BUXORO DAVLAT UNIVERSITETI
HUZURIDAGI ILMIY DARAJALAR BERUVCHI
DSc.03/2025.27.12.Fil.08.07 RAQAMLI ILMIY KENGASH

BUXORO DAVLAT UNIVERSITETI

JUMAYEVA DILNOZA BAXSHULLOYEVNA

TINISH BELGILARNING KORPUSDA BERILISHI VA PUNKTUATSION XATOLARNI ANIQLOVCHI KOMPYUTER DASTURINI YARATISH TEXNOLOGIYASI

10.00.11 – Amaliy va kompyuter lingvistikasi

FILOLOGIYA fanlari doktori (DSc) dissertatsiyasi
AVTOREFERATI

Buxoro-2026

UO‘K: 811.512.133'322:004.43

Filologiya fan doktori (DSc) dissertatsiyasi avtoreferati mundarijasi

Оглавление автореферата доктора наук (DSc) по филологическим наукам

Content of dissertation abstract of Doctor of Science (DSc) on philological sciences

Jumayeva Dilnoza Baxshulloyevna
Tinish belgilarning korpusda berilishi va punktuatsion xatolarni aniqlovchi kompyuter dasturini yaratish texnologiyasi 3

Жумаева Дилноза Бахшуллоевна
Технология создания компьютерной программы для переноса знаков пунктуации в корпусе и выявления пунктуационных ошибок 35

Jumaeva Dilnoza Bakhshulloevna
Technology for developing a computer program for identifying punctuation errors and representing punctuation marks in a corpus 69

E’lon qilingan ishlar ro‘yxati
Список опубликованных работ
List of published works 74

BUXORO DAVLAT UNIVERSITETI
HUZURIDAGI ILMIY DARAJALAR BERUVCHI
DSc.03/2025.27.12.Fil.08.07 RAQAMLI ILMIY KENGASH

BUXORO DAVLAT UNIVERSITETI

JUMAYEVA DILNOZA BAXSHULLOYEVNA

TINISH BELGILARNING KORPUSDA BERILISHI VA PUNKTUATSION XATOLARNI ANIQLOVCHI KOMPYUTER DASTURINI YARATISH TEXNOLOGIYASI

10.00.11 – Amaliy va kompyuter lingvistikasi

FILOLOGIYA fanlari doktori (DSc) dissertatsiyasi
AVTOREFERATI

Buxoro-2026

Fan doktori (DSc) dissertatsiya mavzusi Oliy attestatsiya komissiyasida B2024.4.DSc/Fil905 raqam bilan ro‘yxatga olingan.

Doktorlik dissertatsiyasi Buxoro davlat universitetida bajarilgan.

Dissertatsiya avtoreferati uch tilda (o‘zbek, rus, ingliz (rezyume)) Buxoro davlat universiteti (/www.buxdu.uz/) va “Ziyonet” axborot-ta’lim portalining (www.ziyonet.uz) manzillariga joylashtirilgan.

Ilmiy maslahatchi:	**Toirova Guli Ibragimovna** filologiya fanlari doktori (DSc), professor
Rasmiy opponentlar:	**Gulyamova Shaxnoza Kaxramonovna** filologiya fanlari doktori (DSc), professor
	Elov Botir Boltayevich texnika fanlari doktori (DSc), dotsent
	Xolmuxamedov Baxtiyor Farxodovich filologiya fanlari doktori (DSc), dotsent
Yetakchi tashkilot:	**Berdaq nomidagi Qoraqalpoq davlat universiteti**

Dissertatsiya himoyasi Buxoro davlat universiteti huzuridagi ilmiy darajalar beruvchi DSc.03/2025.27.12 Fil.08.07 raqamli Ilmiy kengashning 2026-yil “5” sentyabr soat 9:00 dagi majlisida bo‘lib o‘tadi. (Manzil: 200118, Buxoro shahri, M.Iqbol ko‘chasi, 11-uy. Tel.: +99865 221-29-14; faks: +99865 221-27-07, e-mail: buxdu_rektor@buxdu.uz).

Dissertatsiya bilan Buxoro davlat universiteti Axborot-resurs markazida tanishish mumkin.

(998 raqami bilan ro‘yxatga olingan). (Manzil: 200118, Buxoro shahri, M.Iqbol ko‘chasi, 11-uy. Tel.: +99865 221-29-14).

Dissertatsiya avtoreferati 2026-yil “20” august kuni tarqatiladi.

(2026-yil “20” avgust dagi 27 - raqamli reestr bayonnomasi).

B.E.Kilichev
Ilmiy darajalar beruvchi ilmiy kengash raisi, filologiya fanlari doktori (DSc), professor

M.Y.Xojiyeva
Ilmiy darajalar beruvchi ilmiy kengash kotibi, filologiya fanlari bo‘yicha falsafa doktori (PhD), dotsent

G.I.Toirova
Ilmiy darajalar beruvchi ilmiy kengash qoshidagi Ilmiy seminar raisi, filologiya fanlari doktori (DSc), professor

KIRISH (fan doktori (DSc) dissertatsiyasi annotatsiyasi)

Dissertatsiya mavzusining dolzarbligi va zarurati. Jahon tilshunosligida kompyuter lingvistikasi va tabiiy tilni qayta ishlash (NLP) texnologiyalari jadal rivojlanib, matnlarni avtomatik tahlil qilish, grammatik hamda punktuatsion xatolarni aniqlash va tuzatishga mo‘ljallangan intellektual tizimlar ta’lim, nashriyot va raqamli axborot tizimlarida keng qo‘llanilmoqda. Shu bilan birga, bunday tizimlarning aksariyati resurslarga boy tillar uchun yaratilgan bo‘lib, o‘zbek tilida punktuatsion xatolarni yuqori aniqlikda aniqlash va avtomatik tuzatish, korpus ma’lumotlariga tayangan intellektual modellarni ishlab chiqish bilan bog‘liq muammolar etarlicha hal etilmagan. Mazkur muammolarni bartaraf etish uchun o‘zbek tili punktuatsiyasining korpus va lingvistik asoslarini takomillashtirish, tinish belgilarini avtomatik aniqlash hamda tiklash algoritmlarini ishlab chiqish, sun’iy intellektga asoslangan punktuatsiya tahlili va avtomatik tuzatish tizimlarini yaratish hamda ularni ta’lim va raqamli axborot tizimlariga joriy etish dolzarb amaliy vazifa hisoblanadi.

Dunyo tilshunosligida kompyuter lingvistikasi va tabiiy tilni qayta ishlash (NLP) yo‘nalishida til korpuslarini avtomatik tahlil qilish, grammatik hamda punktuatsion hodisalarni sun’iy intellekt asosida modellashtirish bo‘yicha ilmiy tadqiqotlar jadal rivojlanmoqda. Bunda punktuatsion birliklarni korpus asosida tavsiflash, avtomatik annotatsiyalash va tahlil qilish ustuvor ilmiy yo‘nalishlardan biri hisoblanadi. Shu bilan birga, o‘zbek tili kabi resurslari cheklangan tillarda punktuatsion birliklarning korpusdagi tavsifi, ularni avtomatik aniqlash va tuzatishning lingvistik hamda algoritmik asoslari yetarli darajada ishlab chiqilmagan. Shu bois o‘zbek tili punktuatsiyasining korpus modelini yaratish, punktuatsion xatolarni avtomatik aniqlash va tiklashning lingvistik hamda algoritmik asoslarini ishlab chiqish dolzarb ilmiy vazifa hisoblanadi.

Mamlakatimizda keyingi yillarda raqamli texnologiyalar, sun’iy intellekt va kompyuter lingvistikasini rivojlantirish, o‘zbek tilining milliy korpusini yaratish hamda tabiiy tilni qayta ishlash bo‘yicha ilmiy-amaliy tadqiqotlar izchil rivojlanmoqda. Xususan, lingvistik korpuslar, til ma’lumotlari bazalari, avtomatlashtirilgan lingvistik tahlil va mashinaviy o‘qitish texnologiyalarini yaratishga qaratilgan tadqiqotlar kengaymoqda. “Bugungi kunda dunyoda sun’iy intellekt iqtisodiyot, ta’lim, sog‘liqni saqlash va xavfsizlik kabi yo‘nalishlarda asosiy tayanch vositaga aylanmoqda”[1]. Mazkur vazifalar ijrosini ta’minlash nuqtayi nazaridan o‘zbek tili punktuatsiyasining korpus asoslarini takomillashtirish, punktuatsion xatolarni sun’iy intellekt asosida avtomatik aniqlash va tuzatish texnologiyalarini ishlab chiqish hamda ularni ta’lim va raqamli axborot tizimlariga joriy etish dolzarb ilmiy-amaliy vazifa hisoblanadi.

O‘zbekiston Respublikasi Prezidentining 2019-yil 21-oktabrdagi PF-5850-son “O‘zbek tilining davlat tili sifatidagi nufuzi va mavqeyini tubdan oshirish chora-tadbirlari to‘g‘risida”, 2017-yil 7-fevraldagi PF-4947-son “O‘zbekiston

[1] O‘zbekiston Respublikasi Prezidentining 2025-yil 22-oktabrdagi PF–189-son “Sun’iy intellekt texnologiyalarini yanada rivojlantirishga oid qo‘shimcha chora-tadbirlar to‘g‘risida”gi Farmoni. Manba: https://lex.uz/uz/docs/-7789403.

Respublikasini yanada rivojlantirish bo‘yicha Harakatlar strategiyasi to‘g‘risida”gi Farmoni, 2017-yil 17-fevraldagi PQ-2789-son “Fanlar akademiyasi faoliyati, ilmiy-tadqiqot ishlarini tashkil etish, boshqarish va moliyalashtirishni yanada takomillashtirish chora-tadbirlari to‘g‘risida”gi Qarori, O‘zbekiston Respublikasi Vazirlar Mahkamasining 2016-yil 13-maydagi VMQ-152-son “Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti faoliyatini tashkil etish to‘g‘risida”gi Qarori hamda mazkur faoliyatga tegishli boshqa me’yoriy-huquqiy hujjatlarda belgilangan vazifalarni amalga oshirishda ushbu dissertatsion ish muayyan darajada xizmat qiladi.

Tadqiqotning respublika fan va texnologiyalari rivojlanishining ustuvor yo‘nalishlariga bog‘liqligi. Mazkur tadqiqot respublika fan va texnologiyalari rivojlanishining I. “Axborotlashgan jamiyat va demokratik davlatni ijtimoiy, huquqiy, iqtisodiy, madaniy, ma’naviy-ma’rifiy rivojlantirishda innovatsion g‘oyalar tizimini shakllantirish va ularni amalga oshirish yo‘llari” ustuvor yo‘nalishiga muvofiq amalga oshirildi.

Dissertatsiya mavzusi bo‘yicha xorijiy ilmiy tadqiqotlar sharhi[2]. Matn korpuslarida tinish belgilarining berilishi, ularni avtomatik tarzda tahlil qilish, hamda punktuatsion xatolarni aniqlovchi dasturlarni ishlab chiqishga oid ilmiy izlanishlar jahonning ilg‘or ilmiy markazlari va oliy ta’lim muassasalarida, jumladan, Massachusetts Institute of Technology, Stanford University, Harvard University (AQSh), Cambridge University, University of Oxford (Buyuk Britaniya), Max Planck Informatics Institute, Humboldt-Universität zu Berlin (Germaniya), Rossiya Fanlar akademiyasi tilshunoslik instituti, M.V.Lomonosov nomidagi Moskva davlat universiteti, Sankt-Peterburg universiteti, Tatariston Kazan federal universiteti (Rossiya), Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti, O‘zR FA Til, adabiyot va folklor instituti, Buxoro davlat universiteti va Farg‘ona davlat universiteti (O‘zbekiston)da olib borilmoqda.

Punktuatsiya va korpus lingvistikasi yo‘nalishida, jumladan, quyidagi ilmiy natijalarga erishilgan: punktuatsion birliklarning sintaktik va pragmatik funksiyalari, ularning matn semantikasi hamda diskursdagi o‘rni aniqlangan (Stanford universiteti, AQSh); punktuatsion birliklarni avtomatik annotatsiyalash, korpuslarda belgilash va tabiiy tilni qayta ishlash algoritmlari ishlab chiqilgan (Oksford universiteti, Buyuk Britaniya); punktuatsion xatolarni aniqlash va tuzatishning statistik hamda neyron modellariga asos solingan (Berlin Gumboldt universiteti, Germaniya); punktuatsiyaning korpus, mashinali tarjima va nutq texnologiyalaridagi qo‘llanilishi nazariy jihatdan asoslab berilgan va modellashtirilgan (Sorbonna universiteti, Fransiya); punktuatsion birliklarni

[2] Dissertatsiya mavzusi bo‘yicha xorijiy ilmiy-tadqiqotlar sharhi quyidagi manbalar asosida amalga oshirildi: https://ijllnet.com/journals/Vol_2_No_1_March_2015/11.pdf; https://www.academypublication.com/issues/past/tpls/vol01/11/04.pdf:https://www.ccsenet.org/journal/index.php/ijel/article/view/0/42036,;https://scholarlypublications.universiteitleiden.nl/;https://ijllnet.thebrpi.org/;www.pushkin.institute.ru,;www.lingvolive.com/ru;https://www.merriam-webster.com/;https://macmillaneducation.my.salesforce-sites.com/help/;https://www.britannica.com/dictionary;https://www.oxfordlearnersdictionaries.com/; www.en.wikipedia.org, www.learnenglishkids. britishcouncil.org; https://www.multitran.com/.

avtomatik tiklash, matnni sintaktik tahlil qilish hamda sun’iy intellekt asosidagi til modellari yaratilgan (Rossiya fanlari akademiyasi, Rossiya; Tokio universiteti, Yaponiya; Pekin universiteti, Xitoy); punktuatsiya tizimining me’yoriy, funksional va kommunikativ xususiyatlari hamda ularni elektron korpuslarda ifodalash masalalari olib borilmoqda (Buxoro davlat universiteti, O‘zbekiston).

Dunyo tilshunosligida hozirda, jumladan, quyidagi ilmiy yo‘nalishlar ustuvor hisoblanmoqda: punktuatsion birliklarning lingvistik va korpus modelini yaratish, ularning sintaktik, semantik va pragmatik xususiyatlarini sun’iy intellekt hamda tabiiy tilni qayta ishlash texnologiyalari asosida tavsiflash, punktuatsion xatolarni avtomatik aniqlash va tiklash algoritmlarini takomillashtirish, elektron korpuslarda punktuatsion annotatsiyalash tamoyillarini ishlab chiqish, shuningdek, matnni avtomatik tahrirlash, mashinali tarjima, nutqni tanish va generativ sun’iy intellekt tizimlari uchun punktuatsiyaning lingvistik modellarini yaratish.

Muammoning o‘rganilganlik darajasi. Jahon lingvistikasi va tabiiy tilni qayta ishlash (NLP) sohasida Tinish belgilarining avtomatik aniqlanishi, ularning korpusdagi realizatsiyasi va punktuatsion xatolarni tahlil qilish masalalari jahon lingvistikasida, ayniqsa, tabiiy tilni qayta ishlash (Natural Language Processing – NLP) sohasida izchil o‘rganib kelinmoqda. Mazkur yo‘nalishdagi izlanishlar XX asrning oxiri va XXI asrning boshlariga to‘g‘ri keladi. Punktuatsiya tizimining morfosintaktik, semantik va pragmatik omillar bilan o‘zaro bog‘liqligi, uning matndagi kommunikativ-funksional roli M.M.Chodorow, D.Klein, C.Manning, D.Jurafsky, M.Steedman, H.Schmid, E.Bick, R.Garside, A.Kilgarriff, T.Kiss, H.Strunk kabi olimlarning tadqiqotlarida yoritilgan. So‘nggi yillarda punktuatsiyani avtomatik tiklash, xatoliklarni aniqlash va tuzatishga oid algoritmlarni ishlab chiqishga e’tibor kuchaymoqda. Jumladan, Vasile Păiş va Dan Tufiș punktuatsiya va bosh harflar tiklashning tarixiy va zamonaviy modellari bo‘yicha keng qamrovli tahliliy tadqiqot taqdim etganlar. Attila Nagy, Bence Bial va Judit Ács BERT asosida chuqur o‘rganishga tayangan holda ingliz hamda venger tillarida yuqori natijalarga (F_1 ~80–82) ega avtomatik punktuatsiya tiklovchi tizim yaratganlar. Shu bilan birga, Junghyun Min va boshqalar punktuatsiya vositasida matn tuzilmasini tushunishni unsupervised yondashuv orqali takomillashtirishga erishgan[3].

[3] Chodorow M., Leacock C., Gamon M., Tetreault J. Automated grammatical error detection for language learners. Synthesis Lectures on Human Language Technologies. 2010. Vol. 3, № 1. – P. 1-134.; Klein D., Manning C. D. Accurate unlexicalized parsing. // Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. 2003. – P. 423-430.; Steedman M. Surface Structure and Interpretation. Cambridge, MA: MIT Press, 1996. 308 p.; Schmid H. Probabilistic Part-of-Speech Tagging Using Decision Trees. // International Conference on New Methods in Language Processing. 1994. – P. 44-49.; Bick E. The Parsing System “Palavras”: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Århus: Aarhus University Press, 2000. 380 p.; Garside R., Leech G., McEnery T. Corpus Annotation: Linguistic Information from Computer Text Corpora. London: Longman, 1997. 280 p.; Kilgarriff A., Rychlý P., Smrž P., Tugwell D. The Sketch Engine. // Proceedings of the Eleventh EURALEX International Congress. – Lorient, France, 2004. – P. 105-116.; Kiss T., Strunk J. Unsupervised Multilingual Sentence Boundary Detection. // Computational Linguistics. 2006. Vol. 32, № 4. – P. 485-525.; Păiș V., Tufiș D. A survey of punctuation restoration methods. // Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). 2021. – P. 1133-1142.; Nagy A., Bial B., Ács J. Automatic punctuation restoration with BERT models. // Proceedings of the Seventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2021). 2021. – P. 121-132.

So‘nggi o‘n yillik ichida rus tilshunosligida tabiiy tilni qayta ishlash sohasi jadal rivojlanmoqda. Ayniqsa, korpus lingvistikasi, avtomatik sintaktik va semantik tahlil, hamda punktuatsiyani to‘g‘ri qo‘llashga oid masalalar o‘z echimini topmoqda. Masalan, A.A.Kibrik va I.M.Boguslavskiylar rus tilida semantik teglash, sintaktik strukturalarni formal tasvirlash, hamda ma’lumotlarni mantiqiy tahlil qilishga oid loyihalarni amalga oshirishgan. Shuningdek, Rus tili milliy korpusi (Национальный корпус русского языка) asosida kompyuter dasturlari va NLP modellari ishlab chiqilmoqda. Punktuatsiya xato va tuzilishini avtomatik aniqlash bo‘yicha ishlar ham faol rivojlanmoqda. Masalan, Yu.D.Aprasov va Ye.I.Yeroxina tomonidan punktuatsiya qoidalarini formallashtirish va ularga asoslangan tahlil algoritmlari yaratilgan. Bugungi kunda rus tili uchun BERT, Deep Pavlov, va Natasha kabi neyron tarmoqlarga asoslangan NLP platformalar keng qo‘llanilmoqda[4].

