - converting it to minimal HTML format;
2) morphological classification;
3) entering metadata of the text;
4) inclusion of the text in the corpus.

Accomplishing these stages, added information into the original text can be constantly increased. In the first step, information about its structure is entered into the text. Different elements of the text (word, paragraph, section, etc.) are separated and appropriate signs are placed. In the second stage, linguistic information is entered. Morphological classification is carried out using the software described in the second chapter of the dissertation. At the third stage, a «passport» of the text is created. It consists of metadata about the text and is prepared separately from the text. At the last stage, the text is included in the corpus, and on completing it will be possible to conduct linguistic research on it with the help of appropriate programs.

Digital and non-digital texts can be used as text sources for the corpus. Naturally, in the second case, it is necessary to somehow enter the text into the computer: it must be rewritten or scanned. For example, we don't have the *DOCX* format of the Fazil Yoldosh ogli version of the Alpomish epic. That's why we need to digitize this book. In this case, the manuscript is converted to *PDF* format while scanning the source or book, and converted to *DOCX* format, which can be opened in Microsoft Office Word with the help of a converter (for example, Fine Reader) that recognizes the writings in it. Of course, it is very difficult for the converter to convert the scanned e-book to *DOCX* format according to the original state of the book. Therefore, based on manual work, the text in *DOCX* format is brought to the same level as the text in the original state. It corrects errors such as spelling errors, incorrect recognition of signs in the text.

Text formatting. Texts are available in various PDF, in image, document and other formats. Before adding texts to the corpus, it is necessary to convert the existing text files to the *.docx format of Microsoft Office version 2010 and higher. Texts in other formats are converted to *.docx format using special programs, and during the process of converting to *.docx format, the original state of the text may be damaged. In this case, the spelling errors in the text are brought to the same level as the original state of the text manually. After that, the text can be loaded into the corpus. In this study, the JSON format was used to store the texts in the corpus (App. 5).

The second part of the third chapter is entitled «**Creating and Analyzing Concordances»,** which covers the notions of concordance, the use and analysis of concordance in developed software. Like other corpora, a search system is implemented throughout the corpora, which is one of the main tasks of the *uzbekcorpora.uz* system. It provides information about the characteristics of the searched word.

Concordance is a traditional, long-known, but still understudied way of studying text. It provides a complete index of words in immediate and extended context[37].

Providing access to search results in the national corpus of the Uzbek language is a special task, the site should have enough information, but at the same time, it should not be overloaded with unnecessary information. This function is intended for professional philologists applying for research. The general principles of providing information to the user were studied[38] and the initial search system project and software were prepared based on them. The pages are made as simple as possible to make the search engine easier to use. For this reason, the main part of the search page is not overloaded with information. This made it possible to visually expand the workspace. Based on the requirements put forward by philologists, the created information resource should provide the user with several options for obtaining information. It provides two options for getting the context of the word as a result of the search: 1. Getting the sentence in which the searched word is involved as a context. 2. Getting the context of the searched word by specifying the number of words to be retrieved from the front and back. Context is the combination of the chosen word with the surrounding words.

Finding a word from the text: After the request is created and sent to the server as a result of the actions of one of the functions described above, word search is performed using the necessary parameters. The logical expression obtained as a result of the algorithm is replaced with a query and sent for processing. As a result, cells with a word in each are searched through spellings and lines containing the word's position in the text (text number, chapter number, paragraph number, sentence number, and word number). The context is searched

[37]Glushakov, S.V. ПрограммированиеWeб-страниц / S.V.Glushakov, I.A.Jakin, T.S.Xachirov. - Xarkov: «Folio», 2005.-390 s.

[38] Fedorchuk, A. Как создаются Wеб-сайты. Краткийкурс / A.Fedorchuk - SPb.:Piter, 2000. - 224 s.

and displayed according to the given coordinates. The address of the original text is determined on the server. The offset of the beginning of the sentence relative to the beginning of the text is found from the tables. After that, the file will be read and then displayed on the screen.

Context Output: Since we have an exact identifier for a word, we can specify how many preceding and following words surround the word we're looking for at its location in the text, or retrieve the sentence with the word in it. In this case, when displaying the list of contexts found from the text files in the corpus to the user, it outputs a maximum of 10 contexts from each text. To see the complete list of contexts in one text, the number of contexts is also displayed in the field where the name of this text is displayed. If you left-click the mouse button on the number of these contexts, a complete list of contexts of the text will be presented on a separate page.

Word output and its morphological parameters: As we mentioned above, we have the ability to associate each word from the text with its counterpart in the database. This is carried out using a unique identifier. The user left-clicks on a specific word, as a result of which a procedure is called that reads the unique identifier of the word and opens a window displaying the word itself and its properties. For this, the data collection unit and the parameter decoding unit are launched.

Extended context output: After showing the found contexts, the user is prompted to go to the advanced context. Extracts the paragraph in which the context is involved as context to the expandable. It is also possible to see the parameters of each word in the extended context.

The third part of the third chapter is called «**Alpomish epic, alphabetic and frequency dictionaries and statistical analysis»**, in which three options were selected to create a text dictionary of «Alpomish» epic, alphabetic and frequency dictionaries were compiled and statistical analysis was obtained. The first is the most perfect of the versions of the epic, the version of Fazil Yoldosh ogli, published by «Sharq» publishing house in 2010, the version of Abdunazar Poyonov, the version of Mardonakul Avliyokul published by «Fan» publishing house of the Academy of Sciences of the Republic of Uzbekistan in 2018 serve as material. This material allows us to make some complete and reasonable conclusions about the writer's creative thinking, artistic skills, and his unique style of using language.

The second issue is determining the method of creating a dictionary. In the world lexicography, among the thematic, combinatory, ideational, and alphabetic methods of creating a dictionary, the method of creating an alphabetic dictionary is approved, and this method is used more often in the lexicography of writing. It should be remained faithful to this tradition in this place as well. At the same time, it has been enriched with frequency and inverse dictionaries.

