ANDIJON DAVLAT UNIVERSITETI HUZURIDAGI
ILMIY DARAJA BERUVCHI PhD.03/30.12.2019.Fil.60.02
RAQAMLI ILMIY KENGASH
ALISHER NAVOIY NOMIDAGI TOSHKENT DAVLAT O‘ZBEK TILI VA ADABIYOTI UNIVERSITETI

ABDULLAYEVA OQILA XOLMO‘MINOVNA

O‘ZBEK TILINING INTERNET AXBOROT MATNLARI KORPUSINI SHAKLLANTIRISHNING NAZARIY VA AMALIY ASOSLARI

10.00.11 – Til nazariyasi. Amaliy va kompyuter lingvistikasi

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

Andijon – 2022

UO‘K: 811.512.133’322.5
811.512.133’373.004

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

Contents of dissertation of sciences of the doctor of philosophy (PhD) on philological sciences

Оглавление автореферата диссертации доктора философии (PhD)по филологическим наукам

ANDIJON DAVLAT UNIVERSITETI HUZURIDAGI
ILMIY DARAJA BERUVCHI PhD.03/30.12.2019.Fil.60.02
RAQAMLI ILMIY KENGASH
ALISHER NAVOIY NOMIDAGI TOSHKENT DAVLAT O‘ZBEK TILI VA ADABIYOTI UNIVERSITETI

ABDULLAYEVA OQILA XOLMO‘MINOVNA

O‘ZBEK TILINING INTERNET AXBOROT MATNLARI KORPUSINI SHAKLLANTIRISHNING NAZARIY VA AMALIY ASOSLARI

10.00.11 – Til nazariyasi. Amaliy va kompyuter lingvistikasi

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

Andijon – 2022

Falsafa doktori (PhD) dissertatsiyasi mavzusi O‘zbekiston Respublikasi Vazirlar Mahkamasi huzuridagi Oliy attestatsiya komissiyasida B2019.3.PhD/Fil1018 raqami bilan ro‘yxatga olingan.

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

Dissertatsiya avtoreferati uch tilda (o‘zbek, ingliz, rus (rezyume) Andijon davlat universiteti veb-sahifasi (www.adu.uz) hamda “Ziyonet” Axborot ta’lim portali (www.ziyonet.uz)da joylashtirilgan.

Ilmiy rahbar:	**Azimova Iroda Alisherovna** filologiya fanlari nomzodi, dotsent
Rasmiy opponentlar:	**Nabiyeva Diloro Abduxamidovna** filologiya fanlari doktori, professor
	Eshmuminov Asqar Allamurodovich filologiya fanlari bo‘yicha falsafa doktori
Yetakchi tashkilot:	**Samarqand davlat universiteti**

Dissertatsiya himoyasi Andijon davlat universiteti huzuridagi ilmiy darajalar beruvchi PhD. 03/30.12.2019.Fil.60.02 raqamli Ilmiy Kengashning 2022-yil “___”_____ soat ____dagi majlisida bo‘lib o‘tadi. (Manzil:170100, O‘zbekiston Respublikasi, Andijon shahar, Universitet ko‘chasi 129-uy. Telefon/faks: 0 (374) 223 88 30, e-mail: agsu_info@edu.uz).

Dissertatsiya bilan Andijon davlat universitetining Axborot-resurs markazida tanishish mumkin. (____ raqam bilan ro‘yxatga olingan). (O‘zbekiston Respublikasi, Andijon shahar, Universitet ko‘chasi 129-uy. Tel.: 0 (374) 223 88 14).

Dissertatsiya avtoreferati 2022-yil “___”_____ kuni tarqatildi.
(2022-yil “____”_____ dagi _____ raqamli reestr bayonnomasi).

Sh.H.Shaxobitdinova
Ilmiy darajalar beruvchi
Ilmiy kengash raisi, f.f.d., professor

F.F.Usmanov
Ilmiy darajalar beruvchi Ilmiy kengash
kotibi, f.f.b.f.d. (PhD)

M.E.Umarxodjayev
Ilmiy darajalar beruvchi Ilmiy kengash
huzuridagi Ilmiy seminar raisi, f.f.d., professor

KIRISH (falsafa doktori (PhD) dissertatsiyasi annotatsiyasi)

Dissertatsiya mavzusining dolzarbligi va zarurati. Jahon tilshunosligida so‘nggi yarim asr davomida til tadqiqida amaliy tajribalar orqali xulosalar chiqarish, tilga mashina yordamida ishlov berish, turli dasturlar orqali lingvistik tadqiqotlar olib borishda yangicha yondashuvlar, turli yo‘nalishlar paydo bo‘ldi. Xuddi shunday yangi yo‘nalishlardan biri korpus lingvistikasi bo‘lib, bu sohada keng miqyosda ilmiy-nazariy tadqiqotlar olib borilmoqda. Xususan, tillarning milliy korpuslarini yaratishga, mavjud korpuslarni rivojlantirishga jiddiy e’tibor qaratilmoqda.

Dunyo tilshunosligida tilning mavjud imkoniyatlarini kengroq o‘rganish, til grammatikasining muammoli tomonlarini kontekstda aniqlash, tilda grammatik qoliplarni belgilash, ko‘p tarmoqli elektron lug‘atlar yaratish ishini yengillashtirish, tilni o‘rganishda zamonaviy axborot texnologiyalaridan foydalanish samaradorligini oshirish, tilda avtomatik tarjima, qidiruv va kompyuter tahlilini yo‘lga qo‘yish, elektron darsliklar va lug‘atlar tayyorlash kabi masalalarni hal qilish uchun tillarda korpuslar yaratishning nazariy va amaliy asoslarini ishlab chiqish, tilning maxsus sohalar bo‘yicha korpusini qurish zaruratining mavjudligi tadqiqotimizning dolzarbligini belgilaydi.

Mamlakatimizda oxirgi yillarda barcha sohalarda ish samaradorligini oshirish, insonlar mehnatini yengillashtirish maqsadida kompyuter texnologiyalaridan foydalanish bo‘yicha zarur islohotlar olib borilmoqda. Xuddi shuningdek, o‘zbek tilining maqomini oshirish, uning faol qo‘llanilishini ta’minlash yuzasidan qabul qilinayotgan qonun va qarorlar mutaxassislar oldiga bir qator muhim vazifalarni qo‘ymoqda. O‘zbekiston Respublikasi Prezidenti Sh.Mirziyoyev o‘zbek tiliga Davlat tili maqomi berilganligining 31 yillik bayrami tabrik nutqida “…o‘zbek tilining jahon maydonida, xususan, Internet axborot tarmog‘ida munosib o‘rin egallashini ta’minlash, ona tilimizda ko‘plab yangi kompyuter dasturlarini yaratish bo‘yicha ham oldimizda turgan muhim va dolzarb vazifalarni hal qilishimiz zarur” deya ta’kidladi. O‘zbek tiliga e’tibor davlat siyosatining ustuvor yo‘nalishlari darajasiga ko‘tarildi. “Davlat tilini rivojlantirishga oid ilmiy-tadqiqot ishlarini qo‘llab-quvvatlash, bu sohada xalqaro hamkorlikni amalga oshirish[1]” amalda olib borilayotgan har bir tadqiqotning ijtimoiy ahamiyati, amaliy natijadorligi muhimligini isbotlaydi.

O‘zbekiston Respublikasi Prezidentining 2016-yil 13-maydagi PF-4797-son “Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti faoliyatini tashkil etish to‘g‘risida”gi, O‘zbekiston Respublikasi Prezidentining 2017-yil 7-fevraldagi PF-4947-son “O‘zbekiston Respublikasini yanada rivojlantirish bo‘yicha Harakatlar strategiyasi to‘g‘risida”gi, 2019-yil 21-oktabrdagi PF-5850-son “O‘zbek tilining davlat tili sifatidagi nufuzi va mavqeyini tubdan oshirish chora-tadbirlari to‘g‘risida”gi, 2020-yil 20-oktabrdagi PF-6084-son “Mamlakatimizda o‘zbek tilini yanada rivojlantirish va til siyosatini

[1]O‘zbekiston Respublikasi Prezidenti Shavkat Mirziyoyevning “O‘zbek tilining davlat tili sifatidagi nufuzi va mavqeini tubdan oshirish chora-tadbirlari to‘g‘risida”gi farmoni // www.xabar.uz

takomillashtirish chora-tadbirlari to‘g‘risida”gi, 2020-yil 29-oktabrdagi PF-6097-son “Ilm-fanni 2030-yilgacha rivojlantirish konsepsiyasini tasdiqlash to‘g‘risida”gi Farmonlari; 2019-yil 4-oktabrdagi PQ-4479-son “O‘zbekiston Respublikasining “Davlat tili haqida”gi qonuni qabul qilinganligining o‘ttiz yilligini keng nishonlash to‘g‘risida”gi qarori, O‘zbekiston Respublikasi Vazirlar Mahkamasining 2019-yil 12-dekabrdagi 984-son “Davlat tilini rivojlantirish departamenti to‘g‘risidagi Nizomni tasdiqlash haqida”gi qarori hamda mazkur faoliyatga tegishli boshqa me’yoriy-huquqiy hujjatlarda belgilangan vazifalarni amalga oshirishga ushbu tadqiqot muayyan darajada xizmat qiladi.

Tadqiqotning respublika fan va texnologiyalar rivojlanishining ustuvor yo‘nalishlariga mosligi. Mazkur tadqiqot respublika fan va texnologiyalar rivojlanishining I. “Axborotlashgan jamiyat va demokratik davlatni ijtimoiy, huquqiy, iqtisodiy, madaniy, ma’naviy-ma’rifiy rivojlantirish, innovatsion iqtisodiyotni rivojlantirish” ustuvor yo‘nalishi doirasida bajarilgan.

Muammoning o‘rganilganlik darajasi. Dunyo tilshunosligida korpus lingvistikasi va korpuslarni yaratish tadqiqotlari 60-yillardayoq boshlangan. Ilk 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.Sinkler, L.Floverdev, T.Makkenri, A.Hardi, M.Makkarti, V.N.Frensis, G.Kennidi singari olimlarning tadqiqotlarida maxsus o‘rganilgan[2].

Korpuslarning ijtimoiy ahamiyati, korpus lingvistikasining keyingi taraqqiyot bosqichlari, internetdan korpus sifatida foydalanish, internet va korpus o‘rtasidagi o‘xshash va farqli jihatlari A.Kilgarrif, K.Styuart, G.Grefenstette, M.Hundt, N.Nesselhaf kabi tadqiqotchilar tomonidan mufassal yoritilib, masalaning nazariy asoslari tadqiq etilgan[3].

[2]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.ilc.cnr.it/EAGLES/browse.html.; 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 Word-tagging 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 University 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.; Flowerdew L. Corpus Linguistic Techniques Applied to Textlinguistics. 1998. – p. 541-552.; McCarthy M., O’Keefe A. What are corpora and how have they evolved?: The Routledge handbook of corpus linguistics. – London and New York, 2010.

[3] Kilgarriff A., Grefenstette G. 2003. Introduction to the special issue on the Web as corpus // Computational Linguistics 29(3). – p. 333-347.; Kilgarriff A. Web as corpus. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.108.2212&rep=rep1&type=pdf; Stuart K. New perspectives on corpus linguistics.; Grefenstette G. The WWW as a Resource for Example-Based MT Tasks. // Translating and the Computer. – London, 1999.; Grefenstette G. Nioche J. Estimation of English and non-English Language Use on the www. // RIAO (Recherche d’Informations Assistee par Ordinateur). – Paris, 2000.; Hundt M., Nesselhauf N., Biewer C. Corpus linguistics and the web. – Amsterdam-New York, 2007.; Hundt M., Nesselhauf N., Biewer C. Corpus linguistics and the web. – Amsterdam-New York, 2007.

T.Makkenri, R.Ksiao, R.Reppen, D.Biberlarning korpus lingvistikasi sohasidagi tadqiqotlari korpuslarning ta'lim jarayonida tutgan o'rni, til o'qitishda korpuslardan foydalanish prinsiplari, afzalliklari to'g'risida aniq va batafsil tasavvur hosil qilish imkoniyatini yaratadi[4].

Rus tilshunosligida V.Plungyan, V.P.Zaxarov, A.B.Kutuzovlar tomonidan amalga oshirilgan qator tadqiqotlar rus tilshunosligida korpus lingvistikasining alohida soha sifatida shakllanishida, rus milliy korpusining yaratilishida alohida ahamiyat kasb etadi[5].

O'zbek tilshunosligida korpus tilshunosligiga qadar kompyuter lingvistikasi sohasida bir talay muhim tadqiqot ishlari olib borilgan. Xususan, A.Po'latov[6], S.Muhamedovlarning bir guruh olimlar hamkorligida amalga oshirgan tadqiqot ishlari[7], N.Abdurahmonova[8], M.Abjalovalarning[9] ilmiy 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.Hamroyevaning "O'zbek tili mualliflik korpusini tuzishning lingvistik asoslari[10]" mavzusidagi dissertatsiyasi, hamda shu nomdagi monografiyasida[11] 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. Oxirgi yillarda o'zbek tilshunosligida ko'plab tadqiqot ishlari – dissertatsiya, maqola va tezislar e'lon qilinmoqda. Jumladan, A.Eshmo'minovning "O'zbek tili milliy korpusining sinonim so'zlar bazasi" mavzusidagi

[4] McEnery T., Xiao R., Tono Y. Corpus-based Language Studies: An Advanced Resource Book. Routledge, 2006.; Reppen R. Using corpora in the language classroom. - Cambridge: Cambridge University Press, 2010.; Biber D., Conrad S., Reppen R. Corpus Linguistics. Investigating Language Structures and Use. - Cambridge: Cambridge University Press, 1998.

[5]Плунгян В. Перспективы: Корпусная лингвистика и корпус русского языка. https://www.youtube.com/watch?v=OBTLuLx962U; Захаров В.П. Корпусная лингвистика: учеб.-метод. Пособие / В.П. Захаров. – СПб., 2005.; Кутузов А.Б. Корпусная лингвистика, 2015.

[6] Пўлатов А., Мўминова Т., Пўлатова И. Дунёвий ўзбек тили. Ўзбек тилида феъл шакллари ва уларнинг рус, инглиз тилидаги кўринишлари. – Тошкент: Университет, 2003.; Пўлатов А. Компьютер лингвистикаси. – Тошкент, 2011.

[7] Мухамедов С.А., Пиотровский Г.Г. Инженерная лингвистика и опыт системно – статистического исследования узбекских текстов. –Т.: Фан, 1986; Махмудов М.А., Пиотровская А.А., Садыков Т. Система машинного анализа и синтеза тюркской словоформы // Переработка текста методами инженерной лингвистики. – Минск, 1982.

[8] Абдурахмонова Н.З. Инглизча матнларни ўзбек тилига таржима қилиш дастурининг лингвистик таъминоти (содда гаплар мисолида): филол. фан. бўйича фалсафа доктори дисс. автореф. – Тошкент, 2018. – 49 б.; Abdurakhmonova N. Uzbek ontology of Uzbek language as example of adjective // Шестая Международная конференция по компьютерной обработке тюркских языков 363 "TurkLang-2018". (Труды конференции) – Ташкент: Издательско-полиграфический дом "Navoiy universiteti", 2018. – 320 с.

[9] Абжалова М. Ўзбек тилидаги матнларни таҳрир ва таҳлил қилувчи дастурнинг лингвистик модуллари (расмий ва илмий услубдаги матнлар таҳрири дастури учун): Филол. фан. бўйича фалсафа доктори (PhD)…дисс. –Фарғона, 2019.; Абжалова М. Матнларга авто-лингвистик ишлов бериш тизимлари // Шестая Международная конференция по компьютерной обработке тюркских языков "TurkLang-2018". (Труды конференции) – Ташкент: Издательско-полиграфический дом "Navoiy universiteti", 2018. – 320 с.

[10] Ҳамроева Ш. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари: Филол. фан. бўйича фалсафа доктори (PhD)…дисс. – Бухоро, 2018 .–253 б.

[11] Ҳамроева Ш. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари: Монография. – Тошкент, 2020. – 229 б.

dissertatsiyasi[12], D.Axmedovaning “Atov birliklarini o‘zbek tili korpuslari uchun leksik-semantik teglashning lingvistik asos va modellari” mavzusidagi dissertatsiyasi[13], B.Mengliyev[14], G.Toirova[15], H.Ataboyevlarning[16] ilmiy izlanishlari natijalarini misol sifatida keltirish mumkin.

Tadqiqotning dissertatsiya bajarilgan oliy ta’lim muassasasi ilmiy-tadqiqot ishlari rejalari bilan bog‘liqligi. Dissertatsiya Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti Amaliy tilshunoslik va lingvodidaktika kafedrasining ilmiy tadqiqot ishlari rejasining “Kompyuter va korpus tilshunosligi – o‘zbek tilini sun’iy intellekt uchun modellashtirish va o‘zbek tilining milliy va xususiy korpuslarini yaratishning lingvistik bazasini tayyorlash” mavzusi doirasida bajarilgan.

Tadqiqotning maqsadi. O‘zbek tilshunosligida internet axborot matnlari korpusini yaratishning nazariy va amaliy asoslarini ishlab chiqish, korpusda matnlarni lingvistik annotatsiyalashning o‘ziga xos tamoyillarini aniqlash va o‘zbek tili elektron axborot matnlari korpusini yaratishdan iborat.

Tadqiqotning vazifalari:

internet matnlari asosida korpus yaratishning nazariy asoslarini aniqlash;

o‘zbek tili elektron axborot matnlari korpusini yaratishning o‘ziga xos muhim jihatlarini aniqlash;

o‘zbek tili korpusida matnlarni lingvistik annotatsiyalash algoritmini ishlab chiqish, ularning ishlash prinsipini tajriba natijalari orqali ko‘rish, korpusda natijalarni lingvistik tahlil qilish.

Tadqiqotning obyekti sifatida o‘zbek tilidagi internet elektron axborot matnlari tanlandi, bunda kun.uz, daryo.uz, xabar.uz, sof.uz, gazeta.uz, qalampir.uz va UzA saytlaridagi 2020-2021-yillar davr oralig‘idagi barcha matnlar tadqiqot obyekti qilib olindi.

Tadqiqotning predmetini o‘zbek tilidagi internet elektron axborot matnlarining korpus yaratishda muhim bo‘lgan leksik-semantik va morfologik xususiyatlari tashkil etadi.

Tadqiqotning usullari. Tadqiqotni amalga oshirishda tasniflash, tavsiflash, qiyoslash, statistik usul va uning tahlil tamoyillaridan foydalanildi.

Tadqiqotning ilmiy yangiligi quyidagilardan iborat:

o‘zbek tili internet axborot matnlari korpusi platformasining lingvistik ta’minotini ishlab chiqish imkonini beruvchi axborot matnlaridagi lisoniy

[12] Эшмўминов А. Ўзбек тили миллий корпусининг синоним сўзлар базаси: Филол. фан. бўйича фалсафа доктори (PhD) дисс. – Қарши, 2019. – 137 б.

[13] Ахмедова Д. Атов бирликларини ўзбек тили корпуслари учун лексик-семантик теглашнинг лингвистик асос ва моделлари: Филол. фан. бўйича фалсафа доктори (PhD) дисс. – Бухоро, 2020. –247 б.