O‘zbek tilshunosligida NLP sohasi ayniqsa morfologik va korpus asosidagi tahlil bo‘yicha sezilarli yutuqlarga erishmoqda. 2022-2023-yillarda Maqsud Sharipov, Og‘abek Sobirov, Ulug‘bek Salayev hamda Elmurod Kuriyozov va Jamolbek Mattiyev boshchiligidagi mualliflar turli yo‘nalishlarda amaliy algoritmlar ishlab chiqdilar. Xususan, M.Sharipov va O.Sobirov FSM asosida lemmatizatsiya algoritmini taklif qilgan bo‘lsa, U.Salayev infleksional yakunlar orqali yuqori aniqlikdagi (91 %) morfologik analiz modelini tatbiq qildi. Boshqa tomondan, M.Sharipov va jamoasi o‘zbek tilida sintaktik-morfologik tayanch korpus yaratish va POS-teggeri vositasida avtomatik annotatsiya tizimini joriy etdi[5]. Biroq tinish belgilarining elektron korpusda berilishi, ularning statistik va funksional taqsimoti, punktuatsion xatolarning avtomatik aniqlanishi bo‘yicha ilgari olib borilgan tadqiqotlar cheklangan miqyosda amalga oshirilgan bo‘lib, mavjud ishlarda punktuatsiya tizimi asosan, nazariy yoki ta’limiy aspektda tahlil qilingan. Jumladan, aksariyat tadqiqotlarda tinish belgilarining pragmatik qiymati, funksional-stilistik xususiyatlari, shuningdek, kompyuter lingvistikasi bilan integratsiyalashgan holda tahlil etilishi yetarli darajada aks ettirilmagan.

Dissertatsiya ishida esa ilgari e’tibor qaratilmagan masala–o‘zbek tilidagi matnlarda tinish belgilarining elektron korpusda avtomatik tarzda belgilanishi, punktuatsion xatolarni aniqlovchi algoritmlarning yaratilishi, ularning uslubiy,

[4] Кибрик А.А., Богуславский И.М. Формальное описание синтаксических структур и семантическая разметка текста // Вестник Московского университета. Серия 9: Филология. – 2012. – № 6. – С. 3-15.; Бородин А.В. Информационные технологии и корпусная лингвистика: на материале Национального корпуса русского языка // Вестник Санкт-Петербургского университета. Язык и литература. – 2020. – Т. 17, № 1. – С. 55-64.; Апрасов Ю.Д., Ерохина Е.И. Алгоритмизация пунктуационных правил в задачах автоматического анализа текста // Информационные технологии. – 2019. – № 4. – С. 28-35.; Куратов Ю., Архипов М. DeepPavlov: открытая библиотека для обработки естественного языка // Труды конференции по искусственному интеллекту и анализу данных (AINL 2019). – Санкт-Петербург, 2019. – С. 132-136.

[5] Sharipov M., Sobirov O. Development of a rule based lemmatization algorithm through Finite State Machine for Uzbek language. – Arxiv preprint, 2022. arxiv:2210.16006 reddit.comarxiv+6arxiv+6ResearchGate+6; Sharipov M., Kuriyozov E., Yuldashev O., Sobirov O. UzbekTagger: The rule based POS tagger for Uzbek language. – Arxiv preprint, 2023. arxiv:2301.12711 GitHub+3arxiv+3ACL Anthology+3; Sharipov M., Mattiev J., Sobirov J., Baltayev R. Creating a morphological and syntactic tagged corpus for the Uzbek language. – Arxiv preprint, 2022. arxiv:2210.15234 arxiv+1arxiv+1; Salaev U. UzMorphAnalyser: A Morphological Analysis Model for the Uzbek Language Using Inflectional Endings. – Arxiv preprint, 2024.

grammatik va kommunikativ jihatlarini aniqlashga qaratilgan kompleks yondashuv asosida maxsus tadqiqot olib boriladi. Bu jarayonda ilgari o‘rganilgan ishlarga tanqidiy yondashiladi, mavjud bo‘shliqlar ilmiy jihatdan asoslanadi hamda zamonaviy lingvistik texnologiyalar yordami bilan tinish belgilarining korpusda aks ettirilish texnologiyasi ishlab chiqiladi.

Tadqiqotning dissertatsiya bajarilgan oliy ta’lim muassasasining ilmiy-tadqiqot ishlari rejalari bilan bog‘liqligi. Dissertatsiya Buxoro davlat universitetining ilmiy tadqiqot ishlari rejasiga muvofiq “Til nazariyasi va kompyuter lingvistikasining muammolari, yechim hamda istiqbollari” (2022-2026 y.y.) mavzusidagi ilmiy yo‘nalish doirasida bajarilgan.

Tadqiqotning maqsadi o‘zbek tilidagi tinish belgilarining korpusdagi qo‘llanilishi asosida punktuatsion xatolarni aniqlovchi kompyuter dasturini yaratish metodologiyasini ishlab chiqishdan iborat.

Tadqiqotning vazifalari:

punktuatsion xatolarni aniqlovchi kompyuter dasturini yaratish uchun zarur metodlar va algoritmlarni ishlab chiqish, ularning samaradorligini baholash va amaliy tatbiqini ta’minlash;

tinish belgilarni aniqlash va xatolarni tuzatish uchun sun’iy intellekt va mashina o‘rganish yo‘nalishlaridagi yangi yondashuvlarni ilmiy jihatdan izohlash, ularning lingvistik va algoritmik asoslarini aniqlash;

korpus lingvistikasi asosida tinish belgilarning qo‘llanilishini o‘rganish, turli janrlardagi matnlar bo‘yicha qonuniyatlarni aniqlash va statistik hamda funksional xulosalar chiqarish;

tinish belgilarning tilning grammatik va semantik tuzilmasidagi o‘rnini aniqlash va ularning matndagi funksiyalarini ilmiy jihatdan izohlash, matnning kommunikativ va stilistik xususiyatlarini hisobga olish;

tinish belgilarni aniqlash va punktuatsion xatolarni tuzatishning zamonaviy lingvistik va algoritmik yondashuvlari asosida belgilash, ularning integratsiyalashgan tizimini ishlab chiqish va amaliy jihatdan dalillash.

Tadqiqotning obyekti sifatida ToshDO‘TAUda ishlab chiqilgan milliy o‘zbek korpusdan matnlardagi 10 ta tinish belgisining qo‘llanilishi va ular bilan bog‘liq punktuatsion xatolar tanlandi.

Tadqiqotning predmetini tinish belgilarning o‘zbek tilida qo‘llanilishi va ularning punktuatsion xatolarni aniqlashdagi o‘rnini avtomatlashtirilgan dasturlar orqali tahlil qilish tashkil etadi.

Tadqiqotning usullari. Tadqiqotda lingvistik metod, statistik metod, natural tilni tushunish (Natural Language Processing - NLP), grammatik analiz (Grammatical Analysis), avtomatik xatolarni aniqlash (Error Detection) kabi usullaridan foydalanilgan.

Tadqiqotning ilmiy yangiligi quyidagilardan iborat:

tilshunoslikda formalizatsiya tamoyili punktuatsiya tizimiga tatbiq etilib, o‘zbek tilidagi tinish belgilarining grammatik *(uyushiq bo‘laklarni vergul bilan ajratish)*, semantik (*izohlovchi munosabatni ikki nuqta va tire orqali ifodalash*) hamda pragmatik (*his-hayajon va so‘roq ma’nolarini undov va so‘roq belgisi*

orqali ifodalash) funksiyalari funksional-tizimli yondashuv asosida baholanib, ularning matnning strukturaviy va mazmuniy yaxlitligini ta’minlashdagi o‘zaro bog‘liqligini aks ettiruvchi lingvistik-funksional model ishlab chiqilgan;

korpus lingvistikasi metodologiyasi asosida o‘zbek tilidagi tinish belgilarining me’yoriy qo‘llanish qonuniyatlari, janrlar kesimidagi variativligi hamda punktuatsion xatolarni avtomatik aniqlash mezonlari formal qoidalar va kontekstual belgilar asosida modellashtirilib, punktuatsion birliklarni grammatik (vergul, nuqtali vergul, ikki nuqta) hamda kommunikativ-pragmatik (so‘roq, undov, ko‘p nuqta) funksiyalariga ko‘ra tasniflash va POS hamda punktuatsion teglar tizimi asosida avtomatik teglashning lingvistik modeli taklif etilgan;

transformer arxitekturasi hamda zamonaviy NLP va sun’iy intellekt texnologiyalari asosida o‘zbek tiliga moslashtirilgan punktuatsiyani avtomatik tiklash modeli ishlab chiqilib, tinish belgilarini kontekstga bog‘liq holda aniqlash va tiklashning lingvistik hamda algoritmik tamoyillari asoslangan hamda modelning eksperimental sinov natijalari, Accuracy, Precision, Recall va F1-score ko‘rsatkichlariga ko‘ra samaradorligi isbotlangan;

POS-teglash asosida tinish belgilarining so‘z turkumlari ketma-ketligi va sintaktik bog‘lanishlarni aniqlash kabi grammatik hamda predikativ, izohlovchi va sanash munosabatlarini ifodalash kabi semantik xususiyatlari aniqlangan;

o‘zbek tili uchun punktuatsion xatolarni avtomatik aniqlash va tuzatish hamda og‘zaki nutqdan yozma matn hosil qilish jarayonida tinish belgilarini avtomatik tiklash imkonini beruvchi dasturiy ta’minot ishlab chiqilib, *“Bugun havo issiq ertaga yomg‘ir yog‘ishi mumkin”* kabi punktuatsiyasiz matnni “*Bugun havo issiq, ertaga yomg‘ir yog‘ishi mumkin”* ko‘rinishida avtomatik tiklash orqali matnning grammatik to‘g‘riligi va o‘qilish qulayligi ta’minlangan.

Tadqiqotning amaliy natijalari quyidagilardan iborat:

o‘zbek tilidagi punktuatsion xatolarni aniqlash uchun moslashtirilgan algoritmlarning ishlab chiqilishi, bu tilshunoslik sohasidagi tadqiqotlarni rivojlantirishda va avtomatlashtirilgan punktuatsion tahlilning ilmiy asoslarini aniqlashda muhim manba sifatida dalillangan;

o‘zbek tilidagi punktuatsiya masalalari statistik, lingvistik va mashinani o‘rganish yondoshuvlari asosida faktik materiallar orqali tahlil qilinib, ularning samaradorligi va amaliy qo‘llanilishi ilmiy asosda isbotlangan;

punktuatsion xatolarni aniqlash va tuzatish bo‘yicha yaratilgan metodologik yondashuvlar, korpus lingvistikasi va sun’iy intellekt texnologiyalari integratsiyasi orqali o‘zbek tilining tilshunoslik, ta’lim va ilmiy tadqiqotlarda foydalanishga imkon beruvchi asoslangan amaliy vosita ekanligi dalillangan.

Tadqiqot natijalarining ishonchliligi qo‘llanilgan yondashuv, usul va nazariy ma’lumotlarning ishonchli ilmiy manbalardan olingani, keltirilgan tahlillar ratsional-tipologik va kompleks tahlil metodlari vositasida asoslanganligi, nazariy fikr va xulosalarning amaliyotga joriy etilganligi, olingan natijalarning vakolatli tuzilmalar tomonidan tasdiqlanganligi bilan izohlanadi.

Tadqiqot natijalarining ilmiy va amaliy ahamiyati. Tadqiqot natijalarining ilmiy ahamiyati o‘zbek tilida tinish belgilarining grammatik, semantik va pragmatik funksiyalarini aniqlash, ularning korpus lingvistikasidagi ifodasini

lingvistik modellashtirishning ilmiy-nazariy asoslarini ishlab chiqish, punktuatsiya birliklari orqali matn strukturasi, sintaktik bog‘liqlik va intonatsion chegaralarni avtomatik aniqlashning lingvistik hamda algoritmik yondashuvlarini takomillashtirish, o‘zbek tili uchun tabiiy tilni qayta ishlash (NLP) texnologiyalarini rivojlantirish, shuningdek, tadqiqot natijalaridan “Tilshunoslik nazariyasi”, “Kompyuter lingvistikasi”, “Leksikologiya” va “Punktuatsiya” fanlarini o‘qitishda hamda zamonaviy lingvistik korpuslar va intellektual tizimlarni yaratishda nazariy manba sifatida foydalanish imkoniyatlarini kengaytirgani bilan belgilanadi.

Tadqiqot natijalarining amaliy ahamiyati o‘zbek tilidagi matnlarni avtomatik qayta ishlash, punktuatsion xatolarni aniqlash va tuzatishga mo‘ljallangan dasturiy ta’minotni yaratish hamda uni (https://punct-restorer.uznlp.uz/) elektron platformasi orqali amaliyotga joriy etish, mazkur ishlanmalarni ta’lim jarayoni, matnlarni avtomatik tahrirlash, til o‘rgatuvchi dasturlar va onlayn yozishmalar sifatini nazorat qilish tizimlarida qo‘llash, shuningdek, tadqiqot natijalari asosida “Punktuatsion lingvistika” va “Tilshunoslikda sun’iy intellekt” fanlari bo‘yicha seminar, laboratoriya mashg‘ulotlari hamda metodik qo‘llanmalarni ishlab chiqish imkoniyatlarini yaratishi va o‘zbek tilining raqamli axborot-kommunikatsiya muhitidagi qo‘llanilish samaradorligini oshirishga xizmat qilishi bilan belgilanadi.

Tadqiqot natijalarining joriy qilinishi. Tinish belgilarning korpusda ifodalanishi hamda punktuatsion xatolarni aniqlovchi kompyuter dasturini yaratish texnologiyasi bo‘yicha olingan natijalar asosida:

tilshunoslikda formalizatsiya tamoyili punktuatsiya tizimiga tatbiq etilib, o‘zbek tilidagi tinish belgilarining grammatik (uyushiq bo‘laklarni vergul bilan ajratish), semantik (izohlovchi munosabatni ikki nuqta va tire orqali ifodalash) hamda pragmatik (his-hayajon va so‘roq ma’nolarini undov va so‘roq belgisi orqali ifodalash) funksiyalari funksional-tizimli yondashuv asosida baholanib, ularning matnning strukturaviy va mazmuniy yaxlitligini ta’minlashdagi o‘zaro bog‘liqligini aks ettiruvchi lingvistik-funksional modeliga oid nazariy xulosalardan 2022-2024-yillarda amalga oshirilgan REP-25112021/113 raqamli “UzUDT: O‘zbek til tabiiy tilni qayta ishlash uchun universal bog‘liqlik daraxti korpusi va uning semantik tahlili” mavzusidagi ilmiy loyihasida foydalanilgan (Mirzo Ulug‘bek nomidagi O‘zbekiston Milliy universitetining 2025-yil 29-sentyabrdagi 04/11-20706-son ma’lumotnomasi). Natijada o‘zbek tilidagi tinish belgilarining tarixiy-qiyosiy tahlili asosida ularni zamonaviy korpuslar uchun normallashtirish imkoniyati yaratilgan;

korpus lingvistikasi metodologiyasi asosida o‘zbek tilidagi tinish belgilarining me’yoriy qo‘llanish qonuniyatlari, janrlar kesimidagi variativligi hamda punktuatsion xatolarni avtomatik aniqlash mezonlari formal qoidalar va kontekstual belgilar asosida modellashtirilib, punktuatsion birliklarni grammatik (vergul, nuqtali vergul, ikki nuqta) hamda kommunikativ-pragmatik (so‘roq, undov, ko‘p nuqta) funksiyalariga ko‘ra tasniflash va POS hamda punktuatsion teglar tizimi asosida avtomatik teglashning lingvistik modeli taklif etilganligiga oid ilmiy xulosalardan 2024-yilda amalga oshirilgan AL-662204396-raqamli “Buxoro adabiy muhiti ijodkorlarining ilmiy-adabiy merosi bo‘yicha elektron platforma

yaratish” mavzusidagi loyihada foydalanilgan (Buxoro davlat universitetining 2025-yil 10-oktyabrdagi 6202-son ma’lumotnomasi). Natijada til elementlarini formal tasniflashda punktuatsiyani ongli darajada o‘zlashtirish imkonini beruvchi nazariy yondashuv asoslangan;

transformer arxitekturasi hamda zamonaviy NLP va sun’iy intellekt texnologiyalari asosida o‘zbek tiliga moslashtirilgan punktuatsiyani avtomatik tiklash modeli ishlab chiqilib, tinish belgilarini kontekstga bog‘liq holda aniqlash va tiklashning lingvistik hamda algoritmik tamoyillari asoslangan hamda modelning eksperimental sinov natijalari, Accuracy, Precision, Recall va F1-score ko‘rsatkichlariga ko‘ra samaradorligi isbotlanganligiga oid ilmiy natijalar 2020-2023-yillarda bajarilgan AM-Ф3-201908172 raqamli “O‘zbek tilining ta’limiy korpusini yaratish” korpus loyihasi doirasida qo‘llangan (Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universitetining 2025-yil 2-sentyabrdagi 01/4-3886-son ma’lumotnomasi). Natijada tinish belgilarining korpus asosidagi statistik tavsifi hamda ularning kontekstual faollik darajalari lingvistik tahlil asosida aniqlangan;

POS-teglash asosida tinish belgilarining so‘z turkumlari ketma-ketligi va sintaktik bog‘lanishlarni aniqlash kabi grammatik hamda predikativ, izohlovchi va sanash munosabatlarini ifodalash kabi semantik xususiyatlari aniqlanganiga oid ilmiy xulosalar 2023-2024-yillarda axborot texnologiyalari yo‘nalishidagi Navoiy teleradiokompaniyasi “Navoiy” telekanalining “Islohotlar amalda”, “Kelajak yoshlari”, “Inson qadri”, “Diyor 24” ko‘rsatuvlarida foydalanilgan (Navoiy viloyati teleradiokanalining 2025-yil 10-noyabrdagi 108-son ma’lumotnomasi). Natijada sintaktik va morfologik belgiga asoslangan punktuatsion belgilovchi tizim komponentlari mustaqil moduli sifatida amaliyotga tatbiq etilgan;

o‘zbek tili uchun punktuatsion xatolarni avtomatik aniqlash va tuzatish hamda og‘zaki nutqdan yozma matn hosil qilish jarayonida tinish belgilarini avtomatik tiklash imkonini beruvchi dasturiy ta’minot ishlab chiqilganligiga oid ilmiy-amaliy yondashuvlar 70230401 – Kompyuter lingvistikasi ta’lim yo‘nalishi magistrlari uchun yaratilgan “Korpus lingvistikasi” va “Kompyuter lingvistikasi” nomli darsliklarni tayyorlashda qo‘llanilgan (Oliy ta’lim, fan va innovatsiyalar vazirligining 2026-yil 18-fevraldagi 109-U-son buyrug‘i, 109-3 raqamli nashr ruxsatnomasi, 2025-yil 8-apreldagi 220-U-son buyrug‘i, 220-07 raqamli nashr ruxsatnomasi). Natijada darsliklarning ilmiy mazmuni boyitilib, korpus lingvistikasi, tabiiy tilni qayta ishlash hamda punktuatsion tahlil texnologiyalariga oid nazariy va amaliy materiallar takomillashtirilgan.