In the dictionary, the words were presented in the position in which they were found in the text, and it became possible to process them statistically with the help of a computer. The results of the statistical analysis of three variants of the “Alpomish” epic were obtained (Table 1).

1. An electronic dictionary of the version of Fazil Yoldosh ogli of “Alpomish” epic was obtained. The epic consists of a 623 Kb **.txt* format file prepared in MS Word. After this version of the epic was processed with the help of the program, a total of 14413 words were used in it, and their usage was 82106 words. 82106 items were created in the electronic dictionary.

2. An electronic dictionary of the Abdunazar Poyonov version of the “Alpomish” epic was received. The epic consists of a 631 Kb **.txt* format file prepared in MS Word. After this version of the epic was processed using the program, a total of 19,098 words were used in it, and their usage was 78,632 words. 78632 items were created in the electronic dictionary.

3. An electronic dictionary of the Mardonaqul Avliyaqul version of the Alpomish epic was obtained. The epic consists of a 488 Kb **.txt* format file prepared in MS Word. After processing this version of the epic using the software, a total of 13,520 words were used in it, and their usage was 62,100 words. 62100 items were created in the electronic dictionary.

The results of the statistical analysis of the version of Fazil Yoldosh ogli of “Alpomish” epic on the stems and parts of speech were obtained. 12 types of parts of speech were used to combine word groups in the dictionary: verb, noun, adjective, number, pronoun, adverb, conjunction, preposition, particle, modal words, exclamatory words and onomatopoeic words. There are 14,413 word forms in the version of Fazil Yoldosh ogli without repeating each other. Consequently, the dictionary of 3126 word stems was created from the word forms in this dictionary (App. 6).

While working with a computer, the following shortcomings were noticed:

– in the dictionary, some words, for example, the names of characters, are listed in quotation marks. The computer is reading each word in quotation marks separately. If there are some additions after the quotation mark, for example, if it is written as *«qo‘l ushlatar»ini (touching hands)* from *«Qultoy»dan* or *«Qultoy bobo»ning*, then the *ini* and *ning* in it are also recorded as a separate word form. It is quoted without taking into account the quotation marks in the word *Qultoydan*. When searching for the word *Qultoydan* in the text, the computer cannot find it, as the quotation marks prevent it. In the text, all words such as *Qultoyni*, *Qultoyda*, *Qultoyning*, *Qultoyga*, *Qultoydan* are in quotation marks. They are given in the dictionary without quotation marks;

– in the text, letters or suffixes are enclosed in square brackets at the end or in the middle of some words. For example, *rostdi[r], dash[t]i, Qo‘ng‘irotni[ng]*. In this case, when entering the word into the dictionary, two different options are obtained. For example, the word *rostdi[r]* was included in the dictionary in the form of both *rostdi* and *rostdir;*

– the fact that the combination of two words is a compound word is understood only through the human mind. Therefore, it remains a problem to reflect compound words in the dictionary;

– some words, for example, *Ultontoz*, are given in the form of *Ulton*, *Ultonshah*, depending on the personage’s use of the language. They should also be recorded as two words. Otherwise, the «computer principle» of the dictionary will be broken;

– *Yig'in-tudani, damba-dam, ort-sirtidan, qilmading-ku*, etc. are given with dashes in the text, and these additions are recorded as pairs of words in the computer dictionary;

– abbreviations in the text are included in the dictionary;

– some words are written in two ways in the text: *boryapti - borayapti, bo'lmasmikan – bo'lmasmikin, ko'targuli - ko'targulik*. The computer recorded them separately as vocabulary units. In this case, it is necessary to further process the dictionary or to leave them as they are and recognize the two cases as two words. We thought it best to leave them alone;

– the computer did not record the numbers in the text. For example, if the text says *8 ta, 10 ta* yoki *8 tasi, 10 tasi* is indicated as a separate word;

– even if some letter is written in a different font, but does not match the fonts in the general text, the computer does not read the word or records it separately in the dictionary.

CONCLUSION

1. In the rapidly developing social life, such directions of modern linguistics as text linguistics, computer linguistics, and corpus linguistics are functioning in harmony, and national language corpora are being created. Great attention is paid to creation of national language corpus software, ensuring its operation on the Internet, linguistic research of texts with the help of computers, and on the basis of this, the study of the national language of each nation and its specific aspects, linguistic features, their original state preservation and pure delivery to the next generation.

2. Based on the study of literature and existing corpora, it was determined that it is possible to create a text base, compile and edit a corpus dictionary, and markup texts through two parts of the software - programs designed to create a corpus and programs that serve to use the corpus.

3. Based on the analysis of existing corpora, the following conclusions were drawn: 1. There is no single methodology for creating corpora for all languages; 2. Despite the different sets of morphological characteristics, morphological (or grammatical) marking is provided in all national corpora; 3. In any large corpus, the possibility of complete and accurate marking-up is limited, and inaccuracies appear mainly due to homonyms in automatic marking; 4. The national corpus should be able to present as many regularities and features of the language as possible and should be balanced in terms of content; 5. The volume share of one type of texts in the total number of texts in the language should be preserved in the national corpus.

4. When creating a corpus, it is appropriate to take into account the presence or absence of different markups and, if so, its level of accuracy. Corpus software includes parsing programs, which have the following qualities: 1. The user has the ability to determine and change the grammatical or syntactic characteristics of

words or sentences in a sufficiently complete manner; 2. The analysis is maximally convenient and simple, and is performed quickly.

5. When working with the corpus, web resources and the *uzbekcorpora.uz* web program were developed, which allow not only local use, but also online use. This made it possible for several users to enrich the corpus or use the corpus at any distance at the same time.

6. An algorithm for searching words from files has been developed. When searching for a word using a search engine, it first looks up the word in the corpus dictionary, and if the word exists, the word ID is found in the container table.

7. To include the texts in the corpus, it is necessary to digitize it and format it according to the requirements of the corpus. In the process of digitization, more attention should be paid to correcting linguistic errors of the text. This allows the user to obtain linguistic information with high accuracy.