[14] Mengliyev B. O‘zbek tili taraqqiyoti va rivojlanish omillari // “O‘zbek tilini dunyo miqyosida keng targ‘ib qilish bo‘yicha hamkorlik istiqbollari” mavzusidagi xalqaro ilmiy-amaliy anjuman materiallari. – Toshkent, 2020.; Менглиев Б., Бобожонов С., Ҳамроева Ш. Ўзбек тилининг миллий корпуси. http://marifat.uz/marifat/ruknlar/fan/1241.htm

[15] Тоирова Г. Миллий корпус яратишнинг технологик жараёни хусусида // Ўзбекистонда хорижий тиллар. Электрон илмий-методик журнал. – Тошкент. 2020, № 2 (31), –Б.57– 64. https://journal.fledu.uz/uz/ 2-31-2020

[16] Ataboyev N. Problematic issues of corpus analysis and its shortcomings // ISJ Theoretical & Applied Science, 10 (78), 2019.; Ataboyev N. Compiling dictionaries by using corpus analysis and its advantages // International Journal of Progressive Sciences and Technologies (IJPSAT) http://ijpsat.es/index.php/ijpsat/article/view/508

birliklarning leksik-semantik, struktur jihatlarini aniqlash orqali nutq birliklarini izohlash tamoyillari dalillangan;

o‘zbek tili internet axborot matnlari korpusida nutq birliklari morfologik annotatsiyalangan, bunda matnlardagi fe’llarning tasniflovchi kategoriyasi, bo‘lishli-bo‘lishsizlik, harakat tarzi kategoriyasi, fe’lning o‘zgalovchi kategoriyasi, otlarning son kategoriyasi, sifatning daraja tasniflovchi kategoriyasi, sonning lug‘aviy-grammatik guruhi, ravish, taqlid, olmoshlarning morfologik ko‘rsatkichlari va yordamchi so‘zlarning lisoniy imkoniyatlari ishlab chiqilgan;

o‘zbek tili internet axborot matnlari korpusidagi lisoniy birliklar semantik annotatsiyalash imkonini beruvchi matnlardagi nutq birliklarining lug‘aviy-ma’noviy, lug‘aviy-mavzuviy guruhlari, lug‘aviy-mazmuniy maydoni aniqlangan;

annotatsiyalashda kontekstda uchragan nutq birliklarining turkumlarga ajratilishidagi omonim so‘zlar, affiksal omonimiya va omonimiyaga yondosh hodisalar muammosi, qo‘shma fe’llar va ko‘makchi fe’lli so‘z qo‘shilmalarini aniqlash masalasi korpusda tajriba orqali dalillangan.

Tadqiqotning amaliy natijasi quyidagilardan iborat:

o‘zbek tili elektron axborot matnlari korpusi qurilishi uchun texnik topshiriqlar ishlab chiqilib, ish jarayoni bosqichma-bosqich belgilandi;

o‘zbek tili internet axborot matnlari korpusi qurildi, sayt sifatida e’lon qilindi (https://uzkorpus.uz/ manzili ostida);

korpusda nutq birliklarini teglash uchun morfologik va semantik teglar amalda ishlab chiqildi, maxsus standart belgilar tanlandi;

korpusda nutq birliklari leksik-semantik va morfologik xususiyatlariga ko‘ra qo‘l mehnati va yarim avtomatik annotatsiyalandi;

korpus oldiga qo‘yilgan muammolarni tadqiq etish natijasida chiqarilgan xulosa, bilimlar o‘zbek tilshunosligi, xususan, kompyuter lingvistikasi, korpus lingvistikasi uchun muhim, zaruriy ilmiy bilimlar berishi, tilshunoslikda yangi lingvistik tahlil usullarining paydo bo‘lishiga asos bo‘lishi, leksikografiyaning rivojlanishiga hissa qo‘shishi ilmiy jihatdan asoslandi.

Tadqiqot natijalarining ishonchliligi dissertatsiyada tanlangan muammoning aniq qo‘yilganligi, korpus qurilishi boshqa tillar korpuslari bilan uzoq muddatli kuzatuv-tajriba orqali o‘rganilganligi, to‘plangan materiallar tahlili orqali o‘zbek tili korpusi qurilganligi, korpusda lingvistik tahlil amallari o‘tkazilib, ilmiy-nazariy xulosalar chiqarilganligi bilan belgilanadi.

Tadqiqot natijalarining ilmiy va amaliy ahamiyati. Dissertatsiya natijalarining ilmiy ahamiyati kompyuter lingvistikasi fanida yangi tadqiqotlar yaratishda nazariy ahamiyat kasb etishi, o‘zbek tilini formallashtirish, kompyuter tilini yaratish bo‘yicha ilmiy qarashlar orqali boyishi bilan belgilanadi.

Tadqiqotning amaliy ahamiyati undagi materiallar, natija va xulosalar o‘zbek tilshunosligida korpus lingvistikasi fanida, bakalavriat va magistratura ta’lim yo‘nalishlarida maxsus kurs, maxsus seminarlar o‘tishda manba va material vazifasini bajarishi, shuningdek, til yo‘nalishida, uslub bo‘yicha tadqiqotlar olib borishda asos bo‘lib xizmat qiladi.

Tadqiqot natijalarining joriy qilinishi. O‘zbek tilining internet axborot matnlari korpusini shakllantirishning nazariy va amaliy asoslari yuzasidan olingan ilmiy natijalar asosida:

internet axborot matnlari korpusi platformasining lingvistik ta’minotini ishlab chiqish, til korpusi uchun o‘zbek tili birliklarini modellashtirish, lingvistik modellashtirish va belgilashda standart vositalarini ishlab chiqish, korpus yaratishda lingvistik modellardan foydalanishga oid xulosalardan 2017-2020-yillarda bajarilgan “Development of the interdisciplinary master program on Computational Linguistics at Central Asian Universities” nomli ERASMUS CLASS loyihasida foydalanilgan (ERASMUS+ ma’lumotnomasi, National Erasmus+ office – Uzbekistan; Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universitetining 2021-yil 6-sentabrdagi 01/4-1464-sonli ma’lumotnomasi). Loyiha doirasida bajarilgan sillabuslar hamda yozilgan darslik til birliklarini modellashtirish bilan to‘ldirilgan;

o‘zbek tili internet axborot matnlari korpusini qurish texnologiyasi, korpusda nutq birliklarini morfologik va semantik annotatsiyalashga oid xulosalardan “Mahalla va oila” ilmiy-tadqiqot institutining 2020-2021-yillarda bajarilgan JHBL-20-sonli “Oila, mahalla va gender tengligi mavzusidagi badiiy asarlarning elektron korpusini yaratish” loyihasida foydalanilgan (“Mahalla va oila” ilmiy-tadqiqot institutining 2021-yil 8-sentabrdagi 01/09-517-sonli ma’lumotnomasi). Loyiha doirasida bajarilgan sillabuslar hamda yaratilgan loyihada korpus qurilishi bo‘yicha ma’lumotlardan foydalanilgan;

o‘zbek tili internet axborot matnlari korpusida nutq birliklarini morfologik va semantik annotatsiyalash tamoyillarini ishlab chiqish, “Tug‘ro” nomli internet axborot matnlari korpus platformasida lingvistik tahlillar o‘tkazish hamda olingan amaliy natijalardan 2016-2020-yillarda bajarilgan 1-147-sonli “Theoretical issues of Azerbaijani-Uzbek-Turkmen linguistics” nomli fundamental loyihada foydalanilgan (Azərbaycan Respublikası Azərbaycan Milli Elmlər Akademiyası İmadəddin Nəsimi adina Dilçilik institutining 2021-yil 29-oktabrdagi 230-sonli ma’lumotnomasi). Natijada, loyiha yangi ilmiy-nazariy ma’lumotlar bilan boyitilgan;

o‘zbek tili internet axborot matnlaridagi nutq birliklarini annotatsiyalash, lingvistik annotatsiyalashning tamoyillarini belgilash, til birliklarini morfologik va semantik izohlab, korpus platformasida lingvistik tahlillar o‘tkazish hamda amaliy natijani olishga oid umumiy xulosalardan Hoji Bayram Vali universitetida o‘qitiladigan “Araştırma teknikleri ve yayın etiği (LE0101100)” fanining ma’ruza va amaliy mashg‘ulotlarida foydalanilgan (Ankara, Hacı Bayram Veli üniversitesi Edebiyat fakültesi Çağdaş Türk Lehçeleri ve Edebiyatları Bölümüning 2021-yil 25-oktabrdagi ma’lumotnomasi). Natijada, ma’ruza va amaliy mashg‘ulotlar mazmuni korpus platformasida lingvistik tahlillar o‘tkazish haqidagi ma’lumotlar bilan boyigan;

o‘zbek tilshunosligida milliy korpuslarning yaratilishi, o‘zbek tilini ona tili va chet tili sifatida o‘qitishda korpuslardan foydalanishga oid ma’lumotlardan O‘zbekiston Milliy teleradiokompaniyasida “O‘zbekiston” telekanali orqali efirga uzatilgan ilmiy-ma’rifiy ko‘rsatuvlarda, xususan, “Oydin hayot live”, “Ta’lim va

taraqqiyot” ko‘rsatuvlarini tayyorlashda samarali foydalanilgan (O‘zbekiston teleradiokompaniyasining 2021-yil 5-iyuldagi 02-13-1129-sonli ma’lumotnomasi). Natijada, ushbu teleko‘rsatuvlarning mazmuni ilmiy dalillar bilan boyitilgan.

Tadqiqot natijalarining aprobatsiyasi. Tadqiqot natijalari 6ta ilmiy-amaliy anjuman, jumladan, 2ta respublika hamda 4ta xalqaro ilmiy-amaliy anjumanda muhokamadan o‘tkazilgan.

Tadqiqot natijalarining e’lon qilinganligi. Dissertatsiya mavzusi bo‘yicha 11 ta ilmiy ish: O‘zbekiston Respublikasi Oliy attestatsiya komissiyasi tomonidan doktorlik dissertatsiyalari asosiy ilmiy natijalarini chop etish tavsiya etilgan ilmiy nashrlarda 5ta maqola, jumladan, 3tasi respublika, 2ta xorijiy jurnallarda nashr etilgan.

Disseratatsiyaning tuzilishi va hajmi. Dissertatsiya kirish, 3ta bob, xulosa, foydalanilgan adabiyotlar ro‘yxati va ilovadan iborat bo‘lib, umumiy hajmi 158 sahifani tashkil qiladi.

DISSERTATSIYANING ASOSIY MAZMUNI

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

Dissertatsiyaning birinchi bobi “**Internet matnlari asosida korpus yaratishning nazariy asoslari**” deb nomlangan. Bobning “Korpus lingvistikasi tilshunoslik yo‘nalishi sifatida” deb nomlangan birinchi faslida korpusning tilda va tilshunoslik sohasida tutgan o‘rni, korpus lingvistikasining nazariy asoslari, korpuslar qanday mezonlar asosida shakllantirilishi tavsiflangan. Korpus – tilni yoki tilning o‘zgarishini to‘liq aks ettirish uchun tashqi mezonlar asosida tanlangan elektron shakldagi matnlar to‘plamidir. Lingvistik tadqiqotlar uchun ma’lumotlar manbayi sifatida ishlaydi[17].

Korpus har qanday sistematik matn to‘plamiga murojaat qilishi mumkin bo‘lsa-da, u bugungi kunda tor ma’noda ham ishlatiladi, odatda, kompyuterlashtirilgan muntazam matn to‘plamlariga murojaat qilish uchun ishlatiladi[18]. Umuman korpus – til birliklarining xususiyatlarini aniqlash maqsadida qidiruv dasturiga bo‘ysundirilgan matnlar majmuyi, tabiiy tildagi elektron shaklda saqlanadigan yozma yoki og‘zaki, kompyuterlashtirilgan qidiruv tizimiga dasturiy ta’minot asosida joylashtirilgan matnlar jamlanmasi ta’rifini shakllantiradi.

Korpus lingvistikasining xususiyatlari:

– tabiiy matnlardan foydalanishning amaldagi qoliplarini empirik tahlil qiladi;

[17] Кутузов А.Б. Корпусная лингвистика. 2015. http://tc.utmn.ru/files/corpus_5.pdf

[18] Nadja Nesselhauf. Corpus linguistics: a practical introduction, 2005. http://www.as.uni-heidelberg.de/

– tahlil uchun asos sifatida korpusdan foydalanadi;
– lingvistik tahlillar maxsus korpus dasturlarida amalga oshiriladi;
– u sifat va miqdoriy tahlil metodlari asosida ish ko‘radi.

“Korpus lingvistikasining shakllanish tarixi”deb nomlangan 1.1.1-faslda korpus lingvistikasining shakllanish va rivojlanish bosqichlarining o‘ziga xosliklari, davrlashtirish masalasi yoritilgan.

Korpus lingvistikasi tarixiga murojaat qilganimizda ba’zi manbalarda uni 3 davrga ajratilganining guvohi bo‘lishimiz mumkin: erta korpus lingvistikasi davri (1950-yilgacha bo‘lgan davr), N.Chomskiy davri (1950-yillar) va zamonaviy korpus lingvistikasi (1960-yildan shu vaqtgacha bo‘lgan davr)[19].

T.Makkenri tadqiqotlarida korpus tilshunosligini, asosan, ikki bosqichga ajratib kuzatishni taklif qiladi. Korpus lingvistikasining shakllanish va rivojlanish tarixini besh bosqichga ajratish maqsadga muvofiq:

–XX asrning 1950-yillarigacha bo‘lgan davr: soha mavjud bo‘lmasa-da, ishning qog‘oz varianti mavjud davr;

–1950-1980-yillargacha bo‘lgan davr: N.Chomskiy nazariyasi ustunlik qilgan, korpuslar yaratish bo‘yicha tadqiqotlar boshlangan, tandiqiy baholash yuqori bo‘lgan davr;

–1990-2010-yillargacha bo‘lgan davr: dunyoda yetakchilik qilgan tillar miqyosida korpuslar yaratilishi avj olgan, metod va nazariyasi ishlab chiqilib ommalashtirilgan, ta’limda yangi metodikaning yuzaga kelishiga sabab bo‘lgan, kompyuter texnologiyalarining imkoniyatlari yuqori darajada rivojlangan davr;

–2010-yillardan bugungi kungacha bo‘lgan davr: barcha tillarda korpus yaratish tadqiqotlari ilgari surilgan, universal dasturlar imkoniyati kengaytirilgan, sohaning mustaqil institut darajasidagi mavqeyi, korpusdan foydalanish doirasining kengayishi yuz bergan davr;

–beshinchi bosqich: sun’iy intellekt bosqichiga yetish davri.

Birinchi bobning 1.1.2-fasli “Korpus lingvistikasining obyekti, metodi va vazifalari” deb nomlangan. Korpus tilshunosligining markazida korpus konsepti turadi. Namuna olish (sampling), muvozanat (balance) va til vakili (representativeness) bo‘lish korpus tilshunosligining asosiy nazariy tushunchalaridir[20]. A.Kutuzov korpus lingvistikasi kursida an’anaviy tilshunoslik va korpus tilshunosligi o‘rtasidagi farqlarni tushuntirib, korpus lingvistikasi oldiga muhim vazifalarni ham qo‘yadi:

– an’anaviy tilshunoslik tilni o‘rganishga e’tibor qaratsa, korpus tilshunosligi diqqat markazida nutqni o‘rganish turadi;

– an’anaviy tilshunoslikning maqsadi tilni tavsiflash, tushuntirish, korpus tilshunosligining maqsadi esa tilni maxsus tanlangan matnlar to‘plami shaklida nutqda qanday namoyon bo‘lgan bo‘lsa, xuddi shunday tasvirlashdir;

– tilshunoslik o‘z tadqiqotida nazariyadan tushuntirishga va nutqiy faktlarni tasdiqlashga o‘tadi, korpus tilshunosligi matn korpusining ma’lumotlariga tayanadi;

[19]Allan, K. The Oxford handbook of the history of linguistics. – Great Britain, 2013. – P. 343.
[20]Research methods in linguistics. Litosseliti. Continuum. 2010. https://www.academia.edu/29875281/Research_Methods_in_Linguistics_Litosseliti_Continuum_2010_pdf

– an’anaviy tilshunoslikda sifat bilan bog‘liq metodlar afzal ko‘rilsa, korpus tilshunosligida miqdoriy metodlar afzal;

– tilshunoslik til universaliyalarini tadqiq etsa, korpus tilshunosligida til universaliyalarining matnda uchrashi tahlil qilinadi;

– tilshunoslikda nutq materialini tanlashda, ularni tadqiq etishning empirik materiallarini aniqlashda til materiallariga asoslanilsa, korpus tilshunosligi xulosalarida korpuslardagi matnlar to‘plamida mavjud nutqiy faoliyatni kuzatishga asoslaniladi;

– tilshunoslik taqqoslashlar, baholashlarga asoslangan kashfiyotlarga ishonsa, korpus tilshunosligi empirik ma’lumotlarni qayta ishlashga asoslangan ilmiy kashfiyotga asoslanadi[21].

A.Kutuzov keltirgan taqqoslashlar orqali bir nechta muhim xulosalarni olish mumkin, jumladan, korpus tilshunosligi amaliy soha bo‘lib, tilga oid muhim xulosalarini faqat kontekstda uchraydigan nutq birliklarining xususiyatiga ko‘ra beradi, korpus tahlillari tilshunoslikda mavjud tahlil va usullarni to‘ldiradi.

Korpus lingvistikasida to‘rt xil tahlil usulini farqlash mumkin:

1. Kalit so‘zlar tahlili (Analysis of keywords). Bu AvaB korpuslarini solishtirganda uchraydigan so‘zlar tahlili. Bu miqdoriy analiz usuli bo‘lib, u til korpusida berilgan nutq birliklarini topish yoki topa olmaslik ehtimolini o‘rganadi. Bu metod korpuslardagi birliklarning statistikasi, uchrash chastotasi miqdorini beradi;

2. Birikmalarni tahlil qilish (Analysis of collocations). Birikmalar– so‘zlar oralig‘ida topilgan so‘zlardir (–/+ so‘zlar chap va o‘ng tomonga). Ushbu tahlil berilgan korpusdagi kontekstda so‘zni topish ehtimolini tekshiradigan statistik testlarga asoslangan;

3. Kollektiv tahlil (Colligation analysis). Bunga so‘z, so‘zlar qatori boshqa so‘zlar bilan qo‘shilib ketishga moyil bo‘lgan sintagmatik qoliplarni tahlil qilish kiradi. Qoliplashtirish leksik element va grammatik kontekst orasidagi aloqaga, uning fraza va gaplardagi sintaktik vazifasi, joylashuviga urg‘u beradi. Ehtimol, har bir so‘z o‘ziga xos kollektiv tahlilni taqdim etadi;

4. N-gramm. N-gramm tahlili hisoblash usuliga tayanadi, unda so‘zlarning satrlari 2,3,4,5 yoki 6 so‘zdan iborat guruhlarga guruhlangan va ularning chastotasi ko‘rib chiqilgan[22]. O‘zbek tili korpusini qurish jarayonida yuqorida sanab o‘tilgan 4 usul olingan natijalarda namoyon bo‘lishini kuzatdik. Mazkur usullarni korpus lingvistikasidagi eng oddiy va muhim bo‘lgan usullar sifatida baholash mumkin.

