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

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

NORBEKOVA MADINA SHUXRAT QIZI

O‘ZBEK TILI MILLIY KORPUSI UCHUN ILMIY-TEXNIK MATNLAR BAZASINI YARATISH

10.00.11– Til nazariyasi. Amaliy va kompyuter lingvistikasi

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

Toshkent – 2025

UO‘K: ***811.512.133’28***

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

Contents of Dissertation Abstract of the Doctor of Philosophy (PhD) in philological sciences

Оглавление автореферата диссертатсии доктора философии (PhD) по филологическим наукам

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

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

NORBEKOVA MADINA SHUXRAT QIZI

O‘ZBEK TILI MILLIY KORPUSI UCHUN ILMIY-TEXNIK MATNLAR BAZASINI YARATISH

10.00.11 – Til nazariyasi. Amaliy va kompyuter lingvistikasi

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

Toshkent – 2025

Falsafa doktori (PhD) dissertatsiyasi mavzusi O‘zbekiston Respublikasi Oliy ta’lim, fan va innovatsiyalar vazirligi huzuridagi Oliy attestatsiya komissiyasida B2024.4.PhD/Fil5438 raqam bilan ro‘yxatga olingan.

Dissertatsiya Alisher Navoiy nomidagi Toshkent davlat ozbek tili va adabiyoti universitetida bajarilgan.

Dissertatsiya avtoreferati uch tilda (o‘zbek, ingliz, rus (rezyume)) O‘zbekiston Milliy universitetini veb-sahifasining www.nuu.uz. hamda “Ziyonet” axborot-ta’lim portalining www.ziyonet.uz manzillariga joylashtirilgan.

Ilmiy rahbar:	**Raupova Laylo Rahimovna** filologiya fanlari doktori, professor
Rasmiy opponentlar:	**Toirova Guli Ibragimovna** filologiya fanlari doktori, professor
	Abduraxmanova Nilufar Zaynobiddin qizi filologiya fanlari doktori, professor
Yetakchi tashkilot:	Andijon davlat chet tillari instituti

Dissertatsiya himoyasi O‘zbekiston Milliy universiteti huzuridagi DSc.03/25.08.2021.Fil.01.16 raqamli Ilmiy kengashning 2025-yil “____” _____ soat___dagi majlisida bo‘lib o‘tadi (Manzil: 100174, Toshkent shahar, Olmazor tumani, Universitet ko‘chasi, 4-uy. Tel.: (99871) 246-02-24; faks: (99871) 246-02-24, e-mail: devonxona@nuu.uz.

Dissertatsiya bilan O‘zbekiston Milliy universitetining Axborot-resurs markazida tanishish mumkin (____ raqam bilan ro‘yxatga olingan). (Manzil: 100174, Toshkent shahar, Olmazor tumani, Universitet ko‘chasi, 4-uy. Tel.: (99871) 246-02-24; faks: (99871) 246-02-24, e-mail: devonxona@nuu.uz.

Dissertatsiya avtoreferati 2025-yil “____” _____ kuni tarqatildi.
(2025-yil “___” _____dagi _____ raqamli reyestr bayonnomasi).

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

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

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

KIRISH (falsafa doktori (PhD) dissertatsiyasi annotatsiyasi)

Dissertatsiya mavzusining dolzarbligi va zarurati. Jahon tilshunosligida tabiiy tilni qayta ishlash sohalari jadal rivojlanib, kompyuter lingvistikasi yutuqlari, jumladan, tilni tadqiq etish, lingvodidaktika, leksikografik tahlil hamda tarjima faoliyati kabi muhim yo‘nalishlarni turli aspektlarda o‘rganishga alohida e’tibor qaratilmoqda. Ayniqsa, kompyuter texnologiyalari va avtomatlashtirilgan tizimlardan foydalanish, tilning tuzilish va ma’no jihatlarini aniqlash, ta’lim jarayonini takomillashtirish, milliy korpuslar va elektron lug‘atlar yaratish, shuningdek, tarjima jarayonlarini samarali tashkil etishda ham muhim ahamiyat kasb etadi.

Dunyo tilshunosligida XXI asrda korpus lingvistikasini ilmiy va nazariy jihatdan tadqiq etish jarayonlari kengayib, turli yo‘nalishlarda muhim natijalarga erishish bo‘yicha izlanishlar olib borilmoqda. Jumladan, avtomatik tarjimani takomillashtirish, tillarni lingvistik modellash, so‘zlarni lemmalash tizimlarini yaratish, korpusli tahlil asosida sintaktik va semantik tuzilmalarni aniqlash, shuningdek, turli tillar uchun xos bo‘lgan milliy-madaniy meros xususiyatlarini raqamlashtirish va yoritishga alohida diqqat qaratilmoqda.

Mamlakatimizda amalga oshirilayotgan jadal islohotlar sharoitida o‘zbek tiliga zamonaviy axborot texnologiyalari asosida ishlov berish, uning milliy korpusini yaratish va lingvistik xususiyatlarini ilmiy-amaliy jihatdan o‘rganishga katta e’tibor berilmoqda. O‘zbek tiliga oid barcha ilmiy, nazariy va amaliy ma’lumotlarni o‘zida jamlagan elektron ko‘rinishdagi o‘zbek tili milliy korpusini yaratish zarur. Shu ma’noda, o‘zbek tili milliy korpusining yaratilishi, milliy-madaniy merosimizni raqamlashtirish va ta’lim jarayonlarida zamonaviy elektron platformalardan foydalanish, shuningdek, o‘zbek tili milliy korpusi uchun ilmiy-texnik matnlar bazasini yaratish bo‘yicha ilmiy-uslubiy strategik vazifalarni yoritib berish yuzasidan ilmiy tadqiqotlarni yanada chuqurlashtirish zarurati mavjud.

O‘zbekiston Respublikasi Prezidentining 2016-yil 13-maydagi PF-4997-son “Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universitetini tashkil etish to‘g‘risida”, 2017-yil 7-fevraldagi PF-4947-son “O‘zbekiston Respublikasini yanada rivojlantirish bo‘yicha Harakatlar strategiyasi to‘g‘risida”, O‘zbekiston Respublikasi Prezidentining 2019-yil 21-oktyabrdagi PF-5850-son “O‘zbek tilining davlat tili sifatidagi nufuzi va mavqeini tubdan oshirish chora-tadbirlari to‘g‘risida”gi Farmonlari, 2017-yil 17-fevraldagi PQ-2789-son “Fanlar akademiyasi faoliyati, ilmiy tadqiqot ishlarini tashkil etish, boshqarish va moliyalashtirishni yanada takomillashtirish chora-tadbirlari to‘g‘risida”, O‘zbekiston Respublikasi Vazirlar Mahkamasining 2019-yil 12-dekabrdagi 984-son “Davlat tilini rivojlantirish departamenti to‘g‘risidagi Nizomni tasdiqlash haqida”gi qarorlari hamda sohaga oid boshqa me’yoriy-huquqiy hujjatlarda belgilangan vazifalarni amalga oshirishda ushbu dissertatsiya muayyan darajada xizmat qiladi.

Tadqiqotning respublika fan va texnologiyalari rivojlanishining ustuvor yo‘nalishlariga mosligi. Tadqiqot respublika fan va texnologiyalar rivojlanishining “Axborotlashgan jamiyat va demokratik davlatni ijtimoiy, huquqiy, iqtisodiy,

madaniy, ma’naviy-ma’rifiy rivojlantirish, innovatsion iqtisodiyotni rivojlantirish” ustuvor yo‘nalishiga muvofiq bajarilgan.

Muammoning o‘rganilganlik darajasi. Jahon tilshunosligida korpus lingvistikasiga oid tadqiqotlar XX asrning 60-yillaridayoq boshlangan. Dastlabki tadqiqot ishlari va korpuslar ingliz tilida yaratilgan. 90-yillarga kelib esa dunyoning juda ko‘p tillarida korpuslar yaratildi. An’anaviy va kompyuter tilshunosligida korpus lingvistikasining o‘rni, metodologiyasi, obyekti va vazifalari G.Lich, R.Garsid, J.Sink1er, L.Floverdev, T.Makkenri, A.Hardi, M.Makkarti, V.N.Frensis, G.Kennidi kabi olimlarning tadqiqotlarida o‘rganilgan[1]. Korpuslarning ijtimoiy ahamiyati, korpus lingvistikasining keyingi taraqqiyot bosqichlari, internetdan korpus sifatida foydalanish, internet va korpus o‘rtasidagi o‘xshash hamda farqli jihatlar A.Kilgarrif, K.Styuart, G.Grefenstette, M.Hundt, N.Nesselhaf kabi tadqiqotchilar tomonidan mufassal yoritilib, masalaning nazariy asoslari tadqiq etilgan[3]. Rus tilshunosligida V.Plungyan, V.P.Zaxarov, A.B.Kutuzovlar tomonidan amalga oshirilgan qator tadqiqotlar korpus lingvistikasining alohida soha sifatida shakllanishida, rus milliy korpusining yaratilishida alohida ahamiyat kasb etadi. O‘zbek tilshunosligida korpus tilshunosligi sohasida ko‘pgina muhim tadqiqot ishlari olib borilgan. Xususan, Sh.Xamroyeva A.Eshmo‘minov, N.Abdurahmonova, G.Toirova, M.Abjalova, O.Abdullayeva, A.Raxmanova, N.G‘ulomova, R.Karimov, M.Xolova, E.Xonnazarov, Z.Xusainova, A.Turdaliyev, M.Tursunovlarning[2] ilmiy

[1] Leech G. Corpus Annotation Schemes. In Literary and Linguistic Computing. Vol. 8. No. 4. – Oxford University Press, 1993. – P. 275-281; Leech G., Wilson A. Recommendations for the morphosyntactic annotation of corpora / EAGLES Document EAG-TCWG-MAC/R, 1994 // www.i1c.cnr.it/EAGLES/browse.htm1; Leech G., Garside R., Steven E.A. The Automatic Grammatical Tagging of the LOB Corpus // ICAXE Ncwo, 1983. – P. 13-33 // https://www.researchgate.net/publication/238760957; Garside R., Leech G., Sampson G. The CLAWS Wordtagging System / The Computational Analysis of English: A Corpus-based Approach. – London: Longman, 1987; McEnery T., Wilson A. Corpus Linguistics (1st ed.). – Edinburgh: Edinburgh University Press, 1996; McEnery T.; Hardie A. Corpus linguistics: Method, theory and practice. – Cambridge: Cambridge Univcrsily Press, 2011; Francis W.N., Johansson S. Problems of Assembling and Computerizing Large Corpora // Computer Corpora in English Language Research. – Bergen: Norwegian Computing Centre for the Humanities, 1982; Francis W.N., Svartvik J. Language Corpora B.C. Directions in Corpus Linguistics, Proceedings of Nobel Symposium 82. – Stockholm, 1991; Kennedy G. An Introduction to Corpus Linguistics. – Harlow: Addison Wesley Longman, 1998; Sinclair J. Corpus, Concordance, Collocation. – Oxford: Oxford University Press, 1991; Flowcrdcw L. Corpus Linguistic Techniques Applied to Textlinguistics, 1998. – P. 541-552; McCarthy M., O'Keefe A. What arc corpora and how have they evolved? The Routledge handbook of corpus linguistics. – London and New York, 2/1.

[2] Хамроева Ш.М. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари. Филол. фан. б. фалс. док. … дисс. автореф. – Қарши, 2018. – 54 б.; Хамроева Ш. Ўзбек тили морфологик анализаторининг лингвистик таъминоти. Филол. фан. д-ри ... дисс. – Тошкент, 2021. – 268 б.; Эшмўминов А.А. Ўзбек тили миллий корпусининг синоним сўзлар базаси. Филол. фан. б. фалс. док. ... дисс. автореф. – Қарши, 2019. – 46 б.; Abdurahmonova N. O‘zbek tili elektron korpuslarining kompyuter modellari. Monografiya. – Toshkent, 2021. – 201 b.; Тоирова Г.И. Ўзбек тили миллий корпусини яратишнинг назарий ва амалий масалалари. Филол. фан. д-ри ... дисс. автореф. – Бухоро, 2021. – 72 б.; Abjalova M. Korpus lingvistikasi uslubiy qo‘llanma. – Toshkent: Bookmany print, 2022; Abdullayeva O. O‘zbek tilining internet axborot matnlari korpusini shakllantirishning nazariy va amaliy asoslari. Filol. fan. b. fals. dok. ... diss. – Toshkent, 2022. – 158 b.; Рахманова А. Ўзбек тили миллий корпусини яратишда компьютер усуллари. Филол. фан. б. фалс. док. ... дисс. – Тошкент, 2022. – 158 б.; G‘ulomova N. Alisher Navoiy mualliflik korpusi va uninng semantik teglari bazasini yaratish. Filol. fan. b. fals. dok. … diss. – Toshkent, 2022. – 189 b.; Karimov R. O‘zbek-ingliz parallel korpusini tuzishning lingvistik va dasturiy masalalari. Filol. fan. b. fals. dok. … diss. – Buxoro, 2022; Xolova M. O‘zbek milliy shevalari korpusi tadqiqi (Boysun tumani “j”lovchi shevalari misolida). Monografiya. – Termiz, 2022. – 144 b.; Xonnazarov E. O‘zbek tili korpusi uchun zamonni ifodalovchi grammatik shakllarni annotatsiyalash: Filol. fan. b. fals. dok. ... diss. – Toshkent, 2023. – 159 b.; Xusainova Z. O‘zbek tili birliklarini tokenlash, stemlash, lemmalashning lingvistik asoslari va dasturiy ta’minoti. Filol. fan. b. fals. dok. … diss. – Toshkent, 2023. – 158 b.; Turdialiyev A. Denov shevasini areal o‘rganish va til korpusiga joylashtirish. Filol. fan. b. fals. dok. … diss. – Toshkent, 2024. – 173 b.; Tursunov M. O‘zbek lingvistik

natijalari shular jumlasidandir. Korpus lingvistikasi o‘zbek tilshunosligida yangi va oxirgi yillarda jadal rivojlanayotgan soha bo‘lganligi bois monografik planda amalga oshirilgan ishlar juda kam. Xususan, Sh.Xamroyevaning o‘zbek tili mualliflik korpusini tuzishning lingvistik asoslari, N.G‘ulomovaning Alisher Navoiy mualliflik korpusini yaratishga bag‘ishlangan dissertatsiyasi[3] hamda shu nomdagi M.Abjalova hammuallifligida yozilgan monografiyasi[4]da korpus lingvistikasining shakllanishi, taraqqiyoti va nazariy asoslari, mualliflik korpusini tuzishning o‘ziga xos nazariy va amaliy jihatlari, shuningdek, mualliflik korpusi tuzishning umumiy va xususiy lingvistik asoslari bayon etilgan[5]. Oxirgi yillarda o‘zbek tilshunosligida ko‘plab tadqiqot ishlari e’lon qilinmoqda. Bunday tadqiqotlar amaliy jihatdan o‘zbek korpus lingvistikasida korpuslarning paydo bo‘lishiga asos bo‘ldi. O‘zbek tili milliy korpusi[6], O‘zbek tilining ta’limiy korpusi[7], O‘zbek tili korpusi[8], Alisher Navoiy korpusi[9] fikrimiz dalili bo‘lib xizmat qiladi. Dissertatsiyani yozish jarayonida nomlari ko‘rsatilgan va boshqa bir qator o‘zbek hamda jahon tilshunoslarining ilmiy izlanishlari e’tiborga olindi. Tadqiqotimizda mazkur yo‘nalishda bajarilgan ishlardan farqli ravishda, o‘zbek tili milliy korpusi uchun ilmiy-texnik matnlar bazasini yaratish monografik tarzda tekshirilgan.

Tadqiqotning dissertatsiya bajarilgan oliy ta’lim muassasasining ilmiy tadqiqot ishlari bilan bog‘liqligi. Dissertatsiya Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti ilmiy tadqiqot rejasining 2022-2024-yillarga mo‘ljallangan “Tilning ijtimoiy, tarixiy va zamonaviy taraqqiyoti” mavzusi doirasida bajarilgan.

Tadqiqotning maqsadi. O‘zbek tili milliy korpusi uchun ilmiy-texnik matnlar bazasini shakllantirishning nazariy va dasturiy asoslarini ishlab chiqish, ilmiy-texnik matnlardagi leksik birliklarni lingvistik teglash va o‘zbek tilidagi ilmiy-texnik matnlar bazasini tasniflashdan iborat.

Tadqiqotning vazifalari:

- korpus lingvistikasining nazariy va metodologik asoslarini tahlil qilish, uning shakllanish bosqichlari va rivojlanish tendensiyalarini yoritish, shu bilan birga, milliy korpus yaratishda muhim hisoblangan ilmiy qarash va yondashuvlarni asoslash;
- o‘zbek tili ilmiy-texnik matnlar bazasini yaratishda zarur bo‘lgan lingvistik mezonlarni belgilash, bu sohadagi nazariy maqsad va vazifalarni ilmiy asosda tushuntirish, ilmiy-texnik matnlarning fanlararo xususiyatlarini hisobga olgan holda dalillash;

---

korpusini yaratish tajribasidan (konkordans, tokenayzer, lemmatayzer, razmetkalash dasturlari asosida). Filol. fan. b. fals. dok. … diss. avtoref. – Qo‘qon, 2024. – 60 b.

[3] G‘ulomova N. Alisher Navoiy mualliflik korpusi va uning semantik teglari bazasini yaratish (“Badoye’ ul-vasat” devoni asosida). Filol. fan. b. fals. dok. … diss. – Toshkent, 2023. – 190 b.

[4] Abjalova M., G‘ulomova N. Mualliflik korpuslari: Alisher Navoiy mualliflik korpusi. Monografiya / M.A.Abjalova, N.S.G‘ulomova. – Navoiy, 2023. – 216 b.

[5] Хамроева Ш.М. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари. Филол. фан. б. фалс. док. ... дисс. автореф. – Қарши, 2018. – 54 б.

[6] http://uzbekcorpora.uz/ijrochi

[7] https://uzschoolcorpara.uz/

[8] https://uzbekcorpus.uz/

[9] http://v1.alishernavoicorpus.uz/korpus_haqida/

- o‘zbek tili milliy korpusi uchun ilmiy-texnik matnlar bazasini yaratishda zamonaviy kompyuter-texnologik yondashuvlarni tanlash, korpus tuzish jarayonida qo‘llaniladigan dasturiy vositalarni tahlil qilish va ularni tizimlashtirish;
- matnlarni lingvistik annotatsiyalash algoritmini ishlab chiqish, teglash jarayonining nazariy-amaliy jihatlarini yoritish, uning samaradorligini tajriba natijalari orqali belgilash.

Tadqiqotning obyekti sifatida o‘zbek tili milliy korpusi uchun ilmiy-texnik matnlar bazasini yaratishga doir o‘zbek tilidagi turli ilmiy va texnik matnlar tanlangan.

Tadqiqotning predmetini o‘zbek tilida ilmiy-texnik matnlar bazasini shakllantirish, ularni teglash muammolari tashkil etadi.

Tadqiqot usullari. Tadqiqot mavzusini yoritishda tasniflash, tavsiflash, qiyoslash, statistik tahlil metodlaridan foydalanildi.

Tadqiqotning ilmiy yangiligi:

korpus va uning lingvodidaktika hamda lingvistika sohalaridagi ma’lumotlarni markazlashtirish va qayta ishlash imkoniyatlari, shuningdek, korpusdan samarali foydalanishning ahamiyati asoslangan;

o‘zbek tili milliy korpusidagi ilmiy-texnik matnlarni teglashda ta’lim, sport, sog‘liqni saqlash, siyosat, madaniyat, ob-havo, iqtisodiyot, texnologiya sohalaridagi ilmiy matnlarning LDA, LSA, NMF usullari va algoritmlari vositasida http://topicmodel.uz/ dasturiy ta’minoti ishlab chiqilganligi isbotlangan;

o‘zbek tili ilmiy-texnik matnlarini teglash usullaridan matnni tushunish, ma’lumot olish, hissiyotlarni tahlil qilish, imloni tekshirish, hujjatlarni umumlashtirish va mashina tarjimasi kabi NLP ilovalarini ishlab chiqishda foydalanish mumkinligi dalillangan;

matnlar bazasini shakllantirishda ilmiy-texnik matnga qo‘yiladigan talablar aniqlangan, ilmiy-texnik matnlarni korpusga joylashtirishda foydalaniladigan teglar ishlab chiqilgan va o‘zbek tili milliy korpusida semantik annotatsiyalash imkonini beruvchi ilmiy-texnik matnlardagi o‘ziga xosliklar tahlillar asosida aniqlangan.

Tadqiqotning amaliy natijalari:

- o‘zbek tili milliy korpusi uchun ilmiy-texnik matnlar bazasini yaratish maqsadida, til va ma’noga oid olib borilgan tadqiqotlar hamda tajribalar orqali, korpusning ilmiy va texnik aspektlariga oid alohida lingvistik usullar ishlab chiqilgan;
- ko‘p sonli ilmiy-texnik matnlarning ma’lumotlar bazasini yaratish jarayonida, leksik va grammatik strukturalar, shuningdek, terminologiyaning aniqlashtirilishi, yuqori darajadagi aniqlik va standartlashtirishga yo‘naltirilgan yangi metodikalar yaratilgan;
- ko‘p turli sohalardagi ilmiy-texnik matnlarning to‘g‘ri teglanishi va annotatsiyalanishi orqali, ularni korpusda yangi ilmiy tadqiqotlar uchun xizmat qiluvchi manbaga aylantirishning muhim ahamiyati asoslangan;
- o‘zbek tili milliy korpusi uchun ilmiy-texnik matnlar bazasining tashkil etilishi, uning yangi kompyuter-texnologiyalari orqali samarali ishlashini ta’minlovchi algoritmlar ishlab chiqilishi, shuningdek, korpusdagi ma’lumotlarning xato tahlili, faol aniqlash va optimizatsiya qilinishi aniqlangan.