8. It was determined that it is preferable to use the JSON format over the XML and JSON formats for storing the texts in the corpus and their linguistic information. Advantages of JSON format over XML format: 1. JSON does not use end tag; 2. JSON is shorter; 3. JSON reads and writes faster; 4. JSON is characterized by the fact that it can use arrays.

9. The rules of creating concordances in the corpus search system were studied. It is possible to create concordances for the entire corpus or a specific text for the searched word. Methods of presenting the contexts of the searched word to the user in different ways have been developed.

10. Alphabetical and frequency dictionaries of texts were compiled and statistical analyzes were obtained. In the software for creating statistical dictionaries of the text, some shortcomings that are allowed in the complete and accurate compilation of the dictionary are shown on the basis of evidence, and recommendations are given.

11. Linguistic editing of the text using a computer is of practical importance for research in the field of philology. It helps identify the quantity of the lexical units and grammatical signs used in the text as well as distinguish the peculiarities of the text language and style. This ensures the development of the fields of corpus linguistics and computer linguistics of modern linguistics, and is of practical importance in raising these fields to the quality level.

НАУЧНЫЙ СОВЕТ PHD.03/04.06.2021.FIL.132.01 ПО ПРИСУЖДЕНИЮ УЧЁНЫХ СТЕПЕНЕЙ ПРИ КОКАНДСКОМ ГОСУДАРСТВЕННОМ ПЕДАГОГИЧЕСКОМ ИНСТИТУТЕ

САМАРКАНДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ТУРСУНОВ МУХАММАДСОЛИХ САДИН УГЛИ

ИЗ ОПЫТА СОЗДАНИЯ УЗБЕКСКОГО ЛИНГВИСТИЧЕСКОГО КОРПУСА (НА ОСНОВЕ ПРОГРАММ КОНКОРДАНСА, ТОКЕНИЗАТОРА, ЛЕММАТИЗАТОРА, РАЗМЕТКИ)

10.00.11 – Теория языка, прикладное языкознание и компьютерная лингвистика

АВТОРЕФЕРАТ
диссертации доктора философии (PhD) по филологическим наукам

Коканд – 2024

Тема диссертации доктора философии (PhD) зарегистрирована в Высшей аттестационной комиссии при Кабинете Министров Республики Узбекистан за № B2023.3.PhD/Fil3991

Диссертация выполнена в Самаркандском государственном университете

Автореферат диссертации на трёх языках (узбекский, английский, русский (резюме)) размещен на веб-странице Научного Совета (www.kspi.uz) и в Информационно-образовательном портале «Ziyonet» www.ziyonet.uz.

Научный руководитель: **Каримов Суюн Амирович**
доктор филологических наук, профессор

Официальные оппоненты: **Абдурахманова Нилуфар Зайнабуддин кизи**
доктор филологических наук, профессор

Хамроева Шахло Мирджоновна
доктор филологических наук,

Ведущая организация: **Джизакский государственный педагогический университет**

Защита диссертации состоится «3» августа 2024 года в 9⁰⁰ часов на заседании Научного совета PhD.03/04.06.2021.Fil.132.01 при Кокандском государственном педагогическом институте по адресу: 150700, г.Коканд, ул.Турон, 23. Тел: (99873) 542-38-38; Факс: (99873) 542-11-43; e-mail: quqondpi@umail.uz.

С диссертацией можно ознакомиться в Информационно-ресурсном центре Кокандского государственного педагогического института (зарегистрирована под №28) Адрес: 150700, г.Коканд, ул.Турон, 23. Тел.(99890) 508-64-42; e-mail:qdpi_arm@umail.uz)

Автореферат диссертации разослан «9» июля 2024 года.
(протокол рассылки №21 от «9» июля 2024 года)

М.Х.Хакимов
Председатель научного совета по присуждению ученых степеней, доктор филол.наук, профессор

А.Турахожаева
Ученый секретарь научного совета по присуждению ученых степеней, доктор философии, доцент

Д.Жамолиддинова
Заместитель председатель научного семинара при ученом совете по присуждению ученых степеней, доктор филол.наук,профессор.

Введение (Аннотация диссертации доктора философии (PhD))

Цель исследования – создание веб-программного обеспечения (на основе программ конкорданса, токенизатора, лемматизатора, разметки) для создания национального корпуса узбекского языка.

Объектом исследования выбраны варианты эпоса «Алпомиш» Фазиль Ёлдош угли (Ташкент: Шарк, 2010); Абдуназара Поёнова (Ташкент: Академнашр, 2018); Мардонагул Авлиягула (Ташкент: Фан, 2018).

В качестве **предмета исследования** были взяты модели, алгоритмы и программное обеспечение национального корпуса узбекского языка.

Научная новизна исследования заключается в следующем:

определены возможности создания национального корпуса посредством программного обеспечения Национального корпуса узбекского языка, обеспечения его работы в сети Интернет, лингвистической классификации текстов с помощью компьютера, изучения на основе этого национального языка каждого народа и его особенностей, и сохранения его чистоты;

через создание программного обеспечения *uzbekcorpora.uz,* предназначенного для создания и управления национальным корпусом узбекского языка, доказано, что осуществление поиска слов и словосочетаний по всему корпусу (конкорданс), определение разметки (морфологию слов), лемматизация, токенизация и создание частотных словарей являются его неотъемлемыми составляющими;

с лингвистической точки зрения доказано, что программы, предназначенные для создания корпуса, а также, служащие для его использования, выполняют такие задачи, как формирование текстовой базы, создание и редактирование корпусного словаря, классификация текстов.

доказано, что корпусное программное обеспечение обладает способностью в достаточной степени идентифицировать и изменять грамматические и стилистические особенности слов или предложений в любом тексте.

Внедрение результатов исследований. На основе научных результатов, полученных из опыта создания узбекского лингвистического корпуса (на основе программ конкорданса, токенизатора, лемматизатора, классификации):

выводы по вопросам создания Национального корпуса с использованием программного обеспечения Национального корпуса узбекского языка, лингвистической классификации текстов с помощью компьютера использованы в фундаментальном грантовом проекте ОТ-Ф1-030 по теме «Издание многотомной монографии (7 томов) «История узбекского языка»»» (Справка № 01/4-1716 от 9 сентября 2023 года Ташкентского государственного университета узбекского языка и литературы имени Алишера Навои). В результате была осуществлена оцифровка текстов, включение в корпус и улучшение исследований.