Birinchi bobning 1.2-fasli “Internet matnlari asosida korpus yaratish” deb nomlangan. Bu faslda internet matnlarining til korpusi sifatidagi o‘rni va internet matnlari asosida korpus yaratishning nazariy asoslari aniqlangan. So‘nggi yillarda barcha tadqiqotchilar internetda mavjud maqolalar to‘plamidan, xususan, qo‘llanmalar, dissertatsiyalar, ilmiy maqolalar yozishda yoki muayyan sohalar bo‘yicha natijaviy xulosalarga ega bo‘lishda foydalanishmoqda. Grefenstette,

[21] Кутузов А.Б. Корпусная лингвистика. – 2015. http://tc.utmn.ru/files/corpus_5.pdf

[22] Using corpus linguistics: research methods. http://www.perezparedes.es/research-methods-corpus-linguistics.

Nioke[23] va Jons hamda Ganilar[24] veb-resurslarning imkoniyatlarini elektron resurslar kam bo‘lgan tillar uchun til korpuslarining manbayi deya ta’kidlashgan, Resnik[25] esa ikki tilli parallel korpuslar manbayi sifatida o‘rgangan. Fuji va Ishika va ensiklopediya yozuvlarini yaratish uchun internetdan foydalanishganini yozishgan[26]. Grefenstette leksik ma’lumot manbayi sifatida internetga oid istiqbollar va tajribalarni taqdim etgan[27], chunki veb ko‘plab tillar uchun minglab kontekstli misollarni taqdim etadi, tillar uchun leksik yozuvlarni empirik dalillardan avtomatik ravishda topish imkoniyatlarini yaratadi. Kuzatuvlarimiz natijasida internetning quyidagi afzalliklarini belgiladik:

1. Til korpuslarini internetda mavjud matnlarning tayyor elektron variantidan qurish afzal hisoblanadi. Veb amaliy muammolar va cheksiz yangi ma’lumotlar oqimi sifatida qulay[28]. Internetda paydo bo‘lgan til birliklari, neologizmlar tez ommalashadi. Eng oxirgi yangi sohalar yuzasidan matnlar paydo bo‘ladi.

2. Internetdan olingan ma’lumotlar uchun URL manzili berilsa, mualliflik huquqi buzilmaydi. Korpuslardan foydalanish har doim ham ochiq bo‘lmaydi. Ba’zi til korpuslariga kirish uchun ro‘yxatdan o‘tish va muayyan miqdordagi to‘lov talab qilinadi. Internet esa barcha uchun ochiq va tekin ma’lumot olishda qulay manba hisoblanadi.

3. Internetda foydalanuvchi dunyo aholisining ko‘pchilik qatlamiga notanish bo‘lgan til namunalarini ham topishi mumkin. Hali korpusi tuzilmagan tillar yuzasidan ma’lumotlar olish imkoniyatiga ega.

Internetning korpus sifatida baholanishi bo‘yicha tarafdor qarashlar mavjud bo‘lsa-da, bir qancha e’tirozli munosabatlarni kuzatish mumkin.

– internet matn turlari va mavzu sohalariga nisbatan muvozanatli emas[29]. Internet matnlarning boy to‘plamiga ega bo‘lsa-da, ularda janr, uslub masalasi hal qilinmaydi.

– internetda matnlarning hajmi, ularning sifatlilik darajasi haqida ma’lumotga ega emasmiz. Matnlarda imloviy, uslubiy xatoliklar uchrash ehtimoli yuqori, chunki internetda filtrlash funksiyasi mavjud emas.

– internetdagi hujjatlar faqat matndan iborat emas, ro‘yxat, jadval, indeks va rasmlardan iborat bo‘ladi,[30] bu esa til korpusida o‘tkaziladigan lingvistik analiz va sintez tahlillar va ulardan olinadigan qimmatli xulosalarning to‘g‘riligiga shubha

[23] Grefenstette G., Nioche J. Estimation of English and non-English Language Use on the WWW. In proceedings of RIAO (Recherche d’Informations Assistee par Ordinateur), Paris, 2000.
[24] Jones R., Ghani R. Automatically building a corpus for a minority language from the web. 38th Meeting of the ACL, Proceedings of the Student Research Workshop. - Hong Kong. October 2000, pp. 29-36.
[25] Resnik P. Mining the web for bilingual text In proceedings of the 37th Meeting of ACL. - Maryalnd, USA, June 1999, pp. 527-534.
[26] Fujii A., Ishikawa T. Utilizing the world wide web as an encyclopedia: Extracting term descriptions from semi-structured text. In proceedings of the 38th Meeting of the ACL, Hong Kong, October 2000, pp. 488-495
[27] Grefenstette G. The WWW as a Resource for Example-Based MT Tasks. Invited Talk, ASLIB ‘Translating and the Computer’ conference. - London. October, 1999.
[28]Kilgarriff A. Web as corpus. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.108.2212&rep=rep1&ty
[29]Agirre E., Olatz A., Hovy E. and Martinez D. Enriching very large ontologies using the WWW. ECAI 2000, Workshop on Ontology Learning. – Berlin, 2000.
[30]Agirre E., Olatz A., Hovy E., Martinez D. Enriching very large ontologies using the WWW. ECAI 2000, Workshop on Ontology Learning. – Berlin, 2000.

uyg‘otadi. Shuningdek, tahlilga tortilgan vebsahifalarning muddati tugashi yoki bloklanishi mumkin[31]. Aniq va to‘g‘ri natijaga mukammal qurilgan korpus orqali erishish mumkin.

– ma’lumotlarni yig‘ish va korpus qurish jarayonida rejasiz, nazoratsiz va tahrir qilinmagan matnlar to‘plami tufayli korpusni qurishga mas’ul shaxslar yakunlanadigan barcha manbalarni to‘liq nazorat qila olmaydi[32].

Internet va til korpusining o‘xshash hamda farqli jihatlarini tahlilga tortishimiz natijasida quyidagi xulosalarga keldik: internet ma’lumotlari asosida til korpuslarini qurish arzon, kam vaqt sarflashga olib keladi. Internet yuklab olinadigan va keyin qayta ishlanadigan matnli ma’lumotlar manbayi sifatida baholanishi maqsadga muvofiq. Til korpusining vebdan farqli jihati uning oldiga qo‘yilgan maqsadiga bog‘liq.

Birinchi bobning uchinchi fasli “Internet matnlari asosida korpus yaratishda qo‘llanuvchi dasturlar va ularning tasnifi” deb nomlangan. Mazkur faslda korpus yaratilishida va korpus tahlillarida foydalaniluvchi dasturlar tasniflangan, yutuq va kamchiliklari aniqlangan. Korpus dasturlarini bajaradigan vazifasiga ko‘ra bir qancha guruhlarga ajratdik: matn yig‘uvchi dasturlar, izohlovchi (annotation) dasturlar, konkordans dasturlari, statistik tahlil qiluvchi, teglovchi (tagger), grammatik tahlil qiluvchi (parser), tokenlovchi, semantik tahlil qiluvchi, audio matnlarni tahlil qiluvchi va aralash tahlilni (mixed) amalga oshiruvchi dasturlar. AntConc, Sketch Engine va BootCat dasturlarining ishlash prinsiplari tahlil qilindi, “Tug‘ro” o‘zbek tili internet axborot matnlari korpusining dasturiy ta’minoti platformada alohida ishlab chiqildi (https://uzkorpus.uz/).

Dissertatsiya ishining **“O‘zbek tilidagi internet axborot matnlari korpus uchun asos sifatida”** deb nomlangan ikkinchi bobning birinchi fasli “O‘zbek tilidagi elektron axborot matnlarining struktur-semantik xususiyatlari”deb ataladi. Bu faslda elektron axborot matnlarining tili va uslubi o‘rganildi, internet axborot matnlarining imkoniyatlari, unga qo‘yilgan talablar, leksik-semantik strukturasi tavsiflandi. “Tug‘ro” til korpusida tadqiqot obyektiga aylantirilgan xabar matnlarida narrativ, deskriptiv, argumentli, didaktik matn tiplarining barchasini uchratish mumkin. Xabar matnlari strukturasi tahlilida otli birikmalar ko‘pchilikni tashkil qildi. Masalan, “O‘zbekiston Respublikasi”, “Butun Xitoy xalq vakillari kengashi”, “Ichki ishlar vaziri”, “Jinoyat kodeksi”, “Mustaqil davlatlar hamdo‘stligi ijroiya qo‘mitasi” kabi murakkab til birliklari kontekstda uchraganda korpusda mazmuniy xususiyatiga ko‘ra bir butunlikda izohlandi.

Nutqimizda mavjud, ammo o‘zbek tili lug‘at boyligiga kirishga ulgurmagan qo‘shma fe’llar izohlandi. Korpusning teglangan 3000ta so‘zshakllarda qo‘shma fe’llar statistikasi 195tani tashkil etdi: bu raqamlar 100ta qo‘shma fe’lning takroriy qo‘llanishi bilan hisoblangan. Ko‘makchi fe’lli so‘z qo‘shilmasi 27ta: 16 marta uchragan KFSQning takroriy uchrash soni ham hisoblangan (diss. 5-ilova). Qo‘shma fe’llar statistikasi o‘rganilganda, internet xabar matnlarida eng ko‘p uchragan *tashkil etmoq* (12 marta), *joriy etmoq* (11 marta), *taqdim etmoq* (7 marta)

[31]Kilgarriff A. Web as corpus. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.108.2212&rep=rep1&ty
[32]Bouzou X. G. The Web as a Corpus A Multilingual Multipurpose Corpus. – Barcelona, 2018.

so‘zlari hisoblanadi. Kontekstda qo‘shma fe’llar fe’l so‘z turkumining turli nisbat va mayl shakllarida berilgan. Korpusda otli so‘z birikma, qo‘shma fe’l va ko‘makchi fe’lli so‘z qo‘shilmasining statistikasi ham berildi.

O‘zbek tili lug‘ati va darsliklarida ayrim til birliklarining qaysi turkumga mansubligi masalasi ochiq qolganligi ko‘rindi. Masalan, *baribir, mazkur, shu bilan birga* nutq birliklari kontekstual mazmuni orqali turkumlandi. O‘zbek tili lug‘atida *bari* jamlovchi olmosh deb berilgan, *bari bir* esa “*hammasi birdek, hech qanday farqi yo‘q*” ma’nolari berilib, turkumga mansublik masalasi ochiq qolgan. Ammo xabar matnlarida *bari bir* so‘zshakli – *baribir* shaklida uchraydi. Lug‘atda esa bu haqida ma’lumot uchramaydi[33].

Bobning ikkinchi fasli “O‘zbek tilidagi elektron axborot matnlari asosida korpus yaratish uchun tanlangan dastur va uning o‘ziga xos xususiyatlari” deb nomlanib, unda o‘zbek tili internet axborot matnlari korpusining qurilish jarayoni yoritildi, bosqichlari, belgilash uchun teglarning tanlanishi, teglash jarayonining muammolari yoritildi.

O‘zbek tili internet axborot matnlari korpusida (https://uzkorpus.uz/) matnlarda uchragan nutq birliklarining morfologik va semantik xususiyatlari belgilangan teglar orqali annotatsiyalandi.

2.1-jadval. Fe’l so‘z turkumi teglarining belgilanishi

	O‘zbekcha qisqartma	**Xalqaro qisqartma**
Yordamchi fe’l	Yor	AUXVerb
O‘timli fe’llar	O‘tl	TranVerb
O‘timsiz fe’llar	O‘ts	IntrVerb
Ravishdosh	Rav	GerVerb
Sifatdosh	Sif	PartVerb
Harakat nomi	Harn	VerbN
Fe’lning aniq nisbat shakli	Aniqn	ActVerb
Fe’lning o‘zlik nisbat shakli	O‘zn	RFLVerb
Fe’lning orttirma nisbat shakli	Ortn	CauVerb
Fe’lning birgalik nisbat shakli	Birn	RcpVerb
Fe’lning majhul nisbat shakli	Majn	PassVerb
Fe’lning bo‘lishli shakli	Bo‘l	PosVerb
Fe’lning bo‘lishsiz shakli	Bo‘lsiz	NegVerb
Buyruq-istak mayli	Buymay	ImpVerb
Shart mayli	Shar	CndVerb
Maqsad mayli	Maq	NecVerb
Xabar mayli	Xab	GenVerb
O‘tgan zamon	O‘tz	PastVerb
Hozirgi zamon	Hozz	PresVerb
Kelasi zamon	Kelz	FutVerb
Ko‘makchi fe’lli so‘z qo‘shilmasi	KFSQ	VP

[33]O‘zbek tilining izohli lug‘ati. – Toshkent, 2006.

Birinchi shaxs	1sh	1Conj
Ikkinchi shaxs	2sh	2Conj
Uchinchi shaxs	3sh	3Conj

Yuqoridagi jadvalda berilgani kabi barcha so‘z turkumlarining grammatik kategoriyalari va lisoniy imkoniyatlari tanlangan teglar bilan belgilandi.

Teglar leksik yozuvlardan farq qiladi. Biz leksik qoidalarda mustahkamlangan birliklarning qisqartma nomlarini belgilashga harakat qildik. Lekin tadqiqotimizda grammatik shakl uchun qaysi tegning belgilanishi emas, balki korpus platformasida teglash masalasining hal qilinganligi, lingvistik tahlillar amalga oshirilganda muhim natijalarning olinishi bilan belgilanadi.

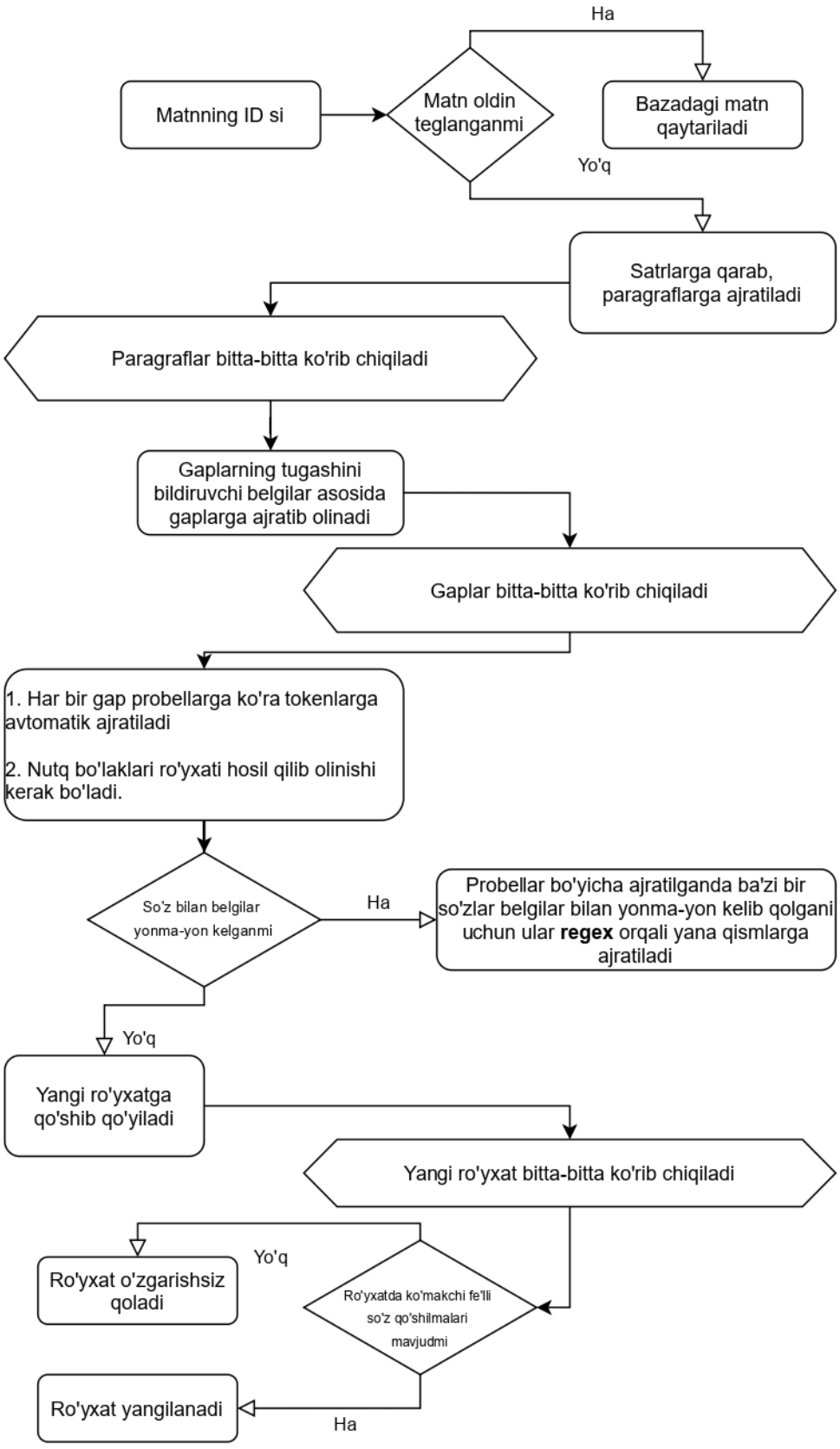
Nutq birliklarini annotatsiyalash jarayonida belgilangan teglar foydalanuvchiga noqulaylik tug‘dirmasligi uchun saytda qo‘llanma berilgan. Qo‘llanmada tegning to‘liq shakli va ingliz tilidagi standart belgisi ham berilgan. Korpusga yuklangan har bir matndagi nutq birligi teglar bilan izohlandi. Masalan, kontekstda *dorilarning* so‘zi lemma: dori <morf ot+Ko‘p+Qark, sem AnOT>, *keldim* so‘zi lemma: kel, <morf fe’l+O‘tz+Xab+1sh, sem Jisf>, *yozuvchining asari* so‘z birikmasi lemma: yozuvchi <morf ot+Bir+Qark+Yas, sem AOT>, lemma: asar <morf ot+Bir+3shE sem AnOt> sifatida teglanadi. Korpusning hozirgi imkoniyatlarida o‘zbek tilidagi matnlar morfologik va semantik xususiyatlariga ko‘ra annotatsiyalandi. Nutq birliklarida omonimlik muammosi yuzaga kelmasligi uchun, kontekst o‘rganilib, dasturda maxsus belgi orqali ajratildi va bu belgi filtr vazifasini bajaradi. Omonimlik xususiyatiga ega nutq birliklari uchun filtrda barcha grammatik shakllari belgilanadi, teglash jarayonida kontekst mazmunidan kelib chiqib, izohlanadi.