Tadqiqot natijalarining ishonchliligi muammoning aniq qo‘yilgani, chiqarilgan ilmiy xulosalarning tavsifiy, qiyosiy tahlillar bilan asoslangani va ularning amaliyotga joriy etilgani, tahlil usullari vositasida asoslangani, nazariy fikr va xulosalarning amaliyotga joriy qilingani, olingan natijalarning vakolatli tashkilotlar tomonidan tasdiqlangani bilan izohlanadi.

Tadqiqot natijalarining ilmiy ahamiyati. Tadqiqot natijalarining ilmiy ahamiyati o‘zbek tili milliy korpusi uchun ilmiy-texnik matnlar bazasini yaratish, uning leksik-semantik va grammatik strukturasini aniqlash, shuningdek, ma’lumotlarni ishlab chiqish va avtomatlashtirishdagi axborot texnologiyalarining samaradorligini baholashga oid nazariy xulosalardan tilshunoslik hamda axborot texnologiyalari yo‘nalishlaridagi ishlarda manba sifatida foydalanish mumkinligi bilan belgilanadi.

Tadqiqot natijalarining amaliy ahamiyati. Tadqiqot natijalarining amaliy ahamiyati o‘zbek tili milliy korpusi uchun ilmiy-texnik matnlar bazasini yaratish, uning semantik va grammatik strukturasini aniqlash, shuningdek, matnni avtomatik ishlash va tahlil qilish usullarini ishlab chiqishda, “Kompyuter lingvistikasi”, “Matematik modellashtirish”, “Matnshunoslik nazariyasi” “Leksikografiya” fanlaridan mashg‘ulotlar olib borishda, o‘quv qo‘llanma, darslik, majmualar yaratishda hamda ilmiy va texnik lug‘atlar, ma’lumotnomalar tuzishda foydalanish mumkinligi bilan asoslanadi.

Tadqiqot natijalarining joriy qilinishi. O‘zbek tili milliy korpusi uchun ilmiy-texnik matnlar bazasini yaratish bo‘yicha olingan ilmiy natijalar asosida:

- ilmiy texnik matnlarning tarjima variantlaridan foydalanishda bu tur matnlarning informatsion, yaxlitlik, izchillik, terminologik xususiyatlariga e’tibor qaratish xususidagi xulosalardan PF-201912258 – “O‘zbek adabiyotining ko‘p tilli (o‘zbek, rus, ingliz tillarida) elektron platformasini yaratish” mavzusidagi amaliy loyihada (2021-2023) foydalanilgan (Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universitetining 2024-yil 6-dekabrdagi 04/1-4044-raqamli ma’lumotnomasi). Natijada elektron platformaga oid ilmiy-texnik matnlar bazasi yangi ma’lumotlar bilan boyitilgan;
- dissertatsiyani tayyorlash jarayonida shakllantirilgan dasturiy ta’minot ma’lumotlar bazasidan, teglash tizimidan, annotatsiyalash tamoyillarining xulosalaridan AM-Ф3-201908172 raqamli “O‘zbek tilining ta’limiy korpusini yaratish” mavzusidagi amaliy loyihada (2020-2023) foydalanilgan (Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universitetining 2024-yil 10-dekabrdagi 04/1-4085-raqamli ma’lumotnomasi). Natijada tadqiqotda keltirilgan ilmiy-texnik materiallar, tematik modellashtirish algoritmlari, yashirin semantik tahlil haqidagi takliflar korpus mazmuni va amliy jihatini boyitishga xizmat qilgan;
- tadqiqot natijalarining amaliy ahamiyati o‘zbek tili milliy korpusi uchun ilmiy-texnik matnlar bazasini yaratish, uning semantik va grammatik strukturasini aniqlash, shuningdek, matnni avtomatik ishlash va tahlil qilish usullarini ishlab chiqishdagi xulosalardan 2024-yilning 24-dekabrdagi 06-28-845-sonli ma’lumotnoma asosida O‘zbekiston milliy teleradiokompaniyasining “O‘zbekiston tarixi” telekanalining “Hamma uchun” dasturi ssenariysini yozishda foydalanilgan.

Tadqiqot natijalarining aprobatsiyasi. Ushbu tadqiqot natijalari yuzasidan 3 ta xalqaro, 2 ta respublika ilmiy-amaliy konferensiyasida ma’ruza qilingan.

Tadqiqot natijalarining e’lon qilinganligi. Dissertatsiya mavzusi bo‘yicha jami 10 ta ilmiy ish chop etilgan. Shulardan, uchtasi O‘zbekiston Respublikasi Oliy attestatsiya komissiyasining doktorlik dissertatsiyalari asosiy ilmiy natijalarini chop etish tavsiya etilgan ilmiy nashrlarda, uchtasi xorijiy jurnallarda va tezislarda ifodalangan.

Dissertatsiyaning tuzilishi va hajmi. Dissertatsiya kirish, uch bob, xulosa, foydalanilgan adabiyotlar ro‘yxatidan iborat bo‘lib, umumiy hajmi 158 sahifani tashkil etadi.

DISSERTATSIYANING ASOSIY MAZMUNI

Tadqiqotning kirish qismida dissertatsiya mavzusining dolzarbligi va zarurati asoslangan, tadqiqotning maqsad va vazifalari, obyekti, predmeti, respublika fan va texnologiyalari rivojlanishining ustuvor yo‘nalishlariga bog‘liqligi ko‘rsatilgan, ilmiy yangiligi va amaliy ahamiyati yoritilgan, tadqiqotning joriylanishi, nashr etilgan ishlar va dissertatsiya tuzilishi haqida ma’lumot berilgan.

Dissertatsiyaning **I bobi “*O‘zbek tili milliy korpusida ilmiy-texnik matnlar bazasini berishning lingvistik asoslari*”** deb nomlanib, 4 bo‘limdan iborat.

1.1-bo‘lim. Jahon tilshunosligida matn hodisasining tadqiq etilishi va taraqqiyotiga bag‘ishlangan. Jahon tilshunosligida XX asrning 70-yillardan boshlab matn tilshunosligi yo‘nalishi asta-sekin rivojlana boshlagan bo‘lsa ham matn masalasi bilan bog‘liq qarashlar 40-yillarda o‘rtaga chiqqan. Dastlab g‘arb tilshunosligida matn va uning tabiatini o‘rganish masalalariga jiddiy e’tibor qaratildi. Rus tilshunosligida matn (“tekst”) tushunchasining paydo bo‘lishi o‘tgan asrning 40-yillariga to‘g‘ri keladi. 1947-yilda A.I.Belich[10] o‘zining tilshunoslik fanlarining tasnifiga bag‘ishlangan maqolasida til faktlarining grammatik tavsifida ma’no umumiyligi asosida bog‘langan va muayyan sintaktik-semantik yaxlitlik tarzida namoyon bo‘ladigan gaplarning butun zanjiriga alohida o‘rin berilishi lozimligiga hamda bu “matn” tushunchasining yuzaga kelishi uchun hal qiluvchi ahamiyatga molik ekanligiga e’tiborni qaratgan, ana shunday gaplar zanjiridagi o‘zaro munosabat va aloqalarni tilshunoslikning sintaksis bo‘limida o‘rganishning maqsadga muvofiqligini ta’kidlagan. A.I.Belich bilan bir qatorda nemis tilshunos olimlari ham 40-yillarda matn haqida o‘z mulohazalarini aytib o‘tadi. Ba’zi adabiyotlarda I.A.Boduen de Kurtene asarlari matn borasidagi qarashlarning poydevori bo‘lib, u keyinchalik M.M.Baxtin[11], V.V.Vinogradov[12], L.V.Sherba[13] kabi bir qator olimlarning tadqiqot markaziga aylanganligi aytiladi. Biroq uzoq vaqt davomida matn tadqiqotchilarni, birinchi navbatda, mazmun (semantik) nuqtayi nazaridan qiziqtirdi (A.M.Peshkovskiy, L.V.Sherba, V.V.Vinogradov, A.Vayl, V.Mathesius, S.Balli, Z.Xarris). XX asrning 60-yillarida matn tilshunosligida,

[10] Белич А.И. К вопросу о распределении грамматического материала по главным грамматическим дисциплинам / Вестник МГУ, 1947. – № 7. – С. 24.

[11] Бахтин М.М. Проблема текста в лингвистике, филологии и других гуманитарныхнауках. Опыт философского анализа / М.М.Бахтин // Литературнокритические статьи. – М.: Художественная литература, 1986.

[12] Виноградов В.В. Стилистика. Теория поэтической речи. Поэтика / В.В.Виноградов. – М.: АН СССР, 1963.

[13] Щерба Л.В. Избранные работы по русскому языку / Л.В.Щерба. – М., 1957.

asosan, matnning izchilligi va tushunarliligini saqlash yo‘llari, shaxs hamda subyekt afzalligini bildirish usullari (anaforik tuzilmalar, pronominalizatsiya, lug‘aviy takrorlashlar, zamon zanjiri va boshqalar), gapning mavzu va rememasiga muvofiq taqsimlanishi o‘rganilgan. 1960-yillarda G‘arbiy Yevropada chex maktabining bevosita va bilvosita ta’siri ostida matnning lingvistik nazariyasi shakllana boshladi. Bu an’anada matn dastlab strukturaviy usullar bilan, ko‘proq tanish til obyektlari bilan o‘xshashlik asosida tahlil qilingan. Matn nazariyasi, matn tavsifi, matn lingvistikasining umumiy shakllanishi va rivojlanishida chex (Praga lingvistik to‘garagi vakillari), nemis, fransuz, ingliz, Amerika, golland, polyak, rus va boshqa tilshunoslik maktablari vakillarining xizmatlari dunyo tilshunosligida e’tirof etilgan hamda doimiy ravishda ilmiy tadqiqotlarda tilga olinadi. Rus tilshunosligida matn nazariyasi va lingvistikasi muammolari V.V.Odinsov, I.R.Galperin, O.I.Moskalskaya, L.M.Loseva, Y.M.Lotman, Z.Y.Turayeva, N.D.Zarubina, E.V.Sidorov, O.L.Kamenskaya, A.I.Gorshkov, N.S.Valgina[14] kabi ko‘plab tilshunoslar tomonidan o‘rganilgan. Matn va uni tashkil etuvchi unsurlari, omillari, xususiyatlari turli nuqtai nazardan tadqiq etilgan. Bu soha atrofida o‘ziga xos, aytish mumkin bo‘lsa, juda katta munozaralar paydo bo‘ldi. Hatto ayrim mutaxassislar matn lingvistikasini tilshunoslikning alohida sohasi emas, balki umuman, tilshunoslikning poydevori, ya’ni muhim bazasi deb hisobladilar. Tilshunoslikda bu yo‘nalishdagi tadqiqotlarni atroflicha tahlil qilgan va tom ma’noda matn lingvistikasi tadqiqi hamda bu tadqiqotlarning nechog‘li ahamiyatliligini tushuntirgan tilshunos O.I.Moskalskaya o‘tgan asrning 60-70-yillariga kelib matnni lingvistik o‘rganishga bo‘lgan qiziqish kuchayganini, dunyo tilshunosligida matn lingvistikasi bo‘yicha juda ko‘p miqdorda tadqiqotlar yuzaga kelganligini va matn lingvistikasi mustaqil tilshunoslik fani sifatida to‘la e’tirof etilganligini ta’kidlaydi.

1.2-bo‘lim. O‘zbek tilshunosligida matn hodisasining tadqiq etilishi va tarixiy taraqqiyoti deb nomlangan. O‘zbek tilshunosligida matn va uning turlari tadqiqi bir necha bosqichda olib borilganini ko‘rish mumkin. Matn tadqiqi sohasidagi tadqiqotlar, asosan, matnlarning semantik, sintaktik, kommunikativ vazifalari, strukturalarini, til va adabiyotshunoslikda rolini o‘rganishga qaratilgan. Ularni quyidagicha bosqichlarda ko‘rib chiqishni lozim topdik. 1. Matn tushunchasining aniqlanishi. Matn turlarini aniqlash va ularni vazifalariga ko‘ra ajratishda matnning o‘ziga xos xususiyatlariga, matn qanday omillarga ko‘ra shu maqomga ega ekaniga ahamiyat qaratish lozim. Matnshunoslikning boshlang‘ich bosqichi matn tushunchasining aniqlanishidan boshlanadi. Bu bosqichda olimlar matnning o‘ziga xos xususiyatlarini, uning strukturaviy va lingvistik elementlarini tahlil qilishga kirishadilar. Bunda muammolar “Matn nima?” va “Matnni qanday turlarga ajratish

[14]Лотман Ю.М. Структура божественного текста. – М.: Искусство, 1970; Лингвистика текст. Материалы научной конференции. Ч. я, II. – М., 1974; Лосева Л.М. Как строится текст. – М.: Просвещение, 1980; Одинцов В.В. Стилистический текст. – М.: Наука, 1980; Москальская О.И. Грамматика текста. – М.: Высшая школа, 1981; Гальперин И.Р. Показанные работы; Зарубина Н.Д. Текст: лингвистические и методологические аспекты. – М.: Русский язык, 1981; Тураева З.Я. Лингвистика – это текст. – М.: Просвещение, 1986; Сидоров Е.В. Проблема в речевой систематичность. – М.: Наука, 1987; Каменская О.Л. Текст и общение. – М.: Высшая школа, 1990; Горшков А.И. Русская стилистика. – М.: Астрель-АСТ, 2001. – С. 63-250; Валгина Н.С. Теория в тексте. – М.: Логос, 2004 и др.

mumkin?” degan savollar orqali yechimini topadi. Bu bosqichda matnning tilshunoslikda tutgan o‘rni va ahamiyati aniqlanadi. 2. Matn turlarining tasnifi**.** Ikkinchi bosqichda matnlar tasniflanadi. Bunda matnlarning turlari, ularning xususiyatlari va kommunikativ vazifalari aniqlanadi. Bu bosqichda matnlarni turlarga ajratishning me’yoriy va metodologik asoslari ishlab chiqiladi. 3. Matnshunoslikning amaliy tadqiqotlari**.** Matn turlarini amaliy jihatdan o‘rganish bosqichida, tilshunoslar matnlarning til va stilistik xususiyatlarini tahlil qilishadi. Bu bosqichda matnlar: grammatik va leksik jihatdan o‘rganiladi. Bunda matnning og‘zaki va yozma shakldaligiga ko‘ra kommunikativ vazifasi (maqsadi) hamda auditoriyasi hisobga olinadi. Matnlarni tarjima qilish, tahlil qilish va stilistik o‘ziga xosliklarini aniqlash, bog‘lovchi vositalar tahlili amaliy ishlarda qo‘llaniladi. 4. Matn turlarining o‘zgarishi va rivojlanishi. Til ijtimoiy jarayon hosili ekan o‘z-o‘zidan matnlar ham o‘zgaruvchan birlikdir. Matn turlarining rivojlanishi va o‘zgarishi ham o‘rganiladi. Bu bosqichda, olimlar matnlarning madaniy, ijtimoiy va tilshunoslikdagi o‘zgarishlar sababli qanday yangi shakllarga kirishini tahlil qiladilar. Masalan, aytaylik, internetda paydo bo‘lgan yangi matn turlari (bloglar, forum postlari, veb-sahifalar). Yangi texnologiyalar, media va kommunikatsiyalar bilan bog‘liq matn shakllari. 5. Matnshunoslikning integratsiyalashgan yondashuvi. Matnshunoslikda yangi yondashuvlar va metodlar ishlanadi. Bu bosqichda matnlar faqat lingvistik emas, balki psixologik, sotsial, va kulturologik nuqtayi nazardan ham o‘rganiladi. Shuningdek, matn turlarini boshqa sohalar bilan, masalan, adabiyotshunoslik**,** madaniyatshunoslik va axborot kommunikatsiya sohalari bilan bog‘lashda yangi yo‘nalishlar paydo bo‘ladi. 6. Matnshunoslikning zamonaviy metodlari. Bu bosqichda matnshunoslikda zamonaviy texnologiyalar va metodologiyalar qo‘llaniladi. Korpus lingvistika va kompyuter tahlili (matnlarni avtomatik ravishda tahlil qilish) kabi yangi yo‘nalishlarda ish olib boriladi. Matnning intertekstual tahlili, semiotik yondashuvlar va boshqa zamonaviy tahlil metodlari ishlatiladi. Bu bosqichlar bo‘yicha olib borilayotgan ilmiy izlanishlarda muhim ilmiy natijalarga erishilayotganini guvohi bo‘lmoqdamiz. Matn hodisasi A.Mamajonov, N.Mahmudov, M.Hakimov, M.Yo‘ldoshev, N.Turniyozov, B.Yo‘ldoshev, M.Abdupattoyev, M.Qurbonova kabi olimlar tomonidan yuqorida qayd etilgan aspektlarda o‘rganilgan. Matnning turlari, aynan ilmiy matnning sintaktik va pragmatik muammolari yuzasidan ish olib borgan o‘zbek tilshunosi M.Hakimov hisoblanadi. U nomzodlik dissertatsiyasida ilmiy matn va uning birliklari orasidagi mazmuniy munosabatni ifoda etuvchi bog‘lovchilar, ularning o‘ziga xos xususiyatlari va vazifalarini yoritib bergan. Ilmiy matnda muallifning xususiy munosabatini yoritish hamda uning turlarini ajratish bilan birga, ilmiy matnning sintagmatik va pragmatik xususiyatlarini faktik ma’lumotlari asosida tadqiq etgan[15]. O‘zbek tilshunosligida matnni kompyuter dasturlari orqali tahlil qilish sohasida olib borilgan ishlar bir qator ilmiy tadqiqotlar va texnologik yondashuvlar asosida amalga oshirilgan. Bu sohada, asosan, o‘zbek tilining sintaktik, morfologik, semantik tahlilini kompyuter yordamida avtomatlashtirishga

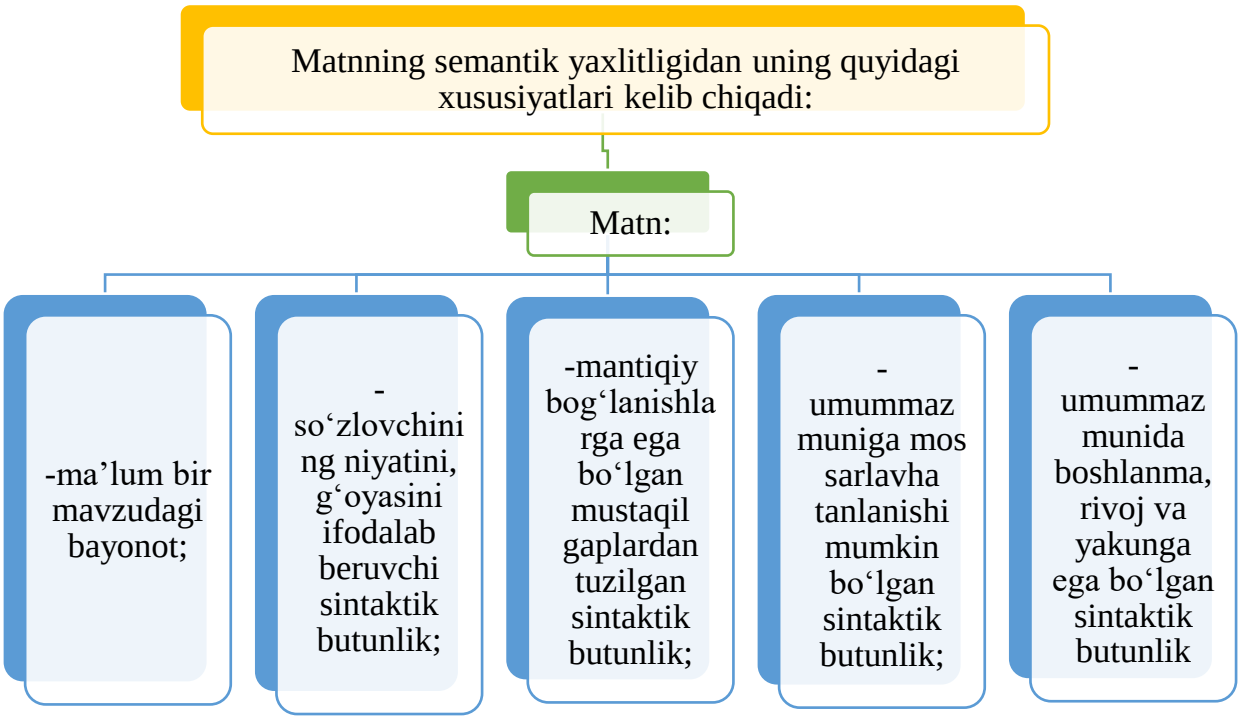
[15] Ҳакимов М. Ўзбек илмий матнининг синтагматик ва прагматик хусусиятлари. Филол. фан. номз. ... дисс. – Тошкент, 1993.

qaratilgan tadqiqotlar olib borilgan. Masalan, Sh.Xamroyevaning “O‘zbek tili morfologik analizatorining lingvistik ta’minoti” mavzusidagi doktorlik dissertatsiyasida til kategoriyalarini avtomatik morfologik tahlilini amalga oshirishning umumiy tamoyillari belgilab berilgan. N.Abdurahmanovaning “O‘zbek tili elektron korpuslarining kompyuter modellari” nomli monografiyasi o‘zbek tilini kompyuter yordamida tahlil qilish va qayta ishlash bo‘yicha olib borilgan ilmiy tadqiqlarni o‘z ichiga oladi. Ushbu monografiyada, asosan, o‘zbek tilining elektron korpuslarini yaratish, ularning kompyuter modellari yordamida ishlash imkoniyatlari va metodologiyasi haqida so‘z boradi.