Tadqiqot natijalarining aprobatsiyasi. Ushbu tadqiqot natijalari bo‘yicha 4 ta xalqaro va 5 ta respublika ilmiy-amaliy anjumanlarida muhokamadan o‘tkazilgan.

Tadqiqot natijalarining e’lon qilinganligi. Dissertatsiya mavzusi bo‘yicha jami 25 ta ilmiy ish chop etilgan, shulardan, 1 ta monografiya, O‘zbekiston Respublikasi Oliy attestatsiya komissiyasi tomonidan doktorlik dissertatsiyalari asosiy natijalarini chop etish tavsiya etilgan ilmiy nashrlarda 20 ta ilmiy maqola (shu jumladan, 10 tasi respublika va 4 tasi xorijiy jurnallarda) nashr etilgan.

Dissertatsiyaning tuzilishi va hajmi. Dissertatsiya kirish, to‘rt bob, xulosa, foydalanilgan adabiyotlar ro‘yxatidan iborat bo‘lib, umumiy hajmi 232 sahifani tashkil etadi.

DISSERTATSIYANING ASOSIY MAZMUNI

Kirish qismida mavzuning dolzarbligi va zarurati asoslangan, tadqiqotning maqsad va vazifalari, obyekti va predmeti tavsiflangan, respublika fan va texnologiyalari rivojlanishining ustuvor yo‘nalishlariga mosligi ko‘rsatilgan, tadqiqotning ilmiy yangiligi va amaliy natijalari bayon qilingan, olingan natijalarning ilmiy va amaliy ahamiyati ochib berilgan, tadqiqot natijalarini amaliyotga joriy etish, nashr etilgan ishlar hamda dissertatsiyaning tuzilishiga doir ma’lumotlar keltirilgan.

Dissertatsiyaning birinchi bobi “**Tinish belgilarining tarixi va ularning dolzarb muammolari**” deb nomlanadi. Ushbu bobning “Jahon tilshunosligida punktuatsiya tizimining nazariy asoslari” nomli birinchi faslida tinish belgilarining nazariy, tarixiy va muammoli jihatlari har tomonlama ilmiy izohlangan. Unda jahon tilshunosligida punktuatsiya tizimining nazariy asoslari, o‘zbek tilshunosligida tinish belgilarining tarixi va uni davrlashtirish, tinish belgilarining dolzarb muammolari tahlil qilinadi. Tinish belgilarining funksional-lisoniy mohiyati, yozma nutqdagi grammatik va stilistik roli, hamda ularning til tizimidagi kommunikativ vazifalari yuzasidan ilgari surilgan qarashlar ilmiy manbalar asosida umumlashtirilgan.

Tinish belgilarining qo‘llanishiga oid qoida va prinsplar tilshunoslikda mustaqil bo‘lim – punktuatsiya (lot. *punctum* – nuqta) doirasida o‘rganiladi. Tarixan bu bo‘lim “interpunktsiya” nomi bilan ham yuritilgan bo‘lib, yozma nutqda tinish belgilarining sintaktik tuzilmani aniqlashtirish, gap va gap bo‘laklarini ajratish, shuningdek, yozma va og‘zaki nutq o‘rtasidagi tafovutni belgilash kabi funksiyalarini nazariy asosda tartibga solgan.

Interpunktsiya tizimining ildizlari qadimgi Rim davriga borib taqaladi, biroq u bilan bog‘liq dastlabki yondashuvlar qadimgi yunon grammatika maktablari faoliyati bilan chambarchas bog‘liq. Mazkur davrda tinish belgilarining vazifasi, asosan, notiq nutqni talaffuzga moslashtirish, gap oxiridagi nuqta va bosh harflarning qo‘llanishiga oid me’yorlarni belgilashdan iborat bo‘lgan.

Yunonlar faqat nuqta belgisining uch xil pozitsiyadagi shaklidan foydalanganlar. Interpunktsiya rivojida Nikanor 8 ta belgidan foydalangan bo‘lsa-da, bu urinishlar sintaktik va mantiqiy tuzilmalarning chalkashligiga sabab bo‘lgan. XV asrgacha tinish belgilariga oid yagona qoidalar mavjud bo‘lmagan. Faqat venetsiyalik nashriyotchi Ald Manutsiy tomonidan qo‘llanilgan belgilarning tizimlashtirilgan me’yori punktuatsiya tarixida burilish yasagan bo‘lib, u interpunktsiya tizimining asoschilaridan biri sifatida e’tirof etiladi. Keyinchalik, Yevropa va rus tilshunosligida nuqtali vergul, ikki nuqta, ko‘p nuqta, so‘roq, undov, tire, qo‘shtirnoq, qavs kabi belgilar punktuatsion tizimga qo‘shilgan.

Bobning **“O‘zbek tilshunosligida tinish belgilarining tarixi va uni davrlashtirish”** deb nomlangan ikkinchi faslida tilda tinish belgilarining

shakllanish jarayoni, ularning dastlabki ko‘rinishlari va ifodalanish usullari xronologik tartibda tahlil qilingan. Xususan, qadimiy yozma manbalarda tinish belgilarining yo‘qligi yoki ularning ramzlar orqali shartli ifodalanishi, shuningdek, keyingi davrlarda arab, fors va rus tilshunosligi ta’sirida o‘zbek tilidagi punktuatsiya tizimidagi o‘zgarishlar va rivojlanish bosqichlari ko‘rsatib o‘tilgan. O‘zbek punktuatsiyasi tarixi turli bosqichlarga bo‘lingan holda shakllangan bo‘lib, bu jarayonda Sh.Shoabdurahmonov va K.Nazarov kabi olimlarning ilmiy tadqiqotlari muhim o‘rin tutadi. I–VIII asrlarda oromiy va runik yozuvlarda tinish belgilarining dastlabki ko‘rinishi intervallar shaklida bo‘lib, ma’no ifodalamasdan, faqat ajratish vazifasini bajargan. IX–XIII asrlarda arab grafikasidagi yozuvlarda belgilar semantik vazifalarni bajara boshlagan, ammo arab yozuvining o‘zbek tiliga mos kelmasligi tufayli tinish belgilarining rivoji cheklangan. XIII asrning ikkinchi yarmidan XX asr oxirigacha bo‘lgan davrda matbaachilik va rus tili ta’siri orqali tinish belgilarining shakllanishi, qoidalarning ishlab chiqilishi va punktuatsiyaning nazariy asoslanishi yuz berdi. XX asrning 90-yillaridan buyon esa o‘zbek tilining davlat tili maqomiga ko‘tarilishi bilan punktuatsiyaga nisbatan talablar kuchaydi, uning ma’rifiy va ijtimoiy ahamiyati yanada oshdi. Saidrizo Alizodaning “Murosalot” asarida tinish belgilarining tasnifi va qo‘llanilishi davriy nuqtayi nazardan tahlil qilingan. A.Fitrat, H.G‘oziyev, G‘.Abdurahmonov, R.A.Alimuhammedov, K.Nazarov, H.G‘afurov, S.Karimov, N.Mahmudov, B.Baxriddinova, F.T.Karimov, O‘.J.Esonov, N.M.Qambarov va D.Turayeva kabi olimlarning asarlarida o‘zbek punktuatsiyasining tarixiy rivoji, nazariy asoslari va zamonaviy qoidalari ilmiy jihatdan tahlil qilingan[6].

O‘zbek tilidagi tinish belgilarining shakllanishi va amaliy qo‘llanishi uzoq tarixiy jarayon natijasida yuzaga kelgan. Ular asosan arab, fors, turkiy va rus yozuv an’analari ta’sirida rivojlangan. Eng qadimiy belgilardan biri nuqta bo‘lib, u dastlab yozuv belgisi sifatida arabcha manbalarda, xususan Abdul Malik xalifaligi (hijriy I asr) davrida qo‘llanilgan. O‘zbek tilida tinish belgisi sifatida XIX asrning ikkinchi yarmida paydo bo‘ldi. A.Fitrat nuqtani “to‘qtash belgisi” deb ta’riflaydi. So‘roq belgisi esa o‘z manbasini lotincha quæstio so‘zidan olib, o‘zbekcha matnlarda 1885-yildan so‘ng uchray boshlagan. Fitrat bu belgini “so‘rash belgisi” deb ta’riflagan. Undov belgisi o‘zbek tiliga rus tili orqali kirib kelib, 1901-yildan boshlab bosma matnlarda muntazam qo‘llana boshlangan. U lotincha Io so‘zidan kelib chiqqan bo‘lib, his-tuyg‘u va undov ifoda etish uchun ishlatiladi.

[6] Фитрат А. Танланган асарлар (дарслик ва ўқув қўлланмалар, илмий мақола ва тадқиқотлар). Нашрга тайёрловчи Ҳ.Болтабоев. – Тошкент: Маънавият, 2009. – 336 б.; Ғозиев Ҳ. Ўзбек пунктуациясининг тарихий тараққиёти (пунктуацион белгилар, уларнинг вазифалари ва қўлланиш асослари). –Тошкент: Фан, 1979. – Б.156-187.; Абдураҳмонов Ф.А. Пунктуация ўргатиш методикаси. – Тошкент: Ўқитувчи, 1968. – 157 б.; Алимуҳаммедов Р.А. Қадимги туркий битикларда пунктуация ва матнни англаш учун ишлатилган ишоратлар ҳақида // Хорижий филология. – 2016. – №1, – Б. 49-53.; Назаров К. Тиниш белгилари ва ёзма нутқ. –Тошкент: Фан, 1974. – 68 б.; Ғафуров Ҳ. Атов гапларда интонация ва пунктуация масалалари. – Тошкент, 1963. – 98 б.; Каримов С.А. Ўзбек тилида тиниш белгилари синонимияси // https://primedu.jspi.uz/index.php/primedu/.; Маҳмудов Н. “Ўзбек тили пунктуациясининг асосий қоидалари”. – Тошкент: Ўзбекистон, 2015. – 88 б.; Бахриддинова Б.М. Замонавий ўзбек пунктуацияси асослари. Монография. – Тошкент: Академнашр, 2015. – 125 б.; Каримов Ф.Т. Пунктуация белгиларининг лингвопсихологик хусусиятлари. Research and education. – П, 2022. №1(3), 2022. Б. 163-167 https:///researchedu.org/index. php/raye/article/viyesh/1179.

Umuman olganda, tinish belgilarining o‘zbek tilida shakllanishi va mustahkamlanishi ma’lum bosqichlarda amalga oshgan bo‘lib, ularning har biri Fitrat, Zohiriy, Cho‘lpon kabi ma’rifatparvarlar tomonidan uslubiy jihatdan asoslab berilgan.

Ushbu bobning uchinchi fasli **“Tinish belgilarini ishlatilishidagi dolzarb muammolar”** deb nomlanib, zamonaviy tilshunoslikda punktuatsiya masalalariga doir bir qator nazariy va amaliy muammolarni qamrab oladi. Bugungi kunda tinish belgilarini qo‘llashdagi noaniqliklar nafaqat o‘zbek tilida, balki ko‘pgina boshqa tillarda ham dolzarb ilmiy masala sifatida kun tartibida turibdi. Ayniqsa, kontekstual omillar ta’sirida bir xil tuzilmali gaplarda turlicha punktuatsiya belgilarining ishlatilishi, yoki aksincha, bir xil belgining turli funksiyalarda namoyon bo‘lishi tilni o‘rganuvchilar uchun ham, tilshunos olimlar uchun ham jiddiy tahlilni talab qiladi. Bu muammolar o‘ziga xos uslubiy, sintaktik va ma’noiy ko‘pma’noliliklar tufayli yuzaga kelmoqda. Shu bilan birga, zamonaviy tilshunoslikda avtomatlashtirilgan matn tahlili jarayonida tinish belgilarini aniqlash, tahlil qilish va to‘g‘ri joylashtirish borasida ham lingvistik va texnik to‘siqlar mavjud. Masalan, mashinali tarjima, avtomatik sintaktik parsing, nutqni yozma shaklga keltirish (speech-to-text) kabi jarayonlarda punktuatsiya qoidalariga amal qilish ko‘p hollarda murakkablik tug‘diradi. Ayniqsa, spontan nutqni yozma shaklga keltirishda, ovoz ta’sirida yuzaga keladigan undov so‘zlar, intonatsion to‘xtashlar, ovoz taqlidlari va his-tuyg‘ularni ifoda etuvchi birliklarni punktuatsiya bilan ifodalash hamon yetarlicha hal etilmagan muammolardan hisoblanadi. Bo‘limda yozma va og‘zaki nutq o‘rtasidagi farqlar, ularda prosodik xususiyatlarining punktuatsiya orqali to‘liq ifodalanmasligi, ayniqsa, nutqning ekspressiv, emotiv yoki funksional uslubiy holatlarida tinish belgilarining chegaralangan qo‘llanishi alohida tahlil qilinadi. Prosodik elementlarning tinish belgilar orqali aks ettirilishida kuzatiladigan me’yoriy va imkoniyat doirasi tadqiq etiladi. Bu esa punktuatsiyaning faqat grammatik emas, balki uslubiy va kommunikativ vazifalarini ham to‘liq qamrab olish zarurligini ko‘rsatadi. Bu bo‘limda punktuatsiyani faqat an’anaviy grammatik me’yorlar bilan cheklash emas, balki uni uslubiy, pragmatik va kognitiv vosita sifatida o‘rganish zarurligi ta’kidlanadi. Ushbu yondashuvdan kelib chiqib, tinish belgilarini o‘rgatish va ularni ta’lim jarayonida samarali tatbiq etish maqsadida interaktiv o‘qitish texnologiyalari, lingvodidaktik modellar, multimedia elementlari, elektron dasturlar va masofaviy ta’lim vositalaridan keng foydalanish imkoniyatlari ko‘rib chiqiladi. Shuningdek, zamonaviy elektron ta’lim platformalarida punktuatsiya mashqlarini avtomatlashtirilgan tekshirish tizimiga joriy etish bo‘yicha ilmiy-amaliy ishlanmalarga ehtiyoj ortib bormoqda.

Bo‘limda shuningdek, spontan nutqning yozma ifodaga ko‘chirilishida yuzaga keladigan punktuatsiyaviy muammolar, undov so‘zlar va ovoz taqlid qiluvchi birliklarning sintaktik vazifalarida tinish belgilarining o‘rni alohida yoritiladi. Bu muammolarni yechish uchun mavjud me’yorlarni qayta ko‘rib chiqish, jumladan, punktuatsiyaning faqat formal strukturalarga bog‘liq emas, balki nutqning kommunikativ maqsadi, gap irodasi, subyektning munosabati kabi omillarga asoslangan tahlili zarur ekaniga e’tibor qaratilgan.

Dissertatsiyaning ikkinchi bobi **“Tilshunoslikda formalizatsiya – nazariy tadqiqot usuli** sifatida” deb nomlanib, to‘rt fasldan iborat. Bobning dastlabki fasli **“Tilshunoslikda formalizatsiya tushunchasi”** deb nomlanadi. Mazkur faslda formalizatsiya tushunchasining mazmuni, uning tilshunoslikdagi tarixiy ildizlari, shakllanish bosqichlari hamda hozirgi zamonaviy lingvistikada tutgan o‘rni tahlil qilinadi. Formal yondashuvlarning semiotik, matematik va informatsion nazariyalar bilan uzviy bog‘liqligi, shuningdek, bu yondashuvlar orqali til birliklarini modellashtirish imkoniyatlari yoritiladi. Ushbu bo‘limda til tizimini aniq, izchil va takrorlanuvchi shaklda ifodalashga xizmat qiluvchi lingvistik modellarning nazariy asosi ochib beriladi. Formalizatsiya masalasi – faqat tilshunoslik falsafasiga oid bo‘lmay, balki umumilmiy metodologik muammo sifatida qaraladi. U zamonaviy lingvistikaning metodologik poydevorini tashkil etib, ayniqsa, matematik mantiq, sun’iy tillar va informatika bilan bevosita bog‘liqdir[7]. Formalizatsiya til jarayonlarini aniq va qat’iy shakllarda modellashtirish, ularni sun’iy belgilar tizimi orqali ifodalash imkonini beradi. Bu uslub lingvistik bilimlarning izchilligini, qayta ishlanishini va avtomatlashtirishni ta’minlaydi.Formal tizimlar – aniq belgilar (alfavit), formulalarni tuzish qoidalari (sintaksis) va mazmun berish tizimi (semantika)ga asoslanadi. Ular yordamida tilning fonologik, grammatik va semantik darajadagi birliklari modellashadi. Shu boisdan, formalizatsiya tabiiy tilning noaniqliklarini kamaytirish, algoritmik tahlil va dasturlash jarayonlarida keng qo‘llaniladigan sun’iy tillarni yaratishda muhim vosita hisoblanadi.

Lingvistik manbalarda formalizatsiya tushunchasi turli yondashuvlar – generativ grammatika, strukturalizm, metatillar, matematik va kompyuter lingvistikasi doirasida keng talqin qilinadi[8]. Bunday yondashuvlar orqali formalizatsiya ilmiy nazariyalarni aniq strukturalarga solish, lingvistik obyektlar o‘rtasidagi bog‘liqliklarni tizimlashtirish imkonini yaratadi. Shu sababli, u zamonaviy tilshunoslikda ilmiy-tadqiqot metodologiyasining ajralmas qismi sifatida qaraladi.