Ba’zi olimlar aynan nutq qismlarini teglash jarayoni tabiiy tilni qayta ishlashning muhim qismi deb hisoblashadi. Chunki teglash jarayonida til birligining nafaqat nutqda bajarayotgan vazifasi va o‘zi mansub kategoriyasiga ko‘ra teglansa, boshqa tomondan nutq qismining kontekstdagi semantikasi va pragmatikasini ham tushunish va hisobga olish zarur hisoblanadi. Chunki bir o‘rinda “ot” izohi ostida kelgan nutq birligi boshqa o‘rinda “fe’l” sifatida teglanishi mumkin. Hozir korpusda teglash va qidiruv jarayonini test sifatida bajarmoqdamiz. Nutq birliklarini teglayotganimizda so‘zning asosiy grammatik shakllarini olishga asosiy e’tiborni qaratdik. Masalan, ot-so‘z shaklni tavsiflash uchun Sh.Hamroyeva tadqiqotida ham ta’kidlanganidek egalik, kelishik, son kategoriyasi asosiy o‘rinni egallaydi[34].

Teglarning tanlanishi bo‘yicha aniq standart qoida ishlab chiqilmagan bo‘lsa-da, lekin bu yaratilayotgan har bir korpus oldiga qo‘yilgan maqsadga bog‘liq. Nutq qismlari izohlanishida morfologik annotatsiya tarkibiga til birliklarining lemmalari aniqlanishi ham kiritiladi. Izohlashgacha bo‘lgan jarayon bosqichma-bosqich amalga oshiriladi.

Korpusda amallarni bajarish jarayonining algoritmi 1-rasmda berilgan.

[34] Ҳамроева Ш. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари: Филол.фан.бўйича фалсафа доктори (PhD)…дисс. –Бухоро, 2018 .–253 б.



2.1.-rasm. Korpusda amallarni bajarish jarayonining ketma-ketligi

Bobning uchinchi fasli “O‘zbek tilidagi elektron axborot matnlari korpusini yaratishdagi o‘ziga xosliklar” deb nomlanib, mazkur faslda korpus qurilishidagi tokenlash, lemmalash va morfologik tahlil masalalari hal qilindi. O‘zbek tilidagi

elektron axborot matnlari korpusida tokenlarga ajratish avtomatik amalga oshirildi, nutq birligining lemmasini aniqlash qo‘l mehnati va yarim avtomatik bajarildi.

Masalan, *Yoshlar bandligi masalalari bo‘yicha yig‘ilishda davlat rahbari yoshlarni qishloq xo‘jaligiga jalb qilish masalasiga to‘xtaldi* (https://kun.uz/uz/news/2021/04/13/) gapi korpusda avtomatik tokenlarga ajratiladi. Bu gapda 15 ta token mavjud; gapdagi tinish belgilar ham token sifatida olinadi. Korpusda avtomatik tokenlarga ajratilgan birliklar ro‘yxat ko‘rinishida paydo bo‘ladi. Morfologik tahlil korpusdagi shaklni lemmalashdan boshlanadi, ya’ni uning leksemasi berildi.

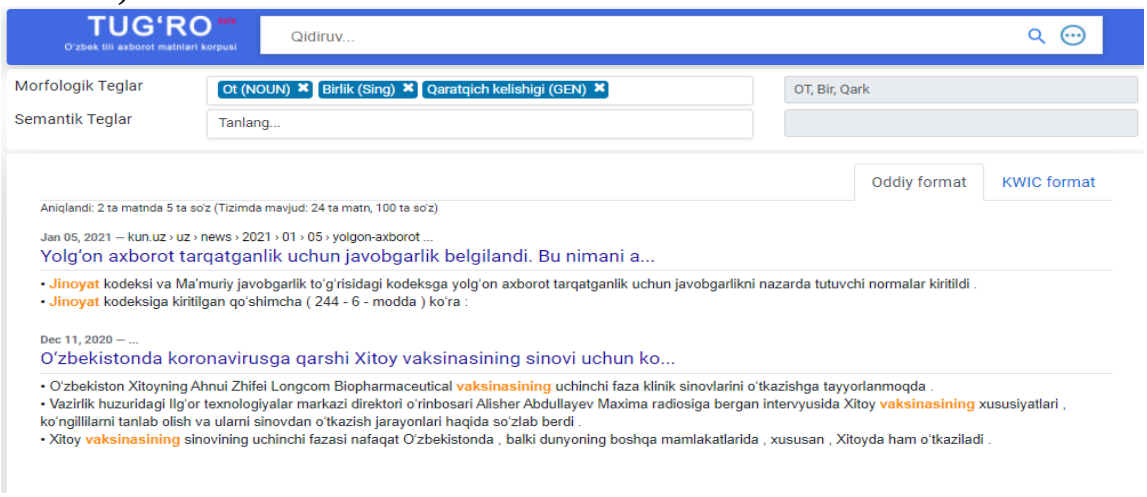
Morfologik tahlil quyidagicha amalga oshiriladi:

1) yoshlar <yoshlar>[morf. OT+ Ko‘p, sem. AOT+JamOT]
2) bandligi <bandlik> [morf. OT+ Bir+3shE+YAS, sem. MavOT]
3) masalalari <masala> [morf. OT+ Ko‘p+3shE, sem. TOT]
4) bo‘yicha <bo‘yicha> [morf. K, sem. BoshqN]
5) yig‘ilishda <yig‘ilish> [morf. OT+ Bir +Harn+ O‘rink, sem. JamOT]
6) davlat <davlat> [morf. OT+ Bir, sem. TashN]
7) rahbari <rahbar> [morf. OT+Bir+3shE, sem. AOT]
8) yoshlarni <yoshlar> [morf. OT+Ko‘p+Tushk, sem. AOT+JamOT]
9) qishloq <qishloq> [morf. OT, sem. TOP]
10) xo‘jaligiga <xo‘jalik> [morf. OT+Bir+3shE+Jo‘nk, sem. TashN]
11) jalb
12) qilish
13) jalb qilish <jalb qil> [morf. FL+Qfl+ Harn, sem. Jismoniyf]
14) masalasiga <masala> [morf. OT+Bir+3shE+ Jo‘nk, sem. TOT]
15) to‘xtaldi <to‘xta> [morf. FL+ Majn+ Xab+ O‘tz, sem. Nutqf]
16) .<.> [morf. TB, sem. BoshqN]

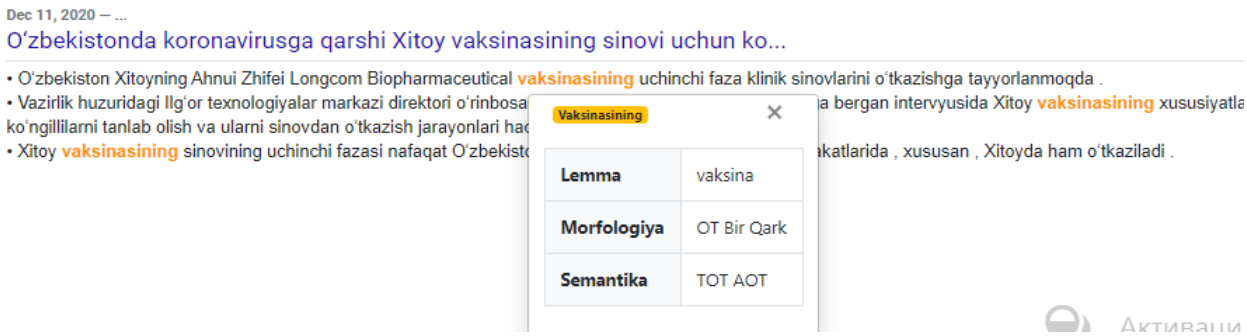
Dissertatsiyaning uchinchi bobi “**O‘zbek tilidagi internet axborot matnlari korpusining lingvistik annotatsiyasi**” deb nomlangan. Uning “Til korpuslarida lingvistik annotatsiya va uning prinsiplari” deb nomlangan birinchi faslida korpusda lingvistik izohlash – annotatsiyalashning mohiyati, prinsiplari, turlari, annotatsiyalash zaruratini keltirib chiqaradigan sabablar belgilangan. Korpus sof matnlardan tashqari, izohlash deb nomlanuvchi qo‘shimcha lingvistik ma’lumotlar bilan ta’minlanadi. Til birliklarini morfosintaktik, semantik izohlash uchun maxsus teglar belgilanadi.

“O‘zbek tilidagi internet axborot matnlari korpusining morfologik annotatsiyasi” deb nomlangan ikkinchi faslda “Tug‘ro” nomli internet axborot matnlari korpusida morfologik annotatsiyalash algoritmi ko‘rsatildi. Morfologik izohlashda kontekstga ko‘ra belgilandi. Til korpusi orqali barcha birliklar teglanib, qo‘shma fe’llar va KFSQlarning maxsus statistikasi berildi. *Shu*_OLKo‘r *tariqa*_OT *ayni*_OLKo‘r *kungacha*_RAV *mamlakatda*_OTO‘rink *koronavirus*_OTQark *infeksiyasiga*_OTJo‘nk *qarshi*_K *emlangan*_FLMajnSif *fuqarolar*_OTKo‘p *soni*_SON3shE *458 555*_RAQSONSanoq *nafarga*_SON *yetdi*_FLAniqnO‘tz ._TB (https://daryo.uz/2021/04/21/). Bu gap faqat morfologik izohlangan. O‘zbek tili axborot matnlari korpusida tadqiqotchilar foydalanuvchi interfeysi orqali nutq birliklari yuzasidan tahlil amalini bajarganda, tahlilga

tortilgan nutq birligi ishtirok etgan kontekst oynada ko‘rinadi (https://uzkorpus.uz manzili ostida). Masalan:



Foydalanuvchi interfeysida tahlilga tortilgan leksemaning morfologik va semantik ko‘rsatkichlari “popup”da aks etadi. Masalan:



“O‘zbek tilidagi internet axborot matnlari korpusining sintaktik annotatsiyasi” deb nomlangan uchinchi faslda korpuslarda sintaktik annotatsiyalashning ahamiyati va zaruriyati xususida so‘z yuritiladi. Sintaktik izohlash grammatika qoidalari asosida matnlarni avtomatik tahlil qilish bilan bog‘liq jarayon hisoblanadi[35]. Texnik jihatdan, bu matnning sintaktik tuzilishini aniqlash amaliyotiga murojaat qilish uchun ishlatiladi[36]. Bu korpus lingvistikasida parsing yoki grammatik tahlil qilish deyiladi. O‘zbek tili internet axborot matnlari korpusida sintaktik annotatsiyalash hali ish jarayonida turibdi.

Bobning to‘rtinchi fasli “O‘zbek tilidagi internet axborot matnlari korpusining semantik annotatsiyasi” deb nomlanib, korpusda so‘zlarning semantik xususiyatlari va semantik maydonlari teglar yordamida annotatsiyalandi. Korpus orqali tilshunoslikda omonimiya mavzusi bo‘yicha yangi fikrlar berilishi, omonimlar lug‘ati tayyorlanishi mumkin. O‘zbek tili korpusida so‘z-shakl va qo‘shimchalar qo‘shilishi jarayonidagi omonimlik masalasi mazkur so‘zlarning lemmasini alohidalash orqali hal etildi. Masalan, “*Talabalar o‘qishga kelishdi*” yoki “*Mahsulot narxini tushirishga kelishdi*” jumlalari kontekstda uchraganda, har bir so‘zshakl semantik teglar orqali izohlandi. Mazkur kontekstdagi “*kelishdi*” fe’li omonimlik muammosini yuzaga keltirmaydi, chunki birinchi gapdagi so‘zning

[35] Barnbrook G. Language and Computers. – Edinburgh: Edinburgh University Press, 1998. – p. 170.
[36] McEnery T., Wilson A. Corpus Linguistics. – Edinburgh: Edinburgh University Press, 1996. – p. 178.

lemmasi “*kel*”, ikkinchi gapdagi birlikning lemmasi “*kelish*” tarzida teglanadi. Xuddi shunday *tortma, eslatma, qo‘llanma, yo‘qlama* shakllarida yoki *unga berdi*, *unga soldi* kabi gaplardagi *unga* birliklaridagi qo‘shimchalar kombinatsiyasidagi muammolar hal etildi. O‘zbek tili internet axborot matnlari korpusi annotatsiyasida 40 ta semantik teg yaratildi. Semantik annotatsiya tasnifi asosiy semantik xususiyatlarni belgilaydigan muhim leksik bo‘linmalarga ko‘ra tanlandi.

3.2-jadval. Semantik teglarning belgilanishi

	O‘zbekcha qisqartma	**Xalqaro qisqartma**
Ibora	IB	IDIOM
Juft so‘z	JUFT	ECH
Takror so‘z	TAK	RDP
Atoqli otlar	AOT	PROPN
Antroponim (kishi ismi)	ANT	PRS
Toponim (joy nomi)	TOP	TOP
Gidronim (suv havzalari nomi)	GID	HYD
Familiya	Fam	FamN
Otasining ismi	OtaIsm	PatrN
Tashkilot nomi	TashN	COM
Mahsulot nomi	MahN	PRO
Geografik hudud nomi	GeoN	GEO
Millat nomi	MillN	NAT
Hayvon nomi	HayN	AnimN
Boshqa	BoshqN	OTH
Turdosh ot	TOT	ConN
Mavhum ot	MavOT	AbstN
Jamlovchi ot	JamOT	CollN
Aniq ot	AnOT	ConcN
Miqdor son	Miq	Qty
Tartib son	Tar	OrdNum
Kasr son	Kasr	FracNum
Chama son	Cham	Approx
Jamlovchi son	Jam	CollNum
Dona son	Dona	Integer
Yumush fe’llari	Yumushf	Move
Tafakkur fe’llari	Tafakkurf	Ment
Sezgi fe’llari	Sezgif	Perc
Ruhiy-holat fe’llari	Ruhiyf	Psych
Nutq fe’llari	Nutqf	Speech
Ishora fe’llari	Ishoraf	Semelf
Jismoniy holat fe’llari	Jismoniyf	Physiol
Tabiiy holat fe’llari	Tabiiyf	Changest
Ko‘rish fe’llari	Ko‘rishf	Noncaus

Xususiyat sifatlari	Xususiyats	Humq
Rang-tus sifatlari	Tuss	Physq:color
Ma’za-ta’m sifatlari	Ta’ms	Physq:taste
Hajm-o‘lchov sifatlari	Hajms	Physq:form
Hid sifatlari	Hids	Physq:smell
Makon-zamon belgisini bildiruvchi sifatlar	Makzams	Place:time

Korpusda nutq birliklari avval morfologik annotatsiyalanadi, keyin semantik annotatsiyalash amalga oshiriladi. Masalan, *Andijondagi Muruvvat uyida 6,5 yoshli siam egizaklari tarbiyalanmoqda* (https://daryo.uz/2020/09/28) jumlasidagi nutq birliklarini vertikal formatda izohlanishini ko‘ramiz.

1) *Andijondagi* <Andijon> [morf. OT+Otlug‘sh., sem. TOP]
2) *muruvvat* <muruvvat> [morf. OT+Bir, sem. MavOT+TashN]
3) *uyida* <uy> [morf. OT+3shE+O‘rink, sem. TOT]
4) *6,5* <6.5> [morf. SONSanoq, sem. Miq]
5) *yoshli* <yoshli> [morf. SIF+Yas, sem. Xususiyats]
6) *siam* <siam> [morf. OT+Bir, sem. AOT+ TOP]
7) *egizaklari* <egizak> [morf. OT+Ko‘p+3shE, sem. AOT]
8) *tarbiyalanmoqda* <tarbiyala> [morf. FL+ Majn+ Hozz, sem. Yumushf]
9) . <.> [morf. TB, sem. BoshqN]

Yuqoridagi namunadan ko‘rishimiz mumkinki, ishimizda nutq birliklari annotatsiyasida morfologik izohlardan keyin semantik izohi berilmoqda. Hamma so‘z yoki bir xil so‘zlar har doim ham bitta semantik maydonga tushmaydi. Masalan, yuqoridagi gapda *Siam* so‘zining semantik maydondagi xususiyatiga ko‘ra atoqli ot va toponim teglari belgilandi. Ba’zi nutq birliklari uchun bir vaqtning o‘zida bir nechta lug‘aviy ma’noni ko‘rsatishga xizmat qiladigan teglarni belgilash holati uchraydi. Bu kabi birliklarni ko‘plab uchratish mumkin.

XULOSA

O‘zbek tilining internet axborot matnlari korpusini shakllantirishning nazariy va amaliy asoslari bo‘yicha olib borilgan ilmiy tadqiqot asosida quyidagi xulosalarga kelindi:

1. Korpus – elektron shaklda saqlanadigan va kompyuterlashtirilgan qidiruvni tashkil etishga imkon beradigan har qanday tildagi og‘zaki yoki yozma matnlardan iborat bo‘lgan tabiiy matnlar to‘plami.

2. Korpus lingvistikasi sohasida olib borilgan tadqiqotlarni o‘rganish, tahlil qilish shuni ko‘rsatdiki, bu yo‘nalish ko‘p o‘lchovli, keng qamrovli, zamonaviy mustaqil yo‘nalishlardan biri. Korpus lingvistikasining shakllanish va rivojlanish bosqichlarini quyidagicha tasnif qilish mumkin:

– XX asrning 50-yillarigacha bo‘lgan davr: birinchi bosqich – soha mavjud bo‘lmasa-da, korpuslarning qog‘oz varianti mavjud davr;

– 50-80-yillar – Xomskiy nazariyasi ustunlik qilgan, korpuslar yaratish bo‘yicha tadqiqotlar boshlangan, tandiqiy baholash yuqori bo‘lgan davr;

– 1990-2010-yillargacha bo‘lgan davr – dunyoda yetakchilik qilgan tillar miqyosida korpuslar yaratilishi avj olgan, metod va nazariyasi ishlab chiqilib ommalashtirilgan, ta’limda yangi metodikaning yuzaga kelishiga sabab bo‘lgan, kompyuter texnologiyalarining imkoniyatlari yuqori darajada rivojlangan davr;

– 2010-yillardan bugungi kungacha bo‘lgan davr – barcha tillarda korpus yaratish tadqiqotlari ilgari surilgan, universial dasturlar imkoniyatlari kengaytirilgan, sohaning mustaqil institut darajasidagi mavqeyi, korpuslardan foydalanish doirasining kengayishi yuz bergan davr;

– so‘nggi – beshinchi bosqich: sun’iy intellekt bosqichi.

3. Internetning korpus sifatida tutgan o‘rnini baholaydigan bo‘lsak, maksimal monitor korpusga tenglashtirish hamda webCorp deb belgilash mumkin. Kuzatishlarimiz korpuslarning qurilishi tartibli va tizimli, internetning o‘lchamlari noma’lum va doimiy ravishda o‘zgarib turishi natijasida til korpusi beradigan aniq va to‘g‘ri natijalarni berish, natijani qayta ko‘rsatish imkoniyati yuqori emasligini ko‘rsatdi. Shu sababli to‘g‘ri yondashilgan va annotatsiyalangan tahlil natijalariga ega bo‘lish uchun o‘zbek tili elektron axborot matnlari korpusini qurish zarurati mavjud.