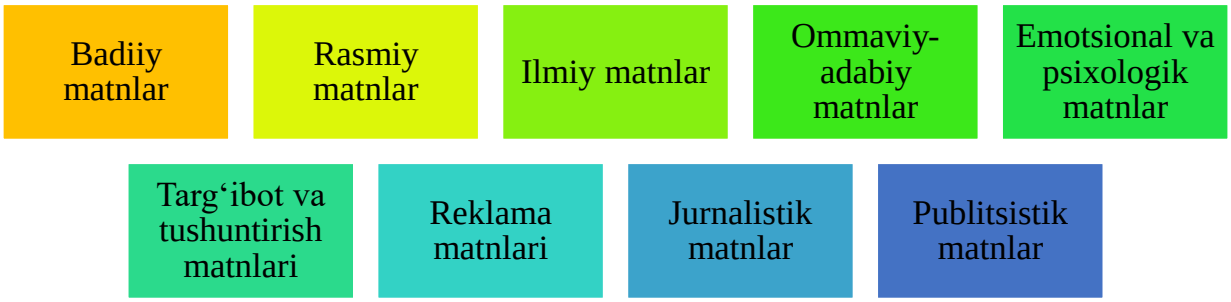
Kompyuter lingvistikasi texnologiyalari matnlarni avtomatik ravishda tahlil qilish, tasniflash va ularning strukturasini aniqlash imkonini beradi. Bu, o‘z navbatida, matn turlarini o‘rganish va ularning xususiyatlarini tushunishda yangi metod hamda yondashuvlarni yaratishga yordam beradi.

1.3-bo‘lim. Matn va uning turlari tavsifi va talqini deb nomlangan.

Matn tavsifi – bu matnning mazmuni, tuzilishi, til va stilistik xususiyatlarini tahlil qilish hamda izohlash jarayonidir. Matn tavsifi matnning o‘ziga xos xususiyatlarini, uning tarkibini va tuzilishini tushunishga yordam beradi. Bu jarayon matnning sifatini, uning o‘quvchiga qanday ta’sir qilishini, qanday til vositalaridan foydalanilganini, matnning maqsadi va funksiyasini aniqlashga qaratilgan.



Matn turlari turli me’yorlar asosida shakllanadi va har bir turda o‘ziga xos til hamda uslub vositalari, maqsadlar va struktura mavjud. Shu asoslarga ko‘ra, matn turlari tasnifi quyidagicha:



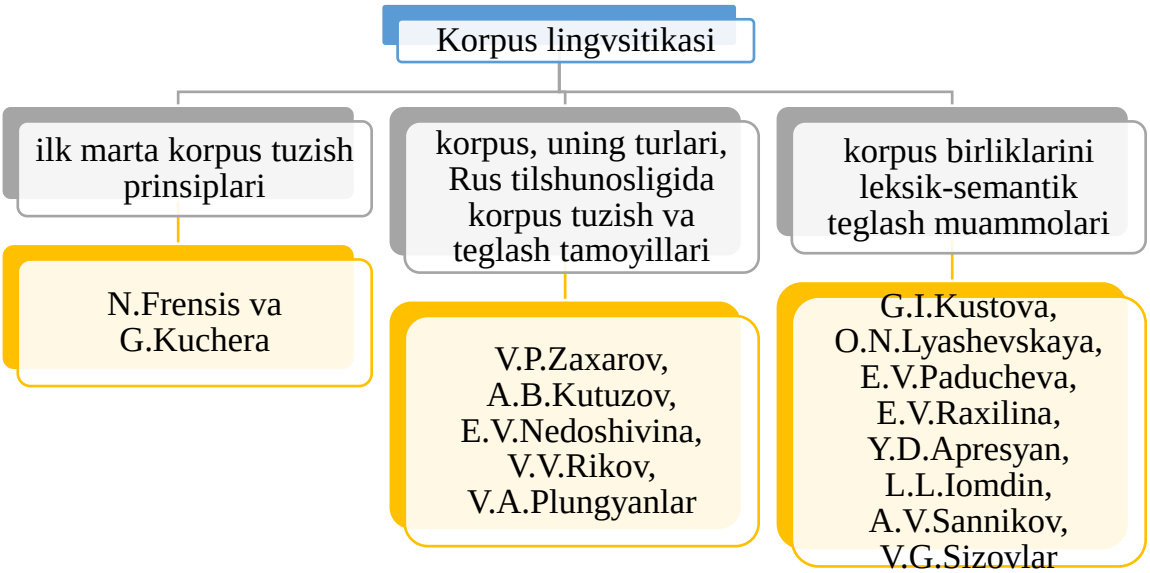
Ularning barchasi xabar funksiyasi va ilmiy ma’lumotlarni, asosan, yozma shaklda mantiqiy izchil, obyektiv va dalillarga asoslangan holda taqdim etishga yo‘naltirilganligi bilan tavsiflanadi. Har bir matn turi o‘zining maxsus vazifasini bajaradi va ma’lum bir auditoriya, guruh, shaxsga murojaat qiladi. Ushbu matn turlari o‘ziga xos uslub, tuzilma va mazmunga ega bo‘lib ma’lumot berish, taqdim etish, ko‘rsatma berish, ijodiy ifoda uchun, tasvirlash uchun, baholash va tahlil qilish maqsadlarida foydalaniladi.

*1.4-bo‘lim. Ilmiy matn va uni tushunish masalalari*ga bag‘ishlangan. Ilmiy matnda ish yuritishga oid, publitsistik yoki badiiy matndan farqli o‘laroq, nutqning funksional turlari qo‘llaniladi (tavsif, bayon, fikrlash, dalil va boshqalar). Ilmiy matnlar maxsus soha tilida yoziladi, bu esa ommaviy o‘qish jarayonini qiyinlashtirishi mumkin. Har bir sohada o‘ziga xos terminlar hamda tushunchalar mavjud va ularni to‘liq tushunmaslik matnni yomon anglashga olib kelishi mumkin. Vaholanki, har bir sohaga oid terminlar lug‘ati mukammal bazasi hali mavjud emas. Masalan, matematika, IT yoki kimyoviy sohalarda ishlatiladigan terminologiya kundalik til bilan farq qiladi. Buning yechimi uchun ilmiy matnni o‘qish va undan foydalanish jarayonida yangi terminlarni aniqlash hamda ularni izohlash uchun lug‘at yoki sohaga oid ilmiy manbalardan foydalanish maqsadli bo‘ladi. Bu, albatta, yangi va keng hajmli lug‘atni talab qiladi. Lug‘atlarni o‘zbek tili milliy korpusida mavjud ilmiy matnlar bazasidan foydalanish yordamida tuzish mumkin. Ilmiy matnlar bir nechta manbalarning xulosaviy natijalarini, tajribalarni yoki tadqiqot samaralarini o‘z ichiga oladi. Bunday matnni tushunish uchun ma’lumotlarni birlashtirish, solishtirish va tahlil qilish talab etiladi. Bu, ayniqsa, ilgari o‘qilgan boshqa matnlar bilan bog‘lanishni talab qiladigan ta’limiy jarayonlar uchun muhimdir. Matnni o‘qib bo‘lgandan keyin, asosiy g‘oya yoki fikr, mazmunni boshqa ma’lumotlar bilan taqqoslash, izlanish va qo‘shimcha manbalarni o‘rganish kerak bo‘ladi hamda bu orqali matnning analiz-sintez jarayoni samaradorligi oshadi. Ushbu jarayonda ham o‘zbek tili ilmiy matnlarini oson topish va foydalanish uchun o‘zbek tili milliy korpusi bazasi zarur hisoblanadi. Ilmiy matnlar odatda tuzilish jihatdan murakkab bo‘ladi. Ular ko‘pincha uzun jumlalar, gaplar, tahlillar, formulaviy fikrlar va statistik tahliliy ma’lumotlar bilan to‘ldirilgan bo‘ladi. Buning natijasida matnni to‘liq, mukammal tushunish qiyinlashadi. Muammoni matnni kichik qismlarga bo‘lib o‘qish va har bir qismni yaxshilab tushunishga harakat qilish orqali bartaraf etish mumkin. Har bir bo‘limni o‘qib bo‘lgandan so‘ng, asosiy maqsad va fikrni qayta umumiy izchillikda ko‘rib chiqish foydali bo‘ladi. Bilamizki,

ilmiy matnlarda boshqa tadqiqotlarga, ilmiy manbalarga yoki maqolalarga havolalar keltiriladi. Ushbu manbalarni ham kuzatishda ularni tushunmaslik, matnning to‘liq ma’nosini anglashni qiyinlashtiradi. Havolalarga ega bo‘lgan manbalarni izlash va ularni tekshirish maqsadga muvofiq. Matndagi qo‘shimcha izlanishlarga e’tibor qaratish orqali masala yuzasidan umumiy tasavvur hosil qilinadi. Ko‘p hollarda sohasiga ko‘ra, ilmiy matnlarda statistik ma’lumotlar, grafiklar va diagrammalar mavjud bo‘ladi, ularni to‘liq tushunish va tahlil qilish uchun statistik tahlil hamda raqamli ma’lumotlarni o‘qish qobiliyati talab qilinadi. Statistik diagramma va grafiklarni tahlil qilishda statistik tahlil metodlarni o‘rganish, shuningdek, tahlil uchun NLP algoritmlaridan foydalanib tuzilgan dasturiy ta’minotlardan foydalanish maqsadlidir. Ilmiy tilda muallifning masalaga yondashishi va tahlillari borasida ba’zi bir noaniqliklar mavjud bo‘lishi mumkin. Masalan, biror hodisa yoki fenomen haqida gapirganda, “ehtimol”, “taxminan”, “umuman olganda” kabi so‘zlar ishlatiladi, bu esa o‘quvchini matnni tushunishida noaniqliklarga olib borishi mumkin. Ilmiy matnlarni o‘qiyotganda, noaniqliklarni va ehtimollarni qo‘llanish maqsadini tahlil qilish lozim. Matnni ulardan holi tahlilga tortish aniq xulosalar olishda yordam beradi. Ilmiy matnlar o‘ziga xos va yangicha fikrlarni taqdim etadi. Bu fikrlarni tanqidiy tahlil qilish va ularni boshqa ilgari o‘qilgan ma’lumotlar bilan solishtirish zarur. Buning uchun o‘quvchidan tanqidiy fikrlashni rivojlantirish, matndagi asosiy g‘oyalarni va tadqiqotlarning ishonchliligini xolis baholash talab qilinadi. Chunki biror ilmiy matn natijasi ikkinchi bir ilmiy matnning yuzaga kelishida debocha, vazifasini o‘tashi mumkin. Ilmiy matn turlari: Ilmiy maqola. Dissertatsiya. Monografiya. O‘quv qo‘llanmalar va darsliklar. Ilmiy referatlar. Tadqiqot hisobotlari. Ilmiy konferensiyalar materiallari. Ilmiy taqdimotlar. Ilmiy-texnik matnlarni ommalashtirish hamda ilm dunyosi va foydalanuvchiga oson yetib borishini ta’minlashning asosiy omili korpus bazasiga qo‘shishdir. Ushbu yondashuv ilmiy-texnik matnni bir vaqtning o‘zida tadqiqot manbasi sifatida xizmat qilishi va bilim beruvchi omil ko‘rinishida gavdalantiradi. Ilmiy-texnik matnlarning bazasi bilimlarni namoyish etish bilan birga, uning yaratilish davri, muallifi va ixtisosligidan ham darak beradi.

II bob “O‘zbek tili milliy korpusida ilmiy-texnik matnlar bazasini yaratishning nazariy masalalari ”ga bag‘ishlanib, 3 bo‘limdan iborat.

2.1-bo‘lim. Jahon tilshunosligida korpus lingvistikasiga doir tadqiqotlar tahlili deb nomlangan. Jahon tilshunosligida korpus sohasidagi maqsadli tadqiqotlar XX asrning 40-yillarida Blumfild, Friz va Bondjerslar tomonidan boshlangan.



Kompyuter lingvistikasi sohasi bo‘yicha ilmiy va amaliy ishlar doirasida chunonchi, ingliz tilini olaylik, XX asrning 60-yillarida AQSHning Braun universiteti olimlari N.Frensis va G.Kucheralar mashinalar vositasida birinchi yirik matn korpusini yaratishgan. Ushbu korpus 500 ga yaqin ingliz tilining amerikancha variantini o‘z ichiga olib, 15 dan ortiq amerika nasrini qamrab olgan birinchi yirik korpus sifatida e’tirof etiladi. Bugungi kunda ko‘plab til korpuslari mavjud bo‘lib, turli maqsadlar doirasida yaratilganligi bois ular til korpuslarining maxsus turlari va turli imkoniyatlarini namoyon etadi. Olimlar, korpusga kiritiladigan til materiallarining hajmi oshgani sayin, lisoniy hodisalarni o‘rganish va tadqiq qilish imkoniyatlari kengayishini anglab yetdilar. Chunki har bir korpus sifat va miqdor jihatidan ma’lum talablarni qondirishi zarur[16]. Britaniya milliy korpusidan[17] ilhomlanib yaratilgan rus milliy korpusi o‘z ichiga nafaqat zamonaviy, balki tarixiy qatlamga oid so‘zlarni ham kiritgani bilan ajralib turadi.

Bu davrda jahon tilshunosligida tilni kompyuter texnologiyalari va statistik modellar asosida tadqiq etish zarurati ortib bordi. Bu jarayon tabiiy tilni qayta ishlash (Natural Language Processing – NLP) doirasida korpus lingvistikasining shakllanishi va mustaqil yo‘nalishga aylanishiga olib keldi. Til hodisalarini katta hajmdagi matnlar asosida o‘rganish imkonini beruvchi korpus lingvistikasi metodologiyasi, ilk bor ingliz tilshunosligida eksperimental ravishda qo‘llanila boshlagan bo‘lib, bugungi kunga kelib ko‘plab tillar uchun asosiy ilmiy tahlil usullaridan biriga aylangan. Korpus lingvistikasining nazariy poydevori G.Lich, J.Sinkler, R.Garsid, T.Makkenri, A.Hardi, M.Makkarti, V.Fransis, G.Kennedi kabi yetuk tadqiqotchilar tomonidan yaratilgan. Ularning ilmiy izlanishlarida korpuslar:

- til birliklarining funksional taqsimoti;
- statistik chastotalar;
- frazematik birliklarning tahlili,

[16] Грудева Е.В. Корпусная лингвистика. Учеб. пособие / Е.В.Грудева. 2-е изд., стер. – М.: ФЛИНТА, 2012. – 165 с.

[17] Британия Миллий корпуси (British National Corpus – BNC)га тенг равишда, инглиз тили корпусларидан яна: инглиз тили Халыаро корпуси (International Corpus of English – ICE), инглиз тили лингвистик Банки (Bank of English), Ҳозирги замон Америка инглиз тили Корпуси (Corpus of Contemporary American English – COCA) ва б.ларни ҳам келтириб ўтиш мумкин.

- til o‘rganish va ta’limdagi qo‘llanilishi kabi masalalarni ilgari surdi.

Ayniqsa, J.Sinkler boshchiligidagi COBUILD (Collins Birmingham University International Language Database) loyihasi ingliz tili zamonaviy korpus tahlilining amaliy modeliga aylandi. U korpus lingvistikasida “usage-based” yondashuvni asoslab berdi, ya’ni tilni o‘rganishda matnlar vositasida real qo‘llanishni asosiy tahlil obyekti sifatida qaradi. Korpuslar yordamida sintaktik va semantik strukturalarni tahlil qilish imkonini beruvchi avtomatlashtirilgan vositalar – lemmalash, teglash, morfologik razmetka, tematik modellashtirish (LDA, NMF, LSA) – tilshunoslikka yangi turtki berdi. Shuningdek, korpusdan foydalangan holda avtomatik tarjima tizimlarini yaratish, grammatik xatolarni aniqlovchi dasturlar ishlab chiqish hamda lug‘at tuzishning ilg‘or yondashuvlari rivojlandi. Zamonaviy tilshunoslikda internet bilan uzviy bog‘liq bo‘lgan yangi korpuslar – veb-korpuslar, blog va tvit korpuslar, parallel korpuslar yaratilmoqda.

Korpus lingvistikasi nafaqat ingliz tili, balki rus, nemis, fransuz, ispan, xitoy va boshqa yirik tillarda ham izchil rivojlanmoqda. Masalan, rus tilshunosligida V.Plungyan, A.B.Kutuzov, V.P.Zaxarovlarning ilmiy ishlari rus milliy korpusining yaratilishi va takomillashtirilishiga muhim hissa qo‘shdi. Ular korpusdagi annotatsiyalash tizimi, matnlarning sathiy tahlili, kontekstual modellashtirish bo‘yicha nazariy asoslarni yaratdilar. Bugungi kunda korpus lingvistikasi multidisiplinar soha sifatida tilshunoslik, informatika, psixolingvistika, tarjimashunoslik, sotsiolingvistika va pedagogika bilan uzviy bog‘lanib, keng ko‘lamli ilmiy tadqiqotlar platformasiga aylangan. Bu yondashuv nafaqat nazariy, balki amaliy tilshunoslik uchun ham muhim ahamiyat kasb etmoqda.

*2.2-bo‘lim. O‘zbek tilshunosligida korpus lingvistikasining rivojlanishi*ga bag‘ishlangan. O‘zbek korpus lingvistikasi ham, ayni vaqtda, taraqqiyot bosqichiga ko‘tarilib ulgurdi. Buning isboti sifatida O‘zbek tili milliy korpusi[18], O‘zbek tilining ta’limiy korpusi[19], O‘zbek tili korpuslari[20] amaliy natijalarini keltirish mumkin. Korpus lingvistikasi korpus yaratish hamda korpus metodlaridan foydalanib, tilning nazariy va amaliy muammolarini o‘rganishga doir ikki yo‘nalish asosida ish olib boradi. O‘zbekistonda korpus lingvistikasining shakllanishi 2018-yillarda nazariy tadqiqotlar bilan boshlanib, 2021-yilda o‘zbek tilining ta’limiy korpusi yaratilishi bilan rivojlandi. O‘zbek tilshunoslari Sh.Xamroyeva, N.Abduraxmonova, M.Abjalova, A.Eshmo‘minov, D.Axmedova, O‘.Xoliyorov, G.Toirova, D.O‘rinboyeva, A.Raxmanovalar tomonidan til korpuslari hamda uning turlari tadqiq qilingan. Buning natijasida esa mualliflik korpuslarini tuzishning lingvistik asoslari[21], tabiiy tilni qayta ishlash[22], sinonim so‘zlarni semantik teglash[23], atov

[18] http://uzbekcorpora.uz/ijrochi
[19] https://uzschoolcorpara.uz/
[20] https://uzbekcorpus.uz/
[21] Xamroyeva Sh. O‘zbek tili mualliflik korpusini tuzishning lingvistik asoslari. Filol. fan. b. fals. dok. ... diss. aftoref. – Qarshi, 2018. – 53 b.
[22] Abjalova M. Tahrir va tahlil dasturlarining lingvistik modullari. Monografiya. – Toshkent, 2020. – B. 176.
[23] Eshmo‘minov A.A. O‘zbek tili milliy korpusining sinonim so‘zlar bazasi. Filol. fan. b. fals. dok. ... diss. – Qarshi, 2019. – 140 b.