“Formalizatsiya” so‘zi o‘zbek tilida “tizimga solish”, “standartlash” ma’nolariga teng keladi. U – nazariy tushunchalar, nazariy modellar yoki matematik qoidalarni shakllantirish orqali nazariy tizimlarni aniqlashtirish va formalashtirish jarayoni bo‘lib, bu jarayonda ularning ifodalash shakliga katta ahamiyat berilgan bo‘lib, shu bilan birga, ushbu tizimlar qat’iy nazariy talablarga muvofiq ravishda tasniflanadi yoki qayta ishlanadi.Shu bilan birga, matematik logika va kibernetika falsafiy va gnoseologik savolni dolzarbligini belgiladi, formalizatsiyaning turli fanlarda shu jumladan, lingvistikada ham ahamiyatlidir[9].

[7] Ветров А.А. Методологические проблемы современной лингвистики. – Москва, 1973. – С. 4.; Шрейдер Ю.А. О понятии математическая модель языка. – Москва: Знание. 1971. – 40 с.; Большая советская энциклопедия. в 30-ти т.. – 3-е изд. – Москва, 1977. – С. 540-541.; Новейший философский словарь / сост. и гл. науч. ред. А.А. Грицанов. – Минск: Интерпрессервис : Кн. Дом, 2001. – 772 с.

[8] Рузавин Г.И. Логика и аргументация: Учебное пособие. – Москва, 1997. – С. 35-36.; Рузавин Г.И. Математизация научного знания. – Москва, 1984. – 160 с; Лингвистический энциклопедический словарь / Глав. ред. В.НЛрцева. – Москва: Советская энциклопедия, 1990. – С. 98-99.

[9] Бирюков Б.В.Кибернетика и методология науки. – Москва: Нука, 1974. – 170 с.

Yuqorida keltirilgan manbalar va maxsus adabiyotlarga tayanib, “formalizatsiya” tushunchasi ilmda falsafiy, logik, matematik va kibernetik nuqtayi nazardan tadqiq qilingani aniq bo‘ldi. Formalizatsiya, mazmunni aniqlashtirish jarayoni bo‘lib, uning shaklini aniqlash orqali amalga oshiriladi va muayyan jarayonni tasvirlashni o‘z ichiga oladi, bu esa matematik tasvirlashni qurishga asos bo‘ladi. Bu metod lingvistikada ham qo‘llaniladi. Lingvistikalarda formalizatsiya, ya’ni “shakldan chiqadigan, ifodalanish xususiyatlariga asoslangan, shakl sifatida xizmat qiladigan” tushuncha sifatida ta’riflanadi. Tilni formalizatsiyani tasvirlash, qoidalar to‘plami bo‘lib, ularni mexanik ravishda qo‘llash imkonini beradi va lingvistik bilish natijalarini amaliyotda qo‘llashga imkon yaratadi. Shuningdek, formalizatsiya algoritmlash, dasturlash va til materialini kompyuterlashtirish jarayonlariga asos bo‘lib xizmat qiladi.

Bobning keying fasli **“Formalizatsiyaning yangi tendensiyalari”** deb nomlanadi. Unda XXI asr lingvistikasida yuzaga kelgan yangi ilmiy qarashlar va tendensiyalar, xususan, kognitiv lingvistika, korpus lingvistikasi va kompyuter lingvistikasida formalizatsiyaga nisbatan yangicha yondashuvlar tahlil qilinadi. Shuningdek, nutq jarayonining dinamikasini modellashtirish, funksional-semantik strukturalarni formal ramkalarda aks ettirish tendensiyalari tavsiflanadi.

Kompyuter formalizatsiyasi hozirda jadal rivojlanayotgan kompyuter modellashtirishining nazariy asosini tashkil etadi. U matematik modellashtirish va tizimli dasturlash sohasidagi yutuqlarni birlashtirib, kelajakda bilimlarni formallashtirish imkoniyatlarini sezilarli darajada kengaytirishga zamin yaratadi[10] . Matn tahlilida elektron hisoblash mashinalaridan (EHM) foydalanish g‘oyasi vazifalar va materiallarni formalizatsiya qilishga asoslangan matematik va amaliy lingvistika usullarini ishlab chiqishga olib keldi. Hozirgi davrda lingvistik tadqiqotlarni avtomatizatsiya qilish jarayonining eng muhim xususiyatlaridan biri bu tabiiy tilni qayta ishlashning avtomatlashtirilgan tizimlarini yaratish hisoblanadi[11].

Matnni avtomatik tahrirlash jarayoni – avtomatlashtirilgan qayta ishlash turlaridan biri bo‘lib, unda EHM xotirasida saqlanayotgan matnga tegishli o‘zgartirishlar va qo‘shimchalar kiritiladi. Matn tahrirlovchilari esa EHM uchun dasturlar, dasturiy hujjatlar, ilmiy hisobotlar va boshqa ma’lumotlarni tayyorlash va kiritish vositasi sifatida qo‘llaniladi.

Formallashtirilmagan vazifalar, formallashtirilgan vazifalardan farqli ravishda, algoritmik yoki tartibli (aniq) yondashuvni nazarda tutmaydi va quyidagi xususiyatlardan biri yoki bir nechtasi bilan tavsiflanadi:

1. Vazifani yechish algoritmi noma’lum yoki kompyuter resurslarining cheklanganligi sababli qo‘llanilishi mumkin emas.
2. Masala sonli shaklda ifodalanishi mumkin emas.

[10] Юрченко В.В. Функциональный подход к формализации знаний. // Искусственный интеллект и проблемы организации знаний. – Москва, 1991. – С. 1-32.

[11] Асиновский А.С., Кузнецова Е.Ж. и др. К вопросу об автоматизации лингвистических исследований // Вопросы языкознания. 1986. – № 4 . – С. 86.

3. Maqsad va vazifalar aniq belgilangan maqsadli funksiyalar terminlarida bayon etilishi imkoniyatsiz[12].

Formalizatsiya matnni avtomatik qayta ishlash jarayonini ta’minlaydi. Bu jarayonda formalizatsiyaning yangi usuli – kompyuter formalizatsiyasi, kompyuter modellashtirishning asosi sifatida namoyon bo‘ladi. Nashr ishlarida avtomatlashtirilgan tizimlarning mavjudligi shunga bog‘liqki, kompyuterlar berilgan algoritmlarga ko‘ra matn ma’lumotlarini qayta ishlaydi.

Shunday qilib, mazkur punktuatsiya me’yorini avtomatlashtirish uchun til materialini formallashtirish, modellashtirish, matematik asoslash va algoritmlashtirish zarurligi aniqlandi.

“Formalizatsiyaning tilshunoslikdagi ahamiyati va punktuatsion belgilarning tahlili” deb nomlangan faslda formalizatsiya metodining tahliliy afzalliklari til tizimidagi turli qatlamlar – morfologik, sintaktik, semantik va pragmatik strukturalarni o‘rganishda qanday rol o‘ynashi atroflicha yoritiladi. Ayniqsa, punktuatsion belgilar misolida formal tasnif, grammatik struktura bilan uyg‘unlik, kontekstga asoslangan modellashtirish usullari asosida puxta ilmiy tahlil beriladi.

Tilshunoslikda formalizatsiya – nazariy tadqiqot usuli sifatida, tilning o‘zgaruvchan va dinamik xususiyatlarini tushunish hamda talqin qilishda muhim o‘rin tutadi. Bu uslub tilni ma’lum me’yor va qoidalarga muvofiq modellashtirishni nazarda tutadi. Formalizatsiya tilshunoslikda tadqiqotlar yo‘nalishini aniq belgilab, til birliklarini strukturaviy va funksional jihatdan tahlil qilish imkonini yaratadi. Ayni paytda, bu yondashuv tilning ko‘p funksiyali va pragmatik xususiyatlarini aniqlash, ularning modelga asoslangan tahlilini tashkil etish imkonini beradi.

O‘zbek tilshunosligida punktuatsiya masalalariga yetarlicha e’tibor qaratilgan bo‘lsa-da, bu yo‘nalishdagi chuqur va tizimli tadqiqotlarga ehtiyoj hanuz mavjud. Formalizatsiya usuli ayniqsa til strukturasi va ishlatilish qoidalarini aniqlashda muhim rol o‘ynaydi. Tinish belgilarining o‘ziga xosligini aniqlash, ularning qo‘llanish kontekstini aniqlashtirish tilshunoslikning dolzarb tadqiqot yo‘nalishlaridan biridir. Ushbu yondashuv orqali tilning formal va funksional jihatlari aniqlanib, ular umumiy tizim asosida muvofiqlashtiriladi.

Shu sababli, tinish belgilarining tadqiqi tilshunoslikda formalizatsiyaning yangi tendensiyalari bilan chambarchas bog‘liq bo‘lib, ularni izohlash va chuqur tushunish nazariy izlanishlarning ustuvor vazifalaridan biri hisoblanadi. Bu jarayonda tilning grammatik, orfografik, sintaktik va punktuatsion me’yorlari birgalikda ishlab chiqilishi, ularning barcha aspektlari integratsiyalashgan holda o‘rganilishi talab etiladi. Punktuatsiya tizimining yangi tadqiqot yo‘nalishlari esa qoidalarning o‘zaro aloqasi, belgilarning amaliy muammolari va ularni avtomatlashtirish imkoniyatlarini aniqlashga xizmat qiladi.

Tinish belgilarini formal teglar orqali belgilash – nafaqat tahrir jarayonining samaradorligini oshiradi, balki o‘quv jarayonida metodikani takomillashtirish, bilimni chuqurlashtirish hamda tahlilni avtomatlashtirish imkoniyatini yaratadi.

[12] Корнеев В.В., Гареев А.Ф. и др. Базы данных. Интеллектуальная обработка информации. – Москва: Нолидж, 2000, 127 с.; Попов Э.В. Экспертные системы: Решение неформализованных задач в диалоге с ЭВМ. – Москва: Наука, 1987. – С. 8.

Ushbu yondashuv lingvistik bilimlarni amaliy sohalar bilan bog‘lashda ko‘p funksional yechimlar taklif etib, ayniqsa, kompyuter lingvistikasi doirasida ta’lim va tahrir faoliyatining qo‘llanilish imkoniyatlarini kengaytiradi.

Formalizatsiya til birliklarini texnik, hisoblash mumkin bo‘lgan strukturalarga aylantirish jarayoni, ya’ni lingvistik tahlil, tabiiy tilni qayta ishlash (NLP) va korpusli tadqiqotlarda katta imkoniyatlar yaratadi. Biroq, bu jarayon ma’lum xavf-xatarlarni ham o‘z ichiga oladi. Jumladan, tilning tabiiyligi, individual xususiyatlari va uslubiy erkinligining inkor qilinishi ehtimoli mavjud. Zero, har qanday til – bu faqat qoidalarga asoslangan struktura emas, balki u shaxsiy uslub, ijodiy izlanish va kontekstga bog‘liq ma’nolarni ham o‘z ichiga olgan tirik tizimdir.

“Tinish belgilarini modellashtirishda kompyuter lingvistikasi imkoniyatlari” deb nomlangan faslda kompyuter lingvistikasi vositalaridan foydalangan holda punktuatsiyani tahlil qilish va avtomatlashtirish borasidagi metodik yondashuvlar keng yoritiladi.

O‘zbek tilida tinish belgilarning avtomatik aniqlash tizimini yaratishda bir qator vazifalarni aniq belgilab olish zarur. Ushbu vazifalarning har biri lingvistik va texnologiyaviy jihatdan yuqori aniqlikni ta’minlashga qaratilgan bo‘lib, ular quyidagilarni o‘z ichiga oladi:

1. Tinish belgisini aniqlash va turini belgilash.

Avtomatik tizim, avvalambor, matnda tinish belgisining mavjudligini aniqlashi va uning qaysi turga mansubligini (nuqta, vergul, nuqtali vergul, qo‘shnuqta va boshqalar) to‘g‘ri belgilay olishi lozim. Bu jarayon so‘zlar tartibi, sintaktik struktura va kontekstual axborotga asoslanishi bilan murakkab hisoblanadi.

2. Belgilarning nutqiy maqsadga mosligini baholash.

Tinish belgisi faqat grammatik mezonlar asosida emas, balki nutqning maqsadidan kelib chiqqan holda ham qo‘llaniladi. Shuning uchun avtomatik aniqlash tizimi belgilangan tinish belgisining kontekstdagi maqsadga muvofiqligini tahlil qila olishi zarur. Bu, ayniqsa, turli uslubdagi matnlarda tinish belgilarining turlicha qo‘llanilishida muhim ahamiyat kasb etadi.

3. Ixtiyoriy, normal va standart belgilarni farqlash.

Tinish belgilarining ba’zi turlari normal yoki standart qoidalarga asoslangan holda qo‘llanilsa, ayrim hollarda ular ixtiyoriy ravishda, yozuvchining uslubiy tanlovi sifatida ishlatiladi. Shuningdek, amaliy tilda ko‘p uchraydigan, lekin grammatik me’yorlarga to‘liq javob bermaydigan holatlar ham mavjud. Avtomatik tizim ushbu uch xil holatni farqlay olishi va tegishli darajada teglashni amalga oshirishi lozim.

O‘zbek tilida tinish belgilarini avtomatik tarzda aniqlashda qator lingvistik va texnik muammolar mavjud bo‘lib, ular quyidagicha guruhlanadi:

1. Jumla tuzilishida erkinlik va parchalanish.

O‘zbek tilida so‘z tartibining nisbatan erkinligi, jumladagi gap bo‘laklarining joylashishidagi variantlilik avtomatik sintaktik tahlilni qiyinlashtiradi. Shuningdek, ko‘p hollarda mualliflar tomonidan jumla tuzilishida parchalangan konstruksiyalar

(ya’ni, tinish belgisi bilan ajratilmagan qo‘shimcha izohlar, qaytariqlar) ishlatiladi, bu esa tinish belgisining avtomatik tanlanishini noaniqlashtiradi.

2. So‘zlar tarkibidagi nuqtalar va qisqartmalar.

Masalan, “t.” (tomondan), “sh.” (shahar), “d.” (daqiqa) kabi qisqartmalarda nuqta alohida tinish belgisi emas, balki so‘z shaklining bir qismi hisoblanadi. Bunday holatlarni punctuation tagger nuqta sifatida qabul qilib, noaniqlikka olib kelishi mumkin. Ushbu muammoni hal qilish uchun tokenlash jarayonida maxsus qoidalar yoki regex qo‘llash tavsiya etiladi .

3. Korpusdagi adabiy, publitsistik va internet matnlar orasidagi farq. Stil va janrlar orasidagi farq ham tinish belgilarining avtomatik tahlilida jiddiy ta’sir ko‘rsatadi. Masalan, internet matnlarida (forum, ijtimoiy tarmoqlar) tinish belgilarini o‘rinsiz yoki ortiqcha qo‘llash holatlari ko‘p uchraydi. Adabiy matnlarda esa uslubiy maqsad bilan tinish belgilardan cheklangan holda foydalanish mumkin. Bu holatlarni hisobga olmagan avtomatik tizimda turli janrlarga xos xususiyatlar to‘g‘ri aks etmasligi mumkin.

Tinish belgilarni avtomatik aniqlash tizimlarini baholashda, odatda, precision, recall va F1-score kabi standart metrikalar qo‘llaniladi. Ushbu metrikalar avtomatik tizim tomonidan aniqlangan tinish belgilarning to‘g‘riligi va to‘liqligini baholash imkonini beradi:

1. Precision (aniqlik) – avtomatik aniqlangan tinish belgilar orasida nechtasi haqiqatda to‘g‘ri ekanini ko‘rsatadi.

2. Recall (qamrov) – barcha haqiqiy tinish belgilardan nechta avtomatik aniqlanganini o‘lchaydi.

3. F1-score – precision va recall ni uyg‘unlashtiruvchi muvozanatli ko‘rsatkich.

Avtomatik tizimlar uchun teglangan korpuslar nafaqat katta hajmda, balki standartlashtirilgan formatda bo‘lishi lozim. Shu sababli, o‘zbek tili uchun maxsus <punct>, <comma>, <period> kabi taglar bilan ishlovchi XML/TEI formatli teglash standartlarini yaratish dolzarb hisoblanadi.

Dissertatsiyaning uchinchi bobi **“Korpus asosida tinish belgilarini formallashtirish va xatolarni aniqlash texnologiyasi”** deb nomlanib, uch fasldan iborat. Bobning birinchi bo‘limi **“Tinish belgilarining me’yoriy va uzual funksiyalarini korpusda formal tasniflash”** deb nomlanadi. Ushbu bo‘limda punktuatsiya vositalarining sintaktik va grammatik me’yorlariga asoslangan funksional vazifalari hamda ularni avtomatik aniqlash va teglash mexanizmlari ilmiy yondashuv asosida ko‘rib chiqiladi.

Tinish belgilarining grammatik vazifalari ilk navbatda sintaktik tuzilish, intonatsiya, fikrni to‘liq va aniq bayon qilishda namoyon bo‘ladi. Masalan, nuqta, vergul, nuqtali vergul, qo‘shtirnoq, tire, ikki nuqta kabi belgilar muayyan grammatik me’yorlarni ifoda etadi. Bu belgilar nutqni mantiqiy bog‘lash, ajratish, tartiblash, tiniq va shiddatsiz o‘qish imkoniyatini beradi. Tinish belgilarini avtomatik aniqlash va korpusda teglash vazifasi tilning kompyuterlashtirilishi bilan yaqinidan bog‘liq. Bu jarayonda korpus lingvistikasida punktuatsiya teglari

(masalan, <.>, <,>, <:>, <...> kabi) orqali belgilar shakl va funksiya jihatidan tasniflanadi. Avtomatlashtirilgan aniqlashda belgining pozitsiyasi, o‘rtadagi so‘z turkumlari, kontekstual funksiyasi hisobga olinadi. Bu esa me’yoriy punktuatsiyaning algoritmlashtirilishiga imkon yaratadi.

Me’yoriy funksiyalar poetik maqsad uchun ham xizmat qilishi mumkin. Ayniqsa, badiiy matnda nuqta, vergul, qavs, tire, so‘roq va undov belgilarining qo‘llanishi orqali ijodkorning hissiyoti, intonatsiyasi, to‘xtalish va dramatizm kabi estetik funksiyalari yuzaga chiqadi. Masalan:

Seni ko‘zg u qilib quyoshga tutgan
Mening xudoxohish yuragim uchun.
Har bir lahza uchun qon yutayotgan,
“Seni… Seni…” – degan tilagim uchun… (Z.Mirzo) →
<ellipsis_emotive>

Bu joyda ortiqcha hissiyot tufayli qahramon so‘zni ayta olmayapti. Bu lirik hayajonning bevosita grafik obrazidir.