взгляды по созданию национального корпуса узбекского языка, поиск слов и словосочетаний в корпусе (конкордансности) с использованием программного обеспечения *uzbekcorpora.uz,* определению разметки (морфологии слов), лемматизации, токенизации и созданию частотных словарей использованы в практическом проекте JHBL-20 «Создание электронного корпуса художественных произведений на тему семьи, соседства и гендерного равенства» НИИ «Сообщество и семья» (справка 01-09/1608 от 4 сентября 2023 г. НИИ «Сообщество и семья»). В результате это послужило научно-практическим ресурсом при формировании электронного корпуса литературных произведений.

Материалы по задачам создания программ, предназначенных для создания корпуса и служащих для использования корпуса, создания и редактирования корпусного словаря, разметки текстов, в которых можно определять и изменять грамматические и стилистические особенности слов или предложений в любом тексте использовались при подготовке сценариев таких программ, как «Вести Края» областного телеканала «Самарканд» (справка № 01-07/138 от 20 апреля 2023 года Самаркандской областной телерадиокомпании). В результате предоставляемая информация обогатилась научными данными, а содержание передач совершенствовалось с образовательной и практической точек зрения.

Структура и объем диссертации. Диссертация состоит из введения, трех основных глав, аннотации, списка использованной литературы и приложений, ее общий объем составляет 120 страниц.

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

I bo‘lim (часть I; Part I)

1. M.S.Tursunov, O‘zbek tili milliy korpusini yaratishda dastlabki ma’lumotlar // Ilmiy axborotnoma, SamDU, 2022-yil 6-son (136), 82-88 betlar. [10.00.00. №6]
2. M.S.Tursunov, Alpomish dostoni matnlarining alfavit va chastotali lug‘atlari hamda statistik tahlillari // “O‘zbekiston: til va madaniyat” jurnali, TDO‘TAU, 2022-yil 3-son, 48-62 betlar. (O‘zbekiston Respublikasi Oliy attestatsiya komissiyasi (OAK) Rayosatining 2021-yil 30-oktabrdagi №308/6-sonli qarori)
3. A.B.Qarshiyev, S.A.Karimov, M.S.Tursunov, Development of a Modern Corpus of Computational Linguistics // Conference: 2020 International Conference on Information Science and Communications Technologies (ICISCT), DOI: 10.1109/ICISCT50599.2020.9351376, 2021. (SCOPUS №3)
4. M.S.Tursunov, Software of the national corpus of the Uzbek language // International Journal of Advance Scientific Research https://sciencebring.com/index.php/ijasr/article/view/457 Issue: Vol. 3 No. 10 (2023): Volume 03 Issue 10, Crossref DOI: https://doi.org/10.37547/ijasr-03-10-31. – P. 190-199. SJIF (2023) – 7.874 (№23)
5. M.S.Tursunov, Software development for Uzbek text corpus // The IX International Conference on Computer Processing of Turkic Languages “TurkLang 2021”, Russia, 21-23 september, 2021.
6. M.S.Tursunov, “O‘zbek tili matnlarini razmetkalash” // “kompyuter linvistikasi: muammolar, yechim va istiqbollar” mavzusidagi III an’anaviy xalqaro ilmiy-amaliy konferensiya, Toshkent:TDO‘TAU, 28.04.2023, 169-173 betlar.
7. M.S.Tursunov, O‘zbek tili milliy korpusini ishlab chiqish // “Raqamli texnologiyalar va suniy intellektni rivojlantirishning zamonaviy holati va istiqbollari” mavzusidagi respublika ilmiy-amaliy konferensiya, Guliston: GulDU 22-23.12.2022, 292-295 betlar.
8. A.B.Qarshiyev, S.A.Karimov, M.S.Tursunov, B.F.Xolmuxamedov, O‘zbek tili milliy korpusi uchun matnlarni formatlash dasturlari // “O‘zbek amaliy filologiyasi istiqbollari” mavzusidagi respublika ilmiy-amaliy konferensiya matyeriallari, Toshkent: ToshDO‘TAU, 26.10.2022, 179-184 betlar.

II bo‘lim (часть II; part II)

1. A.B.Qarshiyev, Sh.B.Maxmidov, M.S.Tursunov, O‘zbek tili morfologiyasini modellashtirish va dasturiy modullarni yaratish // journal of innovations in social sciences, https://sciencebox.uz/index.php/jis/article/view/7516, 2023, 3(6), 44–47 betlar. [ISSN 2181-2594]

2. A.B.Qarshiyev, M.S.Tursunov, B.F.Xolmuxamedov, Uzbekcorpora.uz: программное обеспечение Корпуса узбекского языка // «Родные языки и культуры в современном изменяющемся мире» электронный сетевой нуичный журнал, Тувинский государственный университет, 2022. 71-78 st.

3. M.S.Tursunov, Разработка программного обеспечения для текстового Корпуса на узбекском языке // «Родные языки и культуры в современном изменяющемся мире» электронный сетевой нуичный журнал, Тувинский государственный университет, №1 2022, 62-70 st.

4. A.B.Qarshiyev, S.A.Karimov, M.S.Tursunov, O‘zbek tili milliy korpusi uchun matnlarni formatlash // Raqamli texnologiyalarning nazariy va amaliy masalalari xalqaro jurnali, TATU Samarqand filiali, №1(1)2022, Samarqand. 58-63-c.

5. M.S.Tursunov, Description of the management system programs of the national corpus of the uzbek language // International journal of engineering mathematics, https://iejemta.com/index.php/em/article/view/9/9, volume 4 ISSUE 1, 2022, 32-42 p. [ISSN 1687-6156, impact factor 7.65]

6. A.B.Qarshiyev, S.A.Karimov, M.S.Tursunov, O‘zbek tili korpusining dasturiy ta’minotini yaratishning dastlabki natijalari // “Al-Xorazmiy avlodlari” jurnali, 2021 yil, 1-soni.