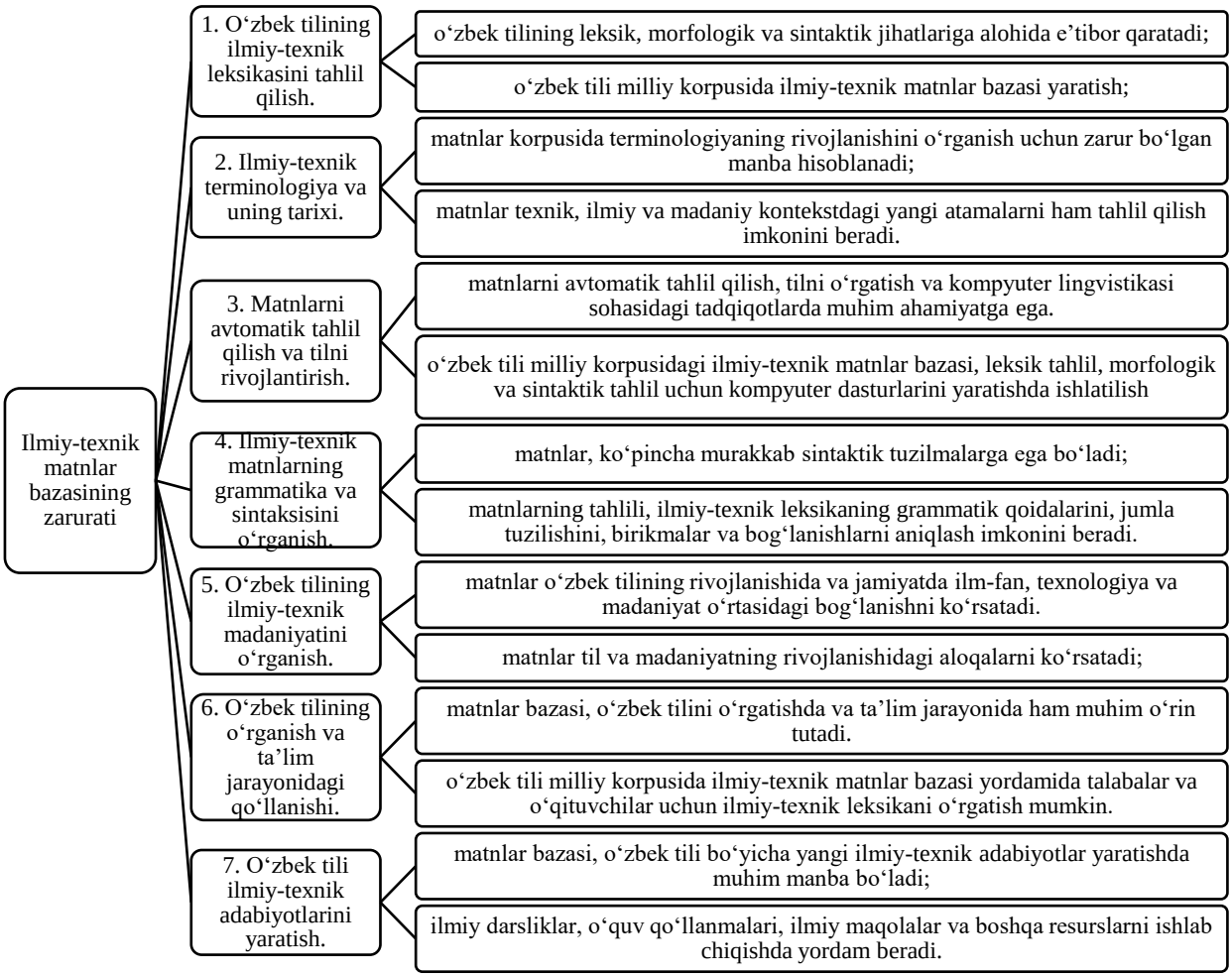
birliklarini leksik-semantik teglash[24], o‘zbek tilining ta’limiy korpusini yaratish[25], elektron korpuslar modellari, korpus tuzishning kompyuter usullari[26] tadqiq etildi. Keyingi yillarda o‘zbek tilining milliy korpusini yaratishga harakatlar paydo bo‘lishi natijasida, bir necha fragment holidagi korpuslar ishga tushdi. Bunga misol qilib Toshkent davlat axborot texnologiyalari universiteti Samarqand filiali Samarqand davlat universiteti jamoasi bilan hamkorlikda bajargan loyiha doirasida “o‘zbek tili milliy korpusi” (“Alpomish” dostoni materiallari asosida)ni, N.Abduraxmonova boshchiligida yaratilgan “o‘zbek tili korpusi”ni hamda N.G‘ulomovaning “Alisher Navoiy mualliflik korpusi”, O.Abdullayevaning “Tug‘ro” nomli o‘zbek tili axborot matnlari korpusini aytish mumkin. Bu amaliy ishlanmalarga bir necha yillar oldin yuqorida qayd etganimiz monografik tadqiqotlar bilan poydevor qo‘yilgandi. Dissertatsiyaning bu bo‘limida yana bugungi kunda faoliyat ko‘rsatayotgan o‘zbek tilining ta’limiy korpusi, o‘zbek tilining milliy korpusi, o‘zbek tili korpuslarining ahamiyatli jihatlari o‘rganilib, fikr-mulohazalar bildirilgan. Shuningdek, korpus lingvistikasiga doir yaqin yillarda amalga oshirilgan tadqiqotlarning nazariy va amaliy ahamiyatlariga doir fikrlar ham o‘rin olgan. Hozirgi kunda amaliy tilshunosligimizning eng dolzarb vazifalaridan biri – o‘zbek tilining mukammal milliy korpusini yaratishdir. Chunki milliy korpus, aslida, milliy tilning boy xazinasini anglatadi.

2.3-bo‘lim. O‘zbek tili milliy korpusi uchun ilmiy-texnik matnlar bazasining zarurati. O‘zbek tili milliy korpusi uchun ilmiy-texnik matnlar bazasini yaratish, o‘zbek tilining zamonaviy rivojlanishini, uning grammatik, leksik va stilistik xususiyatlarini chuqur o‘rganish uchun katta ahamiyatga ega. Ilmiy-texnik matnlar bazasining zarurati bir nechta omillarga asoslanadi:

[24] Ahmedova D. Atov birliklarini o‘zbek tili korpuslari uchun leksik-semantik teglashning lingvistik asos va modellari. Filol. fan. b. fals. dok. ... diss. – Buxoro, 2020. – 156 b.

[25] Raupova L., Elov B., Abjalova M., Alayev R. O‘zbek tilining ta’limiy korpusi va uning imkoniyatlari // O‘zbekistonda til va madaniyat. – Toshkent: ToshDo‘TAU, 2021. – № 4. – B. 60 75; Абжалова М. Синонимайзер (синонимизатор) в образовательной корпусе узбекского языка // TurkLang – 2021: Turkiy tillarni kompyuterda qayta ishlash IX xalqaro konferensiyasi.

[26] Raxmanova A. O‘zbek tili Milliy korpusini yaratishda kompyuter usullari. Filol. fan. b. fals. dok. ... diss. – Farg‘ona, 2022. – 52 b.



O‘zbek tili milliy korpusi uchun ilmiy-texnik matnlar bazasini yaratish tilshunoslik, leksikografiya, grammatika, sintaksis va boshqa ko‘plab sohalarda ilmiy tadqiqotlar uchun muhim ahamiyatga ega. Shuningdek, o‘zbek tilining zamonaviy rivojlanishini o‘rganish va ilmiy-texnik adabiyotlarni yaratish uchun mustahkam asos bo‘lib xizmat qiladi.

III bob “O‘zbek tili milliy korpusida ilmiy-texnik matnlar bazasini teglash va uning dasturiy ta’minoti”ga bag‘ishlangan bo‘lib, 3 bo‘limdan iborat.

*3.1-bo‘lim. Korpus turlari, xususiyatlari va matnlarni razmetkalash usullari*ga bag‘ishlangan. Tabiiy tilni qayta ishlash (NLP)ga asoslangan dasturiy ta’limotlarni ishlab chiqish uchun NLP tizimini mavjud ma’lumotlarni o‘rganishini ta’minlash kerak. Ushbu amalni til korpusi vositasida amalga oshirish mumkin. Til korpusi elektron shaklda taqdim etiladigan *katta hajmli* va *strukturlangan matnlar to‘plami* sifatida qaraladi. Til korpusi *yozma* yoki *og‘zaki* materialni ifodalab, NLP tizimini mavjud resurslarni o‘rganishi uchun **lingvistik tahlilni** amalga oshirish lozim. Til korpusi tabiiy tilni qayta ishlash tizimining asosidir. Unda gazetalar, romanlar, retseptlar, radio eshittirishlardan tortib teleko‘rsatuvlar, filmlar va tvitlargacha bo‘lgan turli shakldagi ma’lumotlar bo‘lishi mumkin. NLPning ko‘plab vazifalarini hal qilishda til korpusidan foydalaniladi: **mashinali o‘qitish**

modellarini o‘rgatish, tilni tushunish, diskurs tahlili, tarjimashunoslik, qoidalarga asoslangan tizimlar, lug‘at va semantika, statistik tahlil, sohaviy bilimlar. NLP tizimini ishlab chiqish uchun foydalanish mumkin bo‘lgan har xil turdagi korpuslar mavjud. NLPda til korpuslari mazmun, maqsad yoki manba kabi turli mezonlar asosida har xil turlarga bo‘linadi.

Korpus matnlari turini tanlash.

Matn korpusini yaratishda muhim masala – unda barcha turdagi yozma matnlar mavjudligini aniqlashdan iborat. Til korpuslarining aksariyati standart yozuvlarning yozma matnlaridan tashkil topgan. Korpus mazmunini samarali shakllantirish uchun inson bilimining barcha mumkin bo‘lgan sohalaridan olingan matnlar to‘plamini korpusda ifodalash lozim. Matn tanlash va to‘plashda manba, materiallarni (masalan, kitoblar, gazetalar, jurnallar va boshqalar) yig‘ish uchun turli nashriyotlarning kataloglari hamda nashrlari ro‘yxatiga murojaat qilish kerak. Misol sifatida, o‘zbek til korpusida adabiyot – 25 %, tasviriy san’at – 3 %, ijtimoiy fanlar – 8 %, tabiiy fanlar – 15 %, tijorat – 10 %, ommaviy axborot vositalari – 35 % va tarjima – 4 % matnlar mavjud. Bunda har bir turkumda bir nechta kichik toifalar mavjud. Masalan, **adabiyotga** romanlar, hikoyalar, insholar va boshqalar kiradi; **tasviriy san’atga** rasm, chizmachilik, musiqa, haykaltaroshlik va boshqalar kiradi; **ijtimoiy fan** falsafa, tarix, ta’lim va boshqalarni o‘z ichiga oladi; **tabiiy fanlarga** fizika, kimyo, matematika, geografiya va boshqalar kiradi; **ommaviy axborot vositalariga** gazetalar, jurnallar, plakatlar, e’lonlar va boshqalar kiradi; **tijorat** buxgalteriya, bank ishi va boshqalarni o‘z ichiga oladi hamda **tarjima** tabiiy tilga tarjima qilingan barcha mavzularni o‘z ichiga oladi.

3.2-bo‘lim. Korpusda ilmiy-texnik matnlarni tematik modellashtirish masalasi.

So‘nggi yillarda korpus ilmiy-texnik matnlarini lingvistik teglash usullarining ishlab chiqilishi, shuningdek, teglangan ma’lumotlarni yaratish va saqlashni qo‘llab-quvvatlash uchun teglash vositalarining ko‘payishi, Amazon Mechanical Turk kabi NLP ilovalarining ishlab chiqilishiga xizmat qildi. Bugungi kunda til korpusidagi ilmiy-texnik matnlarni teglashning **tematik modellashtirish** NLP usulidan foydalanilmoqda. **Tematik modellashtirish** – bu ilmiy-texnik matnlar to‘plami uchun klaster so‘zlarini aniqlash uchun matn ma’lumotlarini avtomatik ravishda tahlil qiladigan mashinali o‘rganish usuli. Bu usul avvaldan odamlar tomonidan tasniflangan teglar yoki o‘quv ma’lumotlarining oldindan belgilangan ro‘yxatini talab qilmaydi. Tematik modellashtirish – til korpusidagi ilmiy-texnik matnlar to‘plamida uchraydigan mavhum “mavzular”ni aniqlash uchun statistik modellashning bir turi. Ilmiy-texnik matnlar odatda bir nechta mavzularga mansub bo‘lib, turli nisbatlarda taqsimlanadi. Shunday qilib, matnda *shashka* – **10 %** va *shaxmat* – **90 %** haqida ma’lumot bo‘lgan matnda shashkaga oid so‘zlarga qaraganda taxminan **9** baravar ko‘p shaxmatga oid so‘zlar bo‘lishi mumkin. Tematik modellashtirish usuli yordamida aniqlangan “mavzular” o‘xshash so‘zlarning klasteridir. Tematik model ushbu klasterni matematik nuqtayi nazardan qamrab oladi. Bu ilmiy-texnik matnlar to‘plamini o‘rganish va har biridagi so‘zlarning statistik ma’lumotlariga asoslanib, mavzularni aniqlashga va har bir matnning mavzular balansi qandayligini bilib olishga imkon beradi. Korpusdagi ilmiy-texnik matnlardan so‘zlarni ajratib olish ko‘proq vaqt talab etadi va bu jarayon

matndagi berib o‘tilgan mavzulardan ajratib olishdan ko‘ra ancha murakkabroqdir. Masalan, har bir 100000 ta matndan va har bir matnda o‘rtacha 500 tadan so‘z mavjud til korpusini tahlil qilamiz. Demak, ushbu korpusni qayta ishlash uchun 500*100000=50000000 ta amalni bajarish kerak bo‘ladi. Shunday qilib, ma’lum mavzularni o‘z ichiga olgan matnni tahlil qilishda, agar o‘rtacha 5 ta mavzudan iborat bo‘lsa, ishlov berish 5*500 so‘z = 2500 ta *amal (thread)ni* tashkil qiladi. Bu butun bir matnni qayta ishlashdan ko‘ra soddaroq ko‘rinadi, shuning uchun tematik modellashtirish shu kabi muammolarni hal qilishda qo‘l keladi va jarayonni vizualizatsiya qilishni osonlashtiradi. NLPda matnni qayta ishlashni osonlashtiradigan matnga boshlang‘ich ishlov berish bosqichlarini amalga oshirish lozim:

- *Nomuhim so‘zlar va tinish belgilarini olib tashlash.*
- *Stemming.*
- *Lemmatizatsiya.*
- *Countvectorizer yoki Tf-Idf usuli yordamida statistik qiymatlarni shakllantirish.*

Bugungi kundagi tematik modellashtirishning ommabop algoritmlariga *yashirin semantik tahlil (Latent Semantic Analysis, LSA), yashirin semantik indeksatsiya (Latent Semantic Indexing, LSI), iyerarxik Dirixle jarayoni (Hierarchical Dirichlet Process, HDP)*, yashirin Dirixle taqsimoti (Latent Dirichlet Allocation, LDA) va *manfiy bo‘lmagan matritsa faktorizatsiyasi (Non-negative Matrix factorization, NMF)* kabi usullarini misol sifatida keltirish mumkin. Tematik modellar bizga ilmiy-texnik matnlarning butun matnida o‘xshash ma’noga ega so‘zlar birikmasi va har bir berilgan matnda mavzular birikmasi sifatida yashiringan potensial mavzularni aniqlashga yordam beradi.

Ma’lumotlarni klasterlash – bu ma’lumotlar to‘plamida *kichik guruhlarni aniqlash* yoki *klasterlash* uchun nazoratsiz mashinani o‘rganish usuli. *Klasterlashning asosiy g‘oyasi quyidagicha:* ma’lumotlar to‘plamidagi kuzatuvlarni turli xil guruhlarga bo‘lish orqali har bir guruh ichidagi kuzatuvlar bir-biriga juda o‘xshash va turli guruhlardagi kuzatuvlar bir-biridan mutlaqo farq qilishi lozim. Kuzatishlar o‘rtasidagi o‘xshashlikni aniqlash uchun mos mezonlarni tanlash kerak. Bugungi kunda klaster tahlili uchun ko‘plab usullar ishlab chiqilgan.

Manfiy bo‘lmagan matritsalarni faktorizatsiya qilish (Non-Negative Matrix Factorization, NMF) usuli. NMF – bu matritsani ikkita matritsaga ajratish usuli bo‘lib, uchta matritsada ham manfiy elementlar mavjud emas. NMF usuli, asosan, *tavsiya tizimlari, signallarni qayta ishlash* va *bioinformatika* sohalarida qo‘llanadi.

Yashirin Dirixle taqsimoti (Latent Dirichlet Allocation, LDA) usuli. NLPda LDA usuli yordamida tematik modellashtirish ilmiy-texnik matnlardagi so‘zlar asosida mumkin bo‘lgan mavzularni aniqlash orqali matnlar to‘plamida **yashirin (latent)** mavzularni aniqlashga imkon beradi. Yashirin Dirixle taqsimoti usulida har bir matn va korpusdagi *har bir so‘z, har bir mavzu* va *so‘zlar* o‘rtasidagi munosabatlar yashirin o‘zgaruvchilar yordamida modellashtiriladi. Korpusdagi har bir matn yashirin o‘zgaruvchilar (mavzular) bo‘yicha **Dirixle taqsimoti** yordamida

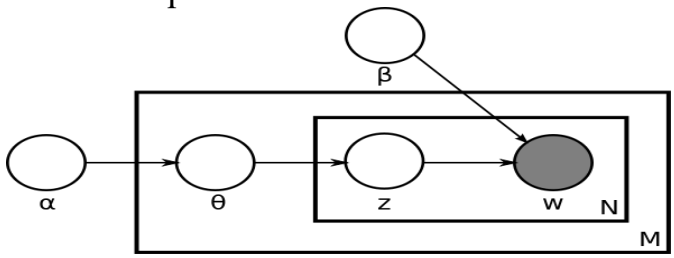
taqdim etiladi va har bir mavzu barcha ilmiy-texnik matnlardagi so‘zlar bo‘yicha boshqa Dirixle taqsimoti orqali hisoblanadi.

3.3-bo‘lim. Korpusda ilmiy-texnik matnlarni teglashning dasturiy ta’minoti

Korpusda ilmiy-texnik matnlarni teglashni amalga oshirish uchun LDA usulidan foydalanilganda, biz kirish matritsasi sifatida so‘zlar to‘plamini olamiz, chunki bu ehtimollik modeli hisoblanadi. Keyingi qadamda LDA algoritmi matritsani ikkita kichik matritsaga ajratadi:

– *hujjatdan mavzu matritsasi;*
– *so‘zdan mavzu matritsasi.*

LSI usulida har bir ilmiy-texnik matnni boshqalardan ajratilgan holda ko‘rib chiqish o‘rniga, munosabatlarni aniqlash uchun barcha ilmiy-texnik matnlarni va ulardagi shartlarni ko‘rib chiqiladi. SVDdan quyi darajali matritsaga yaqinlashish masalasini hal qilish uchun foydalanish mumkin. So‘ngra ushbu matritsadan *termin-hujjat matritsalarini* aniqlash mumkin. Buning uchun quyidagi uch bosqichli amallarni bajarish lozim: 1. Berilgan **C** uchun **SVD** matritsasini $\boldsymbol{C = U \sum V^T}$ ko‘rinishda hosil qilamiz. 2. $\sum$ diagonalidagi **r−k** eng kichik singulyar qiymatlarni nolga almashtirish natijasida hosil bo‘lgan $\boldsymbol{\sum k}$ matritsasini hosil qilish. 3. $\boldsymbol{C_k = U \sum kV^T}$ni **C** ga **k**-darajada yaqinlashtirish. Bu yerda, **C** – *termin-hujjat matritsasi*, $\boldsymbol{U}$, $\sum$ va $\boldsymbol{V^T}$ qiymatlar **SVD** matritsalari. Yuqorida keltirilgan nazariy ma’lumotlar asosida Python tili vositalari yordamida **LSI** uchun **scikit-learning** modulidan foydalanamiz. **Scikit-learning** modulida LSI usuli uchun o‘lchamlarni kamaytirish SVD metodi yordamida amalga oshiriladi. LSI usulining bitta asosiy zaif tomoni bor – **noaniqlik**. Masalan, Microsoft Office yoki siz ishlayotgan ofis haqida gapirayotganingizni tizim qanday aniqlashi mumkin. Ushbu holda LDA usulidan foydalanish mumkin. Misol uchun, agar kuzatishlar ilmiy-texnik matnlarga yig‘ilgan so‘zlar bo‘lsa, har bir ilmiy-texnik matn oz sonli mavzularning aralashmasi ekanligini va har bir so‘zning mavjudligi ilmiy-texnik matn mavzularidan biriga tegishli ekanligini ta’kidlaydi. LSI usuli hujjatda ishlatiladigan so‘zlarni tekshiradi va ularning boshqa so‘zlar bilan munosabatlarini aniqlaydi. LSI usuli tizimga hujjatning o‘zida foydalanilmagan bo‘lsa ham, ilmiy-texnik matn tegishli bo‘lishi mumkin bo‘lgan so‘zlarni aniqlash imkonini beradi.

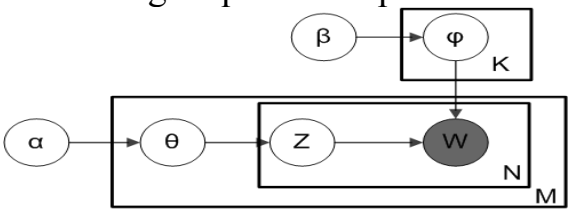


1-rasm

Yuqoridagi 1-rasmda LDA usuli arxitekturasi keltirilgan bo‘lib, bu yerda

– α *– har bir ilmiy-texnik matn uchun mavzular bo‘yicha Dirixle taqsimoti parametri;* β *– mavzular bo‘yicha Dirixle so‘z taqsimoti parametri;* θ_m *– m ilmiy-texnik matn bo‘yicha mavzular taqsimoti;* φ_k *– k mavzu bo‘yicha so‘zlar taqsimoti;* z_{mn} *– m ilmiy-texnik matndagi n-so‘z uchun mavzu;* w_{mn} *– so‘z.* **W** ning kulrangda bo‘lishi shuni anglatadiki, $\mathbf{w_{ij}}$ so‘zlari kuzatilishi mumkin bo‘lgan yagona

o‘zgaruvchilar, qolgan o‘zgaruvchilar esa yashirin o‘zgaruvchilardir. “Mavzu-so‘z” juftliklarning taqsimlanishini modellashtirish uchun siyrak Dirixle taqsimotidan foydalanish mumkin. Chunki mavzudagi so‘zlar bo‘yicha ehtimollik taqsimoti siqilganligi sababli, faqat kichik so‘zlar to‘plami yuqori ehtimolga ega bo‘ladi. Ushbu model bugungi kunda LDA usulining eng keng tarqalgan variantidir. Ushbu modelning rasmli shakli quyida keltirilgan bo‘lib, bu yerda **K** – mavzular sonini va $\varphi_1,\ldots,\varphi_k$ **V** o‘lchovli vektorlardir. φ_i - **V** o‘lchovli vektorlar – mavzu va so‘zlar bo‘yicha Dirixle tomonidan taqsimlangan parametrlarni saqlaydi. $\boldsymbol{\theta}$ va $\boldsymbol{\varphi}$ bilan ifodalangan obyektlarni modellashtirilayotgan ilmiy-texnik matn korpusini ifodalovchi asl “hujjat-so‘z” matritsasini parchalash orqali yaratilgan matritsalar deb hisoblash mumkin. $\boldsymbol{\theta}$ – ilmiy-texnik matnlar bilan belgilangan satrlar va mavzular bo‘yicha belgilangan ustunlardan, $\boldsymbol{\varphi}$ – esa so‘zlar bilan belgilangan mavzular va ustunlardan iborat. Shunday qilib, $\boldsymbol{\varphi_1,\ldots,\varphi_k}$ qatorlar yoki vektorlar to‘plamini bildiradi va $\boldsymbol{\varphi_i}$ – so‘zlar bo‘yicha taqsimot hisoblanadi. $\boldsymbol{\theta_1,\ldots,\theta_k}$ esa har biri mavzular bo‘yicha taqsimot bo‘lgan qatorlar to‘plamini bildiradi.