Siz – qimmat bahosiz, Men – arzon, arzon.
Sizga, balki, faqat yuragim arzir. (U.Azim)

Bu misolda tire ega va kesim o‘rtasida qiyoslash va zid qo‘yish funksiyasini bajarmoqda. Logik ta’kid bilan aytilgan bu gaplar korpusda ***<emph_eq_taqqos>*** tegi bilan belgilanishi mumkin.

Kim darg‘azab, Kim hayron,
Chiqib keldilar qator: Dante, Shiller va
Bayron, Firdavsiy, Balzak, Tagor.
(E.Vohidov)

Bu misolda ikki nuqta jumla oxirida kelib, umumlashtiruvchi so‘z bo‘lmagan holda nomlarni ta’kidlash, ta’riflash vazifasini bajaradi. Bunday holatda ***<inverted_colon_list>*** tegi qo‘llanilishi mumkin.

Qavsda berilgan konstruksiyalar, ayniqsa, she’riyatda tasviriylik, simvolika, emotsional qo‘shimcha orqali matnning badiiy kuchini oshirishga xizmat qiladi:

Sevgilim, sevgini sevmagan zotlar,
Qaro aylasalar oq asotirni,
(Darg‘azab pishqirsa quturgan otlar)
Go‘rdan chiqarsalar Qaro botirni… (I.Mirzo)

Bu holatda qavs tasvirni to‘ldirish, xayol va ehtirosni mustahkamlash uchun qo‘llangan bo‘lib, ***<poetic_parenthesis>*** tegi orqali belgilanishi mumkin.

Ikkinchi bo‘lim **“Zamonaviy o‘zbek adabiyotida punktuatsion shablonlarni korpus asosida aniqlash”** deb nomlanib, unda yozma matnlarda turg‘unlashgan, odatiy bo‘lsa-da, me’yoriy chegaradan chekinmagan qo‘llanmalarning identifikatsiyasi, shuningdek, poetik va ekspressiv kontekstlarda qo‘llaniladigan punktuatsion variantlar tahlil qilinadi.

Okkazional funksiyalarni avtomatik aniqlash uchun korpusdan olingan ma’lumotlarning statistik, sintaktik va kontekstual tahlili talab etiladi. Buning uchun korpusga asoslangan aniqlovchi algoritmlar quyidagi bosqichlarda ishlaydi:

1. Normativ funksiyalarni belgilovchi qoidaviy shablonlarni shakllantirish: Masalan, ko‘p nuqta qachon pauza, tugatilmagan fikr yoki ro‘yxat uchun qo‘llaniladi – bu asosiy model qilib belgilanadi.

2. Shablondan chetlanishlarni belgilash: Agar ko‘p nuqta sintaktik yoki standart joyda emas, balki gap o‘rtasida, gap oxirida hissiyot bilan bog‘liq holda kelsa, bu “okkazional” holat sifatida belgilanadi.

3. Intonatsion markerlar bilan bog‘liq NLP teglari: Masalan, intensity_high, interruption, mental_state kabi teglar bilan aloqador punktuatsion elementlarni belgilovchi funksiyalar.

4. N-gram kontekst tahlili: Tinish belgisiga yaqin joylashgan so‘zlar (ayniqsa, hissiyotga oid so‘zlar: “vayy”, “ha”, “o-ox”, “havas”, “afsus” va h.k.) orqali punktuatsiyaning funksiyasi aniqlanadi.

5. Bo‘shliq va ritmni ifodalovchi qurilmalar: Algoritm ko‘p nuqta va tire kabi belgilarni takrorlanish chastotasiga asoslanib ekspressivlik darajasini baholaydi.

Uchinchi fasl **“Aniqlangan funksiyalar va korpus ma’lumotlari asosida punktuatsion xatolarni aniqlovchi dastur yaratish”** deb nomlanadi. Unda avtomatik tahlilga asoslangan funksional model ishlab chiqiladi, me’yoriy, uzual va noodatiy belgilarni xato sifatida tanib olish jarayonidagi mexanizmlar ko‘rsatib o‘tiladi hamda korpus ma’lumotlari asosida yaratilgan dastur algoritmi asoslab beriladi. Shu tarzda, uchinchi bobda nafaqat punktuatsion vositalarning nazariy tasnifi, balki ularni amaliy tizimda avtomatlashtirilgan tarzda qayta ishlash imkoniyatlari ham yoritiladi.

Tinish belgilarini avtomatik tahlil qilishda ularni teglash (tagging) orqali aniqlash muhim bosqich hisoblanadi. Bu jarayonda grammatik va sintaktik ma’lumotlar bilan belgilarni bog‘lash orqali ularning funksiyalarini avtomatik ravishda aniqlash imkoni tug‘iladi. Aksar NLP tizimlari POS-teglar (part-of-speech), sintaktik strukturalar va dependency parsing elementlaridan foydalanadi.

POS-teglash jarayonida tinish belgilarga alohida teg (misol uchun, “PUNCT”) biriktiriladi. Sintaktik tahlilda esa ularning jumladagi pozitsiyasi va bog‘lanish nuqtalari aniqlanadi. Dependency parsing vositalari yordamida tinish belgining qaysi so‘z bilan bog‘liq ekani va qaysi sintaktik munosabatni ifoda etayotgani belgilab beriladi[13].

N-gram modellari so‘zlar yoki belgilarni birikmalarini ehtimoliy hisoblash orqali ma’lum kontekstda qaysi belgi kelishi mumkinligini (yoki kelgan belgining funksiyasi nimaligini) taxminlashda samarali hisoblanadi. Bu modelda nuqta (“.”), undov belgisi (“!”), ko‘p nuqta (“...”) va savol belgisi (“?”) kabi punktuatsiya vositalari bir xil formal tuzilishga ega bo‘lsa-da, ularning kontekstga bog‘liq vazifalari turlicha bo‘lishi mumkin. Masalan:

Nuqta – xabar gapning tugashi, holatni tasvirlash, fikrni yakunlash;

Savol belgisi – axborot so‘rash, ritorik savol, shubha;

Undov belgisi – kuchli hissiyot, hayrat, buyruqning emosional formasi;

Ko‘p nuqta – so‘zning uzilishi, fikrning ochiq qoldirilishi, dramatik pauza.

[13] Chen D., Manning C. D. A Fast and Accurate Dependency Parser using Neural Networks // Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). – Doha: ACL, 2014. – P. 740-750. https://doi.org/10.3115/v1/D14 – 1082.

Amaliy tahlillarda ko‘p hollarda ma’lum belgi oldidan va keyin qanday so‘z yoki so‘z turkumi kelayotganini tahlil qilib, belgilangan funksiyalar aniqlanadi. Masalan, [noun/pronoun + verb + .] modeli xabar gapning tugashini anglatadi, [interjection + !] modeli hissiy reaksiyani, [unfinished phrase + ...] modeli esa so‘z uzilishi va dramatik pauzani bildiradi.

Avtomatik punktuatsiya tahlilini amalga oshirishda Python tilining imkoniyatlari keng hisoblanadi. Unda turli NLP kitobxonalari (spaCy, NLTK), ma’lumotlar tahlili uchun pandas, va tinish belgilarni aniqlash uchun regex (regulyar ifodalar) keng qo‘llaniladi.

1. **if–elif** bloklari orqali belgi funksiyasini baholash. Jumladagi tinish belgisini funksiyasiga qarab baholash uchun shartli operatorlardan foydalaniladi. Masalan:

```
def evaluate_punctuation(punct, context):
    if punct == '?':
        if 'ми' in context or context.strip().endswith('ми'):
            return 'Savol belgisi – to‘g‘ri'
        else:
            return 'Savol belgisi – noto‘g‘ri yoki noadatiy'
    elif punct == '...':
        return 'Ko‘p nuqta – ekspressiv yoki okkazional holat'
    elif punct == ',':
        if 'agar' in context.lower():
            return 'Vergul – shartli bog‘lovchi oldida – to‘g‘ri'
        else:
            return 'Vergul – ehtimoliy xato'
    else:
        return 'Aniqlanmagan belgi'
```

“3.3.1.-rasm”. Algoritmning arxitekturasi.

2. regex orqali konstruksiyalar tanish. Regulyar ifodalar (re moduli) orqali ma’lum tuzilmalarni tanish mumkin. Masalan, so‘roq shakldagi konstruksiyani aniqlash:

```
import re

text = "U keladimi?"
pattern = r'\b\w+ми\?'

if re.search(pattern, text):
    print("Savol konstruktsiyasi to‘g‘ri")
else:
    print("Ehtimoliy xato")
```

“3.3.2.-rasm”. Algoritmning arxitekturasi.

3. pandas yoki csv bilan korpus ma’lumotini taqqoslash. Dastur korpusdagi me’yoriy qo‘llanish bilan kiritilgan matnni taqqoslashi kerak. Masalan:

```
import pandas as pd

# Корпус маълумотини юклаш
df = pd.read_csv("corp_norms.csv")  # устунлар: context, punctuation, status

# Фойдаланувчи матни
user_text = "U ketdi..."

# Таққослаш
matches = df[df['context'].str.contains("ketdi", case=False)]
if not matches.empty:
    print("Ushbu kontekstda '...' ekspressiv deb tan olingan")
else:
    print("Korpusda mos holat topilmadi – ehtimoliy xato")
```

“3.3.3.-rasm”. Algoritmning arxitekturasi.

Dastur natijalarini baholash (Precision, Recall, F1-score)

Tinish belgilarini avtomatik aniqlashda dasturning samaradorligini baholash uchun klassik NLP metrikalar – aniqlik (Precision), qamrov (Recall) va F1-score hisoblanadi.

To‘rtinchi bob **“Til modellari asosida o‘zbek tilida tinish belgilarini tiklashning zamonaviy usullari”** deb nomlanib, to‘rt fasldan tashkil topgan. Bobning birinchi bo‘limi **“Tabiiy til modellari”** deb nomlanadi. Tabiiy tilni avtomatik qayta ishlash (Natural Language Processing, NLP) jarayonida til modellari muhim o‘rin tutadi. Til modellari belgilar yoki so‘zlar ketma-ketligiga ehtimollik belgilash orqali matnni tahlil qilish va yangi matn hosil qilish imkoniyatini beradi. So‘nggi yillarda statistik yondashuvlar bilan bir qatorda sun’iy neyron tarmoqlar asosidagi modellarning rivojlanishi turli NLP vazifalarining samaradorligini oshirdi. Ushbu maqolada statistik til modellari (N-gram, Markov modeli)ning nazariy asoslari, ularning baholash mezonlari hamda o‘zbek tili milliy korpusi asosida UzbNlp paketidan foydalangan holda amaliy tadqiqot natijalari tahlil qilinadi.

Til modeli (Language Model, LM) matnning ehtimollik tuzilishini tavsiflaydi. Til modellarini ikki asosiy turga bo‘lish mumkin:

1. Statistik til modellari (Statistical Language Models, SLM) – N-gram modellariga asoslanadi.

2. Neyron til modellari (Neural Language Models, NLM) – chuqur o‘rganish metodlari yordamida yaratiladi.

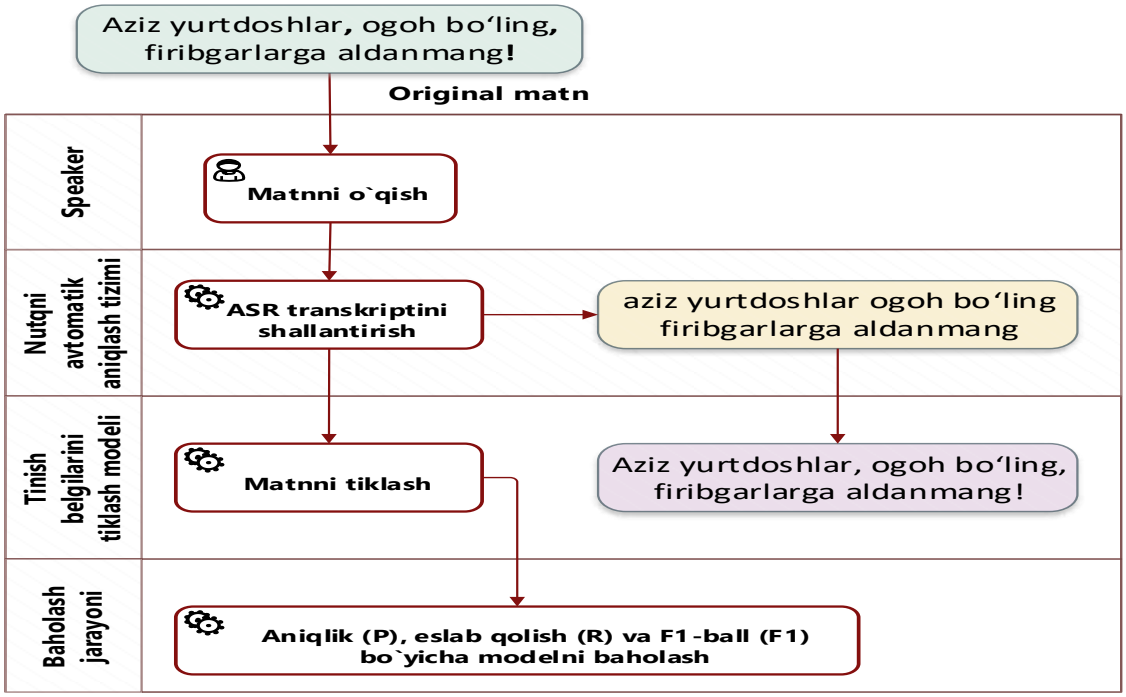
N-gram modellari ketma-ketlikdagi keyingi so‘zning ehtimolini oldingi so‘z(lar) asosida hisoblaydi. Markov modeli esa faqat ma’lum uzunlikdagi kontekstni hisobga oladi (bigram, trigram va h.k.).

Til modellari samaradorligi ichki va tashqi baholash usullari yordamida aniqlanadi. Ichki baholashda chalkashlik (perplexity) asosiy mezon hisoblanadi. Masalan, o‘zbek tili korpusida bigram modelining perplexity ko‘rsatkichi unigram modeliga nisbatan ancha past bo‘lib, modelning aniqligi yuqoriligi qayd etiladi.

Til modellaridagi muammolar: Umumlashtirish va sezgirlik – model o‘quv korpusiga qattiq bog‘liq; Noma’lum so‘zlar (OOV) – korpusda uchramagan birliklar; Kontekstga sezgirlik – silliqlash va diskontlash texnikalarini talab etadi.

Bobning ikkinchi fasli **“Og‘zaki nutqni tanib olish tizimida matndagi tinish belgilarini tiklash usullari”** deb nomlanadi. Avtomatik nutqni tanish (Automatic Speech Recognition – ASR) tizimlari so‘nggi yillarda turli sohalarda keng qo‘llanila boshlandi. Xususan, matnni avtomatik transkripsiya qilish, multimodal tahlil, subtitr tayyorlash, nutq sintezi va interaktiv ilovalarda ASR muhim ahamiyat kasb etmoqda. Biroq ASR tizimlari orqali olingan transkripsiyalar grammatik jihatdan to‘liq emas: ularda katta-kichik harflar saqlanmaydi, xatboshi tuzilishi yo‘qoladi va eng asosiysi, tinish belgilarining yo‘qligi matnning o‘qilishini va ma’noga aniqlik kiritishni qiyinlashtiradi.

Matnda tinish belgilarini joylashtirish NLP vazifasining maqsadi og‘zaki nutqni tanib olish (ASR)da tinish belgilarini tiklashdan iborat(“4.1-rasm”ga qarang).



“4.1-rasm”. Matnda tinish belgilarini joylashtirish NLP vazifasining arxitekturasi.

Ko‘rinib turibdiki, tinish belgilarini tiklash nafaqat matnning o‘qilishi va mazmunini yaxshilaydi, balki keyingi NLP vazifalarining (sintaktik tahlil, mashina tarjimasi, savol-javob tizimlari va h.k.) samaradorligiga ham ijobiy ta’sir ko‘rsatadi.

UzbPunct – ToshDO‘TAU kompyuter lingvistikasi va raqamli texnologiyalar kafedrasi xodimlari tomonidan o‘zbek tili korpusi matnlarining *kraudsorsingli*

matn va audio ma'lumotlar to'plami (crowdsourced text and audio data set). Ma'lumotlar to'plami ikki qismga bo'lingan: **suhbat (UzbTalks)** va **axborot (UzbNews)**. Audio komponentni ishlab chiqarishda yuzdan ortiq xodimlar, ilmiy tadqiqotchilar, magistrantlar va talabalar ishtirok etdi. Audio ma'lumotlarning umumiy uzunligi (hajmi) test to'plamini o'z ichiga olgan holda deyarli o'ttiz olti soatga yetadi. Audio ma'lumotlarni shakllantirishda erkak va ayol nisbatini muvozanatlash uchun choralar ko'rildi. UzbPunct 32000 dan ortiq matn va 1200 ta audio faylga ega bo`lib, ulardan 1000 tasi o'quv to'plamida va 200 tasi test to'plamida joylashtirilgan. Har bir matn uchun *avtomatik tanib olingan og'zaki nutq* va *qo'lda tekislangan matnning* **transkripti** mavjud.

UzbPunct ma'lumotlar to'plamining statistikasi quyida keltirilgan:

– **Matn:** o'zbek tili korpusidan olingan 32000 ta matn;

– **Audio:**

▪*Tanlash tartibi:*

▪tasodifiy tanlangan UzbNews (80%, ya'ni o'quv to'plami uchun 800 ta yozuv) so'zlar soni [150..300] oralig'ida;

▪tasodifiy tanlangan UzbTalks (20%), so'zlar soni [150..300] oralig'ida va kamida bitta so'roq belgisidan iborat gap;

▪*Ma'lumotlar to'plamining bo'linishi:*

▪O'quv ma'lumotlari (Training data): 1000 ta yozuv;

▪Test ma'lumotlari (Test data): 274 ta yozuv;

▪*Spikerlar:*

▪Erkak: 51 ma'ruzachi, 16,7 soat nutq;

▪Ayol: 54 ma'ruzachi, 19 soat nutq;

Strukturlanmagan (qayta ishlanmagan) matn uchun tinish belgilari statistikasi ("4.1-jadval"ga qarang) keltirilgan.

"4.1-jadval".

Strukturlanmagan matn uchun tinish belgilari statistikasi.

Tinish belgisi	yozilishi	Mean	Median	max	Sum	included
fullstop	**.**	12.44	7.0	1129.0	404378	Yes
comma	**,**	10.97	5.0	1283.0	356678	Yes
question mark	**?**	0.83	0.0	130.0	26879	Yes
exclamation mark	**!**	0.22	0.0	55.0	7164	Yes
hyphen	-	2.64	1.0	363.0	81190	Yes
colon	**:**	1.49	0.0	202.0	44995	Yes
ellipsis	**...**	0.27	0.0	60.0	8882	Yes
semicolon	**;**	0.13	0.0	51.0	4270	No
quote	**"**	3.64	0.0	346.0	116874	No
Jami so'zlar soni		**169.50**	**89.0**	**17252.0**	**5452032**	-

Yuqorida keltirilganidek, ma'lumotlar to'plami ikki qismdan iborat: **suhbat (UzbTalks)** va **axborot (UzbNews)**.