7. A.B.Qarshiyev, S.A.Karimov, M.S.Tursunov, The algorithm of morphological and lexical analysis of uzbek texts // Science, Research, Development #31, Warshava, 30-31.07.2020.

8. Karshiev A.B., Karimov S.A., Tursunov M.S., “uzbekcorpora.uz: collection, digitalization and formatting of texts for the national corpus of the uzbek language” // The 8th international conference “actual problems of applied mathematics and information technologies” - al-Khwarizmi 2023, Samarkand: SamSU, 25-26.09.2023, 258-259 pages.

9. S.A.Karimov, M.S.Tursunov, “uzbekcorpora.uz: o‘zbek tili milliy korpusida matn bilan ishlash” // “kompyuter lingvistikasining zamonaviy texnologiyalari - CTCL.2023” mavzusida vazirlik miqyosidagi ilmiy-amaliy anjuman, Toshkent:UzMU, 14.04.2023, 13-21 betlar.

10. A.B.Qarshiyev, S.A.Karimov, M.S.Tursunov, uzbekcorpora.uz: konkordans tuzish va uni tahlil qilish // “O‘zbek tilining milliy korpusi: muammolar va vazifalar” mavzusidagi xalqaro ilmiy-amaliy konferensiya, Samarqand:TATU SF, 27.03.2023, 30-36 betlar.

11. Tursunov M.S., Umirova S.M., Xolmuxamedov B.F., Ubaydullayev M.Sh., O‘zbek tili korpusida alpomish dostonining alfavit va chastotali lug‘atlari hamda statistik tahlillari // “O‘zbek tilining milliy korpusi: muammolar va vazifalar” mavzusidagi xalqaro ilmiy-amaliy konferensiya, Samarqand: TATU SF, 27.03.2023, 57-62 betlar.

12. A.B.Qarshiyev, S.A.Karimov, M.S.Tursunov, Software of the national corpus of the uzbek language // Республиканской научно-технической конференции “Современное состояние и перспективы развития цифровых технологий и искусственного интеллекта”, Samarkand, 26-27.10.2022, 1-qism, 114-121 betlar.

13. A.B.Qarshiyev, S.A.Karimov, M.S.Tursunov, uzbekcorpora.uz: matnlarni razmetkalash dasturlari // “O‘zbek tilining milliy korpusi: muammo va vazifalar” mavzusidagi xalqaro ilmiy-amaliy konferensiya materiallari, Toshkent: ToshDO‘TAU, 31.05.2022, Vol. 1 № 01 (2022), 206-210 betlar.

14. A.B.Qarshiyev, M.S.Tursunov, Sh.B.Maxmidov, O‘zbek tili milliy korpusini loyihalash // “Kompyuter lingvistikasi: muammolar, yechim, istiqbollar” mavzusidagi xalqaro ilmiy-amaliy konferensiya materiallari, Toshkent: ToshDO‘TAU, 22.04.2022, Vol. 1 № 01 (2022), 94-97 betlar.

15. A.B.Qarshiyev, S.A.Karimov, M.S.Tursunov, O‘zbek tili milliy korpusining dasturiy ta’minot strukturasi va vazifalari // “Kompyuter lingvistikasi: muammolar, yechim, istiqbollar” mavzusidagi xalqaro ilmiy-amaliy konferensiya materiallari, Toshkent: ToshDO‘TAU, 22.04.2022, Vol. 1 № 01 (2022), 82-88 betlar.

16. A.B.Qarshiyev, M.S.Tursunov, Software and its use for the uzbek language corpus // “Contemporary mathematics and its application” mavzusidagi xalqaro konferensiyasi, Toshkent, O‘zbekiston, 19-21.11.2021.

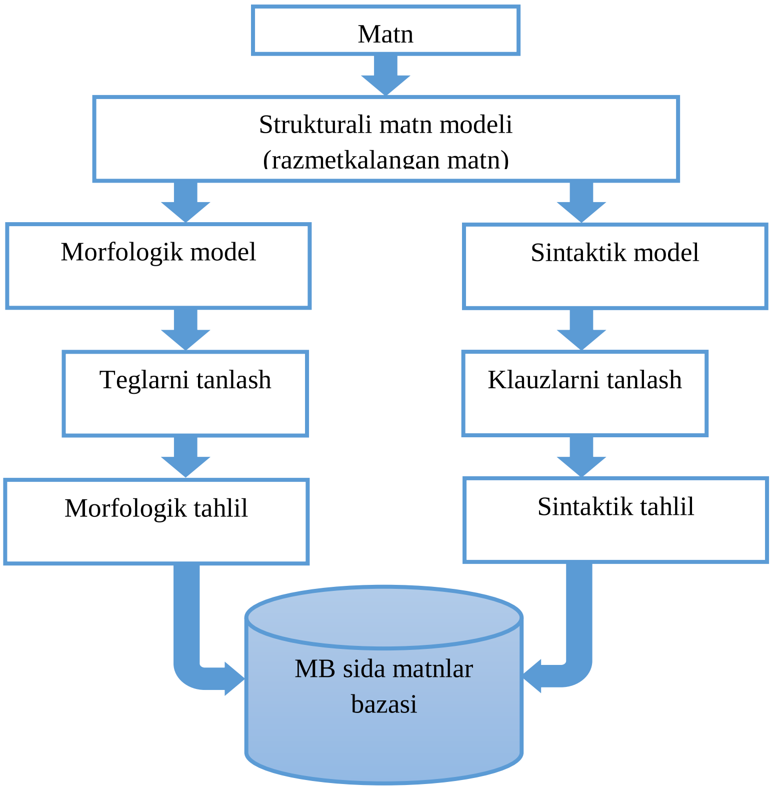
17. A.B.Qarshiyev, M.S.Tursunov, B.F.Xolmuxamedov, Uzbekcorpora.uz: uzbek language corpus software // The IX International Conference on Computer Processing of Turkic Languages “TurkLang 2021”, Russia, 21-23 september, 2021.