4. Korpus imkoniyatlarini belgilashda AntConc, Corpus.byu.edu, Sketch Engine, Web BootCat dasturlari ahamiyatli. Korpusning yuqorida keltirilgan bu menejerlari korpus qurish; korpusdan lingvistik tahliliy natija va xulosalarga ega bo‘lish; turli soha vakillarining foydalanishi uchun zarur. AntConc dasturi universial, undan tijoriy maqsad ko‘zlanmagan. Ammo tilning grammatik xususiyatlarini tahlil qilolmasligi tufayli korpus materialini annotatsiyalash imkoniga ega emas. Sketch Engine va Web BootCat korpus menejerlari – faqat korpus tuzish uchun material yig‘uvchi dastur: ular vositasida tuzilgan korpus annotatsiyalanmaganligi, maxsus lingvistik belgilar bilan teglanmaganligi sababli ishlov berilmagan matn to‘plamidan farq qilmaydi.

5. O‘zbek tili internet axborot matnlari korpusi uchun maxsus dasturiy ta’minot yaratilib, tajriba sifatida tokenlashtirish avtomatik amalga oshirildi, til birliklari – lemmani aniqlash va teglash jarayoni qo‘l mehnati bilan amalga oshirildi. Shu asosda o‘zbek tili internet axborot matnlari korpusida lingvistik ta’minoti ish jarayoni 6 bosqichda amalga oshirish lozimligi nazariy jihatdan aniqlandi:

– o‘zbek tili grammatikasini o‘rganish va ma’lumotlar yig‘ish;

– ma’lumotlarning taxminiy sxemasini ishlab chiqish, maxsus belgilar – teglar tizimini ishlab chiqish;

– elektron axborot matnlarini korpus menejeriga yuklash;

– yuklab olingan matnlarni kodlash, metama’lumotlar kiritish;

– korpus birliklarini annotatsiyalash;

– lingvistik qidiruv tizimini ishlab chiqish.

6. O‘zbek tili internet axborot matnlari korpusini yaratishning muhim jarayonlaridan biri morfologik tahlil va morfologik ishlov berish tamoyillarini ishlab chiqishdir. Shu tamoyillar asosida o‘zbek tili internet axborot matnlari korpusini tuzishning dastlabki bosqichida morfologik va semantik annotatsiyalash amalga oshirildi. Annotatsiyalashda morfologik va leksik-semantik teglar uchun

belgilar tanlash, korpusda har bir belgini kodlash ishi bajarildi. Korpus birliklarini morfologik annotatsiyalashda belgilarni kodlash turidan foydalanish yaxshi samara beradi, ishda polisemiya va omonimiyani farqlash uchun filtr vazifasini bajaruvchi algoritm tavsiya etildi.

7. Korpusda annotatsiyalash uchun xalqaro standart tegger dasturlari tajribasiga tayangan holda, o‘zbek tili birliklarining leksik-grammatik xususiyatlariga ko‘ra 64 ta morfologik, lug‘aviy ma’noviy guruhlariga ko‘ra 40 ta semantik teg aniqlandi: fe’l <FL>, ot <OT>, sifat <SIF>, son <SON>, olmosh <OL>, ravish <RAV>, taqlid so‘z <Taq>, ko‘makchi <K>, bog‘lovchi <B>, yuklama <Y>, undov so‘z <U>, modal so‘z <M>. Morfologik annotatsiyalashda til birliklarining boshqa grammatik ko‘rsatgichlarini izohlash uchun birlikning qisqartmaligi <Qisq>, yasama so‘z ekanligi <Yas>, tub so‘z ekanligi <TUB>, birlik <Bir> yoki ko‘plik <Ko‘p> shakli kabi teglar ham ishlab chiqilib tizimlashtirildi. Teglarni belgilashda o‘zbek tilining standart qoidalari mavjud emasligi sababligi universal teggerlar tahlil qilinib, o‘zbek tili so‘z turkumlari va ularning morfologik kategoriyasiga ko‘ra morfologik teglar va leksik birliklarning LMGlariga ko‘ra semantik teglar tizimi belgilandi.

8. Korpus birligini teglash – tilning o‘ziga xos xususiyatlarini ko‘rsatishda so‘zga maxsus kod biriktirish jarayoni. O‘zbek tili internet axborot matnlari korpusida uchragan muammolar tilning tabiatidan, tilning boshqa tillar bilan o‘zaro ta’siri, shuningdek, korpus materialini yig‘ish uslubidan kelib chiqishi ma’lum bo‘ldi. Annotatsiyalangan o‘zbek tili internet axborot matnlari korpusida tahlillar olib borish o‘zbek tilining tuzilishi, o‘ziga xos grammatik xususiyatlari haqida aniq va ishonchli umumlashmalar chiqarishga asos bo‘ldi.

SCIENTIFIC COUNSIL AWARDING SCIENTIFIC DEGREES
PhD.03/30.12.2019.Fil.60.02 AT ANDIJAN STATE UNIVERSITY
TASHKENT STATE UNIVERSITY OF THE UZBEK LANGUAGE AND LITERATURE NAMED AFTER ALISHER NAVOI

ABDULLAEVA OKILA KHOLMUMINOVNA

THEORETICAL AND PRACTICAL BASIS OF DEVELOPING WEB BASED CORPUS OF UZBEK NEW TEXTS

10.00.11 – Language theory. Applied and computer linguistics

ABSTRACT
OF THE DOCTOR PHILOSOPHY (PhD) ON PHILOLOGICAL SCIENCES

Andijan – 2022

The theme of doctor of philosophy (PhD) thesis was registered by the Supreme Attestation Commission at the Cabinet of Ministers of the Republic of Uzbekistan under No B2019.3.PhD/Fil1018.

The doctor of philosophy thesis was written at Tashkent State University of the Uzbek Language and Literature named after Alisher Navoi.

The abstract of dissertation in three languages (Uzbek, English and Russian (resume)) has been placed on the webpage Andijan State University (www.adu.uz) of Scientific Council and information-educational portal «Ziyonet» (www.ziyonet.uz).

Scientific adviser:	**Azimova Iroda Alisherovna** candidate of Philological Sciences, Associate Professor
Official opponents:	**Nabiyeva Diloro Abduxamidovna** doctor of Philological Sciences, Professor
	Eshmuminov Asqar Allamurodovich doctor of Philosophy (PhD) in Philological Sciences
Leading organization:	**Samarkand State University**

The defense of the dissertation will be held on «__ _»_____, 2022 at_____at the meeting of Scientific Council awarding scientific degrees PhD. 03/30.12.2019.Fil.60.02 at Andijan State University. Address: 129, University str., Andijan city, Uzbekistan, 170100. Tel: 0(374) 223 88 14; fax: 0(374) 223 88 30 e-mail: agsu_info@edu.uz.

The (PhD) dissertation can be reviewed at the Information resource center of Andijan State University (registration number ____). Address: 129, University str., Andijan city, Uzbekistan, 100174. Tel: 0(374) 223 88 14.

The abstract of the dissertation is distributed on «_ __» _____ 2022.
(Protocol of the register № ____ on «___ _» _____ 2022).

Sh.H.Shaxabitdinova
Chairman of the Scientific Council for awarding scientific degrees, Doctor of Philological Sciences, Professor

F.F.Usmanov
secretary of Scientific Council awarding scientific degrees, Doctor of Philosophy (PhD)

M.E.Umarxodjayev
Chairman of the Scientific Seminar at the Scientific Council awarding scientific degrees, Doctor of Philological Sciences, Professor

INTRODUCTION (the abstract of the (PhD) dissertation)

Actuality and necessity of the research theme. During the last half century, new approaches and different directions have emerged in world linguistics in drawing conclusions through practical experiments in language research, machine language processing, and linguistic research through various programs. One of such new directions is corpus linguistics and a large-scale scientific and theoretical research is carried out in this area. In particular, serious attention is paid to the creation of national corpora of languages, the development of existing corpora.

Availability of necessity to study the existing possibilities of language in world linguistics, identifying problematic aspects of language grammar in context, defining grammatical forms in language, facilitating the creation of multidisciplinary electronic dictionaries, improving the use of modern information technology in language learning, automatic translation, search and the development of theoretical and practical bases for the creation of corpora in languages to address issues such as computer analysis, preparation of electronic textbooks and dictionaries, the need to build a corpus of language in specific areas determines the relevance of our research.

In recent years, necessary reforms have been carried out in our country on the use of computer technologies in order to increase the efficiency of work in all spheres, to facilitate human labor. At the same time, the laws and decisions adopted to increase the status of the Uzbek language and ensure its active application set a number of important tasks for specialists. President of the Republic of Uzbekistan Sh. Mirziyoyev in his congratulatory speech on the 31st anniversary of the granting of the status of the state language to the Uzbek language said: “We need to solve important and urgent tasks ahead of us to ensure that the Uzbek language occupies a worthy place in the World Arena, in particular in the Internet Information Network, to create many new computer programs in our native language[37].” Attention to the Uzbek language has increased to the level of priority directions of state policy. The support of research work on the development of the state language, the implementation of international cooperation in this field proves the importance of social significance, practical effectiveness of each research conducted in practice.

This study, to some extent, will serve to carry out the tasks set out in decrees of President of Republic of Uzbekistan including PF-4797 of May 13, 2016 “On the organization of activity of the Tashkent State University of Uzbek Language and Literature named after Alisher Navoi”, PF-4947 of February 7, 2017 “On the Action Strategies for further development of the Republic of Uzbekistan”, PF-5850 of October 21, 2019 “On measures to radically increase the prestige and status of the Uzbek language as the state language”, No. PF-6084 of October 20, 2020 “On measures to further develop the Uzbek language and improve language policy in

[37]O‘zbekiston Respublikasi Prezidenti Shavkat Mirziyoyevning “O‘zbek tilining davlat tili sifatidagi nufuzi va mavqeini tubdan oshirish chora-tadbirlari to‘g‘risida”gi farmoni // www.xabar.uz

our country", PF-6097 of October 29, 2020 "On approval of the Concept for the development of science until 2030"; resolution of PQ-4479 of October 4, 2019 "On the broad celebration of the thirtieth anniversary of the adoption of the Law of the Republic of Uzbekistan "On the State Language", resolution of the Cabinet of Ministers of the Republic of Uzbekistan dated December 12, 2019 № 984 "On approval of the Regulation on the Department of State Language Development" and other regulations related to this activity.

Relevant research priority areas of science and developing technology of the Republic. This dissertation is in line with the priorities of the development of science and technology of the republic I. "Social, legal, economic, cultural, spiritual and educational development of an informed society and a democratic state, the development of an innovative economy".

Comments on scientific researches on the theme of the dissertation. In world linguistics, corpus linguistics and Corpus creation research began since the 1960s.The first research work and corpus were created in English. And by the 1990s, corpora were created in many languages of the world. In traditional and computer linguistics, the role, methodology, objects and functions of Corpus Linguistics have been specially studied in the research of scientists such as G.Leech, R.Garsid, J.Sinclair, L.Flowerdev, T. McEnery, A.Hardy, M.McCarthy, W.N.Francis, G.Kennedy[38].

The social significance of corpuses, the subsequent stages of the development of Corpus Linguistics, the use of the internet as a corpus, similar and different aspects between the internet and corpuses have been covered in detail and the theoretical basis of the issue have been investigated by researchers such as A.Kilgarrif, K.Stewart, G.Grefenstette, M.Hundt, N.Nesselhauf[39].

Researches of T.McEnery, R.Xiao, R.Reppen, D.Biewer in the field of corpus linguistics create an opportunity to obtain a clear and detailed picture of the

[38] 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.ilc.cnr.it/EAGLES/browse.html.; 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 Word-tagging 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 University 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.; Flowerdew L. Corpus Linguistic Techniques Applied to Textlinguistics. 1998. – p. 541-552.; McCarthy M., O'Keefe A. What are corpora and how have they evolved?: The Routledge handbook of corpus linguistics. – London and New York, 2010.

[39] Kilgarriff A., Grefenstette G. 2003. Introduction to the special issue on the Web as corpus // Computational Linguistics 29(3). – p. 333-347.; Kilgarriff A. Web as corpus. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.108.2212&rep=rep1&type=pdf; Stuart K. New perspectives on corpus linguistics.; Grefenstette G. The WWW as a Resource for Example-Based MT Tasks. // Translating and the Computer. – London, 1999.; Grefenstette G. Nioche J. Estimation of English and non-English Language Use on the www. // RIAO (Recherche d'Informations Assistee par Ordinateur). – Paris, 2000.; Hundt M., Nesselhauf N., Biewer C. Corpus linguistics and the web. – Amsterdam-New York, 2007.; Hundt M., Nesselhauf N., Biewer C. Corpus linguistics and the web. – Amsterdam-New York, 2007.

role of corpus in the educational process, the principles and advantages of using corpus in language teaching[40].

A number of researches in Russian linguistics carried out by V.Plungyan, V.P.Zakharov, A.B.Kutuzov plays a special role in the formation of corpus linguistics as a separate field in Russian linguistics, in the creation of the Russian national corpus[41].

In Uzbek linguistics, up to Corpus Linguistics, a lot of important research work has been carried out in the field of computer linguistics. In particular, these include the research work of A.Pulatov[42], S.Muhamedov in collaboration with a group of scientists[43], the scientific results of N.Abdurahmanova[44], M.Abjalova[45]. Since Corpus Linguistics is a new and rapidly developing field in Uzbek linguistics in recent years, the work carried out in the monographic scale is very rare. In Particular, in Sh.Hamroyeva's dissertation on the theme "Linguistic bases of Uzbek language authorship corpus"[46], as well as her monograph[47] of the same name, the formation, development and theoretical basis of corpus linguistics, the specific theoretical and practical aspects of the formation of the corpus linguistics, as well as the general and private linguistic basis for the formation of the corpus linguistics are described. In recent years, in Uzbek linguistics, a lot of research works – dissertations, articles and theses are being published. A.Eshmuminov's dissertation on "Synonymous database of the Uzbek National Corpus"[48], D.Akhmedova's dissertation on "Linguistic bases and models of lesxical-semantic

[40] McEnery T., Xiao R., Tono Y. Corpus-based Language Studies: An Advanced Resource Book. Routledge, 2006.; Reppen R. Using corpora in the language classroom. - Cambridge: Cambridge University Press, 2010; Biber D., Conrad S., Reppen R. Corpus Linguistics. Investigating Language Structures and Use. - Cambridge: Cambridge University Press, 1998.

[41]ПлунгянВ. Перспективы: Корпусная лингвистика и корпус русского языка. https://www.youtube.com/watch?v=OBTLuLx962U; Захаров В.П. Корпусная лингвистика: учеб.-метод. Пособие / В.П. Захаров. – СПб., 2005; Кутузов А.Б. Корпусная лингвистика, 2015.

[42] Пўлатов А., Мўминова Т., Пўлатова И. Дунёвий ўзбек тили. Ўзбек тилида феъл шакллари ва уларнинг рус, инглиз тилидаги кўринишлари. – Тошкент: Университет, 2003; Пўлатов А. Компьютер лингвистикаси. – Тошкент, 2011.

[43] Мухамедов С.А., Пиотровский Г.Г. Инженерная лингвистика и опыт системно – статистического исследования узбекских текстов. –Т.: Фан, 1986; Махмудов М.А., Пиотровская А.А., Садыков Т. Система машинного анализа и синтеза тюркской словоформы // Переработка текста методами инженерной лингвистики. – Минск, 1982.

[44] Абдурахмонова Н.З. Инглизча матнларни ўзбек тилига таржима қилиш дастурининг лингвистик таъминоти (содда гаплар мисолида): филол. фан. бўйича фалсафа докт. дисс. автореф. – Тошкент, 2018. – 49 б.; Abdurakhmonova N. Uzbek ontology of Uzbek language as example of adjective // Шестая Международная конференция по компьютерной обработке тюркских языков 363 "TurkLang-2018". (Труды конференции) – Ташкент: Издательско-полиграфический дом "Navoiy universiteti", 2018. – 320 с.

[45] Абжалова М. Ўзбек тилидаги матнларни таҳрир ва таҳлил қилувчи дастурнинг лингвистик модуллари (расмий ва илмий услубдаги матнлар таҳрири дастури учун): Филол. фан. бўйича фалсафа доктори (PhD)…дисс. –Фарғона, 2019.; Абжалова М. Матнларга авто-лингвистик ишлов бериш тизимлари // Шестая Международная конференция по компьютерной обработке тюркских языков "TurkLang2018". (Труды конференции) – Ташкент: Издательско-полиграфический дом "Navoiy universiteti", 2018. – 320 с

[46]ҲамроеваШ. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари: Филол. фан. бўйича фалсафа доктори (PhD)…дисс. – Бухоро, 2018 .–253б.

[47]Ҳамроева Ш. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари: Монография. – Тошкент, 2020. – 229 б.

[48]Эшмўминов А. Ўзбек тили миллий корпусининг синоним сўзлар базаси: Филол. фан. бўйича фалсафа доктори (PhD) дисс. – Қарши, 2019. – 137 б.

tagging of naming units for Uzbek language corpus"[49], scientific researches of B.Mengliyev[50], G.Toirova[51], H.Ataboyev[52] can be cited as an example.

Relevance of the dissertation research with the plans of the scientific research works of the higher education. The dissertation was carried out within the framework of the scientific research plan of the Applied Linguistics and Linguistics Department of the Tashkent State University of Uzbek language and Literature named after Alisher Navoi named "Computer and Corpus Linguistics – modeling of the Uzbek language for artificial intelligence and preparation of the linguistic base for the creation of national and private corpus of the Uzbek language".

The aim of the research work is to develop the theoretical and practical basis for the creation of the internet information texts corpus in Uzbek linguistics, to identify the specific principles of linguistic annotation of texts in the corpus and to create the corpus of electronic information texts in Uzbek language.

The tasks of the research work:

to determine the theoretical framework for the creation of a corpus on the basis of internet texts;

to identify the specific important aspects of the creation of the electronic information texts corpus of the Uzbek language;

to develop an algorithm of linguistic annotation of texts in the Uzbek language corpus, view the principle of their work through experimental results, analyze linguistically results in the corpus.

The object of the research work is internet electronic information texts in Uzbek language including all texts on kun.uz, daryo.uz, xabar.uz, sof.uz, gazeta.uz, qalampir.uz and UzA between 2020-2021-years period, were selected.

The subject of the research work is the lexical-semantic and morphological features of the Uzbek electronic information texts in the Uzbek language, which are important in the creation of the corpus.

Methods of the research. In conducting the study, the principles of classification, description, comparison, statistical method and its analysis were used.

Scientific novelty of the research work is as follows:

the principles of interpreting speech units are proven by identifying lexical-semantic, structural aspects of linguistic units in information texts, allowing to develop linguistic support of the platform of the Uzbek language internet information texts corpus;

---

[49]Ахмедова Д. Атов бирликларини ўзбек тили корпуслари учун лексик-семантик теглашнинг лингвистик асос ва моделлари: Филол. фан. бўйича фалсафа доктори (PhD) дисс. – Бухоро, 2020. –247 б.