Og‘zaki nutqni tanib olish vazifasini bajarish davomida (“4.2-jadval”ga qarang) keltirilgan tinish belgilari baholanadi:

“4.2-jadval”.

Tinish belgilari va ularning belgilanishi.

№	Tinish belgisi	Punctuation mark	Symbol
1.	nuqta	fullstop	**.**
2.	vergul	comma	**,**
3.	so‘roq belgisi	question mark	**?**
4.	undov belgisi	exclamation mark	**!**
5.	tire	hyphen	**–**
6.	ikki nuqta	colon	**:**
7.	uch nuqta	ellipsis	**...**
8.	nuqtali vergul	semi-colon	**;**
9.	bo‘sh joy	blank (no punctuation)	

Baholash kerak bo‘lgan natija faqat tinish belgilari qo‘shilgan matndir.

Yakuniy natijalar har bir tinish belgisini alohida bashorat qilish uchun aniqlik (precision), eslab qolish (recall) va F1 ballari (F1 scores) baholash ko‘rstakichlari vositasida baholanadi. Taqdim etilgan qiymatlar har bir tinish belgisi uchun F1 ballarining o‘rtacha og‘irligiga nisbatan taqqoslanadi.

Hujjat bo‘yicha ball:

$$\textbf{Aniqlik (precision)} = \frac{TP}{TP + FP};$$

$$\textbf{Eslab qolish (recall)} = \frac{TP}{TP + FN};$$

$$\textbf{F1} = 2 * \frac{\textbf{Precision} * \textbf{Recall}}{\textbf{Precision} + \textbf{Recall}};$$

Har bir tinish belgisi uchun global (p) ball:

$$\textbf{Precision}_P = \textbf{avg}_{micro}\textbf{Precision}_P = \frac{\sum_{d \in Documents} TP}{\sum_{d \in Documents} TP + FP};$$

$$\textbf{Recall}_P = \textbf{avg}_{micro}\textbf{Recall}_P = \frac{\sum_{d \in Documents} TP}{\sum_{d \in Documents} TP + FN};$$

Modelning yakuniy ball ko‘rsatkichi global ballarning o‘rtacha og‘irligi sifatida hisoblanadi.

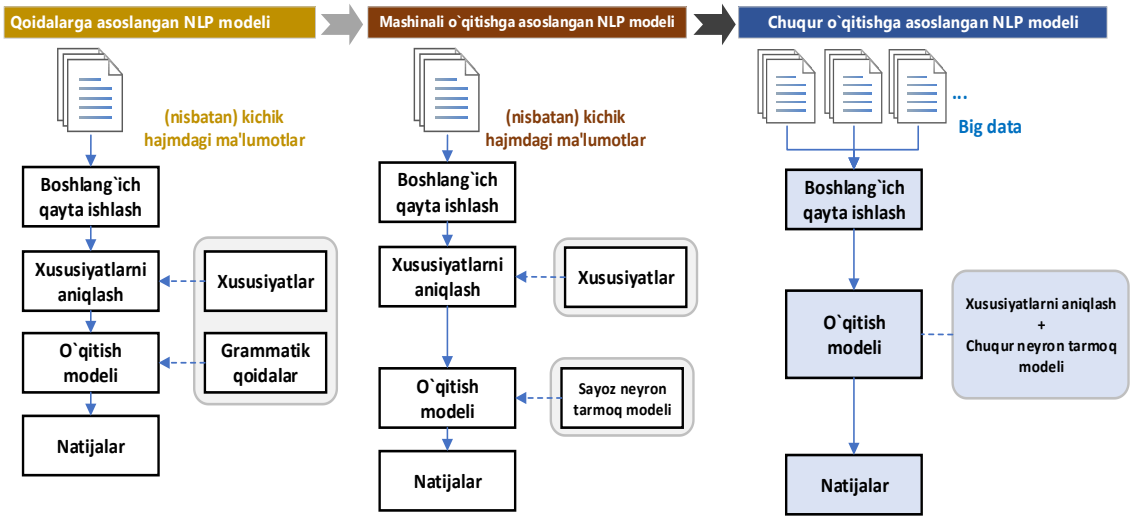
$$\frac{1}{N}\sum_{p \in Punctuation} support(p) * \textbf{avg}_{micro}\textbf{F}_1(p)$$

Bobning uchinchi fasli **“Transformer modellari yordamida o‘zbek tili matnida tinish belgilarini tiklash”**deb nomlanadi. Unda avtomatik nutqni tanish tizimlarida (ASR) o‘zbek tili matnlarida tinish belgilarini tiklash masalasi haqida so‘z yuritiladi. Unda tinish belgilarining matn o‘qilishi va ma’nosini aniqlashdagi ahamiyati ko‘rsatiladi. Shuningdek, ingliz tilida qo‘llanilgan transformer modellarining (BERT, RoBERTa va boshqalar) tajribasi tahlil qilinadi va o‘zbek tili uchun ham maxsus ma’lumotlar to‘plami yaratilgani, UzbBERT modeli asosida

dastlabki muvaffaqiyatli natijalar olingani aytiladi. Ya’ni bu faslda o‘zbek tili uchun tinish belgilarini avtomatik tiklash muammosi, uning yechimlari va dastlabki ilmiy natijalari haqida so‘z yuritiladi.

O‘zbek tili kabi kam resursli tillardagi matnlarda tinish belgilarini tiklashning yangi modellari ToshDO‘TAU kompyuter lingvistikasi va raqamli texnologiyalar kafedrasi olimlari tomonidan ishlab chiqildi. Ular tomonidan ishlab chiqilgan UzbCorpus deb nomlanuvchi til korpusi resurslari asosida BERT modelining turli variantlarida o‘qitildi va model baholash matnlarida tekshirildi. Ushbu paragrafda transformerga asoslangan turli modellar tadqiq qilinadi va ko‘p resursli manba (ingliz) va kam resursli (o‘zbek) tillaridagi matnlarda ushbu vazifani hal qilish usullari keltiriladi. Turli xil BERT modellari natijalari aniqlik (precision, P), eslab qolish (recall, R) va F1-ball (F1) baholar bo‘yicha tahlil qilindi va o‘zbek tili matnlarda tinish belgilarini tiklash bo‘yicha F1-ball bo‘yicha 86.7% qiymatga erishildi. O‘zbek tili uchun ushbu tadqiqot kelajakdagi ishlar uchun asos bo‘lib xizmat qilishi mumkin.

ML yordamida tinish belgilarini tiklash berilgan matndagi to‘g‘ri tinish belgilarini bashorat qilish uchun algoritmlar va statistik modellardan foydalanishni o‘z ichiga oladi. Jarayon odatda uchta asosiy bosqichni o‘z ichiga oladi: **ma’lumotlarni tayyorlash**, **xususiyatlarni ajratib olish** va **modelni o‘qitish**. Quyidagi “4.2-rasm”da tinish belgilarini tiklash modellari keltirilgan:



“4.2-rasm”. Tinish belgilarini tiklash modellari.

Nutqni avtomatik aniqlash (Automatic Speech Recognition, ASR) kabi NLP muammolarida yaratilgan matn odatda tinish belgilarisiz bo‘lib, uni **tanib olish** yoki **tahlil qilish** murakkab jarayon. Kiritilgan matn har bir so‘zning xususiyatlari uchun kontekst muhim bo‘lgan ketma-ketlik sifatida ko‘rib chiqilishi mumkinligi sababli, chuqur neyron tarmoqlarning xarakteristikalari yaxshi usul hisoblanadi.

Transformer modellari yordamida o‘zbek tili matnida tinish belgilarini tiklash maqsadida *o‘quv* (*4000*), *baholash (500)* va *test (200)* to‘plamlarini tayyorlash maqolalar tanlab olindi. Ushbu ma’lumotlar to‘plamlaridagi so‘zlar (tokenlar) soni mos ravishda 1380000, 180000 va 88000 ta so‘zlardan iborat. Shuningdek, modelni baholash uchun *qo‘lda teglangan* va *ASR transkripsiyalaridan* iborat ikkita test ma’lumotlar to‘plami tayyorlanadi. Buning uchun to‘rtta o‘zbekcha

hikoyalardan (ya’ni, monolog o‘qilgan nutq, monologue read speech) olingan 65 daqiqalik nutq parchalarini to‘planadi. Ulardagi tinish belgilar qo‘lda transkripsiya qilinadi. **Google Cloud speech API** ilovasi yordamida audiolar uchun ASR transkripsiyalarini shakllantiriladi. Google Cloud speech API ilovasida o‘zbek tili uchun tinish belgilari bermaydi. Shu sababli, Google Cloud speech API vositasida olingan transkripsiyalarni qo‘lda teglash lozim. Hosil qilingan ASR transkripsiyalardagi so‘zlarning xatolik darajasini (Word Error Rate, WER) qo‘lda yozilgan transkripsiyalar bilan solishtirish orqali hisoblaganda **14.8% WER** natijasni berdi. Qo‘lda teglangan va ASR transkripsiyalaridagi so‘zlar soni mos ravishda *6821* va *6417* so‘zdan iborat. Ushbu faslda o‘zbek tilidagi quyidagi tinish belgilari ko‘rib chiqiladi: **nuqta, vergul, so‘roq** va **Ø**.

Quyidagi “4.3-jadval”da ingliz va o‘zbek tillari uchun teglarning taqsimlanishini taqdim etilgan. Ushbu jadvalda qavs ichida tinish belgilarining foizini keltirilgan.

“4.3-jadval”.

Ingliz va o‘zbek tilidagi ma’lumotlar to‘plamining taqsimoti.

Dataset	Jami	Nutqa(.)	Vergul(,)	So‘roq(?)	Boshqa(Ø)
Ingliz tili					
O‘quv	2102417	132393(6.3%)	158392(7.53)	9905(0.47%)	1801727(85.7%)
Baholash	295800	18910(6.39%)	22451(7.59%)	1517(0.51%)	252922(85.5%)
Test (Ref.)	12626	807(6.39%)	830(6.57%)	46(0.36%)	10943(86.67%)
Test (ASR)	12822	809(6.31%)	798(6.22%)	35(0.27%)	11180(87.19%)
O‘zbek tili					
O‘quv	1379986	98791(7.16)	65235(4.73)	4555(0.33%)	1211405(87.7%)
Baholash	179371	13161(7.34)	7544(4.21%)	534(0.3%)	158132(88.16%)
Test (news)	87721	6263(7.14%)	4102(4.68%)	305(0.35%)	77051(87.84%)
Test (Ref.)	6821	996(17.6%)	279(4.09%)	170(2.49%)	5376(78.82%)
Test (ASR)	6417	887(13.82%)	253(3.94%)	125(1.95%)	5152(80.29%)

Yuqoridagi ma’lumotlar bazasida so‘roq gaplarning taqsimlanishi juda past ekanligini (1% dan kam), datasetning asosan yangiliklar korpusidan olinganligida hisoblanadi. Ammo, bu ko‘rsatkichlar o‘zbek va ASR transkripsiyalarida ancha yuqori qiymatga ega. Buning sababi shundaki, bu matnlar odamlarning o‘zaro suhbatga kirishadigan va bir-biriga savollar beradigan qisqa hikoyalardan tanlangan. Hikoyalarning adabiy uslubi yangiliklardan farq qilishi sababli, o‘zbek va ASR transkripsiyalarida nuqta belgisining taqsimoti ham yuqoriroqdir. Bu esa o‘z navbatida (“4.4-jadval”ga qarang) ma’lumotlar to‘plamidagi o‘rtacha gaplarning uzunligining ancha kichikroq bo‘lishiga olib keladi.

“4.4-jadval”.

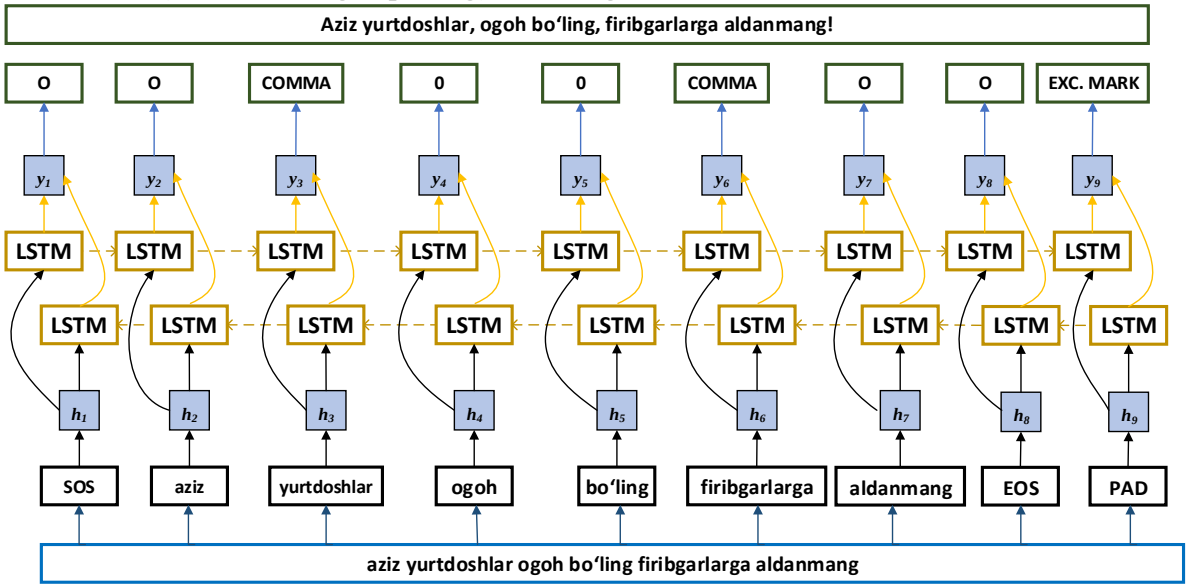
Datasetdagi o‘rtacha gap uzunligi.

Dataset	Ingliz tili		O‘zbek tili	
	Avg.	Std.	Avg.	Std.
O‘quv	13.8	10.8	12.4	7.6
Baholash	13.5	10.7	12.1	7.2
Test (news)	-	-	12.4	7.2
Test (Ref.)	13.8	9.6	4.8	3.2
Test (ASR)	14.2	9.7	5.3	3.6

Og‘zaki nutqdan iborat o‘quv ma’lumotlaridagi nomalum tokenlar 79.1% ni

tashkil qilgan bo‘lsa, test ma’lumotlarida ushbu ko‘rstakich 78.82% ga teng bo‘ldi. Ushbu natijalar asosida transkripsiya qilingan ma’lumotlar taqsimoti tabiiy suhbatlarga ko‘proq o‘xshashdir.

Amalga oshirilgan tajribalar uchun foydalanilgan transformer modelining arxitekturasi (“4.3-rasm”ga qarang) keltirilgan.



“4.3-rasm”. Tajribalarimiz uchun umumiy transformer modeli arxitekturasi.

ASR natijasida yuzaga kelishi mumkin bo‘lgan uchta xatoga mos keladigan uch xil *to‘ldirish* turlari quyidagilar:

1. Birinchidan (ya’ni, almashtirish) biz tokenni boshqa token bilan almashtiramiz. Tajribalarimizdagi tasodifiy ravishda tokenni maxsus noma’lum token bilan almashtiramiz.

2. Ikkinchidan (ya’ni, o‘chirish), biz qayta ishlangan kirish ketma-ketligidan tasodifiy ba’zi tokenlarni o‘chirib tashlaymiz.

3. Nihoyat, kirishning biron bir tasodifiy pozitsiyasiga noma’lum tokenni qo‘shamiz (ya’ni, kiritish).

Yuqoroda keltirilgan uchta xato ham korpus matnlarida bir xil darajada keng tarqalmaganligi sababli, turli xil to‘ldirish amallari model samaradorligiga turlicha ta’sir ko‘rsatadi. Shu sababli, berilgan matnni qayta ishlash uchun biz uchta sozlanishi mumkin bo‘lgan parametrlardan foydalanamiz:

1. **α** – token o‘zgarishi ehtimolini aniqlash parametri;
2. **αsub** – almashtirish ehtimolini boshqarish parametri;
3. **αdel** – o‘chirish ehtimolini nazorat qilish parametri.

Qo‘shish ehtimoli **1-(αsub+αdel)** ifoda orqali aniqlanadi.

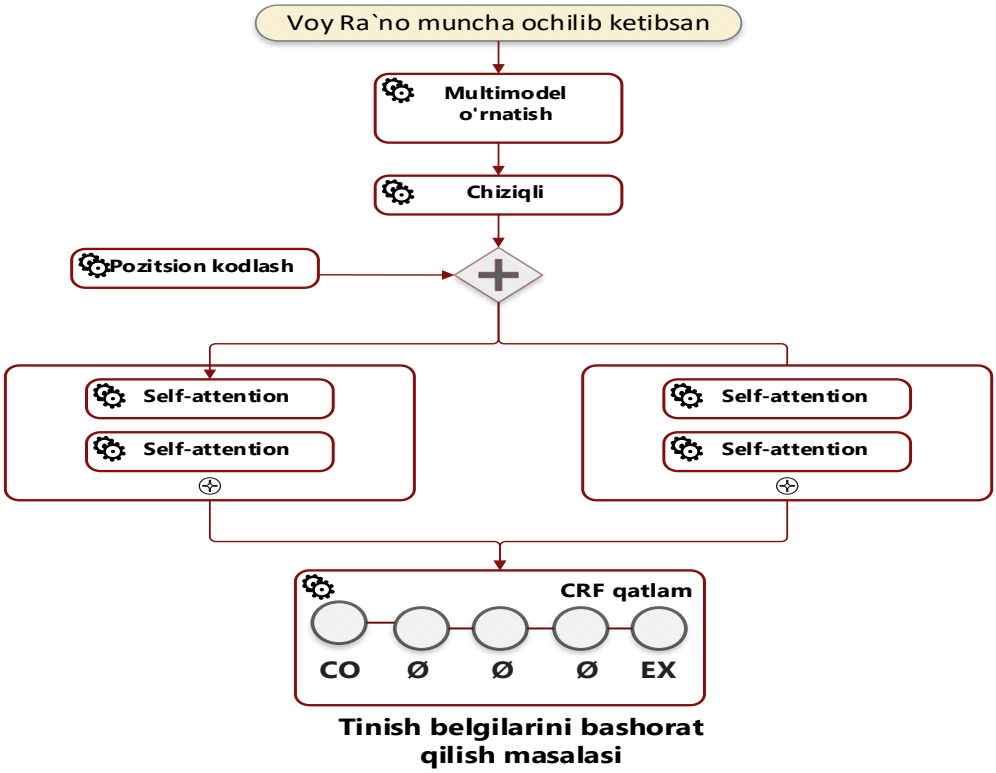
Matnda *o‘zgartirish amalini* qo‘llashda biz o‘sha pozitsiyadagi tokenni noma’lum token bilan almashtirdik va maqsadli tinish belgisini o‘zgarishsiz qoldirdik.