2-rasm.

LDA usuli korpusdagi har bir ilmiy-texnik matn butun korpusda mavjud bo‘lgan mavzular aralashmasini o‘z ichiga oladi deb taxmin qiladi. Mavzu tuzilishi yashirin bo‘lib, biz mavzularning o‘zini emas, balki faqat ilmiy-texnik matnlar va so‘zlarni kuzatishimiz mumkin. Tuzilish yashirin bo‘lgani uchun, LDA usuli ma’lum so‘zlar va ilmiy-texnik matnlarni hisobga olgan holda mavzu tuzilishi haqida xulosa chiqarishga intiladi. *LDA usulida ilmiy-texnik matn tuzilishi.*

LDA usuli ilmiy-texnik matnlarni ma’lum ehtimollar bilan korpus so‘zlari (lug‘ati) orqali mavzular to‘plamini aniqlab beradi. Ilmiy-texnik matnlar quyidagi keltirilgan qoidalar asosida ishlab chiqiladi deb taxmin qilinadi:

– *ilmiy-texnik matnlardagi so‘zlarning sonini (N) aniqlash;*

– *ilmiy-texnik matnlar uchun mavzular to‘plamini aniqlash (belgilangan K mavzular to‘plami bo‘yicha Dirixle ehtimoli taqsimotiga ko‘ra);*

– *ilmiy-texnik matnlardagi har bir so‘zni quyidagicha hosil qilish:*

– *avval mavzu tanlash;*

– *so‘zni yaratish uchun mavzudan foydalanish (mavzuning multinomial taqsimotiga ko‘ra).*

– *korpusda ilmiy-texnik matnlar to‘plami uchun ushbu generativ modelni nazarda tutgan holda, LDA to‘plamni yaratgan bo‘lishi mumkin bo‘lgan mavzular to‘plamini aniqlash uchun korpusda ilmiy-texnik matnlardan orqaga qaytishga harakat qilish.*

Keyingi qadamda LDA usulini Python tili vositasida tadbiq qilish masalasini ko‘rib chiqamiz. Hisob-kitoblar aniqligini oshirish uchun til korpusi matnlaridan barcha tinish belgilarini olib tashlash kerak. *N-gramm til modellari.* Bigrammalar –

korpusda ilmiy-texnik matnlarda birga keladigan ikkita so‘zdir. Trigrammalar – birgalikda uchraydigan 3 ta so‘zdir. Til korpusi matnlaridagi n-grammlarni aniqlash uchun **Gensim** paketining **Gensim.Phrases** usuli yordam beradi. Berilgan til korpusi matnlarini LDA usuli asosida tematik modellashtirish uchun bigrammalarni, trigrammalarni shakllatirishimiz va korpusda ilmiy-texnik matnlar to‘plamini lemmatizatsiya qilishimiz kerak. Lemmatizatsiya – bu so‘zning leksik shakllarini guruhlash jarayoni bo‘lib, ularni so‘zning lemmasi yoki lug‘at shakli bilan aniqlangan yagona element sifatida tahlil qilish mumkin. Til korpusi matnlarini lemmalash uchun quyidagi funksiyalarni shakllantiramiz. Ushbu funksiyalarni namunaviy ma’lumotlar to‘plamiga qo‘llaymiz. Yuqoridagi keltirilgan fikr-mulohazalar asosida bizning korpusda ilmiy-texnik matnlar to‘plamidan DTM (document-term matrix) matritsasini hosil qilamiz.

LDA modeli shakllanirilganidan so‘ng, modelni baholash uchun ko‘rsatkichlarni ishlab chiqamiz. LDA modelini baholashning 2 ta ko‘rsatkichi mavjud:

1) ***ajablanish (Perplexity)***
2) ***muvofiqlik balli (Coherence Score)***

Hosil qilingan LDA modeli asosida har bir mavzu bo‘yicha unga eng muhim terminlarni shakllantiramiz. Keling, mavzu vizualizatsiyasini tahlil qilamiz. Mavzular chap tomonda, so‘zlar esa o‘ng tomonda qanday ko‘rsatilishiga e’tibor bering. Korpusda ko‘p uchraydigan mavzular kattaroq shaklda namoyish qilingan. Bir-biriga yaqinroq mavzular o‘xshashroq, bir-biridan uzoqroq bo‘lgan mavzular kamroq o‘xshash. Mavzuni tanlaganingizda, tanlangan mavzu doirasidagi so‘zlarni ko‘rish mumkin. Bu o‘lchov so‘zning qanchalik ko‘p yoki qanchalik farqlovchi kombinatsiyasi bo‘lishi mumkin. Keyingi qadamda mavzular soni **< 5** bo‘lgan hol uchun muvofiqlik qiymatlarini olish funksiyasini ishlab chiqamiz. Ushbu misol uchun **4** ta mavzu ko‘rib chiqilgan. Chunki bu ma’lumotlar to‘plamidagi maqsadli yorliqlar **4** ga ajratilgan. Aks holda, eng yuqori muvofiqlik balliga ega bo‘lgan ma’qul. LDA usulining maqsadi, shuningdek, korpusda ilmiy-texnik matnlarning qancha qismi qaysi mavzu tomonidan yaratilganligini hisoblashdan iborat. Bobning ushbu qismida o‘zbek tili milliy korpusi uchun yaratilgan ilmiy-texnik matnlar bazasini teglash (ya’ni lingvistik jihatdan belgilash) jarayonini avtomatlashtirishga qaratilgan dasturiy yechimlar tahlil qilindi. Shu asosda **http://topicmodel.uz** platformasi yaratildi, bu dasturiy ta’minot quyidagi imkoniyatlarni taqdim etadi:

- matnni tizimli kiritish va kalit so‘zlar sonini belgilash;
- tematik modellarni shakllantirish;
- matnlarni *ta’lim, sport, sog‘liqni saqlash, siyosat, madaniyat, ob-havo, iqtisodiyot, texnologiya* kabi sohalarga ajratish;

Ishda Web-nashrlardagi (masalan, kun.uz, daryo.uz, xabar.uz) maqolalarni teglash va modellashtirish bo‘yicha amaliy namunalarda sinov o‘tkazildi.

Dasturiy ta’minotning yana bir afzalligi – foydalanuvchi interfeysining qulayligi, vizual statistik taqdimoti va tematik klasterlarni grafik shaklda ko‘rsatish imkoniyatidir. Bu hol, o‘z navbatida, tahlil jarayonining intuitiv va tez amalga oshirilishiga yordam beradi. Yaratilgan dasturiy tizim fanlararo izlanishlarda,

kompyuter lingvistikasi asosidagi modellarni o‘qitishda keng qo‘llanilishi mumkin. Bu tizim yordamida matnlarni:

- semantik tahlil qilish;
- grammatik strukturalarni ajratish;
- mashina tarjimasiga tayyorlash;
- terminologik izlanishlar uchun bazaga aylantirish mumkin

Xulosa

1. Matn o‘rganilishining I bosqichi antik davrga, II bosqichi XVII-XIX asrlarga to‘g‘ri keladi. Bu davrda ko‘plab asarlar tahlili yaratilgan. XX asrga kelib, matnlarni o‘rganishda semantik, struktur, pragmatik yondashuvlar kuzatildi. XXI asrga qadar matnlar tilning boshqa bir birliklari tahlili doirasida o‘rganilgan bo‘lsa, XIX asrning II yarmi hamda XX asrning o‘rtalariga kelib matnni o‘rganish dolzarblik kasb etdi. XX asr matnning “oltin asri”ga aylandi. XXI asr matnning qayta tug‘ilish davri sifatida e’tirof etiladi. Bu davrda matn lingvokulturologik, sotsilolingvistik, kognitiv, psixologik, korpus bazasi elementi aspektlarida tekshirila boshladi.

2. Matnni o‘rganish ishi bir necha yo‘nalishda olib borildi. Birinchi yo‘nalishdagi ishlarda matnning formal-grammatik va ma’no qurilishini o‘rganishga e’tibor qaratildi. Ikkinchi yo‘nalishdagi ishlarda matnni turlicha idrok qilishga sabab bo‘ladigan matn qurilishining formal va konseptual xususiyatlari tahlil qilinadi. Uchinchi yo‘nalishdagi ishlarda matnning o‘zini idrok qilish masalasi yoritiladi.

3. Til korpusi – bu tilning haqiqiy foydalanuvchilari tomonidan ishlab chiqilgan, so‘z, ibora va umuman, til qanday ishlatilishini tahlil qilish uchun ishlatiladigan juda *katta hajmli* va *strukturlangan matnlar to‘plami* sifatida qaralib, *yozma* yoki *og‘zaki* materialni ifodalaydi. Korpus, shuningdek, dasturiy ta’minotni ishlab chiqishda ishlatiladigan turli til ma’lumotlar bazalarini yaratish uchun ishlatiladi, masalan, *bashoratli klaviatura, imloni tekshirish va tuzatish, matn/nutqni tushunish tizimlari, matndan nutqqa modullar, mashina tarjimasi tizimlari* va boshqalar. Til korpusi foydalanuvchilar uchun to‘liq foydali bo‘lishi uchun uni teglash kerak. NLPda til korpuslari mazmun, maqsad yoki manba kabi turli mezonlar asosida har xil turlarga bo‘linadi: *matnli korpuslar, multimodal korpuslar, parallel korpuslar, tarixiy korpus va annotatsiyalangan/teglanlar korpuslar.*

4. Bugungi kunda, jahon kompyuter lingvistikasi sohasida, korpus matnlarini teglashning zamonaviy usullari hisoblangan: gap chegalarini aniqlash, tokenlash, lemmalash, POS teglash, sintaktik tahlil, semantik tahlil, NER obyektlarini aniqlash va koreferensiyani hal qilish kabi yondashuvlari mavjud. Korpus ilmiy-texnik matnlarini teglash usullaridan *matnni tushunish, ma’lumot olish, hissiyotlarni tahlil qilish, imloni tekshirish, hujjatlarni umumlashtirish* va *mashina tarjimasi* kabi NLP ilovalarini ishlab chiqishda foydalanish mumkin.

5. Bugungi kunda til korpusidagi ilmiy-texnik matnlarni teglashning tematik modellashtirish deb nomlanuvchi NLP usulidan foydalanilmoqda. Tematik modellashtirish – bu ilmiy-texnik matnlar to‘plami uchun klaster so‘zlarini aniqlash uchun matn ma’lumotlarini avtomatik ravishda tahlil qiladigan mashinali o‘rganish

usuli. Bu usul “nazoratsiz” mashinali o‘rganish sifatida tanilgan bo‘lib, avvaldan odamlar tomonidan tasniflangan teglar yoki o‘quv ma’lumotlarining oldindan belgilangan ro‘yxatini talab qilmaydi.

6. Korpusdagi ilmiy-texnik matnlarni tasniflash modellari o‘qitishni talab qilganligi sababli, ular “nazorat ostidagi” mashinali o‘rganish usullari sifatida tanilgan. Tematik modellashtirish – til korpusidagi ilmiy-texnik matnlar to‘plamida uchraydigan mavhum “mavzular”ni aniqlash uchun statistik modellashning bir turi. Ilmiy-texnik matnlar odatda bir nechta mavzularga mansub bo‘lib, turli nisbatlarda taqsimlanadi. Korpusdagi ilmiy-texnik matnlardan so‘zlarni ajratib olish ko‘proq vaqt talab etadi va bu jarayon matndagi berib o‘tilgan mavzulardan ajratib olishdan ko‘ra ancha murakkabroqdir. Masalan, har bir 100000 ta matndan va har bir matnda o‘rtacha 500 tadan so‘z mavjud til korpusini tahlil qilindi. Demak, ushbu korpusni qayta ishlash uchun 500*100000=50000000 ta amalni bajarish kerak bo‘ladi.

7. NLPda tematik modellashtirish – bu katta hajmdagi matnlarni avtomatik ravishda umumlashtirish uchun ishlatilishi mumkin bo‘lgan algoritmlar to‘plami. Til korpusi matnlarini tahlil qilishda *o‘lchov, xususiyatlar* soni juda katta bo‘lgan modellarni o‘qitishni qiyinlashtiradi va modellarning samaradorligini pasaytiradi. Nazorat qilinmaydigan mashinali o‘rganish vazifalarida qo‘llanadigan tematik modellashtirish teglash sifatida ko‘rib chiqiladi, birinchi navbatda, til korpusidan zarur ma’lumotlarni olish uchun ishlatiladi hamda so‘rovlarning bajarilish samaradorligining oshishiga yordam beradi.

8. Tematik modellashtirish – bu turli xil sohalarda foydalanish holatlarida qo‘llanadigan ko‘p qirrali algoritm. Tematik modellashtirish qidiruv tizimlarida mavzular bo‘yicha foydalanuvchi qiziqishlarini xaritalashda keng qo‘llaniladi. Bugungi kunda tematik modellashtirish usullari: *hujjatlarni tasniflash, toifalarga ajratish, umumlashtirish* kabi NLP vazifalarini hal qilishda qo‘llanmoqda. Shuningdek, tematik modellashtirish usullari ijtimoiy tarmoqlardagi foydalanuvchilarning his-tuyg‘ularini tahlil qilish imkonini beradi.

9. Bugungi kundagi tematik modellashtirishning ommabop algoritmlariga *yashirin semantik tahlil (Latent Semantic Analysis, LSA), yashirin semantik indeksatsiya (Latent Semantic Indexing, LSI), ierarxik Dirixle jarayoni (Hierarchical Dirichlet Process, HDP)*, yashirin Dirixle taqsimoti (Latent Dirichlet Allocation, LDA) va *manfiy bo‘lmagan matritsa faktorizatsiyasi (Non-negative Matrix factorization, NMF)* kabi usullarini misol sifatida keltirish mumkin. Ular orasida LDA amalda aniqroq hamda samarali natijalarni ko‘rsatgan va shuning uchun keng miqyosda qo‘llanadi.

10. O‘zbek tili korpusidagi ilmiy-texnik matnlarni teglash uchun LDA, LSA, NMF usullari va algortimlari vositasida dasturiy ta’minot ishlab chiqildi hamda korpusdagi ilmiy-texnik matnlarni teglashda quyidagi sohalar qamrab olindi: *ta’lim, sport, sog‘liqni saqlash, siyosat, madaniyat, ob-havo, iqtisodiyot, texnologiya.* Ilmiy tadqiqotda keltirilgan algoritmlar asosida http://topicmodel.uz/ dasturiy ta’minoti ishlab chiqildi.

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

ALISHER NAVO'I TASHKENT STATE UNIVERSITY OF UZBEK LANGUAGE AND LITERATURE

NORBEKOVA MADINA SHUKHRAT QIZI

CREATING A SCIENTIFIC AND TECHNICAL TEXT DATABASE FOR THE NATIONAL CORPUS OF THE UZBEK LANGUAGE

10.00.11 - Theory of Language. Applied and computational linguistics

DISSERTATION ABSTRACT OF DOCTOR OF PHILOSOPHY (PhD) ON PHILOLOGICAL SCIENCES

Tashkent – 2025

UO‘K: ***811.512.133’28***

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

Contents of Dissertation Abstract of the Doctor of Philosophy (PhD) in philological sciences

Оглавление автореферата диссертатсии доктора философии (PhD) по филологическим наукам

The topic of the Doctor of Philosophy (PhD) dissertation is registered with the Higher Attestation Commission under the Ministry of Higher Education, Science and Innovation of the Republic of Uzbekistan under the number B2024.4.PhD/Fil5438.

The doctoral thesis was carried out at Alisher Navo'i Tashkent State University of Uzbek language and Literature.

The dissertation abstract is available in three languages (Uzbek, English, Russian and (resume)) on the website of the National University of Uzbekistan (www.nuu.uz.) and on the "ZiyoNet" Information and Education Portal (www.ziyonet.uz).

Scientific advisor:	**Raupova Laylo** Doctor of Philological sciences, professor
Official opponents:	**Toirova Guli** Doctor of Philological sciences, professor
	Abdurakhmanova Nilufar Doctor of Philological sciences, professor
Leading organization:	**Andijan State Institute of Foreign Languages**

The dissertation defense will take place at the meeting of the Scientific Council DSc.03/25.08.2021.Fil.01.16 at the National University of Uzbekistan on "____" _____2025, at _____ hours. (Address: 4 University Street, Almazar district, Tashkent 100174. Tel.: (99871) 246-02-24; Fax: (99871) 246-02-24; E-mail: devonxona@nuu.uz.

The dissertation can be accessed at the Information Resource Center of the National University of Uzbekistan (registered under No. ____). (Address: 4 University Street, Almazar district, Tashkent 100174. Tel.: (99871) 246-02-24; Fax: (99871) 246-02-24; E-mail: devonxona@nuu.uz.

The dissertation author's abstract is distributed on "____" _____ 2025 (registrey record No. _____ dated "____" _____ 2025).

N.A.Rakhmonov
One-time chairman of the Scientific Council for the award of scientific degrees,
Doctor of Philological sciences, professor

M.B.Xujamkulova
One-time scientific Secretary of the Scientific Council for the award of scientific degrees,
Candidate of Philological Sciences, Associate professor

A.E.Mamatov
One-time Chairman of the Scientific Seminar of the Scientific Council for the award of scientific degrees,
Doctor of Philological sciences, professor

INTRODUCTION (the abstract of the (PhD) dissertation)

Topicality and necessity of the thesis. In the field of world linguistics, natural language processing areas are rapidly developing, with particular attention being paid to the study of computational linguistics achievements in various aspects, including important directions such as language research, linguodidactics, lexicographic analysis, and translation activities. Notably, the use of computer technologies and automated systems plays a crucial role in determining the structural and semantic aspects of language, enhancing the educational process, creating national corpora and electronic dictionaries, as well as in the effective organization of translation processes.

In the field of world linguistics in the 21st century, the scope of scientific and theoretical research in corpus linguistics has expanded, with studies being conducted to achieve significant results in various directions. Particular attention is being paid to improving machine translation, linguistic modeling of languages, creating lemmatization systems, identifying syntactic and semantic structures based on corpus analysis, as well as digitizing and highlighting the characteristics of national-cultural heritage specific to different languages.

In the context of rapid reforms being implemented in our country, significant attention is being given to processing the Uzbek language using modern information technologies, creating its national corpus, and conducting scientific and practical studies of its linguistic features. It is essential to develop an electronic national corpus of the Uzbek language that encompasses all scientific, theoretical, and practical information related to the language. In this regard, there is a need to further intensify scientific research on creating the national corpus of the Uzbek language, digitizing our national and cultural heritage, utilizing modern electronic platforms in educational processes, as well as elucidating scientific and methodological strategic tasks for establishing a database of scientific and technical texts for the Uzbek language national corpus.

This dissertation serves, to a certain extent, in implementing the tasks specified in the following decrees and resolutions: the Decree of the President of the Republic of Uzbekistan No. UP-4997 dated May 13, 2016 “On the Establishment of Alisher Navo’i Tashkent State University of Uzbek Language and Literature”, No. UP-4947 dated February 7, 2017 “On the Action Strategy for Further Development of the Republic of Uzbekistan”, No. UP-5850 dated October 21, 2019 “On Measures to Fundamentally Increase the Prestige and Status of the Uzbek Language as the State Language”, the Resolution No. PP-2789 dated February 17, 2017 “On Measures to Further Improve the Activities of the Academy of Sciences, Organization, Management and Financing of Scientific Research”, the Resolution of the Cabinet of Ministers of the Republic of Uzbekistan No. 984 dated December 12, 2019 “On Approval of the Regulations on the Department for Development of the State Language”, as well as other regulatory legal documents related to this field.

Research corresponds to the priority areas of scientific and technological development in the Republic. The research was conducted in accordance with the priority direction of the republic's science and technology development: “Social,

legal, economic, cultural, spiritual and educational advancement of the information society and democratic state, and development of an innovative economy".