18. A.B.Qarshiyev, S.A.Karimov, M.S.Tursunov, O‘zbek tili korpusi: tokenayzer, lemmatayzer, razmetkalash dasturlarini tavsiflash va ulardan foydalanish // “O‘zbek milliy va ta’limiy korpuslarini yaratishning nazariy va amaliy masalalari” mavzusidagi xalqaro ilmiy-amaliy konferensiya materiallari, Toshkent: ToshDO‘TAU, 07.05.2021, 46-50 betlar.

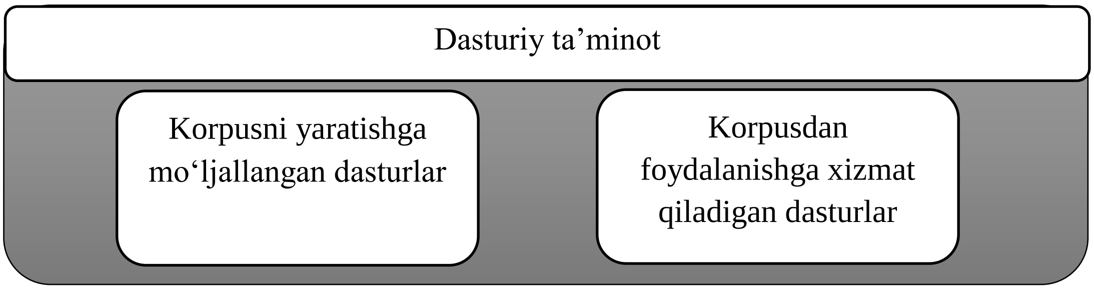
19. A.B.Qarshiyev, S.A.Karimov, M.S.Tursunov, B.F.Xolmuxamedov, O‘zbek tili korpusining dasturiy ta’minoti interfeysi va qidiruv tizimidan foydalanish // “Kompyuter lingvistikasi: muammolar, yechim, istiqbollar” mavzusidagi Respublika ilmiy-amaliy konferensiya, Toshkent: ToshDO‘TAU, 23.04.2021, 5-11 betlar.

Ilovalar
Appendixes
Приложения

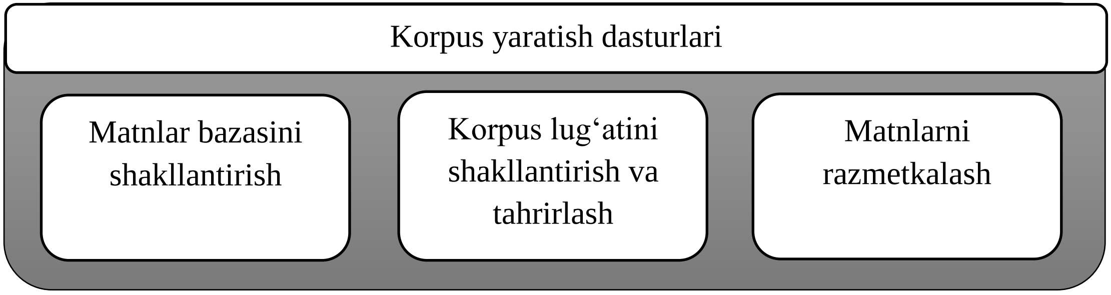
1-ilova. Korpus modelining sxematik tasviri



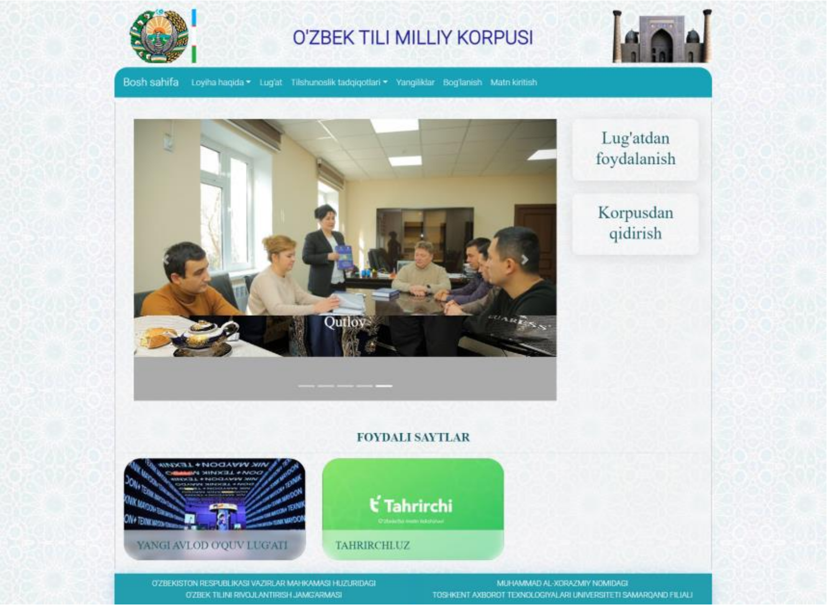
2-ilova. Dasturiy ta'minot tarkibi


Dasturiy ta'minot
Korpusni yaratishga mo'ljallangan dasturlar
Korpusdan foydalanishga xizmat qiladigan dasturlar

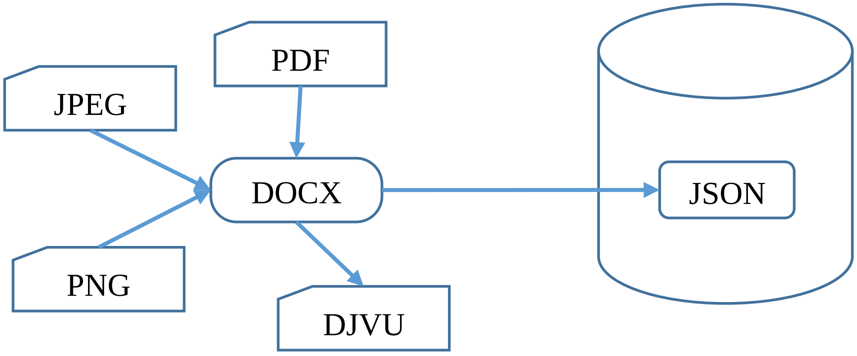
3-ilova. Korpus yaratish dasturlari tarkibiy qismlari