[50]Mengliyev B. O'zbek tili taraqqiyoti va rivojlanish omillari // "O'zbek tilini dunyo miqyosida keng targ'ib qilish bo'yicha hamkorlik istiqbollari" mavzusidagi xalqaro ilmiy-amaliy anjuman materiallari. – Toshkent, 2020.; Менглиев Б., Бобожонов С., Ҳамроева Ш. Ўзбек тилининг миллий корпуси. http://marifat.uz/marifat/ruknlar/fan/1241.htm

[51] Тоирова Г. Миллий корпус яратишнинг технологик жараёни хусусида //Ўзбекистонда хорижий тиллар. Электрон илмий-методик журнал. – Тошкент. 2020, № 2 (31), –Б.57– 64. https://journal.fledu.uz/uz/ 2-31-2020

[52] Ataboyev N. Problematic issues of corpus analysis and its shortcomings // ISJ Theoretical & Applied Science, 10 (78), 2019.; Ataboyev N. Compiling dictionaries by using corpus analysis and its advantages // International Journal of Progressive Sciences and Technologies (IJPSAT) http://ijpsat.es/index.php/ijpsat/article/view/508

speech units are morphologically annotated in the Uzbek language internet information texts corpus, in which developed classification category of verbs in texts, positive-negative category, category of action style, variable category of verb, numerical category of nouns, comparision-classification category of adjectives, lexical-grammatical group of number, adverb, imitation, morphological features of pronouns and linguistic possibilities of auxiliary words;

lexical-semantic, lexical-thematic groups, lexical-conseptual area of speech units in texts that allow semantic annotation of linguistic units in Uzbek language internet information texts corpus is determined;

the problem of identification of homonymous words, affixal homonymy and homonymy-related phenomena in the classification of speech units encountered in the context in the annotation, the problem of determining compound verbs and auxiliary verb conjunctions are proven through experience in the corpus.

Practical results of the research work consist of the followings:

technical instructions for the construction of the Uzbek language electronic information texts Corpus were developed and the work process was determined step by step;

the Uzbek language internet information texts corpus was built and published as a site (https://uzkorpus.uz/);

morphological and semantic tags were developed in practice to tag speech units in the corpus, special standard characters were selected;

speech units in the corpus were manual and semi-automatic annotated according to lexical-semantic and morphological features;

the fact that the conclusions, the knowledge drawn from the study of the problems facing the corpus provide important scientific knowledge of Uzbek linguistics, in particular, computer linguistics, corpus linguistics, become the basis for the emergence of new methods of linguistic analysis in linguistics, contribute to the development of lexicography were scientifically based.

Authenticity of the research results are determined by the fact that the problem identified in the dissertation is clearly stated, that corpus construction has been studied through long-term observations with corporations of other languages that the Uzbek language corpus has been built through the analysis of the collected materials, that linguistic analysis and scientific-theoretical conclusions were made in the corpus.

Scientific and practical value of the research. The scientific significance of the results of the dissertation is determined by its theoretical significance in the creation of new research in the field of computer linguistics, enrichment through scientific views on the formation of the Uzbek language, the creation of computer language.

Practical significance of the study the materials, results and conclusions contained therein serve as the basis for carrying out the function of source and material in Uzbek linguistics in the subject of Corpus Linguistics, a special course in the direction of bachelor’s and master’s education, special seminars, as well as research in the stylistic direction of language.

Implementation of the research results. On the basis of scientific results obtained on the theoretical and practical basis of formation of the internet information texts corpus of the Uzbek language:

to form of linguistic support of the internet information text corpus platform, conclusions on the use of linguistic models in the creation of a language corpus, the development of standard tools for modeling and defining linguistic units of the Uzbek language for the Language corpus have been used in the ERASMUS CLASS project "Development of the interdisciplinary master program on Computational Linguistics at Central Asian Universities", implemented in 2017-2020 (Bulletin of ERASMUS+, National Erasmus+ office – Uzbekistan; bulletin of Tashkent State University of Uzbek language and Literature named after Alisher Navoi dated September 6, 2021 № 01/4-1464.) Syllabuses implemented within the framework of the project are supplemented with modeling of written textbook Language units;

conclusions on technology of construction of the internet information texts corpus of the Uzbek language, morphological and semantic annotation of speech units in the Corpus have been used in the project JHBL-20 "Creation of an electronic corpus of works of art on the theme of family, neighborhood and gender equality", implemented by the Research Institute "Mahalla and Family" in 2020-2021. (Reference of the Research Institute "Mahalla and Family" dated September 8, 2021, No. 01 / 09-517). Syllabuses implemented within the framework of the project and information on the corpus construction were used in the created project;

development of the principles of morphological and semantic annotation of speech units in the Uzbek language internet information texts Corpus, linguistic analysis on the platform of the internet information texts Corpus named "Tugro" and practical results obtained have been used in the fundamental project "Theoretical issues of Azerbaijan-Uzbek-Turkmen linguistics" number 1-147, implemented in 2016-2020. (Reference № 230 of the Institute of Linguistics named after Imadeddin Nasimi of the Azerbaijan National Academy of Sciences of the Republic of Azerbaijan dated October 29, 2021) As a result, the project is enriched with new scientific and theoretical information;

general conclusions on annotation of speech units in the internet information texts of the Uzbek language, setting the principles of linguistic annotation, morphological and semantic interpretation of language units, linguistic analysis on the Corpus platform and obtaining practical results have been used in lectures and practical classes on "Research Techniques and Publishing (LE0101100)" taught at Haji Bayram Wali University. (Reference of Hacı Bayram Veli University, Faculty of Literature, Department of Contemporary Turkish Dialects and Literature, Ankaradated October 25, 2021). As a result, the content of lectures and practical classes was enriched with information on linguistic analysis on the corpus platform;

information on the creation of national corpus in Uzbek linguistics, the use of Corpus in teaching Uzbek as a native language and a foreign language have been effectively used production of scientific and educational programs on the

National Television and Radio Company of Uzbekistan on the TV channel "Uzbekistan", in particular, "Oydin Hayot Live", "Talim va taraqqiyot". (Reference of the Uzbek Television and Radio Company № 02-13-1129 of July 5, 2021). As a result, the content of these TV shows has been enriched with scientific evidence.

Approbation of the research results. The research results were discussed at 6 scientific conferences, including 2 national and 4 international scientific conferences.

Publication of the research results. 11 scientific works on the topic of the dissertation have been carried out: 5 articles in scientific publications recommended by the Higher Attestation Commission of the Republic of Uzbekistan for publication of the main scientific results of doctoral dissertations, including 3 in national and 2 in foreign journals.

The structure and scope of the dissertation. The dissertation consists of an introduction, 3 chapters, a conclusion, a list of references and an appendix, the total volume is 158 pages.

THE MAIN CONTENT OF THE DISSERTATION

The introductory part substantiates the actuality and necessity of the research theme, the degree to which the problem has been studied, the aims and functions of the research, the object and subject of the research, its relevance to the priorities of science and technology, scientific novelty and practical results. Information on the implementation of research results in practice, approbation of research results, published works and the structure of the dissertation.

The first chapter of the dissertation is entitled **"Theoretical foundations of creating a corpus based on Internet texts".** The first part of the chapter, entitled "Corpus linguistics as a field of linguistics", describes the role of the corpus in the field of language and linguistics, the theoretical foundations of corpus linguistics, the criteria by which corpus is formed. A corpus is a collection of electronic texts selected on the basis of external criteria to fully reflect a language or language change. It works as a source of information for linguistic research[53].

Although the corpus can refer to any systematic text set, it is still used in a narrow sense today, typically to refer to computerized regular text sets[54]. In general, a corpus, a set of texts that are subject to a search engine to determine the properties of language units, forms a description of a set of texts placed in a computer-based search engine based on software, written or oral, stored in electronic form in a natural language.

Features of corpus linguistics:

- empirically analyzes the current patterns of use of natural texts;
- uses the corpus as a basis for analysis;

[53]Кутузов А.Б. Корпусная лингвистика. 2015. http://tc.utmn.ru/files/corpus_5.pdf

[54]Nadja Nesselhauf. Corpus linguistics: a practical introduction, 2005.http://www.as.uni-heidelberg.de/

- linguistic analysis is carried out in special corpus programs;

- works on the basis of qualitative and quantitative methods of analysis.

Part 1.1.1, entitled “History of the formation of corpus linguistics”, describes the peculiarities of the stages of formation and development of corpus linguistics, the problem of chronology.

When we look at the history of corpus linguistics, we can see that in some sources it is divided into 3 periods: the period of early corpus linguistics (before 1950), the period of N. Chomsky (1950s) and modern corpus linguistics (period from 1960 to the present)[55].

In his research, T. McEnery proposes to observe corpus linguistics mainly in two stages. It is expedient to divide the history of formation and development of corpus linguistics into five stages:

- the period up to the 1950s: the period when there was a paper version of the work, although the field did not exist;

- the period from 1950 to 1980: the period when N. Chomsky’s theory prevailed, research on the creation of corpus began, and the critical assessment was high;

- the period from 1990 to 2010: the period of creation of corpus in the world's leading languages, the development and popularization of methods and theories, the emergence of new methods in education, the high level of development of computer technology;

- the period from 2010 to the present: the period of research on the creation of the corpus in all languages, the expansion of universal software, the position of the industry as an independent institution, the expansion of the use of the corpus;

- the fifth stage: the stage of reaching the stage of artificial intelligence.

Part 1.1.2 of the first chapter is entitled “Object, method and tasks of corpus linguistics”. At the centre of corpus linguistics is the corpus concept. Sampling, balance, and representativeness are the basic theoretical concepts of corpus linguistics[56]. In the course of corpus linguistics, A. Kutuzov explains the differences between traditional linguistics and corpus linguistics and sets important tasks for corpus linguistics:

- while traditional linguistics focuses on language learning, corpus linguistics focuses on speech learning;

- the purpose of traditional linguistics is to describe and explain language, and the purpose of corpus linguistics is to describe language as it appears in speech in the form of a specially selected set of texts;

- linguistics in its research shifts from theory to explanation and confirmation of spoken facts, corpus linguistics relies on the data of the text corpus;

- qualitative methods are preferred in traditional linguistics, while quantitative methods are preferred in corpus linguistics;

[55]Allan, K. The Oxford handbook of the history of linguistics. – Great Britain, 2013. – P. 343.

[56]Research methods in linguistics. Litosseliti. Continuum. 2010. https://www.academia.edu/29875281/Research_Methods_in_Linguistics_Litosseliti_Continuum_2010_pdf

– if linguistics studies language universals, in Corpus Linguistics the occurrence of linguistic features in the text is analyzed;

- linguistics is based on the selection of speech material, the identification of empirical materials for their study, the conclusions of corpus linguistics are based on the observation of the speech activity present in the set of texts in the corpus;

- linguistics is based on comparisons, evaluation-based discoveries, while corpus linguistics is based on scientific discoveries based on the processing of empirical data[57].

Several important conclusions can be drawn from the comparisons made by A. Kutuzov, in particular, corpus linguistics is a field of practice that provides important linguistic conclusions only in terms of the nature of the units of speech that occur in the context, while corpus analysis complements existing analyzes and methods in linguistics.

Four different methods of analysis can be distinguished in Corpus Linguistics:

1. Analysis of keywords. This is an analysis of words that come across when comparing the corpuses A and B. This is a method of quantitative analysis, which studies the probability of finding or not finding speech units given in the language corpus. This method gives the statistics of units in the corpus, the amount of frequency of occurrence;
2. Analysis of collocations. Compounds are words found between words (- / + words left and right). This analysis is based on statistical tests that examine the probability of finding a word in the context of a given corpus;
3. Colligation analysis. This includes analyzing syntagmatic patterns that tend to blend in with other words along with the word. Formulation emphasizes the relationship between a lexical element and a grammatical context, its syntactic function in phrases and sentences, and its location. Probably each word presents its own colligation analysis;
4. N-gramm. N-gramm analysis is based on a method of calculation in which the rows of words are grouped into groups of 2, 3, 4, 5 or 6 words and their frequencies are considered[58]. During the construction of the Uzbek language corpus, we observed that the above 4 methods are reflected in the results obtained.These methods can be considered as the simplest and most important methods in corpus linguistics.

Part 1.2 of the first chapter is entitled “Creating a Corpus Based on Internet Texts”. This part explores the role of Internet texts as a language corpus and the theoretical foundations for creating a corpus based on Internet texts. In recent years, all researchers have been using the collection of articles available on the internet, in particular, when writing manuals, dissertations, scientific articles, or when obtaining conclusions on specific areas. Grefenstette, Nioke[59] and Johns as

[57]Кутузов А.Б. Корпусная лингвистика. – 2015. http://tc.utmn.ru/files/corpus_5.pdf
[58]Using corpus linguistics: research methods. http://www.perezparedes.es/research-methods-corpus-linguistics
[59]Grefenstette G., Nioche J. Estimation of English and non-English Language Use on the WWW. In proceedings of RIAO (Recherche d’Informations Assistee par Ordinateur), Paris, 2000.

well as Gani[60] stated that the web resource capacity is the source of language corpus for languages where electronic resources are scarce. Resnik[61] learned as a source of bilingual parallel corpuses. Fuji and Ishikawa wrote that they used the internet to create encyclopedia writings[62]. Grefenstette provides perspectives and experiences on the Internet as a source of lexical information[63], because the web offers thousands of contextual examples for many languages, creates opportunities for languages to automatically find lexical records from empirical evidence. As a result of our observations, we have identified the following advantages of the Internet:

1. It is preferable to build language corpus from a ready-made electronic version of texts available on the Internet. The web is handy as a practical problem and an endless stream of new information[64]. Linguistic units and neologisms that appear on the Internet are rapidly gaining popularity. Texts will appear on the latest new fields.

2. Providing a URL for information obtained from the Internet does not infringe copyright. The use of enclosures is not always open. Some language courses require registration and a fee. The Internet is an open and free source of information for all.

3. On the Internet, users can also find language patterns that are unfamiliar to most segments of the world's population. It has access to information on languages that have not yet been created.

While there are differing views on the value of the Internet as a corpus, there are a number of objections.

- the Internet is not balanced in terms of text types and subject areas[65]. Although the Internet has a rich collection of texts, it does not address the issue of genre or style.

- we do not have information about the size of texts on the Internet, their level of quality. Spelling and style errors in texts are more likely to occur because the Internet does not have a filtering function.

- documents on the Internet doesn't only consists of text, but lists, tables, indexes and figures[66], which casts doubt on the accuracy of linguistic analysis and synthesis analysis in the language corpus and the valuable conclusions drawn from them. It is also possible that the websites involved in the analysis will expire or be

[60]Jones R., Ghani R. Automatically building a corpus for a minority language from the web. 38th Meeting of the ACL, Proceedings of the Student Research Workshop. - Hong Kong. October 2000, pp. 29-36.
[61]Resnik P. Mining the web for bilingual text In proceedings of the 37th Meeting of ACL. - Maryalnd, USA, June 1999, pp. 527-534.
[62]Fujii A., Ishikawa T. Utilizing the world wide web as an encyclopedia: Extracting term descriptions from semi-structured text. In proceedings of the 38th Meeting of the ACL, Hong Kong, October 2000, pp. 488-495
[63]Grefenstette G. The WWW as a Resource for Example-Based MT Tasks. Invited Talk, ASLIB 'Translating and the Computer' conference. - London. October, 1999.
[64]Kilgarriff A. Web as corpus. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.108.2212&rep=rep1&ty
[65]Agirre E., Olatz A., Hovy E. and Martinez D. Enriching very large ontologies using the WWW. ECAI 2000, Workshop on Ontology Learning. – Berlin, 2000.
[66]Agirre E., Olatz A., Hovy E., Martinez D. Enriching very large ontologies using the WWW. ECAI 2000, Workshop on Ontology Learning. – Berlin, 2000.

blocked[67]. Accurate and right results can be achieved with a perfectly constructed corpus.

- due to unplanned, unsupervised and unedited collection of texts during data collection and building construction, those responsible for building the corpus will not be able to fully control all the resources that will be completed[68].

After analyzing the similarities and differences between the Internet and the language corpus, we came to the following conclusions: the construction of language corpus on the basis of Internet data is cheaper and less time consuming. The Internet should be considered as a source of textual information that can be downloaded and then processed. The difference between the language corpus and the web depends on its purpose.

Part 3 of the first chapter is entitled "Programs used to create a corpus based on Internet texts and their classification". In this part, the programs used in the creation of the corpus and the analysis of the corpus are classified, the advantages and disadvantages are identified.We have divided corpus programs into several groups according to the task they perform: text collector programs, annotator (annotation) programs, concordance programs, statistical analyzer, tagger, grammatical analyzer (parser), tokenizer, semantic analyzer, audio text analyzer and mixed analysis (mixed) programs.The principles of operation of AntConc, Sketch Engine and BootCat were analyzed, the software of the corpus of Uzbek Internet information texts "Tugro" was developed separately on the platform (https://uzkorpus.uz/).

The first part of the second chapter of the dissertation, named "**Internet information texts in Uzbek as a basis for the corpus**", is entitled "Structural and semantic features of electronic information texts in Uzbek language". In this part the language and style of electronic information texts are studied, the possibilities of Internet information texts, its requirements, lexical-semantic structure are described. In the "Tugro" language corpus, all types of narrative, descriptive, argumentative, and didactic texts can be found in the texts of messages that have been the object of research. In the analysis of the structure of message texts, noun compounds predominated. For example, when complex language units such as "Republic of Uzbekistan", "All-China People's Congress", "Interior Minister", "Criminal Code", "Commonwealth Executive Committee" are encountered in context, they are explained in its entirety according to its semantic properties in the corpus.

The common verbs in our speech, but which have not yet entered the vocabulary of the Uzbek language, have been explained. In the 3,000 corpus-tagged verbs, the number of compound verbs was 195: the number was calculated by repeating 100 compound verbs. The auxiliary verb conjugation is 27. The number of recurrences of auxiliary verb conjugation, which occurred 16 times, was also calculated. When studying compound verb statistics, the most common words in internet message texts are *tashkil etmoq* (12 times), *joriy etmoq* (11 times), and

[67]Kilgarriff A. Web as corpus. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.108.2212&rep=rep1&ty
[68]Bouzou X. G. The Web as a Corpus A Multilingual Multipurpose Corpus. – Barcelona, 2018.

taqdim etmoq (7 times). In the context, compound verbs are given in different voice and mood forms of the verb phrase. The corpus also provides statistics on noun phrases, compound verbs, and auxiliary verbs.

The Uzbek dictionary and textbooks show that the question of which category some language units belong to remains open. For example, speech units *baribir*, *mazkur*, *shu bilan birga* were categorized by contextual content. In the Uzbek dictionary, all of them are cumulative noun, and all of them mean "all the same, no difference", and the question of belonging to the category remains open. But in the text of the message, word "*bari bir*" is encountered as "*baribir*". There is no information in the dictionary about that[69].

The second part of the chapter is entitled "The program selected for the creation of the corpus on the basis of electronic information texts in the Uzbek language and its features". It describes the process of construction of the corpus of Uzbek Internet information texts, the stages, the choice of tags for designation, the problems of the tagging process.

In the Uzbek language internet information texts corpus (https://uzkorpus.uz/) morphological and semantic features of speech units encountered in the texts were annotated by marked tags.