The extent of study of the problem. Research on corpus linguistics in world linguistics began in the 1960s. Initial research papers and corpora were created in English. By the 1990s, corpora had been developed in numerous languages worldwide. The role, methodology, object, and tasks of corpus linguistics in traditional and computational linguistics have been explored in the studies of scholars such as G.Lich, R.Garside, J.Sinclair, L.Flowerdew, T.McEnery, A.Hardie, M.McCarthy, W.N.Francis, and G.Kennedy[27]. The social significance of corpora, the subsequent stages of development in corpus linguistics, the use of the Internet as a corpus, and the similarities and differences between the Internet and corpora have been thoroughly examined by researchers such as A. Kilgarriff, K. Stuart, G. Grefenstette, M. Hundt, and N. Nesselhauf, who have investigated the theoretical foundations of the issue. In Russian linguistics, a series of studies conducted by V. Plungyan, V.P. Zakharov, and A.B. Kutuzov have played a crucial role in establishing corpus linguistics as a distinct field and in the creation of the Russian National Corpus. In Uzbek linguistics, numerous significant research works have been carried out in the field of corpus linguistics. Notably, the scientific contributions of Sh.Khamroyeva, A.Eshmuminov, N. Abdurahmonova, G. Toirova, M.Abjalova, O.Abdullaeva, A.Rakhmanova, N.Gulomova, R.Karimov, M.Xolova, E.Khonnazarov, Z.Khusainova, A.Turdaliyev, and M.Tursunov[28] are among these

[27] Leech G. Corpus Annotation Schemes. In Literary and Linguistic Computing. – Vol. 8, No. 4. Oxford University Press, 1993. – P. 275-281.; Leech G., Wilson A. Recommendations for the morphosyntactic annotation of corpora. / EAGLES Document EAG-TCWG-MAC/R 1994. www.i1c.cnr.it/EAGLES/browse.htm1., Leech G., Garside R., Steven E.A. The Automatic Grammatical Tagging of the LOB Corpus // ICAXE Ncwo, 1983. p. 13-33. https://www.researchgate.net/publication/238760957; Garside R., Leech G., Sampson G. The CLAWS Wordtagging System / The Computational Analysis of English: A Corpus-based Approach. London: Longman, 1987. McEnery T., Wilson A. Corpus Linguistics (1st ed.). – Edinburgh: Edinburgh University Press, 1996; McEnery T.; Hardie A. Corpus linguistics: Method, theory and practice. Cambridge: Cambridge Univcrsily Press, 2011; Francis W.N., Johansson S. Problems of Assembling and Computerizing Large Corpora // Computer Corpora in English Language Research. – Bergen: Norwegian Computing Centre for the Humanities, 1982; Francis W.N., Svartvik J. Language Corpora B.C. // Directions in Corpus Linguistics, Proceedings of Nobel Symposium 82 Stockholm, 1991.; Kennedy G. An Introduction to Corpus Linguistics. Harlow: Addison Wesley Longman, 1998.; Sinclair J. Corpus, Concordance, Collocation. Oxford: Oxford University Press, 1991.; Flowcrdcw L. Corpus Linguistic Techniques Applied to Textlinguistics. 1998. – P. 541-552.; McCarthy M., O'Keefe A. What arc corpora and how have they evolved? The Routledge handbook of corpus linguistics. – London and New York, 2/1.

[28] Хамроева Ш.М. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари. Филол. фан. б. фалс. док. (PhD) дисс. автореф. – Қарши, 2018. – 54 б.; Хамроева Ш. Ўзбек тили морфологик анализаторининг лингвистик таъминоти: Филол. фан. док. ...дис. – Тошкент, 2021. – 268 б.; Эшмўминов А.А. Ўзбек тили миллий корпусининг синоним сўзлар базаси. Филол. фан. б. фалс. док. (PhD) дисс. автореф. – Қарши, 2019. – 46 б.; Abdurahmonova N. O'zbek tili elektron korpuslarining kompyuter modellari. Monografiya. –Toshkent, 2021. – 201 b.; Тоирова Г.И. Ўзбек тили миллий корпусини яратишнинг назарий ва амалий масалалари. Филол.фан.док. (DSc) дисс. автореф. – Бухоро, 2021. –72 б.; Abjalova M. Korpus lingvistikasi uslubiy qo'llanma. – Toshkent: Bookmany print, 2022; Abdullayeva O. O'zbek tilining internet axborot matnlari korpusini shakllantirishning nazariy va amaliy asoslari. Filol. fan. b. fals. dok.(PhD) ...diss. – Toshkent, 2022. – 158 b.; Рахманова А. Ўзбек тили миллий корпусини яратишда компьютер усуллари. Филол. фан. бўй. фалс. док. (PhD) ...дисс. – Тошкент, 2022. – 158 б.; G'ulomova N. Alisher Navoiy mualliflik korpusi va uninng semantik teglari bazasini yaratish. Filol. fan. b. fals. dok. (PhD)… diss. – Toshkent, 2022. – 189 b.; Karimov R. O'zbek-ingliz parallel korpusini tuzishning lingvistik va dasturiy masalalari. Filol. fan. b. fals. dok. (PhD)… diss. – Buxoro, 2022; Xolova M. O'zbek milliy shevalari korpusi tadqiqi (Boysun tumani "j"lovchi shevalari misolida). Monografiya. –Termiz, 2022. 144 b.; Xonnazarov E. O'zbek tili korpusi uchun zamonni ifodalovchi grammatik shakllarni annotatsiyalash: Filol. fan. b. fals. dok. (PhD)… diss. – Toshkent, 2023. – 159 b.; Xusainova Z. O'zbek tili birliklarini tokenlash, stemlash, lemmalashning lingvistik asoslari va dasturiy ta'minoti. Filol. fan. b. fals. dok. (PhD)… diss. – Toshkent, 2023. – 158 b.; Turdialiyev A. Denov shevasini

important works. As corpus linguistics is a new and rapidly developing field in Uzbek linguistics in recent years, there are very few monographic works on the subject. Notably, Sh.Khamroyeva's research on the linguistic foundations of compiling an Uzbek language author corpus, N. Gulomova's dissertation on creating an Alisher Navoi author corpus[29], and a monograph of the same title co-authored by M.Abjalova[30], discuss the formation, development, and theoretical foundations of corpus linguistics. These works also explore the specific theoretical and practical aspects of compiling an author corpus, as well as the general and specific linguistic principles underlying the creation of an author corpus[31]. In recent years, numerous research works have been published in Uzbek linguistics. Such research has practically laid the foundation for the emergence of corpora in Uzbek corpus linguistics. The National Corpus of the Uzbek Language[32], the Educational Corpus of the Uzbek Language[33], the Corpus of the Uzbek Language[34], and the Alisher Navoi Corpus[35] serve as evidence to support our statement. During the process of writing the dissertation, the scientific research of the aforementioned and several other Uzbek and international linguists was taken into consideration. In our study, unlike previous work carried out in this field, we have conducted a monographic investigation into creating a database of scientific and technical texts for the national corpus of the Uzbek language.

The relevance of the research to the work plans of the research institution where the dissertation was completed. The dissertation was carried out within the framework of the research plan of Alisher Navo'i Tashkent State University of Uzbek Language and Literature for 2022-2024, focusing on the topic "Social, historical, and contemporary development of language".

The purpose of this research is to develop theoretical and software foundations for creating a database of scientific and technical texts for the National Corpus of the Uzbek language. This includes linguistic tagging of lexical units in scientific and technical texts and classifying the database of Uzbek scientific and technical texts.

The tasks of the research. Based on the main goal, the following scientific tasks were set before the research:

- Analyze the theoretical and methodological foundations of corpus linguistics, highlight the stages of its formation and development trends, as well as substantiate scientific views and approaches that are crucial in creating a national corpus;

areal o'rganish va til korpusiga joylashtirish. Filol. fan. b. fals. dok. (PhD)… diss. – Toshkent, 2024. – 173 b.; Tursunov M. O'zbek lingvistik korpusini yaratish tajribasidan (konkordans, tokenayzer, lemmatayzer, razmetkalash dasturlari asosida). Filol. fan. b. fals. dok. (PhD)… diss. avtoref. – Qo'qon, 2024. – 60 b.

[29] G'ulomova N. Alisher Navoiy mualliflik korpusi va uning semantik teglari bazasini yaratish ("Badoye' ul-vasat" devoni asosida). Filol. fan. b. fals. dok. (PhD)… diss. – Toshkent, 2023. – 190 b.

[30] Abjalova M., G'ulomova N. Mualliflik korpuslari: Alisher Navoiy mualliflik korpusi. [Matn] : monografiya / M.A. Abjalova, N.S. G'ulomova. – Navoiy, 2023. – 216 b.

[31] Хамроева Ш.М. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари. Филол. фан. б. фалс. док. (PhD) дисс. автореф. – Қарши, 2018. –54 б.

[32] http://uzbekcorpora.uz/ijrochi

[33] https://uzschoolcorpara.uz/

[34] https://uzbekcorpus.uz/

[35] http://v1.alishernavoicorpus.uz/korpus_haqida/

- Define the linguistic criteria necessary for creating a database of scientific and technical texts in the Uzbek language, explain the theoretical goals and objectives in this field on a scientific basis, and justify the interdisciplinary characteristics of scientific and technical texts;

- Select modern computer-technological approaches for creating a database of scientific and technical texts for the national corpus of the Uzbek language, analyze and systematize software tools used in the corpus compilation process;

- Develop an algorithm for linguistic annotation of texts, elucidate the theoretical and practical aspects of the tagging process, and determine its effectiveness through experimental results.

The object of the research is various scientific and technical texts in the Uzbek language related to the creation of a database of scientific and technical texts for the national corpus of the Uzbek language.

The subject of the research is the problems of forming a database of scientific and technical texts in the Uzbek language and their tagging.

Research methods. In covering the research topic, methods of classification, description, comparison, and statistical analysis were used.

The scientific novelty of the research consists from:

The importance of corpus and its capabilities for centralizing and processing data in the fields of linguodidactics and linguistics, as well as the significance of effective corpus utilization, have been substantiated;

It has been demonstrated that the software http://topicmodel.uz/ has been developed using LDA, LSA, and NMF methods and algorithms for tagging scientific and technical texts in the Uzbek National Corpus, covering fields such as education, sports, healthcare, politics, culture, weather, economics, and technology;

It has been proven that the methods of tagging scientific and technical texts in the Uzbek language can be utilized in developing NLP applications such as text comprehension, information extraction, sentiment analysis, spell checking, document summarization, and machine translation;

The requirements for scientific and technical texts in forming a text database have been identified, tags for incorporating scientific and technical texts into the corpus have been developed, and the distinctive features of scientific and technical texts that enable semantic annotation in the Uzbek National Corpus have been determined based on analysis.

The practical results of the research are as follows:

- In order to create a database of scientific and technical texts for the National Corpus of the Uzbek language, separate linguistic methods related to the scientific and technical aspects of the corpus have been developed through research and experiments on language and meaning;

- In the process of creating a database comprising numerous scientific and technical texts, new methodologies have been developed aimed at refining lexical and grammatical structures, as well as terminology, with a focus on achieving a high level of accuracy and standardization;

- The importance of transforming scientific and technical texts from various fields into a valuable resource for new scientific research within the corpus has been substantiated through their proper tagging and annotation;

- The establishment of a scientific and technical text database for the National Corpus of the Uzbek language, the development of algorithms ensuring its effective operation through new computer technologies, as well as error analysis, active detection, and optimization of data within the corpus have been identified.

The reliability of the research results is substantiated by the precise formulation of the problem, the scientific conclusions drawn being based on descriptive and comparative analyses and their practical implementation, the application of analytical methods, the practical implementation of theoretical ideas and conclusions, and the validation of obtained results by competent organizations.

Scientific and practical significance of research results.

The scientific significance of the research results is determined by the possibility of using theoretical conclusions as a source in linguistic and information technology fields. These conclusions pertain to creating a database of scientific and technical texts for the national corpus of the Uzbek language, determining its lexical-semantic and grammatical structure, as well as evaluating the effectiveness of information technologies in data processing and automation.

The practical significance of the research results is explained by the possibility of their application in creating a database of scientific and technical texts for the national corpus of the Uzbek language, determining its semantic and grammatical structure, as well as developing methods for automatic text processing and analysis. The results can be utilized in teaching courses such as “Computer Linguistics”, “Mathematical Modeling”, “Theory of Textology”, and “Lexicography”, in creating textbooks, manuals, and educational complexes, as well as in compiling scientific and technical dictionaries and reference books.

Implementation of research results. Based on the scientific results obtained for creating a database of scientific and technical texts for the National Corpus of the Uzbek Language:

- Conclusions regarding the importance of considering informational, integrity, consistency, and terminological features when using translations of scientific and technical texts were applied in the practical project PF-201912258 – “Creation of a multilingual (Uzbek, Russian, English) electronic platform for Uzbek literature” (2021-2023) (Reference No. 04/1-4044 dated December 6, 2024, from Alisher Navo’i Tashkent State University of Uzbek Language and Literature). As a result, the database of scientific and technical texts for the electronic platform has been enriched with new information.

- The software database, tagging system, and conclusions on annotation principles developed during the dissertation preparation were utilized in the practical project AM-F3-201908172 titled “Creation of an educational corpus of the Uzbek language” (2020-2023) (Reference No. 04/1-4085 dated December 10, 2024, from Alisher Navo’i Tashkent State University of Uzbek Language and Literature). Consequently, the scientific and technical materials, thematic modeling algorithms,

and proposals for latent semantic analysis presented in the study have contributed to enriching the content and practical aspects of the corpus.

- The practical significance of the research results, including the creation of a scientific and technical text database for the Uzbek national corpus, determination of its semantic and grammatical structure, and conclusions on developing methods for automatic text processing and analysis, were used in writing the script for the "Hamma uchun" program on the "O'zbekiston tarixi" TV channel of the National Television and Radio Company of Uzbekistan, as evidenced by certificate No. 06-28-845 dated December 24, 2024.

Approbation of the research results. The results of this research were presented at 3 international and 2 national scientific and practical conferences.

Publication of the research results. A total of 10 scientific works have been published on the topic of the dissertation. Of these, three are published in scientific journals recommended by the Higher Attestation Commission of the Republic of Uzbekistan for publishing the main scientific results of doctoral dissertations, and three are presented in foreign journals and conference abstracts.

Structure and volume of the dissertation. The dissertation comprises an introduction, three chapters, a conclusion, and a list of references, with a total volume of 167 pages.

MAIN CONTENT OF THE DISSERTATION

The introduction substantiates the relevance and necessity of the dissertation topic, defines the research goals and objectives, its object and subject, demonstrates the alignment with priority directions of science and technology development in the republic, outlines the scientific novelty and practical significance, and provides information on the implementation of research results, published works, and the structure of the dissertation.

Chapter I of the dissertation is titled "**Linguistic Foundations for Providing a Scientific and Technical Text Database in the National Corpus of the Uzbek Language**" and consists of 4 sections.

Section 1.1 is dedicated to the study and development of the text phenomenon in world linguistics. Although the field of text linguistics began to gradually develop in world linguistics from the 1970s, views related to the issue of text emerged in the 1940s. Initially, Western linguistics paid serious attention to the study of text and its nature. The emergence of the concept of "text" in Russian linguistics dates back to the 1940s. In 1947, A.I. Belich, in his article on the classification of linguistic disciplines, emphasized that in the grammatical description of language facts, special attention should be given to the entire chain of sentences connected on the basis of semantic commonality and manifested as a certain syntactic-semantic unity. He noted that this was crucial for the emergence of the concept of "text", and stressed the importance of studying the interrelationships and connections in such sentence chains within the syntactic branch of linguistics. Along with A.I. Belich[36], German linguists also expressed their views on text in the 1940s. Some literature suggests that the works of

[36] Белич А.И. К вопросу о распределении грамматического материала по главным грамматическим дисциплинам / Вестник МГУ, 1947, N 7, с. 24

I.A. Baudouin de Courtenay laid the foundation for views on text, which later became the focus of research for a number of scholars such as M.M.Bakhtin[37], V.V.Vinogradov[38], and L.V.Shcherba[39]. However, for a long time, researchers were primarily interested in text from a content (semantic) perspective (A.M. Peshkovsky, L.V.Shcherba, V.V.Vinogradov, A.Weil, V.Mathesius, S.Bally, Z.Harris). In the 1960s, text linguistics mainly studied ways to maintain textual coherence and comprehensibility, methods of expressing personal and subjective preference (anaphoric structures, pronominalization, lexical repetitions, tense chains, etc.), and the distribution of sentences according to theme and rheme. During the 1960s, a linguistic theory of text began to take shape in Western Europe under the direct and indirect influence of the Czech school. In this tradition, the text was initially analyzed using structural methods, primarily based on similarities with more familiar language elements. The contributions of representatives from Czech (members of the Prague Linguistic Circle), German, French, English, American, Dutch, Polish, Russian, and other linguistic schools to text theory, text description, and the general formation and development of text linguistics are widely recognized in global linguistics and are consistently cited in scientific research. In Russian linguistics, the issues of text theory and linguistics have been extensively studied by numerous linguists, including V.V. Odintsov, I.R. Galperin, O.I. Moskalskaya, L.M. Loseva, Y.M. Lotman, Z.Y. Turaeva, N.D. Zarubina, E.V. Sidorov, O.L. Kamenskaya, A.I. Gorshkov, and N.S. Valgina[40]. The text and its constituent elements, factors, and characteristics have been studied from various perspectives. This field has given rise to significant debates, one might say. Some specialists even considered text linguistics not as a separate branch of linguistics, but rather as the foundation or crucial basis of linguistics as a whole. Linguist O.I. Moskalskaya, who comprehensively analyzed research in this direction within linguistics and explained the study of text linguistics in the true sense and the importance of these studies, notes that by the 1960s and 1970s, interest in the linguistic study of text had increased significantly. She emphasizes that a large number of studies on text linguistics emerged in world linguistics, and text linguistics gained full recognition as an independent linguistic science.

Section 1.2. This section is titled “The Study and Historical Development of the Text Phenomenon in Uzbek Linguistics”. The study of text and its types in Uzbek linguistics has been conducted in several stages. Research in the field of text analysis primarily focuses on examining the semantic, syntactic, and communicative functions, structures, and roles of texts in linguistics and literary studies. We have deemed it necessary to consider these stages as follows: 1. Defining the concept of text. When

[37] Бахтин, М.М. Проблема текста в лингвистике, филологии и других гуманитарныхнауках. Опыт философского анализа / М.М. Бахтин // Литературнокритические статьи. - М.:Художественная литература, 1986.

[38] Виноградов, В.В. Стилистика. Теория поэтической речи. Поэтика / В.В. Виноградов. – М.: АН СССР, 1963.

[39] Щерба, Л.В. Избранные работы по русскому языку / Л.В. Щерба. – М., 1957.

[40]Лотман Ю.М. Структура божественного текста. - М.: Искусство, 1970; Лингвистика текст. Материалы научной конференции. Ч. я, II. - М., 1974; Лосева Л.М. Как строится текст. - М.: Просвещение, 1980; Одинцов В.В. Стилистический текст. - М.: Наука, 1980; Москальская О.И. Грамматика текста. –М.: Высшая школа, 1981; Гальперин И.Р. Показанные работы; Зарубина Н.Д. Текст: лингвистические и методологические аспекты. - М.: Русский язык, 1981; Тураева З. Я. Лингвистика – это текст. - М.: Просвещение, 1986; Сидоров Е.В. Проблема в речевой систематичность. - М.: Наука, 1987; Каменская О.Л. Текст и общение. - М.: Высшая школа, 1990; Горшков А.И. Русская стилистика. -М.: Астрель-АСТ, 2001, с. 63- 250; Валгина Н.С. Теория в тексте. -М.: Логос, 2004 и др.

determining text types and distinguishing them according to their functions, it is essential to pay attention to the specific features of the text and the factors that contribute to its status as a text. The initial stage of textology begins with defining the concept of text. At this stage, scholars start analyzing the specific features of the text, its structural and linguistic elements. Here, the problems are addressed through questions such as “What is a text?” and “How can texts be categorized?” This stage also determines the place and significance of text in linguistics. 2. Classification of text types. In the second stage, texts are classified. This involves determining the types of texts, their characteristics, and communicative functions. At this stage, the normative and methodological foundations for text classification are developed. 3. Applied research in textology. During the practical study of text types, linguists analyze the linguistic and stylistic features of texts. At this stage, texts are examined from grammatical and lexical perspectives. The communicative function (purpose) and audience of the text are considered in relation to its oral and written forms. Translation, analysis, and identification of stylistic features, as well as analysis of cohesive devices, are employed in practical work. 4. Changes and development of text types. Since language is a product of social processes, texts are also dynamic units. The development and evolution of text types are studied as well. At this stage, scholars analyze how texts take on new forms due to cultural, social, and linguistic changes. For example, new types of texts that have emerged on the Internet (blogs, forum posts, web pages), and text forms associated with new technologies, media, and communications. 5. Integrated approach in textology. New approaches and methods are being developed in textology. At this stage, texts are studied not only from a linguistic perspective but also from psychological, social, and cultural viewpoints. Additionally, new directions are emerging in connecting text types with other fields, such as literary studies, cultural studies, and information and communication sciences. 6. Modern methods of textual studies. At this stage, modern technologies and methodologies are applied in textual studies. Work is carried out in new areas such as corpus linguistics and computer analysis (automatic analysis of texts). Intertextual analysis of the text, semiotic approaches, and other modern analytical methods are employed. We are witnessing the achievement of significant scientific results in the research conducted at these stages. The phenomenon of text has been studied in the aforementioned aspects by scholars such as A.Mamajonov, N.Mahmudov, M.Hakimov, M.Yuldashev, N.Turniyozov, B.Yuldashev, M.Abdupattoyev, and M.Kurbanova. The Uzbek linguist M. Hakimov is noted for his work on types of text, specifically the syntactic and pragmatic issues of scientific texts. In his candidate dissertation, he highlighted the connectives that express semantic relationships between scientific text and its units, their specific features, and functions. In addition to examining the author's personal stance in scientific texts and distinguishing its types, he studied the syntagmatic and pragmatic features of scientific texts based on factual data[41]. In Uzbek linguistics, the work conducted in the field of text analysis using computer programs has been based on numerous scientific studies and technological approaches. Research in this area has primarily focused on automating the syntactic, morphological, and semantic analysis

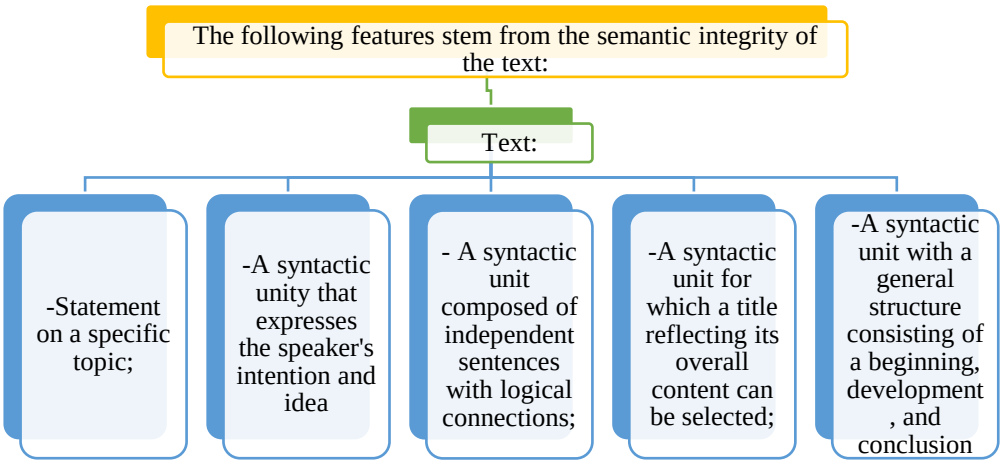
[41] Ҳакимов М. Ўзбек илмий матнининг синтагматик ва прагматик хусусиятлари: Филол. фан. н-ди ...дисс. –Тошкент, 1993.

of the Uzbek language using computers. For instance, Sh.Khamroyeva's doctoral dissertation titled "Linguistic support of the morphological analyzer of the Uzbek language" established the general principles for automatic morphological analysis of language categories. N.Abdurakhmanova's monograph "Computer Models of Electronic Corpora of the Uzbek Language" encompasses scientific research on computer-aided analysis and processing of the Uzbek language. This monograph primarily addresses the creation of electronic corpora of the Uzbek language, as well as the possibilities and methodologies for working with their computer models.