Korpus yaratish dasturlari
Matnlar bazasini shakllantirish
Korpus lug'atini shakllantirish va tahrirlash
Matnlarni razmetkalash

4-ilova. Dasturning asosiy sahifasi


O'ZBEK TILI MILLIY KORPUSI
Bosh sahifa
Loyiha haqida
Lug'at
Tilshunoslik tadqiqotlari
Yangiliklar
Bog'lanish
Matn kiritish
Lug'atdan foydalanish
Korpusdan qidirish
Qutloy
FOYDALI SAYTLAR
YANGI AVLOD O'QUV LUG'ATI
Tahrirchi
TAHRIRCHI.UZ
O'ZBEKISTON RESPUBLIKASI VAZIRLAR MAHKAMASI HUZURIDAGI O'ZBEK TILINI RIVOJLANTIRISH JAMG'ARMASI
MUHAMMAD AL-XORAZMIY NOMIDAGI TOSHKENT AXBOROT TEXNOLOGIYALARI UNIVERSITETI SAMARQAND FILIALI

5-ilova. Matnlarni korpusga saqlash formati

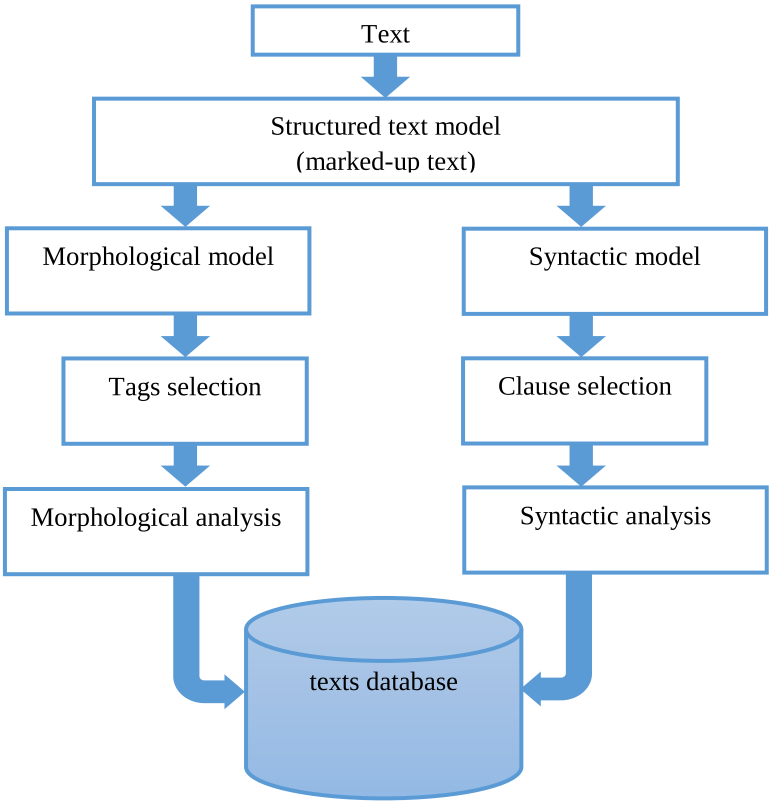


6-ilova. “Alpomish” dostonining Fozil Yo‘ldosh variant matnining so‘z turkumlariga nisbatan so‘zlarning ulushi

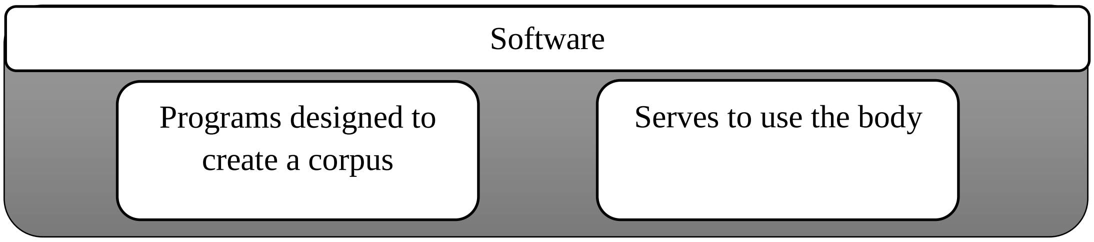
№	**Turkum**	**Matndagi so‘zlarni so‘z turkumlariga nisbatan ko‘rsatkichlari**			
		Asos	**%**	**So‘z shakl**	**%**
1	Fe’l	562	18,0	5168	35,9
2	Ot	1769	56,6	7027	48,8
3	Sifat	461	14,7	1112	7,7
4	Son	37	1,2	135	0,9
5	Olmosh	38	1,2	357	2,5
6	Ravish	160	5,1	412	2,9
7	Bog‘lovchi	14	0,4	24	0,2
8	Ko‘makchi	7	0,2	16	0,1
9	Yuklama	9	0,3	15	0,1
10	Modal so‘z	31	1,0	60	0,4
11	Undov so‘z	28	0,9	68	0,5
12	Taqlid so‘z	10	0,3	19	0,1
Jami		**3126**	**100**	**14413**	**100**

Appendixes

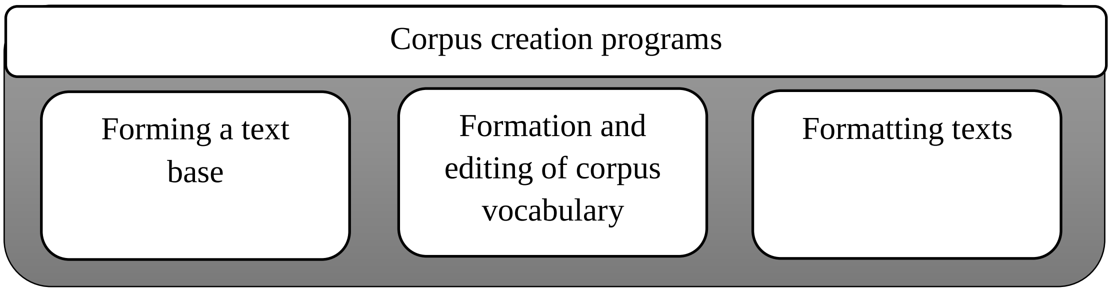
Application 1. Schematic representation of the corpus model