Bobning to‘rtinchi fasli “**POS teglash vositasida o‘zbek tili matnlarida tinish belgilarini qo‘yish”** deb nomlanadi. Tinish belgilarini tiklash masalasi og‘zaki nutqni avtomatik aniqlash (Automatic Speech Recognition – ASR) tizimlari doirasida keng o‘rganilgan NLP muammolaridan biri hisoblanadi. ASR

texnologiyasi inson nutqini kompyuter tomonidan aniqlash va matn ko‘rinishida transkripsiya qilish imkonini beruvchi dasturlash vositasidir. Mazkur masalaning dolzarbligi shundaki, og‘zaki nutq matnlari tinish belgilarisiz shakllanadi va bu holat keyingi bosqichlarda – sintaktik tahlil, tarjima yoki semantik modellashtirish jarayonlarida muayyan qiyinchiliklarni yuzaga keltiradi.

Dekodlash bosqichi va baholash ko‘rsatkichlari

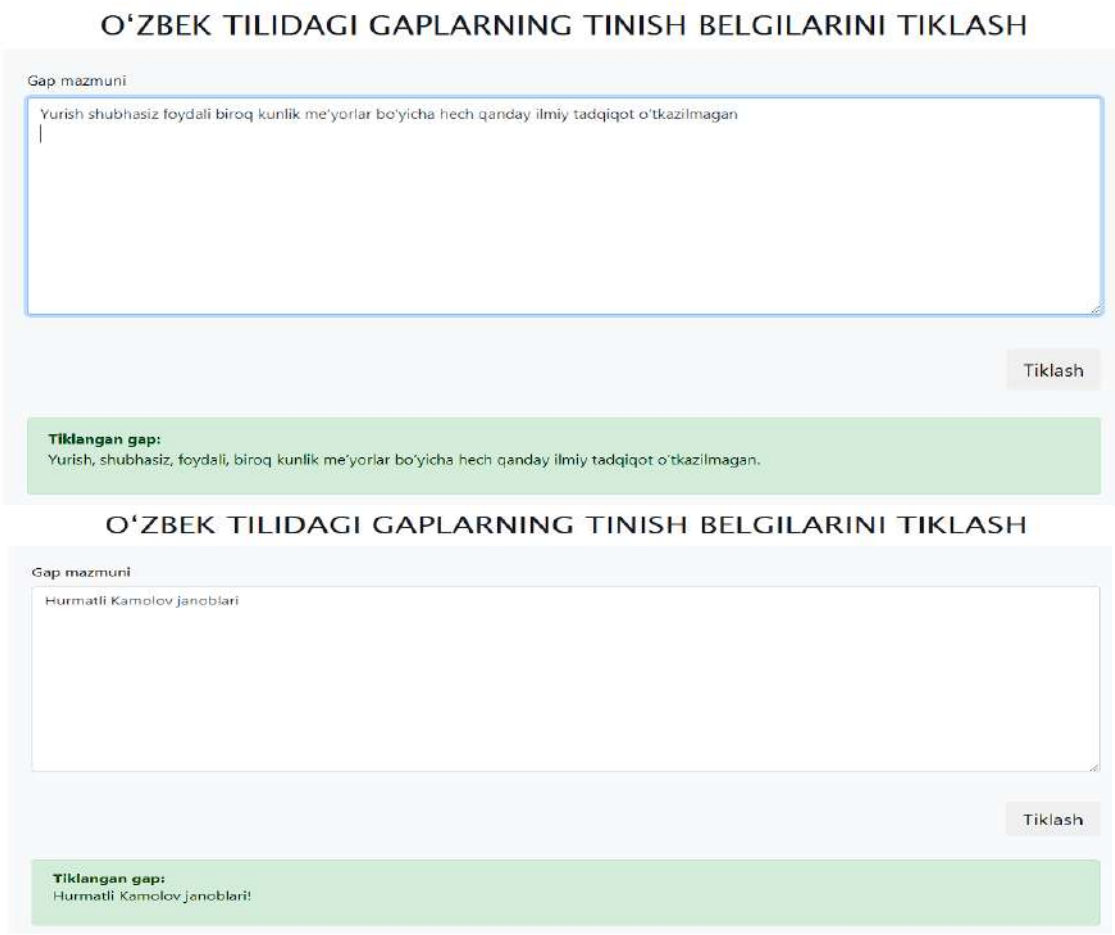
Dekodlash bosqichida POS teglash vazifasi va vazifa tasniflagichi o‘chiriladi. Lekin, ushbu bosqichda tinish belgilarni bashorat qilish vazifasi qo‘llaniladi(“4.4-rasm”ga qarang). Shunday qilib, model qo‘shimcha hisoblashlarini amalga oshirmasdan POS teglash vazifasidagi tinish belgilarni tiklash vazifasining o‘zgarmas sintaktik xususiyatlaridan foydalanishi mumkin.



“4.4-rasm”. Dekodlash bosqichida tavsiya etilgan modelning arxitekturasi.

Tinish belgilari ketma-ketligini yuqori ehtimollik orqali aniqlashda Viterbi algoritmi (Sammut va Webb, 2010)dan foydalanish mumkin. Amalga oshirilgan barcha tajribalardagi modellar **F1-score (F1)** ko’rsatkich bo‘yicha baholanadi. F1-score baholash ko‘rsatkichi **aniqlik (precision, *P*)** va **eslab qolish (recall, *R*)** qiymatlar asosida hisoblanadi. Biz tinish belgilarining matnda to‘g‘ri qo‘llanilganiga e’tibor qaratamiz. Shunday qilib, to‘g‘ri bashorat qilingan “**Ø**” tinish belgisi e’tiborga olinmaydi. Biz faqat uchta tinish belgisining ishlashini baholaymiz: **vergul (CO), nuqta (PE)** va **so‘roq (QU)**. Barcha tajribalar ma’lumotlar to‘plamining ikkita test to‘plamida sinovdan o‘tkazildi: **Ref.** va **ASR**. **O‘zbek tili milliy korpusi (UzCorpus)**dan olingan ma’lumotlar to‘plami tinish belgilarini bashorat qilish va POS teglash vazifalari uchun ishlatiladi.

Ishlab chiqilgan dasturiy ta’minotdan lavhalar (https://punct-restorer.uznlp.uz/) (“4.5-rasm”ga qarang)



“4.5-rasm”. Dasturiy ta’minot.

UMUMIY XULOSA

1. O‘zbek tilida punktuatsiyaning tarixiy taraqqiyotini bosqichma-bosqich tasniflash va davrlar bo‘yicha farqlarni aniqlash tilshunoslikda yangi nazariy manzara yaratishga xizmat qiladi. Bu jarayon milliy punktuatsiya tizimining shakllanishida ichki qonuniyatlar bilan birga tashqi omillar – arab, fors-tojik va rus tilshunosligi ta’sirini ham tadqiq etish imkonini beradi. Tinish belgilarining turli davrlarda egallagan vazifalarini qiyosiy tahlil qilish orqali yozma nutqning grammatik va semantik me’yorlari qanday shakllangani hamda qanday o‘zgarishlarga uchragani nazariy asosda ochib beriladi. Shu bilan birga, punktuatsiya tizimining evolyutsiyasi umumiy turkiy tilshunoslik nuqtayi nazaridan qiyosiy izlanishlarga, jahon tilshunosligidagi umumiy qonun-qoidalar bilan solishtirishga keng yo‘l ochadi. Shunday qilib, tarixiy yondashuv asosida punktuatsiya tizimini o‘rganish milliy tilning tinish belgilari tizimini nafaqat amaliy qoida sifatida, balki tilshunoslik nazariyasining mustaqil va metodologik jihatdan ahamiyatli yo‘nalishi sifatida tadqiq etish zarurligini ko‘rsatadi. Bu yondashuv punktuatsiyaning lingvistik strukturadagi o‘rni, uning funksiyalari hamda tarixiy evolyutsiyasini tizimli tarzda tahlil qilish imkonini beradi.

2. Punktuatsiya yozma nutqning grammatik tartibini ta’minlab, ma’no va uslubiy qirralarni aniqlashda hamda kommunikativ maqsadni ravshanlashtirishda muhim vosita hisoblanadi. O‘zbek tilidagi punktuatsiya tizimi tarixiy rivojlanish va

boshqa tillar ta’sirida shakllangani sababli, uning me’yoriy asoslarini takomillashtirish va funksional xususiyatlarini chuqur tadqiq etish dolzarb vazifadir. Shuningdek, tinish belgilarini avtomatik tarzda qo‘yish masalasi sintaktik va semantik murakkablikka ega bo‘lib, kompyuterli tahlil va NLP texnologiyalarida yuqori samaradorlikka erishish uchun yanada takomillashtirishni talab etadi.

3. Formalizatsiya tilshunoslikda til jarayonlarini modellashtirish va tahlil qilishning metodologik asosi sifatida muhim ahamiyat kasb etadi. U tilning ichki qonuniyatlarini anglash, yozma nutqda sintaktik va semantik aniqlikka erishish hamda kompyuter lingvistikasida avtomatik tahlil, mashinali tarjima va korpus yaratish jarayonlarini takomillashtirishga xizmat qiladi.

4. Tinish belgilarini formalizatsiya orqali tadqiq qilish ularning grammatik, semantik va kommunikativ vazifalarini ochib berish bilan birga, NLP texnologiyalarida punctuation restoration kabi vazifalarni samarali hal etish imkonini beradi. O‘zbek tilining morfologik murakkabligi va erkin so‘z tartibiga qaramay, BERT va BiLSTM kabi modellar tinish belgilarini avtomatik joylashtirishda yuqori natijalar ko‘rsatib, milliy korpus yaratilishi va intellektual tizimlarni rivojlantirishga ilmiy asos bo‘lib xizmat qilmoqda.

5. Tinish belgilar korpusda me’yoriy va uzual qirralari bilan birga poetik va ekspressiv vazifalari orqali ham namoyon bo‘lib, yozma nutqning grammatik aniqligi, kommunikativ maqsadi va badiiy ta’sirini ta’minlaydi. Ular orqali matnda individual uslubiy xususiyatlar ifodalanib, punktuatsiya faqat formal vosita emas, balki ko‘p qirrali semiotik element sifatida tadqiq etilishi zarurligi ko‘rinadi.

6. Korpus lingvistikasi va NLP texnologiyalari asosida tinish belgilarning formal va ekspressiv qo‘llanishini avtomatik tahlil qilish imkoni yaratilib, xatolarni aniqlash, uslubiy o‘zgachiliklarni belgilash va milliy korpusni boyitish kabi vazifalar samarali hal etilmoqda. Bu jarayon tilshunoslik va kompyuter lingvistikasini uyg‘unlashtirgan holda punktuatsiyani yangi metodologik yondashuvlar asosida tadqiq qilish zarurligini ko‘rsatadi.

7. Tinish belgilarini korpus materiallari asosida tizimli o‘rganish hamda ularga nisbatan formal yondashuvni qo‘llash tilning grammatik va semantik qonuniyatlarini yanada chuqurroq anglash imkonini beradi. Bu yondashuv punktuatsiyaning lingvistik strukturadagi o‘rni, funksiyalari va matn kontekstidagi samaradorligini ilmiy asosda tahlil qilishga zamin yaratadi. Punktuatsiya vositalari yozma nutqning me’yoriy tartibini ta’minlovchi qoidaviy element sifatidagina emas, balki matnning ma’noviy yaxlitligi, uslubiy qirralari va kommunikativ samaradorligini shakllantiruvchi muhim semiotik vosita sifatida ham nazariy asoslanadi. Shu bois, punktuatsion xatolarni aniqlovchi kompyuter dasturini yaratish jarayonida lingvistik qonun-qoidalarning formallashtirilishi, korpusdan chiqarilgan statistik ma’lumotlar bilan uyg‘unlashtirilishi va sun’iy intellekt algoritmlari orqali modellashtirilishi nazariy jihatdan mustaqil ilmiy yo‘nalish sifatida shakllanadi.

8. Punktuatsiya xatolarini avtomatik aniqlash va og‘zaki nutqni yozma shaklga keltirish texnologiyalari o‘zbek tili misolida ilk bor sinovdan o‘tkazilgani milliy tilshunoslik uchun muhim qadam hisoblanadi. Bu jarayon yozma nutqdagi

sintaktik va uslubiy aniqlikni ta’minlash, tahrir jarayonlarini yengillashtirish hamda ta’lim jarayonida til qoidalarini samarali o‘zlashtirishga xizmat qiladi. Shuningdek, mazkur texnologiyalar milliy korpusni boyitish, elektron lug‘at va avtomatik tarjima tizimlarini takomillashtirish, shuningdek, media va axborot platformalarida o‘zbek tilidagi matnlarni yuksak aniqlik bilan qayta ishlash imkonini yaratadi.

9. Tilni modellashtirish va ASR tizimlari NLP sohasida asosiy yo‘nalishlardan biri bo‘lib, ularning maqsadi til korpuslariga asoslanib, tabiiy til qonun-qoidalarini o‘rganish va ularni amaliy jarayonlarda qo‘llashdan iboratdir. Statistik va neyron modellar asosida til elementlari o‘rtasidagi ehtimoliy bog‘lanishlarni aniqlash orqali audiodan matnga o‘tkazish, avtomatik tarjima, his-tuyg‘ularni tahlil qilish, savol-javob tizimlari va matn klassifikatsiyasi kabi turli vazifalar samarali hal etilmoqda. Ayniqsa, GPT, BERT, Claude, Falcon kabi zamonaviy til modellari katta korpuslarda o‘qitilib, maskallangan til modellashtirish usuli orqali turli kontekstlarda ma’noli reprezentatsiyalar hosil qilish imkonini bermoqda.

10. ASR tizimlarida eng dolzarb muammolardan biri – tinish belgilarini qo‘yish vazifasi bo‘lib, bu yozma matnning o‘qilishi va grammatik aniqligini ta’minlaydi. Zamonaviy tadqiqotlarda CRF, RNN, CNN, transformerlar kabi modellar punktuatsiyani avtomatik tiklashda qo‘llanilib, BERT asosidagi modellar ketma-ketlikni belgilash orqali yuqori natijalar ko‘rsatmoqda. Shu bilan birga, ASR samaradorligini oshirish uchun xato turlariga og‘irlik berish, izohlovchilar o‘rtasidagi nomuvofiqliklarni hisobga olish, tinish belgilarining siljishi va majburiy izlanish xatolarini inobatga olish zarur. Demak, tilni modellashtirish va ASR sohasidagi izlanishlar nafaqat nazariy, balki amaliy jihatdan ham muhim ahamiyatga ega bo‘lib, ular intellektual dasturlar va kundalik hayotdagi turli ilovalarni takomillashtirishda hal qiluvchi o‘rin tutadi.

НАУЧНЫЙ СОВЕТ DSc.03/2025.27.12 Fil.08.07
ПО ПРИСУЖДЕНИЮ УЧЕНЫХ СТЕПЕНЕЙ ПРИ
БУХАРСКОМ ГОСУДАРСТВЕННОМ УНИВЕРСИТЕТЕ
БУХАРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ЖУМАЕВА ДИЛНОЗА БАХШУЛЛОЕВНА

ТЕХНОЛОГИЯ СОЗДАНИЯ КОМПЬЮТЕРНОЙ ПРОГРАММЫ ДЛЯ ПЕРЕНОСА ЗНАКОВ ПУНКТУАЦИИ В КОРПУСЕ И ВЫЯВЛЕНИЯ ПУНКТУАЦИОННЫХ ОШИБОК

10.00.11 – Прикладная и вычислительная лингвистика.

АВТОРЕФЕРАТ
дисертации доктора (DSc) Филологических наук

Бухара - 2026

Тема докторской диссертации (DSc) зарегистрирована в Высшей аттестационной комиссии под номером B2024.4.DSc/Fil905

Докторская диссертация выполнена в Бухарском государственном университете.

Автореферат диссертации размещен на трех языках (узбекском, русском, английском (резюме)) на сайте Бухарского государственного университета (www.buxdu.uz) и информационно-образовательном портале «Ziyonet» (www.ziyonet.uz).

Научный консультант:	**Тоирова Гули Ибрагимовна** доктор филологических наук (DSc), профессор
Официальные оппоненты:	**Гулямова Шахноза Кахрамоновна** доктор филологических наук (DSc), профессор
	Элов Ботир Болтаевич доктор технических наук (DSc), доцент
	Холмухамедов Бахтиёр Фарходович доктор филологических наук (DSc), доцент
Ведущая организация:	**Каракалпакский государственный университет имени Бердака**

Защита диссертации состоится «5» сентября 2026 года в 9:00 часов на заседании Научного совета DSc.03/2025.27.12 Fil.08.07 по присуждению ученых степеней при Бухарском государственном университете. (Адрес: 200118, г. Бухара, ул. М.Икбола, 11. Тел.: +99865 221-29-14; факс: +99865 221-27-07, e-mail: buxdu_rektor@buxdu.uz).

С диссертацией можно ознакомиться в Информационно-ресурсном центре Бухарского государственного университета. (зарегистрировано под номером 996). (Адрес: 200118, г. Бухара, ул. М.Икбола, 11. Тел.: +99865 221-29-14).

Автореферат диссертации разослан «20» августа 2026 года.

(Протокол реестра рассылки № 27 от «20» августа 2026 года.)

Б.Е.Киличев
Председатель Научного совета по присуждению учёных степеней, доктор филологических наук (DSc), профессор

М.Ю.Хожиева
Секретарь Научного совета по присуждению учёных степеней, доктор философии по филологических наук (PhD), доцент

Г.И.Тоирова
Председатель Научного семинара при Научном совете по присуждению ученых степеней, доктор филологических наук (DSc), профессор

ВВЕДЕНИЕ (аннотация диссертации доктора (DSc) наук)

Актуальность и востребованность темы диссертации. В мировой лингвистике стремительно развиваются технологии компьютерной лингвистики и обработки естественного языка (Natural Language Processing, NLP), обеспечивающие автоматический анализ текстов, выявление и исправление грамматических и пунктуационных ошибок, однако для узбекского языка до настоящего времени недостаточно разработаны интеллектуальные модели и алгоритмы, основанные на корпусных данных, что обусловливает необходимость совершенствования корпусных и лингвистических основ узбекской пунктуации, разработки систем автоматического выявления и восстановления знаков препинания на основе технологий искусственного интеллекта и их внедрения в образовательные и цифровые информационные системы как актуальной научно-практической задачи.