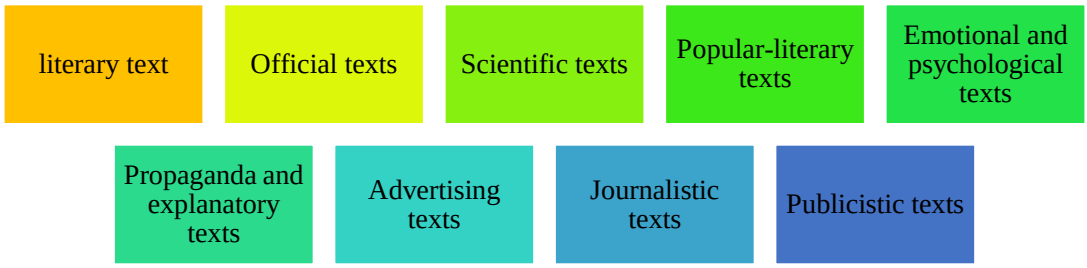
Computational linguistics technologies enable automatic analysis, classification, and determination of text structures. This, in turn, facilitates the development of new methods and approaches for studying text types and understanding their characteristics.

Section 1.3 is titled "Description and interpretation of text and its types".

Text description is the process of analyzing and interpreting the content, structure, linguistic, and stylistic features of a text. The description of a text helps to understand its unique characteristics, composition, and structure. This process aims to determine the quality of the text, its impact on the reader, the linguistic devices employed, and the text's purpose and function.



Text types are formed based on various norms, and each type has its own linguistic and stylistic devices, purposes, and structure. According to these foundations, the classification of text types is as follows:



They are all characterised by a news function and a focus on presenting scientific information, mainly in written form, in a logically consistent, objective and evidence-based manner. Each type of text has its own specific function and is

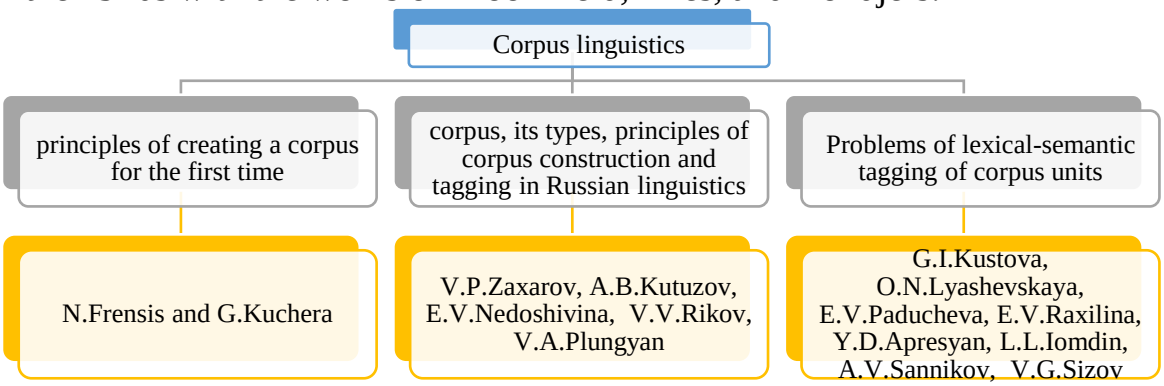
aimed at a particular audience, group or individual. They have their own style, structure and content and are used to inform, present, instruct, express creativity, describe, evaluate and analyse.

Section 1.4 is dedicated to scientific texts and issues related to their comprehension. In scientific texts, unlike business-related, journalistic, or literary texts, functional types of speech are employed (such as description, exposition, reasoning, argumentation, and others). Scientific texts are written in specialized field-specific language, which can complicate the process of mass reading. Each field has its own specific terms and concepts, and incomplete understanding of these can lead to poor comprehension of the text. However, a comprehensive database of terminology dictionaries for each field does not yet exist. For instance, the terminology used in mathematics, IT, or chemistry differs significantly from everyday language. To address this, it is advisable to use dictionaries or field-specific scientific sources to identify and interpret new terms while reading and working with scientific texts. This, of course, requires new and comprehensive dictionaries. Such dictionaries can be compiled using the database of scientific texts available in the national corpus of the Uzbek language. Scientific texts often include summary results, experiments, or research findings from multiple sources. Understanding such texts requires the ability to combine, compare, and analyze data. This is especially important for educational processes that require connecting with previously read texts. After reading the text, it is necessary to compare the main idea or concept and content with other information, conduct research, and study additional sources. This enhances the effectiveness of the text's analysis and synthesis process. In this process, the database of the Uzbek language national corpus is essential for easily finding and using scientific texts in Uzbek. Scientific texts are usually structurally complex. They often contain long sentences, analyses, formulaic statements, and statistical analytical data. Consequently, it becomes challenging to fully comprehend the text. This issue can be addressed by breaking the text into smaller parts and striving to understand each part thoroughly. After reading each section, it is beneficial to review the main goal and idea again in a general sequence. We know that scientific texts include references to other studies, scientific sources, or articles. Failing to understand these sources when examining them also makes it difficult to grasp the full meaning of the text. It is advisable to search for and verify the referenced sources. By paying attention to additional research in the text, a general understanding of the issue is formed. In many cases, depending on the field, scientific texts contain statistical data, graphs, and diagrams, which require the ability to fully comprehend and analyze statistical information and read numerical data. When analyzing statistical diagrams and graphs, it is recommended to study statistical analysis methods and use software developed with NLP algorithms for analysis. In scientific language, there may be some ambiguities regarding the author's approach to the issue and their analyses. For example, when talking about an event or phenomenon, words such as "probably", "approximately", "in general" are used, which can lead the reader to ambiguities in understanding the text. When reading scientific texts, it is necessary to analyze the purpose of using ambiguities and probabilities. Analyzing the text without them helps to draw clear

conclusions. Scientific texts present unique and new ideas. It is necessary to critically analyze these ideas and compare them with other previously read information. This requires the reader to develop critical thinking, objectively assess the main ideas in the text and the reliability of the research. Because the result of one scientific text can serve as a prelude to the emergence of another scientific text. Types of scientific texts: Scientific article. Dissertation. Monograph. Textbooks and textbooks. Scientific abstracts. Research reports. Proceedings of scientific conferences. Scientific presentations. The main factor in popularizing scientific and technical texts and ensuring their easy access to the scientific world and the user is their inclusion in the corpus database.

Chapter II, titled "**Theoretical Issues of Creating a Scientific and Technical Text Database in the National Corpus of the Uzbek Language**", consists of 3 sections.

Section 2.1 is titled "Analysis of Corpus Linguistics Research in World Linguistics". Focused research in the field of corpus linguistics in world linguistics began in the 1940s with the works of Bloomfield, Fries, and Bondjers.



In the field of computational linguistics, particularly for the English language, the first large-scale text corpus was created using machines by scientists N. Francis and G. Kucera from Brown University in the United States during the 1960s. This corpus is recognized as the first major corpus, containing nearly 500 samples of American English and encompassing more than 15 American prose works. Today, numerous language corpora exist, created for various purposes, showcasing special types and diverse capabilities of language corpora. Scientists have come to understand that as the volume of linguistic materials included in a corpus increases, the opportunities for studying and researching linguistic phenomena expand. This is because each corpus must meet certain qualitative and quantitative requirements[42]. The Russian National Corpus, inspired by the British National Corpus[43], is distinguished by its inclusion of words not only from the modern language but also from historical layers. Scholars note that by the 1990s, the number of created corpora had exceeded 600. Analysis of existing corpora reveals that a corpus should present

[42] Грудева Е.В. Корпусная лингвистика: учеб. пособие / Е.В. Грудева. – 2-е изд., стер. – М.: ФЛИНТА, 2012. – 165 с.

[43] Британия Миллий корпуси (British National Corpus – BNC)га тенг равишда, инглиз тили корпусларидан яна: инглиз тили Халыаро корпуси (International Corpus of English – ICE), инглиз тили лингвистик Банки (Bank of English), Ҳозирги замон Америка инглиз тили Корпуси (Corpus of Contemporary American English – COCA) ва б.ларни ҳам келтириб ўтиш мумкин.

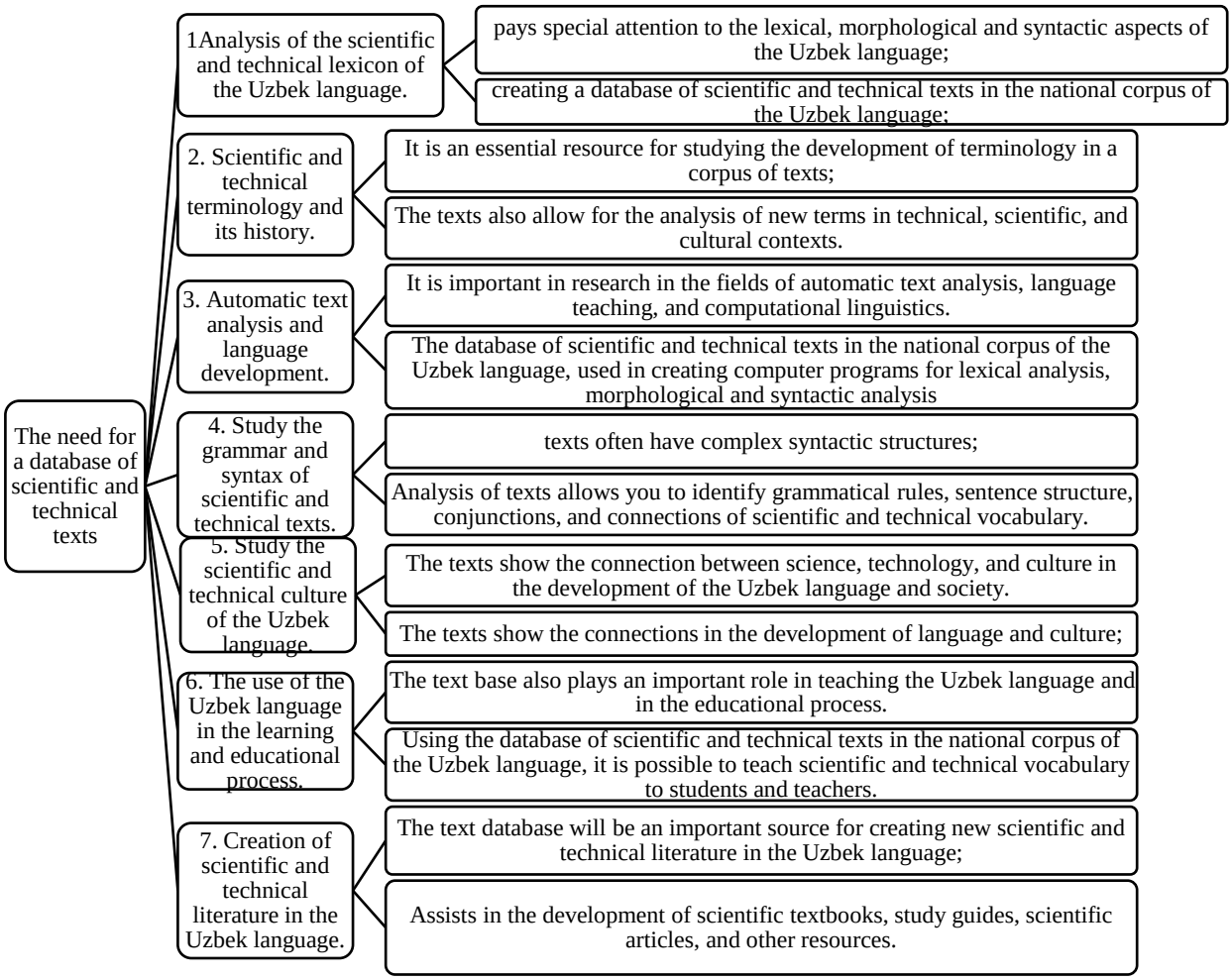
as many linguistic patterns and features as possible while maintaining a balanced composition. The proportion of each text type within the corpus should mirror its volumetric share in the overall body of texts in the language.

Section 2.2. This section is dedicated to the development of corpus linguistics in Uzbek linguistics. Uzbek corpus linguistics has also reached a stage of significant development. As evidence of this progress, one can cite the practical results of the National Corpus of the Uzbek Language[44], the Educational Corpus of the Uzbek Language[45], and the Corpora of the Uzbek Language[46]. Corpus linguistics operates in two main directions: the creation of corpora and the study of theoretical and practical language problems using corpus methods. The formation of corpus linguistics in Uzbekistan began with theoretical research in 2018 and advanced further with the creation of the educational corpus of the Uzbek language in 2021. Uzbek linguists Sh.Khamroyeva, N.Abdurakhmonova, M.Abjalova, A.Eshmuminov, D.Akhmedova, U.Kholiyorov, G.Toirova, D.Urinboeva, and A.Rakhmanova have conducted research on language corpora and their types. As a result, they have investigated the linguistic foundations of compiling author corpora[47], natural language processing[48], semantic tagging of synonymous words[49], lexico-semantic tagging of nominative units[50], creation of an educational corpus of the Uzbek language[51], models of electronic corpora, and computer methods of corpus creation[52]. In recent years, efforts to create a national corpus of the Uzbek language have led to the launch of several fragmented corpora. Examples of this include the “National Corpus of the Uzbek Language” (based on materials from the “Alpomish” epic) developed within the framework of a project carried out by the Samarkand branch of Tashkent State University of Information Technologies in collaboration with the team from Samarkand State University; the “Corpus of the Uzbek Language” created under the leadership of N.Abdurakhmanova; as well as N.Gulomova’s “Alisher Navoi Author’s Corpus” and O.Abdullayeva’s “Tug‘ro” corpus of Uzbek language information texts. The foundation for these practical developments was laid several years ago through the aforementioned monographic studies. In this section of the dissertation, the educational corpus of the Uzbek language, the national corpus of the Uzbek language, and the significant aspects of the Uzbek language corpora currently in operation are examined, and opinions are

[44] http://uzbekcorpora.uz/ijrochi
[45] https://uzschoolcorpara.uz/
[46] https://uzbekcorpus.uz/
[47] Xamroyeva Sh. O‘zbek tili mualliflik korpusini tuzishning lingvistik asoslari: Filol.fan.bo‘yicha falsafa doktori (PhD) dis. aftoref. – Qarshi, 2018. – 53 b.
[48] Abjalova M. Tahrir va tahlil dasturlarining lingvistik modullari. [Matn]: monografiya. – Toshkent, 2020. – B. 176.
[49] Eshmo‘minov A.A. O‘zbek tili milliy korpusining sinonim so‘zlar bazasi: Filol.fan.bo‘yicha falsafa dokt. (PhD) diss. – Qarshi, 2019. – 140 b.
[50] Ahmedova D. Atov birliklarini o‘zbek tili korpuslari uchun leksik-semantik teglashning lingvistik asos va modellari: Filol.fan.bo‘yicha falsafa dokt. (PhD) diss. – Buxoro, 2020. – 156 b.
[51] Raupova L., Elov B., Abjalova M., Alayev R. O‘zbek tilining ta’limiy korpusi va uning imkoniyatlari. // O‘zbekistonda til va madaniyat, – Toshkent: ToshDo‘TAU, 4/2021. – B. 60 75.; Абжалова М. Синонимайзер (синонимизатор) в образовательной корпусе узбекского языка. // TurkLang – 2021: Turkiy tillarni kompyuterda qayta ishlash IX xalqaro konferensiyasi.
[52] Raxmanova A. O‘zbek tili Milliy korpusini yaratishda kompyuter usullari. Filol.fan.bo‘yicha falsafa dokt. (PhD) diss. – Farg‘ona, 2022. – 52 b.

presented. Additionally, views on the theoretical and practical significance of recent research in corpus linguistics are included. At present, one of the most urgent tasks in our applied linguistics is the creation of a comprehensive national corpus of the Uzbek language. This is because the national corpus essentially represents a rich treasury of the national language.

Section 2.3. The necessity of a scientific and technical text database for the National Corpus of the Uzbek Language Creating a database of scientific and technical texts for the National Corpus of the Uzbek Language is of great importance for an in-depth study of the modern development of the Uzbek language, including its grammatical, lexical, and stylistic features. The necessity for a scientific and technical text database is based on several factors:



The creation of a database of scientific and technical texts for the national corpus of the Uzbek language is of great importance for scientific research in linguistics, lexicography, grammar, syntax, and many other fields. It also serves as a solid foundation for studying the modern development of the Uzbek language and creating scientific and technical literature.

Chapter III is devoted to "**Tagging the Scientific and Technical Text Database in the National Corpus of the Uzbek Language and its Software**" and consists of 3 sections.

Section 3.1. is dedicated to the types of corpora, their characteristics, and methods of text markup. To develop software based on Natural Language Processing

(NLP), it is necessary to ensure that the NLP system learns from available data. This can be accomplished using a language corpus. A language corpus is considered a large collection of structured texts presented in electronic form. The language corpus represents written or oral material, and linguistic analysis must be carried out for the NLP system to learn from available resources. The language corpus is the foundation of the natural language processing system. It can contain information in various forms, ranging from newspapers, novels, recipes, and radio broadcasts to television programs, films, and tweets. Language corpus is used to solve many tasks in NLP: training machine learning models, language understanding, discourse analysis, translation studies, rule-based systems, vocabulary and semantics, statistical analysis, and domain knowledge. Various types of corpora are available for developing NLP systems. In NLP, language corpora are divided into different types based on various criteria such as content, purpose, or source.

Select the type of corpus texts.

An important aspect of creating a text corpus is ensuring the presence of all types of written texts within it. Most language corpora consist of written texts in standard formats. To effectively form the corpus content, it is necessary to include a collection of texts from all possible areas of human knowledge. When selecting and gathering texts, one should refer to catalogs and publication lists from various publishers to collect sources and materials (such as books, newspapers, magazines, and others). For example, the Uzbek language corpus contains 25% literature, 3% fine arts, 8% social sciences, 15% natural sciences, 10% commerce, 35% mass media, and 4% translations. Within each category, there are several subcategories. For instance, literature includes novels, stories, essays, and others; fine arts include painting, drawing, music, sculpture, and others; social sciences encompass philosophy, history, education, and others; natural sciences include physics, chemistry, mathematics, geography, and others; mass media comprises newspapers, magazines, posters, announcements, and others; commerce includes accounting, banking, and others; and translation covers all topics translated into natural language.

Section 3.2. “The issue of thematic modeling of scientific and technical texts in the corpus”. In recent years, the development of methods for linguistically tagging scientific and technical texts in corpora, as well as the increase in tagging tools to support the creation and storage of tagged data, has contributed to the development of NLP applications such as Amazon Mechanical Turk. Currently, the NLP method of thematic modeling is being used for tagging scientific and technical texts in language corpora. Thematic modeling is a machine learning method that automatically analyzes text data to identify cluster words for collections of scientific and technical texts. This allows for the study of a collection of scientific and technical texts, and based on the statistical data of words in each text, it enables the identification of topics and determination of the thematic balance in each text. Extracting words from scientific and technical texts in a corpus requires more time, and this process is much more complex than extracting them from given topics in the text. For example, let's analyze a language corpus containing 100,000 texts with an average of 500 words each. Therefore, 500*100000=50,000,000 operations need

to be performed to process this corpus. Thus, when analyzing a text that includes specific topics, if it consists of an average of 5 topics, processing involves 5*500 words = 2,500 operations (threads). This appears simpler than processing an entire text, which is why thematic modeling is useful in solving such problems and facilitates the visualization of the process. In NLP, it is necessary to carry out initial text processing stages that facilitate text processing:

–Remove unimportant words and punctuation marks.