Table 2.1. Determination of the tags of the verb

	Abbreviation in Uzbek	**International abbreviation**
Auxilary verbs	Yor	AUXVerb
Transitive verbs	O‘tl	TranVerb
Intransitive verbs	O‘ts	IntrVerb
Gerund verbs	Rav	GerVerb
Participle verbs	Sif	PartVerb
Infinitive verbs	Harn	VerbN
Active voice	Aniqn	ActVerb
Reflexive voice	O‘zn	RFLVerb
Causative voice	Ortn	CauVerb
Reciprocal voice	Birn	RcpVerb
Passive voice	Majn	PassVerb
Positive verbs	Bo‘l	PosVerb
Negative verbs	Bo‘lsiz	NegVerb
Imperative verbs	Buymay	ImpVerb
Conditional mood	Shar	CndVerb
Subjunctive mood	Maq	NecVerb
General mood	Xab	GenVerb
Past tense	O‘tz	PastVerb
Present tense	Hozz	PresVerb
Future tense	Kelz	FutVerb
Auxiliary verb	KFSQ	VP

[69]O‘zbek tilining izohli lug‘ati. – Toshkent, 2006.

conjugation		
1-person	1sh	1Conj
2-person	2sh	2Conj
3-person	3sh	3Conj

As given in the table above, grammatical categories and linguistic possibilities of all word categories were determined by the selected tags.

Tags are different from lexical records. We have tried to define abbreviation of units reinforced in lexical rules. However, in our study, it is not the designation of which tag for the grammatical form, but the resolution of the problem of tagging on the corpus platform, is determined by the acquisition of important results when linguistic analyzes are performed.

The tags provided during the annotation of speech units are provided on the site so as not to cause inconvenience to the user. The manual also provides the full form of the tag and the Standard English symbol. The unit of speech in each text loaded into the corpus was explained with tags. For example, in the context, the word for *dorilarning* in Uzbek is lemma: dori <morph at + Plural + Genitive Case, sem Definite Noun>, the word *keldim* (came) in Uzbek is lemma: *kel*, <morph verb + Past Simple + inform. + I rd person, sem Physical Action Verb>, *yozuvchining asari* is labeled as lemma: yozuvchi <morph Noun + Singular + Genitive Case + Derivate, sem AOT>, lemma: asar <morph Noun + Singular + Possessive form of III rd person sem Definite Noun>. In the current capacity of the corpus, Uzbek texts are annotated according to their morphological and semantic features. In order to avoid the problem of homonymy in speech units, the context was studied and separated in the program by a special symbol, which acts as a filter. For speech units with homonymous properties, all grammatical forms are defined in the filter, interpreted based on the contextual content during the tagging process.

Some scholars believe that the process of tapping parts of speech is an important part of natural language processing. This is because in the process of tagging, it is necessary to understand and take into account not only the function of the language unit in speech and the category to which it belongs, but also the semantics and pragmatics of the part of speech in context. Because in one place the unit of speech under the definition of “noun” can be called “verb” in another place. We are now performing the tagging and search process on the case as a test. When we tagged on speech units, we focus on getting the basic grammatical forms of the word. For example, in the research of Sh.Khamroyeva[70], the categories of possessive form of noun, noun cases, and singular and plural forms of noun play an important role in describing the noun form.

Although there is no clear standard rule for the selection of tags, it depends on the purpose of each case created. In the interpretation of parts of speech, the morphological annotation also includes the definition of lemmas of language units. The pre-interpretation process is continued gradually.

[70] Ҳамроева Ш. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари: Филол.фан.бўйича фалсафа доктори (PhD)…дисс. –Бухоро, 2018 .–253 б.

The algorithm of the process of performing operations in the corpus is shown in Figure 1.

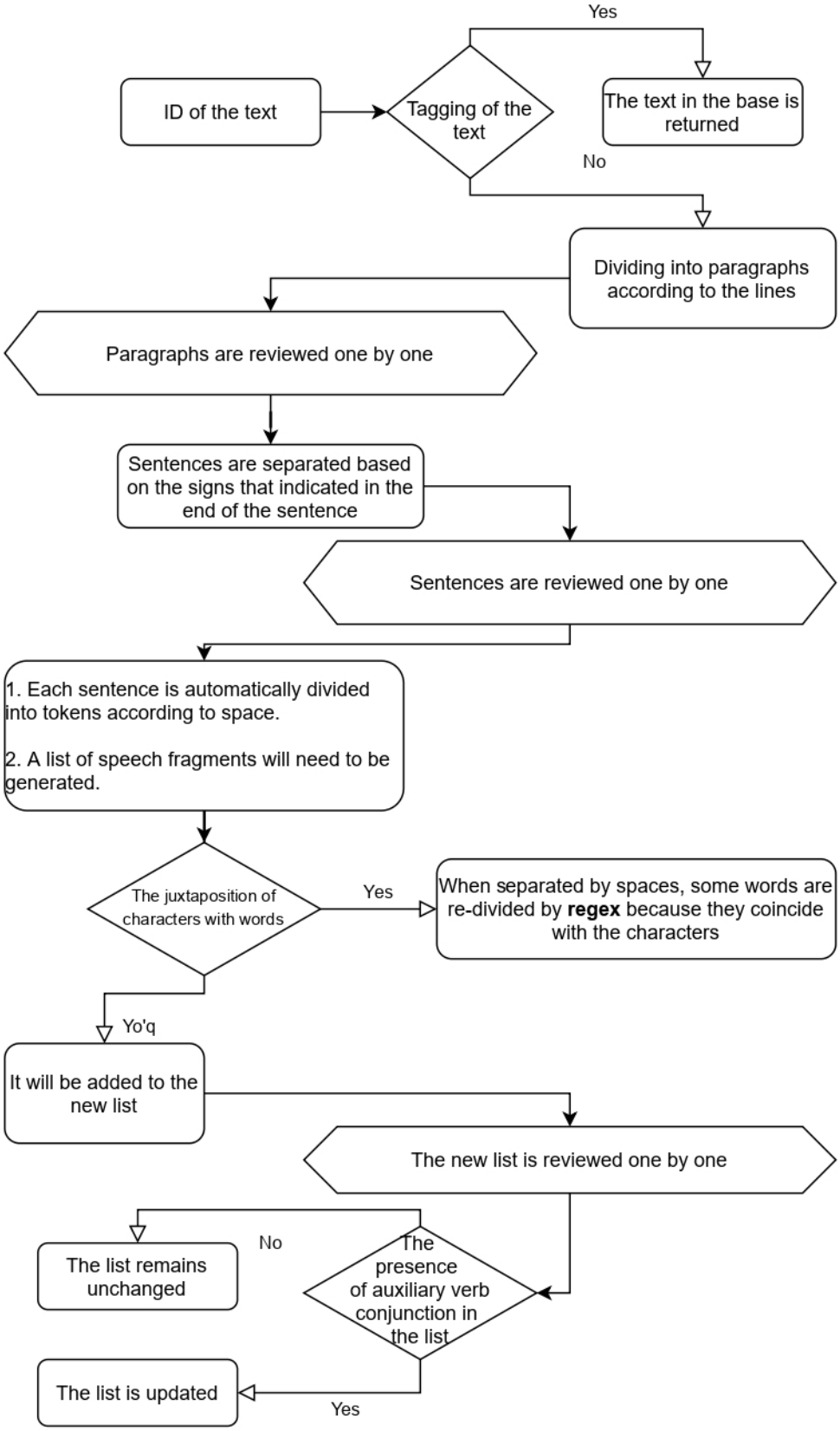


Figure 2.1 Sequence of operations in the corpus

The third part of the chapter is entitled “Peculiarities in the creation of the corpus of electronic information texts in the Uzbek language”, in which the issues

of tokenization, lemmatization and morphological analysis of the corpus structure are solved. In the corpus of electronic information texts in the Uzbek language, the allocation of tokens was carried out automatically, the detection of the lemma of the speech unit was done manually and semi-automatically. For example, the sentence *Yoshlar bandligi masalalari bo'yicha yig'ilishda davlat rahbari yoshlarni qishloq xo'jaligiga jalb qilish masalasiga to'xtaldi* (https://kun.uz/uz/news/2021/04/13/) is divided into automatic tokens in the corpus. There are 15 tokens in this sentence; punctuation in sentences is also taken as a token. Units allocated to automatic tokens in the corpus appear in the form of a list. Morphological analysis begins with the lemma of the form in the corpus, its lexeme is given.

Morphological analysis is performed as follows:

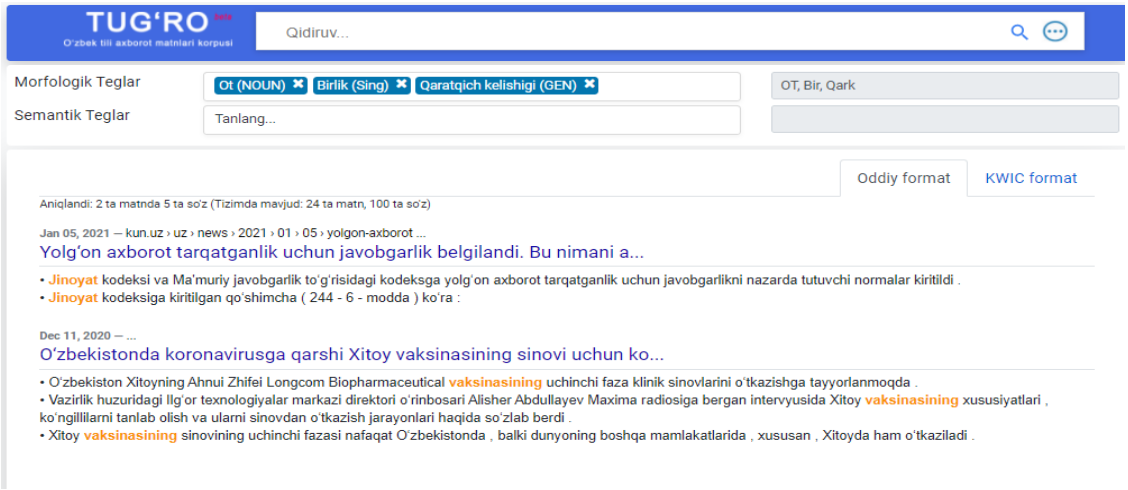
1) yoshlar <yoshlar>[morf. OT+ Ko'p, sem. AOT+JamOT]
2) bandligi <bandlik> [morf. OT+ Bir+3shE+YAS, sem. MavOT]
3) masalalari <masala> [morf. OT+ Ko'p+3shE, sem. TOT]
4) bo'yicha <bo'yicha> [morf. K, sem. BoshqN]
5) yig'ilishda <yig'ilish> [morf. OT+ Bir +Harn+ O'rink, sem. JamOT]
6) davlat <davlat> [morf. OT+ Bir, sem. TashN]
7) rahbari <rahbar> [morf. OT+Bir+3shE, sem. AOT]
8) yoshlarni <yoshlar> [morf. OT+Ko'p+Tushk, sem. AOT+JamOT]
9) qishloq <qishloq> [morf. OT, sem. TOP]
10) xo'jaligiga <xo'jalik> [morf. OT+Bir+3shE+Jo'nk, sem. TashN]
11) jalb
12) qilish
13) jalb qilish <jalb qil> [morf. FL+Qfl+ Harn, sem. Jismoniyf]
14) masalasiga <masala> [morf. OT+Bir+3shE+ Jo'nk, sem. TOT]
15) to'xtaldi <to'xta> [morf. FL+ Majn+ Xab+ O'tz, sem. Nutqf]
16) .<.> [morf. TB, sem. BoshqN]

The third chapter of the dissertation is entitled "Linguistic annotation of the corpus of Internet information texts in the Uzbek language". The first chapter, entitled "Linguistic annotation in language corpus and its principles", identifies the essence, principles, types of linguistic interpretation in the corpus - the reasons for the need for annotation. In addition to the pure text, the corpus provides additional linguistic information called interpretation. Special tags are set for morpho-syntactic, semantic interpretation of language units.

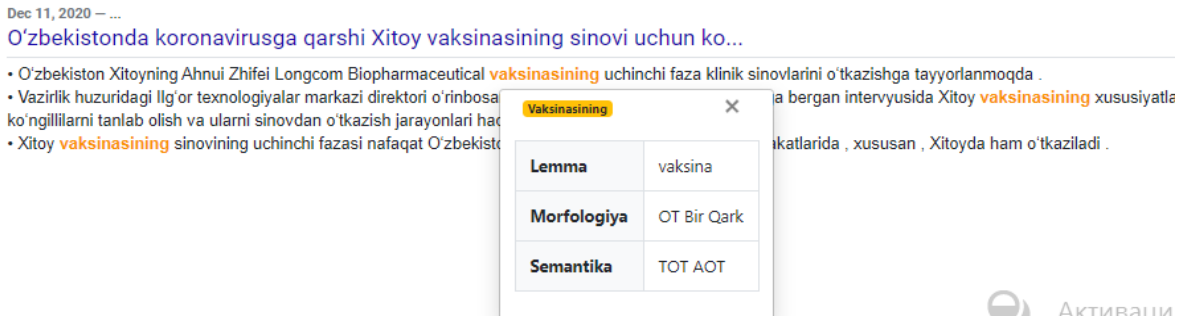
The second chapter, entitled "Morphological annotation of the corpus of Internet information texts in the Uzbek language", shows the algorithm of morphological annotation in the corpus of Internet information texts "Tugro". Morphological interpretation is determined by context. Through the language corpus, all units were tagged and special statistics of compound verbs the auxiliary verb conjugations were provided.

*Shu*_OLKo'r *tariqa*_OT *ayni*_OLKo'r *kungacha*_RAV *mamlakatda*_OTO'rink *koronavirus*_OTQark *infeksiyasiga*_OTJo'nk *qarshi*_K *emlangan*_FLMajnSif *fuqarolar*_OTKo'p *soni*_SON3shE 458 555_RAQSONSanoq *nafarga*_SON *yetdi*_FLAniqnO'tz ._TB

(https://daryo.uz/2021/04/21/). This sentence is only morphologically explained. In the corpus of Uzbek language information texts, when researchers perform an analysis of speech units through the user interface, the context in which the analyzed speech unit is involved appears in the window (under the address https://uzkorpus.uz). For example:



Morphological and semantic indicators of the lexeme, which are drawn to the analysis in the user interface, are reflected in the "popup". For example:



The third part, entitled "Syntactic annotation of the corpus of Internet information texts in the Uzbek language", discusses the importance and necessity of syntactic annotation in corpus. Syntactic interpretation is the process of automatically analyzing texts based on grammatical rules[71]. Technically, this is used to refer to the practice of determining the syntactic structure of a text[72]. This is called parsing or grammatical analysis in corpus linguistics. Syntactic annotation in the corpus of Uzbek language Internet information texts is still in progress.

The fourth part of the chapter is entitled "Semantic annotation of the corpus of Internet information texts in the Uzbek language", in which the semantic features and semantic fields of words are annotated using tags. Through the corpus, new ideas on the topic of homonymy in linguistics can be provided, and a dictionary of homonyms can be prepared. In the Uzbek language corpus, the homonymy in the process of adding words and forms is solved by separating the

[71]Barnbrook G. Language and Computers. – Edinburgh: Edinburgh University Press, 1998. – p. 170.
[72]McEnery T., Wilson A. Corpus Linguistics. – Edinburgh: Edinburgh University Press, 1996. – p. 178.

lemma of these words. For example, when the words "*Talabalar o'qishga kelishdi*" or "*Mahsulot narxini tushirishga kelishdi*" met in context, each word was interpreted with semantic tags. The verb "*kel*" in this context does not pose a problem of homonymy, because the lemma of the first sentence is "*kel*" and the lemma of the unit in the second sentence is "*kelish*". The same problems with the forms of *tortma, eslatma, qo'llanma, yo'qlama,* unit *unga* insentences *unga berdi*, *unga soldi* have been solved. 40 semantic tags have been created in the annotation of the corpus of Uzbek language Internet information texts. The semantic annotation classification was chosen according to the important lexical units that define the basic semantic features.

Table 3.2. Determination of semantic tags

	Abbreviation in Uzbek	**International abbreviation**
Idiom	IB	IDIOM
Even word	JUFT	ECH
Repeat word	TAK	RDP
Proper noun	AOT	PROPN
Personal noun	ANT	PRS
Toponim (name of place)	TOP	TOP
Hydronym (name of water basins)	GID	HYD
Family name	Fam	FamN
Patronymic	OtaIsm	PatrN
Name of organization	TashN	COM
Name of product	MahN	PRO
Name of the geographical area	GeoN	GEO
Name of nationality	MillN	NAT
Name of the animal	HayN	AnimN
Other	BoshqN	OTH
Nominal noun	TOT	ConN
Abstract noun	MavOT	AbstN
Cumulative noun	JamOT	CollN
Concrete noun	AnOT	ConcN
Quantity number	Miq	Qty
Ordinal number	Tar	OrdNum
Fraction number	Kasr	FracNum
Approximate number	Cham	Approx
Cumulative number	Jam	CollNum
Single number	Dona	Integer
Action verbs	Yumushf	Move
Mental verbs	Tafakkurf	Ment
Feeling verbs	Sezgif	Perc

Psychological verbs	Ruhiyf	Psych
Speech verbs	Nutqf	Speech
Gesture verbs	Ishoraf	Semelf
Physical verbs	Jismoniyf	Physiol
Natural Verbs	Tabiiyf	Changest
Visual verbs	Ko‘rishf	Noncaus
Feature adjectives	Xususiyats	Humq
Color adjectives	Tuss	Physq:color
Taste adjectives	Ta’ms	Physq:taste
Dimension adjectives	Hajms	Physq:form
Smell adjectives	Hids	Physq:smell
Adjectives denoting the sign of space-time	Makzams	Place:time

In the corpus, speech units are first morphologically annotated, then semantic annotated. For example, we see that the units of speech in the sentence *Andijondagi Muruvvat uyida 6,5 yoshli siam egizaklari tarbiyalanmoqda* (https://daryo.uz/2020/09/28) are interpreted in a vertical format.

1) *Andijondagi* <Andijon> [morf. OT+Otlug‘sh., sem. TOP]
2) *muruvvat* <muruvvat> [morf. OT+Bir, sem. MavOT+TashN]
3) *uyida* <uy> [morf. OT+3shE+O‘rink, sem. TOT]
4) *6,5* <6.5> [morf. SONSanoq, sem. Miq]
5) *yoshli* <yoshli> [morf. SIF+Yas, sem. Xususiyats]
6) *siam* <siam> [morf. OT+Bir, sem. AOT+ TOP]
7) *egizaklari* <egizak> [morf. OT+Ko‘p+3shE, sem. AOT]
8) *tarbiyalanmoqda* <tarbiyala> [morf. FL+ Majn+ Hozz, sem. Yumushf]
9) .<.> [morf. TB, sem. BoshqN]

As we can see from the above example, in our work, the annotation of speech units is followed by a semantic explanation after morphological explanations. Not all words or the same words always fall into the same semantic field. For example, in the above sentence, the tags of noun and toponyms are marked according to the semantic properties of the word *Siam*. For some speech units, tags are used to indicate multiple lexical meanings at the same time. There are many such units.

CONCLUSION

On the basis of scientific research on the theoretical and practical basis of the formation of the corpus of information texts of the Uzbek language on the Internet, the following conclusions were drawn:

1. A corpus is a collection of natural texts that can be stored electronically and composed of oral or written texts in any language that allows for computerized search.