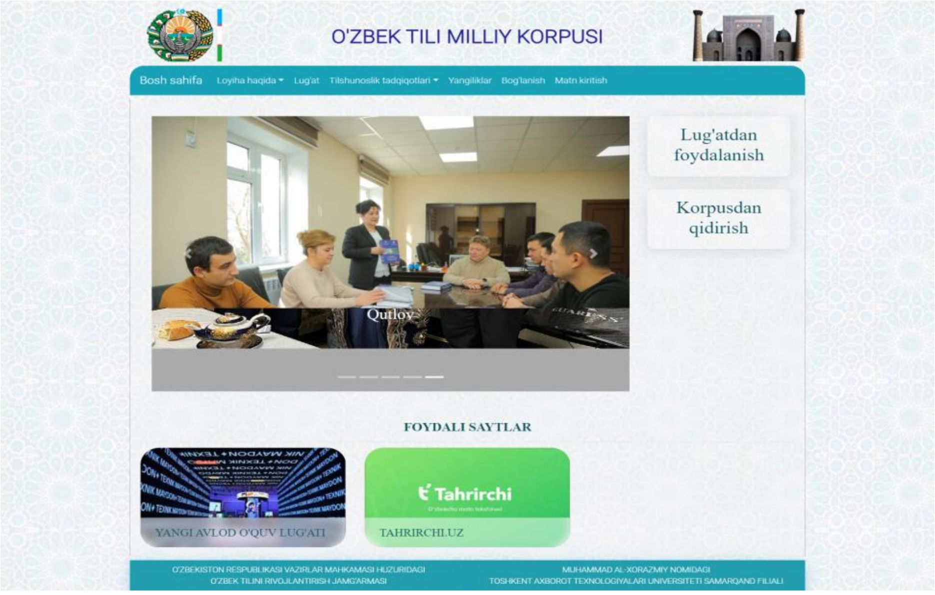
Application 2. Software content



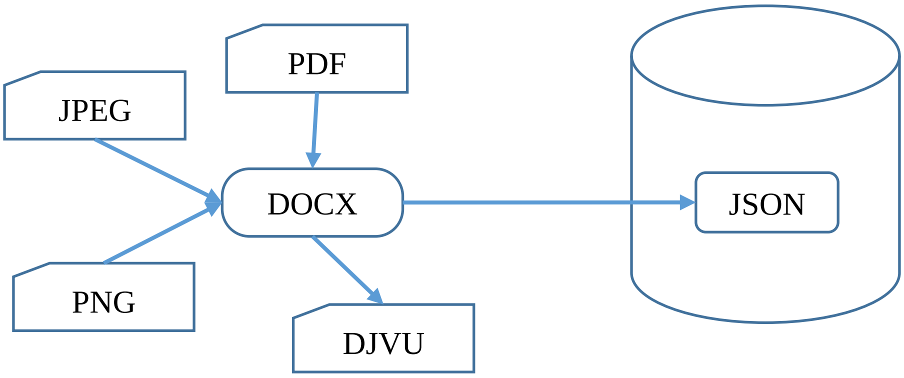
Application 3. Components of corpus building programs



Application 4. The main page of the program


O'ZBEK TILI MILLIY KORPUSI
Bosh sahifa
Loyiha haqida
Lug'at
Tilshunoslik tadqiqotlari
Yangiliklar
Bog'lanish
Matn kiritish
Lug'atdan foydalanish
Korpusdan qidirish
Qutloy
FOYDALI SAYTLAR
YANGI AVLOD O'QUV LUG'ATI
Tahrirchi
TAHRIRCHI.UZ
O'ZBEKISTON RESPUBLIKASI VAZIRLAR MAHKAMASI HUZURIDAGI O'ZBEK TILINI RIVOJLANTIRISH JAMG'ARMASI
MUHAMMAD AL-XORAZMIY NOMIDAGI TOSHKENT AXBOROT TEXNOLOGIYALARI UNIVERSITETI SAMARQAND FILIALI

Application 5. Format for storing texts in a corpus



Application 6. The percentage of words in relation to the word groups of Fazil Yoldosh ogli variant of the text of “Alpomish” epic

№	**Part of speech**	**Indicators of words in the text against word groups**			
		Stem	**%**	**Word form**	**%**
1	Verb	562	18,0	5168	35,9
2	Noun	1769	56,6	7027	48,8
3	Adjective	461	14,7	1112	7,7
4	Numeral	37	1,2	135	0,9
5	Pronoun	38	1,2	357	2,5
6	Adverb	160	5,1	412	2,9
7	Conjunction	14	0,4	24	0,2
8	Preposition	7	0,2	16	0,1
9	Particle	9	0,3	15	0,1
10	Modal word	31	1,0	60	0,4
11	Exclamatory word	28	0,9	68	0,5
12	Onomatopoeic word	10	0,3	19	0,1
Jami		**3126**	**100**	**14413**	**100**

Avtoreferat Qo‘qon davlat pedagogika instituti
“Qo‘qon DPI. Ilmiy xabarlar” tahririyat-nashriyot bo‘limida tahrirdan o‘tkazilgan.

Bosishga ruxsat etildi: 2024 y. Nashriyot bosma tabog‘i – 3,5.
Shartli bosma tabog‘i – 1,6. Bichimi 84x108 1/16. Adadi 100.
«Poligraf Super Servis» MCHJ
150114, Farg‘ona viloyati, Farg‘ona shahar, Aviasozlar ko‘chasi 2-uy.