–Stemming.

–Lemmatization.

–Formation of statistical values using the CountVectorizer or Tf-Idf method.

Examples of popular thematic modeling algorithms today include Latent Semantic Analysis (LSA), Latent Semantic Indexing (LSI), Hierarchical Dirichlet Process (HDP), Latent Dirichlet Allocation (LDA), and Non-negative Matrix Factorization (NMF). Thematic models help us identify potential topics hidden in the entire text of scientific and technical documents as combinations of words with similar meanings and combinations of topics in each given text.

Data clustering is an unsupervised machine learning method used to identify or cluster small groups within a dataset. The main idea of clustering is as follows: by dividing observations in a dataset into different groups, the observations within each group should be very similar to each other, while observations in different groups should be completely different from each other. To determine the similarity between observations, it is necessary to choose appropriate criteria. Currently, many methods for cluster analysis have been developed.

Non-Negative Matrix Factorization (NMF) method. NMF is a method of decomposing a matrix into two matrices, where none of the three matrices contain negative elements. The NMF method is mainly used in recommendation systems, signal processing, and bioinformatics fields.

Latent Dirichlet Allocation (LDA) method. In NLP, thematic modeling using the LDA method allows for the identification of hidden (latent) topics in a collection of texts by determining possible themes based on words in scientific and technical texts. In the Latent Dirichlet Allocation method, each text and each word in the corpus, as well as the relationships between each topic and words, are modeled using latent variables. Each text in the corpus is represented using a Dirichlet distribution over latent variables (topics), and each topic is calculated using another Dirichlet distribution over words in all scientific and technical texts.

Section 3.3. “Software for tagging scientific and technical texts in the corpus”.

When using the LDA method to implement tagging of scientific and technical texts in the corpus, we take a set of words as the input matrix, since this is considered a probabilistic model. In the next step, the LDA algorithm splits the matrix into two smaller matrices:

document-topic matrix;

word-topic matrix.

In the LSI method, instead of examining each scientific and technical text separately from others, all scientific and technical texts and their terms are considered to determine relationships. SVD can be used to solve the problem of

approximation to a lower-rank matrix. Then, from this matrix, it is possible to determine term-document matrices. For this, it is necessary to perform the following three-step process:

1. For the given C, we form the SVD matrix $\boldsymbol{C = U \sum V^T}$. 2. Form the matrix $\sum$ formed by replacing the smallest r−k singular values on the diagonal $\sum$ with zero. 3. Approximate $\boldsymbol{C_k = U \sum kV^T}$ to C in the k-th degree. Here, C is the term-document matrix, and U, $\sum$ and V^T are the SVD matrices. Based on the theoretical information given above, we use the LSI module using Python language tools. In the LSI module, the dimensionality reduction for the LSI method is performed using the SVD method. The LSI method has one main weakness - uncertainty. For example, how can the system determine that you are talking about Microsoft Office or the office where you work? In this case, the LDA method can be used. For example, if the observations are words collected in scientific and technical texts, it emphasizes that each scientific and technical text is a mixture of a small number of topics and the presence of each word refers to one of the topics of the scientific and technical text. The LSI method examines the words used in the document and determines their relationships with other words. The LSI method allows the system to identify words that may be relevant to the scientific and technical text, even if they are not used in the document itself.

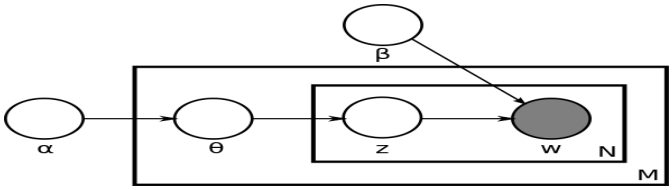
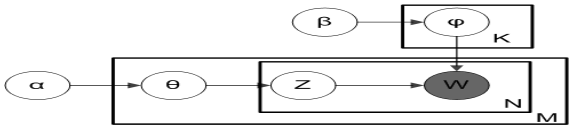


Figure 1

Figure 1 above illustrates the architecture of the LDA method, where α is the Dirichlet distribution parameter for topics in each scientific and technical text; β is the Dirichlet word distribution parameter for topics; θ_m represents the distribution of topics in the m-th scientific and technical text; φ_k denotes the distribution of words for the k-th topic; z_{mn} is the topic for the n-th word in the m-th scientific and technical text; and w_{mn} is the word. The gray color of W indicates that the words $\mathbf{w_{ij}}$ are the only observable variables, while the remaining variables are latent. To model the distribution of "topic-word" pairs, a sparse Dirichlet distribution can be utilized. This is because the probability distribution over words in a topic is concentrated, resulting in only a small set of words having high probability. This model is currently the most widely used variant of the LDA method. The illustrated form of this model is presented below, where K represents the number of topics and $\varphi_1,\ldots,\varphi_k$are V-dimensional vectors. The φ_i - **V** -dimensional vectors contain parameters distributed by Dirichlet over topics and words. Objects represented by θ and φ can be considered as matrices created by decomposing the original "document-word" matrix that represents the corpus of the modeled scientific and technical text. θ consists of rows labeled by scientific and technical texts and columns labeled by topics, while φ consists of topics labeled by rows and words labeled by columns. Thus, $\boldsymbol{\varphi_1,\ldots,\varphi_k}$ denotes a set of rows or vectors, where φ_i represents the distribution over words. Similarly, $\boldsymbol{\theta_1,\ldots,\theta_k}$ denotes a set of rows, each of which represents a distribution over topics.

Figure 2.



The LDA method assumes that each scientific and technical text in the corpus contains a mixture of topics present in the entire corpus. The topic structure is hidden, and we can observe not the topics themselves, but only scientific and technical texts and words. Since the structure is hidden, the LDA method aims to draw conclusions about the topic structure, taking into account certain words and scientific and technical texts. Structure of the scientific and technical text by the LDA method

The LDA method identifies a set of topics through corpus words (vocabulary) of scientific and technical texts with certain probabilities. It is assumed that scientific and technical texts are developed based on the following rules:

- determine the number of words (N) in scientific and technical texts;
- determine the set of topics for scientific and technical texts (according to the Dirichlet probability distribution for the specified set of K topics);
- generate each word in scientific and technical texts as follows:
- first select a topic;
- use the topic to generate the word (according to the multinomial distribution of the topic).
- considering this generative model for the set of scientific and technical texts in the corpus, try to go back through the scientific and technical texts in the corpus to identify the set of topics that could have generated the LDA set.

In the next step, we will consider the implementation of the LDA method using Python. To increase the accuracy of the calculations, it is necessary to remove all punctuation marks from the texts of the language corpus. N-gram language models. Bigrams are two words that occur together in the scientific and technical texts in the corpus. Trigrams are three words that occur together. The **Gensim.Rhrases** method of the **Gensim** rocket helps to identify n-grams in language corpus texts. In order to perform thematic modelling of the given language corpus texts based on the LDA method, we need to form bigrams, trigrams and lemmatize the set of scientific and technical texts in the corpus. Lemmatization is the process of grouping lexical forms of a word, which can be analyzed as a single element defined by the lemma or dictionary form of the word. We formulate the following functions for lemmatizing the texts of a language corpus. We apply these functions to an example data set. Based on the above considerations, we generate a DTM (document-term matrix) from a set of scientific and technical texts in our corpus.

After the LDA model is formed, we will develop indicators for assessing the model. There are 2 indicators for evaluating the LDA model:

1) Perplexity

2) Coherence Score Based on the created LDA model, we form the most important terms for each topic. Let’s analyze the topic visualization. Pay attention to how topics are displayed on the left side and words on the right side. In the corpus,

more frequent topics are presented in a larger form. Closer topics are more similar, while more distant ones are less similar. When you select a topic, you can see the words within the selected topic. This measurement can be a combination of how common or how distinctive a word is. In the next step, we develop a function for obtaining coherence values for the case when the number of topics is < 5. For this example, 4 topics were considered because the target labels in this dataset are divided into 4. Otherwise, it is preferable to have the highest coherence score. The purpose of the LDA method is also to calculate what proportion of the scientific and technical texts in the corpus is created by which topic. Using the presented methods and algorithms, software has been developed, which covers the following areas when tagging scientific and technical texts in the Uzbek language corpus: *education, sports, healthcare, politics, culture, weather, economics, technology*...

CONCLUSION

1. The first stage of text study corresponds to the ancient period, while the second stage spans the 17th-19th centuries. During this period, analyses of numerous works were produced. By the 20th century, semantic, structural, and pragmatic approaches to studying texts emerged. Until the 21st century, texts were studied within the framework of analyzing other linguistic units. However, by the second half of the 19th century and mid-20th century, the study of texts became increasingly relevant. The 20th century became the "golden age" of text studies. The 21st century is recognized as a period of text renaissance. During this era, texts began to be examined from linguacultural, sociolinguistic, cognitive, psychological, and corpus-based perspectives.

2. The study of texts was conducted in several directions. Works in the first direction focused on examining the formal-grammatical and semantic structure of texts. The second direction analyzed formal and conceptual features of text construction that lead to different text perceptions. The third direction addressed the issue of text perception itself.

3. A language corpus is a very large and structured collection of texts, developed by actual users of the language, used to analyze the usage of words, phrases, and language in general, representing written or oral material. Corpora are also used to create various language databases for software development, such as predictive keyboards, spell checkers and correctors, text/speech comprehension systems, text-to-speech modules, machine translation systems, and others. To make the language corpus fully useful for users, it needs to be tagged. In NLP, language corpora are divided into different types based on various criteria such as content, purpose, or source: text corpora, multimodal corpora, parallel corpora, historical corpora, and annotated/tagged corpora.

4. Today, in the field of world computational linguistics, there are modern methods of tagging corpus texts, including: sentence boundary detection, tokenization, lemmatization, POS tagging, syntactic analysis, semantic analysis, Named Entity Recognition (NER), and coreference resolution. Methods of tagging corpus scientific and technical texts can be used in the development of NLP

applications, such as text comprehension, information retrieval, sentiment analysis, spell checking, document summarization, and machine translation.

5. Today, the NLP method known as thematic modeling is used for tagging scientific and technical texts in language corpora. Thematic modeling is a machine learning method that automatically analyzes text data to identify cluster words for collections of scientific and technical texts. This method, known as “unsupervised” machine learning, does not require a predetermined list of tags or training data previously classified by humans.

6. Since the models for classifying scientific and technical texts in corpora require training, they are known as “supervised” machine learning methods. Thematic modeling is a type of statistical modeling used to identify abstract “topics” encountered in collections of scientific and technical texts within a language corpus. Scientific and technical texts usually relate to several topics and are distributed in different proportions. Extracting words from scientific and technical texts in corpora requires more time, and this process is much more complex than extracting topics from the given texts. For example, a language corpus containing 100,000 texts with an average of 500 words per text was analyzed. Therefore, to process this corpus, 500*100000=50,000,000 operations need to be performed.

7. Thematic modeling in NLP is a set of algorithms that can be used for automatic summarization of large volumes of text. When analyzing language corpus texts, the large number of dimensions and features makes it difficult to train models and reduces their effectiveness. Thematic modeling, applied in unsupervised machine learning tasks, is considered as a form of tagging and is primarily used to extract necessary information from the language corpus, thereby helping to increase the efficiency of query execution.

8. Thematic modeling is a versatile algorithm used in various fields. It is widely applied in search engines to map user interests by topics. Today, thematic modeling methods are used to solve NLP tasks such as document classification, categorization, and summarization. Additionally, thematic modeling methods enable the analysis of user sentiments on social networks.

9. Examples of popular thematic modeling algorithms today include Latent Semantic Analysis (LSA), Latent Semantic Indexing (LSI), Hierarchical Dirichlet Process (HDP), Latent Dirichlet Allocation (LDA), and Non-negative Matrix Factorization (NMF). Among them, LDA has demonstrated more accurate and effective results in practice and is therefore widely used.

10. Software for tagging scientific and technical texts in the Uzbek language corpus was developed using LDA, LSA, and NMF methods and algorithms. The following areas were covered in tagging scientific and technical texts in the corpus: education, sports, healthcare, politics, culture, weather, economics, and technology. Based on the algorithms presented in the scientific research, the software http://topicmodel.uz/ was developed.

РАЗОВЫЙ НАУЧНЫЙ СОВЕТ DSc.03/25.08.2021.Fil.01.16 ПО ПРИСУЖДЕНИЮ УЧЁНЫХ СТЕПЕНЕЙ ПРИ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ УЗБЕКИСТАНА ИМЕНИ МИРЗО УЛУГБЕКА

ТАШКЕНТСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ УЗБЕКСКОГО ЯЗЫКА И ЛИТЕРАТУРЫ ИМЕНИ АЛИШЕРА НАВОИ

НОРБЕКОВА МАДИНА ШУХРАТ ҚИЗИ

СОЗДАНИЕ БАЗЫ НАУЧНО-ТЕХНИЧЕСКИХ ТЕКСТОВ ДЛЯ НАЦИОНАЛЬНОГО КОРПУСА УЗБЕКСКОГО ЯЗЫКА

10.00.11 - Теория языка. Прикладная и компьютерная лингвистика

АВТОРЕФЕРАТ ДИССЕРТАЦИИ ДОКТОРА ФИЛОСОФИИ (PhD) ПО ФИЛОЛОГИЧЕСКИМ НАУКАМ

Ташкент – 2025

Тема диссертации доктора философии по филологическим наукам (PhD) зарегистрирована в Высшей аттестационной комиссии при Министерстве высшего образования, науки и инноваций Республики Узбекистан за номером № B2024.4.PhD/Fil5438.

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

Автореферат диссертации на трех языках (узбекском, английском, русском (резюме)) размещен на веб-сайте Национального университета Узбекистана www.nuu.uz и информационно-образовательном портале “Зийонет” www.ziyonet.uz.

Научный руководитель:	**Раупова Лайло Раҳимовна** доктор филологических наук, профессор
Официальные оппоненты:	**Тоирова Гули Ибрагимовна** доктор филологических наук, профессор
	Абдурахманова Нилуфар Зайнобиддин қизи доктор филологических наук, профессор
Ведущая организация:	Андижанский государственный институт иностранных языков

Защита диссертации состоится на заседании Научного совета DSc.03/25.08.2021.Fil.01.16 при Национальном университете Узбекистана “____” _____2025 г. в ____ часов. (Адрес: 100174, г. Ташкент, Алмазарский район, ул. Университетская, 4. Тел.: (99871) 246-02-24; факс: (99871) 246-02-24; e-mail: devonxona@nuu.uz.

С диссертацией можно ознакомиться в Информационно-ресурсном центре Национального университета Узбекистана (зарегистрирована под No ____). (Адрес: 100174, г. Ташкент, Алмазарский район, ул. Университетская, 4. Тел.: (99871) 246-02-24; факс: (99871) 246-02-24; e-mail: devonxona@nuu.uz.

Автореферат диссертации разослан «___» _____ 2025 года.
(Протокол реестра рассылки за № ___ от «___» _____ 2025 года.

Н.А.Раҳмонов
Председатель Научного совета по присуждению ученых степеней, доктор филологических наук, профессор

М.В.Хужамкулова
Ученый секретарь Научного совета по присуждению ученых степеней, кандидат филологических наук

А.Э.Маматов
Председатель Научного семинара при Научном совете по присуждению ученых степеней, доктор филологических наук, профессор

Введение (аннотация диссертации доктора философии (PhD))

Цель исследования разработка теоретических и программных основ формирования базы научно-технических текстов для национального корпуса узбекского языка, лингвистическое тегирование лексических единиц в научно-технических текстах и классификация базы научно-технических текстов на узбекском языке.

Объектом исследования выбраны различные научные и технические тексты на узбекском языке для создания базы научно-технических текстов национального корпуса узбекского языка.

Предметом исследования являются проблемы формирования базы научно-технических текстов на узбекском языке и их разметки.

Методы исследования. При освещении темы исследования использовались методы классификации, описания, сопоставления и статистического анализа.

Научная новизна исследования заключается в следующем:

обоснованы возможности корпуса для централизации и обработки данных в области лингводидактики и лингвистики, а также важность его эффективного использования;

доказана разработка программного обеспечения http://topicmodel.uz/ с использованием методов и алгоритмов LDA, LSA, NMF для разметки научно-технических текстов в национальном корпусе узбекского языка в областях образования, спорта, здравоохранения, политики, культуры, погоды, экономики, технологий;

аргументировано, что методы разметки научно-технических текстов узбекского языка могут быть использованы при разработке NLP-приложений, таких как понимание текста, извлечение информации, анализ эмоций, проверка орфографии, обобщение документов и машинный перевод;

определены требования к научно-техническому тексту при формировании текстовой базы, разработаны теги, используемые при размещении научно-технических текстов в корпусе, и на основе анализа выявлены особенности научно-технических текстов, позволяющие осуществлять семантическое аннотирование в национальном корпусе узбекского языка.

Внедрение результатов исследования. На основе полученных научных результатов по созданию базы научно-технических текстов для Национального корпуса узбекского языка:

Выводы о необходимости учитывать информационные, целостные, последовательные и терминологические особенности научно-технических текстов при использовании их переводных вариантов были применены в практическом проекте ПФ-201912258 “Создание многоязычной (на узбекском, русском, английском языках) электронной платформы узбекской литературы” (2021-2023) (Справка No 04/1-4044 от 6 декабря 2024 года Ташкентского государственного университета узбекского языка и литературы имени

Алишера Навои). В результате база научно-технических текстов электронной платформы была обогащена новыми данными.

База данных программного обеспечения, система тегирования и выводы принципов аннотирования, сформированные в процессе подготовки диссертации, были использованы в практическом проекте АМ-Ф3-201908172 на тему “Создание образовательного корпуса узбекского языка” (2020-2023) (справка Ташкентского государственного университета узбекского языка и литературы имени Алишера Навои No 04/1-4085 от 10 декабря 2024 г.). В результате научно-технические материалы, представленные в исследовании, алгоритмы тематического моделирования и предложения по скрытому семантическому анализу послужили обогащению содержания корпуса и его практического аспекта.

Практическая значимость результатов исследования заключается в том, что выводы по созданию базы научно-технических текстов для национального корпуса узбекского языка, определению его семантической и грамматической структуры, а также разработке методов автоматической обработки и анализа текста были использованы при написании сценария программы “Для всех” телеканала “История Узбекистана” Национальной телерадиокомпании Узбекистана на основании справки No 06-28-845 от 24 декабря 2024 года.

Структура и объем диссертации. Диссертация состоит из введения, трех глав, заключения и списка использованной литературы. Общий объем работы составляет 158 страниц.

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

1. Norbekova M. The Text As A Speech-Creative Process // Asian Journal of Multidimensional Research (ISSN: 2278-4853), 2022. December. Vol. 11. Issue 12. SJIF 2022. – P. 91-95 (Impact Factor – 8.179).
2. Norbekova M. Milliy korpusga qo‘yiladigan zaruriy lingvistik va lingvodidaktik talablar // UzA. – Toshkent, 2022. – № 9 (35). – B. 63-67.
3. Norbekova M. Korpus elektron shaklda saqlanadigan til birliklari majmui sifatida // O‘zMU xabarlari (Jurnal ISSN ISSN 2181-7324). – Toshkent, 2022. – B. 255-258 (10.00.00. № 15).
4. Norbekova M. Ilmiy matnlar tipologiyasida birlamchi matn va unga munosabat bildiruvchi ikkinchi turdagi matnlar haqida / “O‘zbek folklori va shevalari tadqiqotlari: Amaliyot, Metodologiya, Yangicha Yondashuv” mavzusidagi Xalqaro ilmiy-nazariy konferensiya materiallari. – Toshkent, 2024. – B. 350-352.
5. Norbekova M. O‘zbek tilining ta’limiy korpusiga aloqador hodisalar / “O‘zbek tilining milliy korpusi: muammo va vazifalar” mavzusidagi Xalqaro ilmiy-amaliy konferensiya materiallari. – Toshkent, 2022. – B. 355-362.
6. Norbekova M. Ilmiy matn xususiyatlari ilmiy maqola misolida / “O‘zbek filologiyasi: muammo va yechimlar” mavzusidagi Respublika ilmiy-amaliy anjumani to‘plami. – Toshkent, 2024. – B. 364-366.
7. Norbekova M. Ilmiy matnnning asosiy sifat xususiyatlari / “Ilm - fan va innovatsion yutuqlarni rivojlantirishning dolzarb muammolari” mavzusidagi Respublika ilmiy-amaliy konferensiyasi. – T., 2024. – № 11. – B. 127-132.

II bo‘lim (II part; II часть)

8. Norbekova M. Specific Features Of The Scientific Text And Its Necessity In The National Corpus Of The Uzbek Language // Current Research Journal Of Philological Sciences, 2024. – № 3 (12). – P. 60-65.
9. Norbekova M. Ilmiy matnnning asosiy kategoriyalari va mazmun birliklari / “Yangi renessans istiqbolida filologik pedagogik tadqiqotlarning nazariy va amaliy ahamiyati” mavzusidagi Xalqaro ilmiy-amaliy konferensiya materiallari. – Toshkent, 2024. – B. 110-111.
10. Norbekova M. Matnning tavsifi va tasnifi xususida qarashlar // O‘zMU xabarlari. – Toshkent, 2024. – B. 327-330 (10.00.00. № 15).