2. Research and analysis of studies in the field of corpus linguistics has shown that this direction is one of the multidimensional, comprehensive, modern independent directions. The stages of formation and development of corpus linguistics can be classified as follows:

- the period up to the 1950s: the period when there was a paper version of the work, although the field did not exist;
- the period from 1950 to 1980: the period when N. Chomsky's theory prevailed, research on the creation of corpus began, and the critical assessment was high;
- the period from 1990 to 2010: the period of creation of corpus in the world's leading languages, the development and popularization of methods and theories, the emergence of new methods in education, the high level of development of computer technology;
- the period from 2010 to the present: the period of research on the creation of the corpus in all languages, the expansion of universal software, the position of the industry as an independent institution, the expansion of the use of the corpus;
- the fifth stage: the stage of reaching the stage of artificial intelligence.

3. Assessing the role of the Internet as a corpus, the maximum monitor can be equated to a corpus and set to webCorp. Our observations show that the construction of the corpus is orderly and systematic, the size of the Internet is unknown and constantly changing, so the language corpus is not able to give accurate and correct results, to repeat the result. Therefore, it is necessary to build a corpus of electronic information texts in Uzbek language in order to obtain the results of a well-approached and annotated analysis.

4. AntConc, Corpus.byu.edu, Sketch Engine, Web BootCat programs are important in defining the capabilities of the corpus. These managers of the corpus mentioned above are necessary for the construction of the corpus, to have linguistic analytical results and conclusions from the corpus, to use representatives of various industries. AntConc is universal and has no commercial purpose. However, it is not possible to annotate the corpus material due to its inability to analyze the grammatical features of the language. Sketch Engine and Web BootCat corpus managers are software that collects material only for corpus construction: the corpus created by them is no different from a collection of unprocessed text because it is not annotated and is not tagged with special linguistic symbols.

5. Special software has been created for the corpus of information texts of the Uzbek language, automatic tokenization has been carried out as an experiment, and the process of identifying and tagging language units - lemma - has been carried out manually. On this basis, it was theoretically determined that the process of linguistic support in the corpus of Internet information texts of the Uzbek language should be carried out in 6 stages:

- study of Uzbek grammar and data collection;
- development of a schematic diagram of the data, the development of a system of special characters - tags;
- uploading electronic information texts to the corpus manager;
- encoding of downloaded texts, entering metadata;

- annotation of corpus units;
- development of a linguistic search system.

6. One of the important processes in the creation of the corpus of information texts of the Uzbek language is the development of principles of morphological analysis and morphological processing. Based on these principles, morphological and semantic annotations were made at the initial stage of the formation of the corpus of Uzbek language information texts. In the annotation, the selection of characters for morphological and lexical-semantic tags, coding of each character in the corpus was performed. The use of character encoding type in morphological annotation of corpus units gives good results, in the study an algorithm acting as a filter to differentiate polysemy and homonymy was proposed.

7. Based on the international standard tegger programs for annotation in the corpus, 64 morphological tags according to lexical and grammatical features, 40 semantic tags were identified according to the lexical features of Uzbek language units: verb <FL>, noun <OT>, adjective <SIF>, number <SON>, pronoun <OL>, adverb <RAV>, imitation word <Taq>, auxiliary word <K>, conjunction <B>, particle <Y>, exclamation<U>, modal word <M>. In order to explain other grammatical features of language units in morphological annotation tags such as unit abbreviation < Qisq >, artificial word <Yas>, primitive word <TUB>, singular <Bir> or plural <Ko'p> are also developed and systematized. Since there are no standard rules of the Uzbek language in the designation of tagging, universal taggers were analyzed and a system of morphological tags according to Uzbek word groups and their morphological categories and semantic tags according to LMGs of lexical units was determined.

8. Tagging the corpus unit is the process of attaching a special code to a word, indicating the specific features of the language. It turned out that the problems encountered in the corpus of Uzbek language Internet information texts stem from the nature of the language, the interaction of the language with other languages, as well as the method of collecting corpus material. The analysis of the annotated Uzbek language in the corpus of Internet information texts became the basis for making clear and reliable generalizations about the structure and peculiarities of the Uzbek language.

НАУЧНЫЙ СОВЕТ ПО ПРИСУЖДЕНИЮ УЧЕНЫХ СТЕПЕНЕЙ PhD.03/30.12.2019.Fil.60.02 ПРИ АНДИЖАНСКОМ ГОСУДАРСТВЕННОМ УНИВЕРСИТЕТЕ

ТАШКЕНТСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ УЗБЕКСКОГО ЯЗЫКА И ЛИТЕРАТУРЫ ИМЕНИ АЛИШЕРА НАВОИ

АБДУЛЛАЕВА ОКИЛА ХОЛМУМИНОВНА

ТЕОРЕТИЧЕСКОЕ И ПРАКТИЧЕСКИЕ ОСНОВЫ СОЗДАНИЯ КОРПУСА ИНФОРМАЦИОННЫХ ИНТЕРНЕТ ТЕКСТОВ УЗБЕКСКОГО ЯЗЫКА

10.00.11-Теория языка. Прикладная и компьютерная лингвистика

АВТОРЕФЕРАТ ДИССЕРТАЦИИ НА СОИСКАНИЕ СТЕПЕНИ ДОКТОРАФИЛОСОФИИ (PhD) ПО ФИЛОЛОГИЧЕСКИМ НАУКАМ

Андижан – 2022

Тема диссертации доктора философии (PhD) зарегистрирована в Высшей аттестационной комиссии при Кабинете министров Республики Узбекистан B2019.3.PhD/Fil1018.

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

Автореферат диссертации выполнен на трех языках (узбекский, английский, русский (резюме)) размещен на веб-странице по адресу Андижанского государственного университета (www.adu.uz) и на информационно-образовательном портале «ZiyoNet» (www.ziyonet.uz).

Научный руководитель:	**Азимова Ирода Алишеровна** кандидат филологических наук, доцент
Официальные оппоненты:	**Набиева ДилораАбдухамидовна** доктор филологических наук, профессор
	Эшмуминов Аскар Алламуродович доктор философии по филологическим наукам
Ведущая организация:	**Самаркандский государственный университет**

Защита диссертации состоится 2022 года «_____» _____ в _____ часов на заседании научного совета PhD.03/30.12.2019.Fil.60.02 при Андижанском государственном университете (Адрес: 170100, Андижан, улица Университетская, 129-дом. Тел: 0(374) 223 88 14; факс: 0(374) 223 88 30 e-mail: agsu_info@edu.uz).

С докторской диссертации можно ознакомиться в информационно-ресурсном центре Андижанского государственного университета (зарегистрирована за №_____). (Адрес: 170100, Андижан, улица Университетская, 129-дом. Тел: 0(374) 223 88 14; факс: 0(374) 223 88).

Автореферат диссертации разослан: «_____» _____ 2022 года.
(реестр протокола рассылки № ____от «_____»_____ 2022 года).

Ш.Х. Шахобитдинова
Председатель научного совета по присуждению ученых степеней, д.ф.н., профессор

Ф.Ф.Усманов
Ученый секретарь научного совета по присуждению ученых степеней, д.ф.ф.н.

М.Э.Умарходжаев
Председатель научного семинара при научном совете по присуждению ученых степеней, д.ф.н., профессор

ВВЕДЕНИЕ (аннотация к диссертации доктора философии (PhD))

Актуальность и необходимость темы диссертации. В мировой лингвистике за последние полвека появились новые подходы, различные направления в изучении языка, позволяющие делать выводы на основе практического опыта, машинной обработки языка, проведения лингвистических исследований с помощью различных программ. Одним из таких новых направлений является корпусная лингвистика, и в этой области ведутся масштабные научно-теоретические исследования. В частности, серьёзное внимание уделяется созданию национальных корпусов языков, развитию существующих корпусов.

Разработка теоретических и практических основ создания корпусов языков для решения таких вопросов, как более широкое изучение существующих возможностей языка в мировой лингвистике, определение проблемных аспектов грамматики языка в контексте, определение грамматических форм в языке, облегчение работы по созданию многопрофильных электронных словарей, повышение эффективности использования современных информационных технологий в изучении языка, внедрение автоматического перевода, поиска и компьютерного анализа в языке, подготовка электронных учебников и словарей., наличие необходимости построения корпуса языка по специальным областям определяет актуальность нашего исследования.

В последние годы в нашей стране проводятся необходимые реформы по использованию компьютерных технологий с целью повышения эффективности работы во всех сферах, облегчения человеческого труда. В то же время принятые законы и решения по повышению статуса узбекского языка и обеспечению его активного применения ставят перед специалистами ряд важных задач.

Президент Республики Узбекистан Ш. Мирзиёев в своей поздравительной речи по случаю 31-й годовщины присвоения узбекскому языку статуса государственного языка сказал: “Нам предстоит решить важные и неотложные задачи, чтобы узбекский язык занял достойное место на мировой арене, в частности в информационной сети Интернет, создать много новых компьютерных программ на родном языке”. Внимание к узбекскому языку возросло до уровня приоритетных направлений государственной политики. Поддержка научно-исследовательской работы по развитию государственного языка, осуществление международного сотрудничества в этой области доказывает важность социальной значимости, практическую эффективность каждого проведенного исследования на практике. Данное исследование в определенной степени послужит выполнению задач, поставленных в указах Президента Республики Узбекистан, в том числе ПФ-4797 от 13 мая 2016 года “Об организации деятельности Ташкентского государственного университета узбекского языка и литературы имени Алишера Навои”. ПФ-4947 от 7 февраля 2017 года “О Стратегии действий по дальнейшему развитию Республики Узбекистан”,

ПФ-5850 от 21 октября 2019 года “О мерах по кардинальному повышению престижа и статуса узбекского языка как государственного”, №. ПФ-6084 от 20 октября 2020 года “О мерах по дальнейшему развитию узбекского языка и совершенствованию языковой политики в нашей стране”, ПФ-6097 от 29 октября 2020 года “Об утверждении Концепции развития науки до 2030 года”; постановление PQ-4479 от 4 октября 2019 года “О широком праздновании тридцатилетия со дня принятия Закона Республики Узбекистан “О государственном языке”, постановление Кабинета Министров Республики Узбекистан от 12 декабря 2019 года № 984 “Об утверждении Положения о Департаменте развития государственного языка” и другие нормативные акты, связанные с этой деятельностью.

Цель исследования является разработка теоретических и практических основ создания корпуса информационных интернет-текстов в узбекском языкознании, определение конкретных принципов лингвистического аннотирования текстов в корпусе и создание корпуса электронных информационных текстов на узбекском языке.

Объектом исследования были выбраны электронные информационные интернет-тексты на узбекском языке, включая все тексты на сайтах kun.uz, daryo.uz, xabar.uz, sof.uz, gazeta.uz, qalampir.uz и УзА за период 2020-2021 гг.

Научная новизна исследования:

принципы интерпретации речевых единиц доказаны путем выявления лексико-семантических, структурных аспектов языковых единиц в информационных текстах, что позволяет разработать лингвистическое обеспечение платформы корпуса информационных интернет-текстов на узбекском языке;

морфологически аннотированы речевые единицы в корпусе информационных интернет-текстов на узбекском языке, в котором разработана классификационная категория глаголов в текстах, положительно-отрицательная категория, категория стиля действия, переменная категория глагола, числовая категория существительных, сравнительно-классификационная категория прилагательных, лексико-грамматическая группа числа, наречия, имитации, морфологические особенности местоимений и лингвистические возможности вспомогательных слов;

определены лексико-семантические, лексико-тематические группы, лексико-консептуальная область единиц речи в текстах, позволяющие осуществлять семантическое аннотирование языковых единиц в узбекско язычном корпусе информационных интернет-текстов;

проблема идентификации омонимичных слов, аффиксальной омонимии и явлений, связанных с омонимией, в классификации речевых единиц, встречающихся в контексте в аннотации, проблема определения составных глаголов и вспомогательных глагольных союзов подтверждается опытом работы в корпусе.

Внедрение результатов исследований. На основе полученных научных результатов по теоретическим и практическим основам формирования корпуса информационных интернет-текстов узбекского языка:

разработка лингвистического обеспечения платформы корпуса информационных текстов интернета, выводы по использованию лингвистических моделей при создании корпуса языков, разработке стандартных инструментов моделирования и определения лингвистических единиц узбекского языка для корпуса языков были использованы в проекте ERASMUS CLASS “Development of the interdisciplinary master program on Computational Linguistics at Central Asian Universities”, реализуемом в 2017-2020 гг. (Бюллетень ERASMUS+, Национальный офис Erasmus+ - Узбекистан; бюллетень Ташкентского государственного университета узбекского языка и литературы имени Алишера Навои от 6 сентября 2021 года № 01/4-1464.) Силлабусы, выполненные в рамках проекта, а также написанный учебник дополнены моделированием языковых единиц;

Выводы по технологии построения корпуса информационных интернет-текстов узбекского языка, морфологическое и семантическое аннотирование речевых единиц в корпусе были использованы в проекте JHBL-20 “Создание электронного корпуса художественных произведений на тему семьи, соседства и гендерного равенства”, реализуемом НИИ “Махалла и семья” в 2020-2021 гг. (Справка НИИ “Махалла и семья” от 8 сентября 2021 года, № 01 / 09-517). В созданном проекте были использованы силлабусы, реализованные в рамках проекта, и информация о построении корпуса;

разработка принципов морфологического и семантического аннотирования речевых единиц в корпусе информационных интернет-текстов на узбекском языке, лингвистический анализ на платформе корпуса информационных интернет-текстов под названием “Тугро” и полученные практические результаты были использованы в фундаментальном проекте “Теоретические вопросы азербайджано-узбекско-туркменского языкознания” № 1-147, реализуемом в 2016-2020 гг. (Справка Института языкознания имени Имадеддина Насими Национальной Академии Наук Азербайджанской Республики № 230 от 29 октября 2021 г.) В результате проект обогатился новой научно-теоретической информацией;

Общие выводы по аннотированию речевых единиц в информационных интернет-текстах узбекского языка, установлению принципов лингвистического аннотирования, морфологической и семантической интерпретации языковых единиц, лингвистическому анализу на платформе корпус и получению практических результатов были использованы в лекциях и практических занятиях по “Технике научных исследований и издательскому делу (LE0101100)”, преподаваемых в Университете Хаджи Байрам Вали. (Справка Университет Хаджи Байрам Вели, факультет литературы, кафедра современных турецких диалектов и литературы, Анкара от 25 октября 2021 г.). В результате содержание лекций и практических занятий обогатилось информацией о лингвистическом анализе на корпусной платформе;

информация о создании национального корпуса в узбекском языкознании, использование корпуса в преподавании узбекского языка как родного и иностранного были эффективно использованы производстве научно-образовательных программ на национальной телерадиокомпании Узбекистана на телеканале “Узбекистан”, в частности, “Ойдин хаёт лайв”, “Таълим ва тараккиёт”. (Справка Узбекской телерадиокомпании № 02-13-1129 от 5 июля 2021 года).В результате содержание этих телепередач обогатилось научными данными.

Структура и объем диссертации. Диссертация состоит из введения, 3 глав, заключения, списка литературы и приложения, общий объем - 158 страниц.

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

I bo‘lim (I part; I часть)

1. Abdullayeva O.X. Корпус лингвистикаси тилшунослик йўналиши сифатида // Илм сарчашмалари. – Урганч, 2019. № 9–Б. 43-46. (10.00.00 № 3.)
2. Abdullayeva O.X.Programs used to create the language corpus and their principles // ACADEMICIA: An International Multidisciplinary Research Journal. Vol. 10, Issue 6, June 2020. – B. 1778-1783. (Impact Factor: 7.492)
3. Abdullayeva O.X. O‘zbek tili korpusini yaratish bosqichlari va muammoli jihatlari // So‘z san’ati. – Toshkent, 2021. ISSN- 2181-9297, 4-jild, 1-son. – B. 153-158. (Impact Factor: 5.794)
4. Abdullayeva O.X. Korpus lingvistikasida lemmalashtirish, stemming va tokenlashtirish jarayoni // O‘zMU xabarlari. – Toshkent, 2021.№ 1/4/1. – B.-240-243.(10.00.00 № 15.)
5. Abdullayeva O.X. Корпус лингвистикаси алоҳида тадқиқот соҳаси сифатида / “Адабиётшунослик ва таржимашуносликнинг долзарб муаммолари: адабий жараён, қиёсий адабиётшунослик, услубшунослик ва тилшунослик масалалари” республика илмий-амалий анжумани материаллари. – Бухоро, 2019. – Б. 203-206.
6. Abdullayeva O.X. Xorijiy tillarni o‘qitishda korpusning ahamiyati // Global ta’lim va milliy metodika taraqqiyoti mavzusidagi respublika ilmiy-amaliy anjuman materiallari. – Toshkent, 2020. – B. 82-85.
7. Abdullayeva O.X. Korpus lingvistikasi va soha rivojidagi o‘ziga xosliklar // “O‘zbek tilini dunyo miqyosida keng targ‘ib qilish bo‘yicha hamkorlik istiqbollari” mavzusidagi xalqaro ilmiy-amaliy anjuman materiallari. – Toshkent, 19-20-oktabr 2020.
8. Abdullayeva O.X. O‘zbek tili korpusini yaratish bosqichlari va muammoli jihatlari // Foreign language teaching and Applied linguistics mavzusidagi xalqaro ilmiy-amaliy anjuman materiallari. – Samarqand, 21-22-dekabr 2020.

II bo‘lim (II part; II часть)

9. Abdullayeva O.X. Til korpuslarini yaratishda qo‘llaniluvchi dasturlar tasnifi // O‘zbekiston milliy axborot agentligi. – Toshkent, 2020. – B. 194-202.
10. Abdullayeva O.X. Компьютер лингвистикаси учун Ўзбек тили морфемаларининг комбинацион базасини шакллантириш муаммолари / “Ўзбек тили тараққиёти ва халқаро ҳамкорлик масалалари” мавзусидаги халқаро илмий-амалий анжуман материаллари. – Тошкент, 2019. – Б. 434-437.
11. Abdullayeva O.X. O‘zbek tilidagi elektron axborot matnlari korpusini yaratishdagi o‘ziga xosliklar / O‘zbek milliy va ta’limiy korpuslarini yaratishning

nazariy va amaliy asoslari mavzusidagi xalqaro ilmiy-amaliy anjumani materiallari, 7-may 2021-yil. –B.296-299.

Avtoreferatning o‘zbek, ingliz va rus tillaridagi nusxalari
«Oltin bitiklar» jurnali tahririyatida tahrirdan o‘tkazildi.
(02.02.2022-yil)

Bosmaxonaga 2022-yil 05-fevralda berildi. Bosishga
2022-yil 07-fevralda ruxsat etildi. Bichimi 84x108 1/32.
Hajmi 3,5. bosma taboq. Times New Roman garniturasi,
ofset qog‘ozi, ofset usulida chop etildi.
Buyurtma 15 . Adadi 100 dona.

“Step by step print” MChJ bosmaxonasida chop etildi.
Andijon shahar Xrabek ko‘chasi 94-b uy.
O‘zbekiston Respublikasi Prezidenti adminstratsiyasi
huzuridagi Axborot va ommaviy kommunikatsiyalar
agentligining 12.07.2019 dagi 12-3299 raqamli guvohnomasi.

.