В мировой лингвистике интенсивно развиваются научные исследования в области компьютерной лингвистики и обработки естественного языка (NLP), направленные на автоматический анализ языковых корпусов, а также моделирование грамматических и пунктуационных явлений с использованием технологий искусственного интеллекта. Одним из приоритетных научных направлений являются корпусное описание, автоматическая аннотация и анализ пунктуационных единиц. Вместе с тем для языков с ограниченными ресурсами, в частности узбекского языка, недостаточно разработаны лингвистические и алгоритмические основы корпусного описания пунктуационных единиц, их автоматического выявления и исправления. В этой связи разработка корпусной модели узбекской пунктуации, а также лингвистических и алгоритмических основ автоматического выявления и восстановления пунктуационных ошибок представляет собой актуальную научную задачу.

В последние годы в Республике Узбекистан последовательно развиваются научно-практические исследования, направленные на развитие цифровых технологий, искусственного интеллекта и компьютерной лингвистики, создание Национального корпуса узбекского языка, а также технологий обработки естественного языка. В частности, расширяются исследования по созданию лингвистических корпусов, языковых баз данных, технологий автоматизированного лингвистического анализа и машинного обучения. Как отметил Президент Республики Узбекистан Шавкат Мирзиёев, “Сегодня в мире искусственный интеллект становится ключевым инструментом в таких сферах, как экономика, образование, здравоохранение и безопасность”[1]. В этой связи совершенствование корпусных основ узбекской пунктуации, разработка технологий автоматического выявления и исправления пунктуационных ошибок на основе искусственного интеллекта,

[1] O‘zbekiston Respublikasi Prezidentining 2025-yil 22-oktabrdagi PF–189-son “Sun’iy intellekt texnologiyalarini yanada rivojlantirishga oid qo‘shimcha chora-tadbirlar to‘g‘risida”gi Farmoni. Manba: https://lex.uz/uz/docs/-7789403.

а также их внедрение в образовательные и цифровые информационные системы являются актуальной научно-практической задачей.

Диссертационная работа в определённой степени способствует реализации задач, определённых Указом Президента Республики Узбекистан от 21 октября 2019 года № ПФ-5850 “О мерах по коренному повышению роли и статуса узбекского языка как государственного языка”, Указом Президента Республики Узбекистан от 7 февраля 2017 года № ПФ-4947 “О Стратегии действий по дальнейшему развитию Республики Узбекистан”, Постановлением Президента Республики Узбекистан от 17 февраля 2017 года № ПП-2789 “О мерах по дальнейшему совершенствованию деятельности Академии наук, организации, управления и финансирования научно-исследовательских работ”, Постановлением Кабинета Министров Республики Узбекистан от 13 мая 2016 года № 152 “Об организации деятельности Ташкентского государственного университета узбекского языка и литературы имени Алишера Навои”, а также другими нормативно-правовыми актами, регулирующими данную сферу деятельности.

Соответствие исследования приоритетным направлениям развития науки и технологий Республики. Данное исследование проводилось в соответствии с приоритетным направлением развития науки и техники республики I. “Формирование системы инновационных идей и способов их реализации в социально-правовом, правовом, экономическом, культурном, духовном и образовательном развитии информационного общества и демократического государства”.

Обзор зарубежных научных исследований по теме диссертации[2].

Научные исследования по распределению знаков препинания в текстовых корпусах, их автоматическому анализу и разработке программ обнаружения пунктуационных ошибок проводились в ведущих научных центрах и высших учебных заведениях по всему миру, включая Массачусетский технологический институт, Стэнфордский университет, Гарвардский университет (США), Кембриджский университет, Оксфордский университет (Великобритания), Институт информатики им. Макса Планка, Университет имени Гумбольдта в Берлине (Германия), а также Институт лингвистики Российской академии наук, Московский государственный университет им. М.В.Ломоносова, Санкт-Петербургский университет, Казанский федеральный университет Татарстана, Центрально-Азиатский университет имени Алишера Навои, Ташкентский государственный университет узбекского языка и литературы, Институт языка, литературы и фольклора Академии наук Республики Узбекистан, Бухарский государственный университет и Ферганский государственный университет.

[2] Dissertatsiya mavzusi bo‘yicha xorijiy ilmiy-tadqiqotlar sharhi quyidagi manbalar asosida amalga oshirildi: https://ijllnet.com/journals/Vol_2_No_1_March_2015/11.pdf: https://www.academypublication.com/issues/past/tpls/vol01/11/04.pdf;https://www.ccsenet.org/journal/index.php/ijel/article/view/0/42036,;https://scholarlypublications.universiteitleiden.nl/;https://ijllnet.thebrpi.org/;www.pushkin.institute.ru,;www.lingvolive.com/ru;https://www.merriam-webster.com/;https://macmillaneducation.my.salesforce-sites.com/help/;https://www.britannica.com/dictionary;https://www.oxfordlearnersdictionaries.com/; www.en.wikipedia.org, www.learnenglishkids. britishcouncil.org; https://www.multitran.com/.

В области пунктуации и корпусной лингвистики, в частности, достигнуты следующие научные результаты: исследованы синтаксические и прагматические функции пунктуационных единиц, их роль в семантике текста и дискурсе (Стэнфордский университет, США); разработаны методы автоматической аннотации пунктуационных единиц, их разметки в корпусах и алгоритмы обработки естественного языка (Оксфордский университет, Великобритания); заложены основы статистических и нейросетевых моделей выявления и исправления пунктуационных ошибок (Университет имени Гумбольдта в Берлине, Германия); раскрыты возможности применения пунктуации в корпусной лингвистике, машинном переводе и речевых технологиях (Сорбоннский университет, Франция); проведены научные исследования по автоматическому восстановлению знаков препинания, синтаксическому анализу текста и созданию языковых моделей на основе технологий искусственного интеллекта (Российская академия наук, Россия; Токийский университет, Япония; Пекинский университет, Китай); изучаются нормативные, функциональные и коммуникативные особенности системы пунктуации, а также вопросы её представления в электронных корпусах (Бухарский государственный университет, Узбекистан).

В мировой лингвистике в настоящее время, в частности, приоритетными направлениями научных исследований являются: разработка лингвистических и корпусных моделей пунктуационных единиц; описание их синтаксических, семантических и прагматических свойств на основе технологий искусственного интеллекта и обработки естественного языка; совершенствование алгоритмов автоматического выявления и восстановления пунктуационных ошибок; разработка принципов пунктуационной аннотации электронных корпусов; создание лингвистических моделей пунктуации для систем автоматического редактирования текста, машинного перевода, распознавания речи и генеративного искусственного интеллекта.

Степень изученности проблемы. В области мировой лингвистики и обработки естественного языка (ОБЯЗ) постоянно изучаются вопросы автоматического обнаружения знаков препинания, их применения в корпусе и анализа пунктуационных ошибок, особенно в области обработки естественного языка (ОБЯЗ). Исследования в этой области восходят к концу XX и началу XXI веков. Взаимосвязь системы пунктуации с морфосинтаксическими, семантическими и прагматическими факторами, её коммуникативная и функциональная роль в тексте рассматриваются в работах таких учёных, как М.М.Ходоров, Д.Кляйн, К.Мэннинг, Д.Юрафски, М.Стидман, Х.Шмид, Э.Бик, Р.Гарсайд, А.Килгаррифф, Т.Кисс, Х.Странк. В последние годы всё больше внимания уделяется разработке алгоритмов автоматического восстановления пунктуации, обнаружения и исправления ошибок. В частности, Василе Пэиш и Дан Туфиш представили всестороннее аналитическое исследование исторических и современных моделей пунктуации и написания заглавных букв. Аттила Надь, Бенце Бяль и Юдит Ач разработали автоматическую систему распознавания пунктуации на

основе BERT, которая показала высокие результаты (F_1 ~80–82) на английском и венгерском языках с использованием глубокого обучения. Тем временем, Чонхён Мин и др. улучшили понимание структуры текста с помощью пунктуации, используя неконтролируемый подход[3].

За последнее десятилетие в российской лингвистике стремительно развивается область обработки естественного языка. В частности, решаются задачи, связанные с корпусной лингвистикой, автоматическим синтаксическим и семантическим анализом, а также правильным использованием пунктуации. Например, А.А.Кибрик и И.М.Богуславский реализовали проекты по семантической разметке в русском языке, формальному описанию синтаксических структур и логическому анализу данных. Также разрабатываются компьютерные программы и модели НЛП на основе Национального корпуса русского языка. Активно развивается и работа по автоматическому обнаружению пунктуационных ошибок и структур. Например, Ю.Д.Апрасов и Е.И.Ерохина создали формализацию правил пунктуации и основанные на них алгоритмы анализа. Сегодня для русского языка широко используются платформы НЛП на основе нейронных сетей, такие как Берт, Дип Павлов и Наташа [4].

В узбекской лингвистике область обработки естественного языка (NLP) добивается значительных успехов, особенно в морфологическом и корпусном анализе. В 2022-2023 годах авторы под руководством Максуда Шарипова, Огабека Собирова, Улугбека Салаева, Эльмурода Куриёзова и Джамолбека Матиева разработали практические алгоритмы в различных областях. В частности, М. Шарипов и О. Собиров предложили алгоритм лемматизации на основе конечных автоматов (FSM), а У.Салаев реализовал высокоточную (91%) модель морфологического анализа с использованием флективных окончаний. С другой стороны, М.Шарипов и его команда

[3] Chodorow M., Leacock C., Gamon M., Tetreault J. Automated grammatical error detection for language learners. Synthesis Lectures on Human Language Technologies. 2010. Vol. 3, № 1. – P. 1-134.; Klein D., Manning C. D. Accurate unlexicalized parsing. // Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. 2003. – P. 423-430.; Steedman M. Surface Structure and Interpretation. Cambridge, MA: MIT Press, 1996. 308 p.; Schmid H. Probabilistic Part-of-Speech Tagging Using Decision Trees. // International Conference on New Methods in Language Processing. 1994. – P. 44-49.; Bick E. The Parsing System Palavras: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Århus: Aarhus University Press, 2000. 380 p.; Garside R., Leech G., McEnery T. Corpus Annotation: Linguistic Information from Computer Text Corpora. London: Longman, 1997. 280 p.; Kilgarriff A., Rychlý P., Smrž P., Tugwell D. The Sketch Engine. // Proceedings of the Eleventh EURALEX International Congress. Lorient, France, 2004. – P. 105-116.; Kiss T., Strunk J. Unsupervised Multilingual Sentence Boundary Detection. // Computational Linguistics. 2006. Vol. 32, № 4. – P. 485-525.; Păiș V., Tufiș D. A survey of punctuation restoration methods. // Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). 2021. P. 1133-1142.; Nagy A., Bial B., Ács J. Automatic punctuation restoration with BERT models. // Proceedings of the Seventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2021). 2021. – P. 121-132.

[4] Кибрик А.А., Богуславский И.М. Формальное описание синтаксических структур и семантическая разметка текста // Вестник Московского университета. Серия 9: Филология. – 2012. – № 6. – С. 3-15.; Бородин А.В. Информационные технологии и корпусная лингвистика: на материале Национального корпуса русского языка // Вестник Санкт-Петербургского университета. Язык и литература. – 2020. – Т. 17, № 1. – С. 55-64.; Апрасов Ю.Д., Ерохина Е.И. Алгоритмизация пунктуационных правил в задачах автоматического анализа текста // Информационные технологии. – 2019. – № 4. – С. 28-35.; Куратов Ю., Архипов М. DeepPavlov: открытая библиотека для обработки естественного языка // Труды конференции по искусственному интеллекту и анализу данных (AINL 2019). – Санкт-Петербург, 2019. – С. 132-136.

создали корпус синтаксической морфологической базы на узбекском языке и реализовали систему автоматической аннотации с использованием POS-теггера[5]. Однако предыдущие исследования, посвященные представлению знаков препинания в электронных корпусах, их статистическому и функциональному распределению, а также автоматическому обнаружению пунктуационных ошибок, были ограничены, а существующие работы анализировали систему пунктуации в основном с теоретической или образовательной точки зрения. В частности, большинство исследований неадекватно отражали прагматическую ценность знаков препинания, их функционально-стилистические свойства, а также их комплексный анализ с использованием вычислительной лингвистики.

В диссертационной работе проведено специальное исследование, основанное на комплексном подходе к ранее не рассматривавшейся проблеме – автоматическое определение знаков препинания в узбекских текстах в электронном корпусе, создание алгоритмов обнаружения пунктуационных ошибок и выявление их методологических, грамматических и коммуникативных аспектов. В ходе этого процесса критически анализируются ранее изученные работы, научно обосновываются существующие пробелы и разрабатывается технология отражения знаков препинания в корпусе с помощью современных лингвистических технологий.

Связь темы диссертации с научно-исследовательскими работами высшего учебного заведения, в котором выполнена диссертация. Диссертация выполнена в соответствии с планом научно-исследовательских работ Бухарского государственного университета в рамках научного направления Проблемы, решения и перспективы теории языка и компьютерной лингвистики” (2026-2030 гг.).

Цель исследования заключается в разработке методологии создания компьютерной программы для выявления пунктуационных ошибок на основе анализа функционирования знаков препинания в корпусе узбекского языка.

Задачи исследования:

разработать необходимые методы и алгоритмы для создания компьютерной программы, обнаруживающей пунктуационные ошибки, оценить их эффективность и обеспечить их практическое применение;

научно обосновать новые подходы в области искусственного интеллекта и машинного обучения к обнаружению знаков препинания и исправлению ошибок, определить их лингвистические и алгоритмические основы;

изучить использование знаков препинания на основе корпусной лингвистики, выявить закономерности в текстах различных жанров и сделать статистические и функциональные выводы;

[5] Sharipov M., Sobirov O. Development of a rule-based lemmatization algorithm through Finite State Machine for Uzbek language. – Arxiv preprint, 2022. arxiv:2210.16006 reddit.comarxiv+6arxiv+6ResearchGate+6; Sharipov M., Kuriyozov E., Yuldashev O., Sobirov O. UzbekTagger: The rule-based POS tagger for Uzbek language. – Arxiv preprint, 2023. arxiv:2301.12711 GitHub+3arxiv+3ACL Anthology+3; Sharipov M., Mattiev J., Sobirov J., Baltayev R. Creating a morphological and syntactic tagged corpus for the Uzbek language. – Arxiv preprint, 2022. arxiv:2210.15234 arxiv+1arxiv+1; Salaev U. UzMorphAnalyser: A Morphological Analysis Model for the Uzbek Language Using Inflectional Endings. – Arxiv preprint, 2024.

определить место знаков препинания в грамматической и семантической структуре языка и научно объяснить их функции в тексте, учитывая коммуникативные и стилистические особенности текста;

определить знаки препинания и исправить пунктуационные ошибки на основе современных лингвистических и алгоритмических подходов, разработать и практически обосновать их интегрированную систему.

Объектом исследования является использование 10 знаков препинания в текстах из национального узбекского корпуса, разработанного в Ташкентском государственном университете языков и литературы, а также связанные с ними пунктуационные ошибки.

Предмет исследования. Предметом исследования является использование 10 знаков препинания в текстах из национального узбекского корпуса, разработанного в Ташкентском государственном университете языков и литературы, а также связанные с ними пунктуационные ошибки.

Методы исследования. В процессе исследования используются такие методы, как лингвистические методы, статистические методы, обработка естественного языка (NLP), грамматический анализ и автоматическое обнаружение ошибок.

Научная новизна исследования заключается в следующем:

в языкознании принцип формализации применён к системе пунктуации, в рамках функционально-системного подхода определены грамматические (обособление однородных членов предложения с помощью запятой), семантические (выражение пояснительных отношений посредством двоеточия и тире) и прагматические (выражение эмоционально-экспрессивного и вопросительного значений посредством восклицательного и вопросительного знаков) функции знаков препинания узбекского языка, на основе чего разработана лингво-функциональная модель, отражающая их взаимосвязь в обеспечении структурной и смысловой целостности текста;

на основе методологии корпусной лингвистики закономерности нормативного употребления знаков препинания в узбекском языке, их вариативность в разрезе жанров, а также критерии автоматического выявления пунктуационных ошибок смоделированы на основе формальных правил и контекстуальных признаков; предложена лингвистическая модель классификации пунктуационных единиц по их грамматическим (запятая, точка с запятой, двоеточие) и коммуникативно-прагматическим (вопросительный знак, восклицательный знак, многоточие) функциям, а также их автоматической разметки на основе системы POS-тегов и пунктуационных тегов;

на основе архитектуры Transformer, современных технологий NLP и искусственного интеллекта разработана модель автоматического восстановления пунктуации, адаптированная к узбекскому языку, обоснованы лингвистические и алгоритмические принципы контекстно-зависимого определения и восстановления знаков препинания, а эффективность модели подтверждена результатами экспериментальных исследований по показателям Accuracy, Precision, Recall и F1-score;

на основе POS-разметки выявлены грамматические характеристики знаков препинания, связанные с определением последовательности частей речи и синтаксических связей, а также их семантические характеристики, выражающие предикативные, пояснительные отношения и отношения перечисления;

для узбекского языка разработано программное обеспечение, обеспечивающее автоматическое выявление и исправление пунктуационных ошибок, а также автоматическое восстановление знаков препинания в процессе преобразования устной речи в письменный текст. В частности, обеспечена автоматическая пунктуационная разметка текста, не содержащего знаков препинания, например Bugun havo issiq ertaga yomg‘ir yog‘ishi mumkin”, с его преобразованием в нормативный вариант Bugun havo issiq, ertaga yomg‘ir yog‘ishi mumkin”, что способствует повышению грамматической корректности и удобочитаемости текста.

Практические результаты работы заключаются в следующем:

разработка адаптированных алгоритмов для выявления пунктуационных ошибок в узбекском языке, доказавших свою важность как ресурс для развития исследований в области лингвистики и определения научных основ автоматизированного анализа пунктуации;

анализ пунктуационных вопросов в узбекском языке на основе фактических материалов с использованием статистических, лингвистических подходов и методов машинного обучения, а также научно доказана их эффективность и практическое применение;

разработанные методологические подходы для выявления и исправления пунктуационных ошибок, посредством интеграции корпусной лингвистики и технологий искусственного интеллекта, доказали свою эффективность как практический инструмент, позволяющий использовать узбекский язык в лингвистике, образовании и научных исследованиях.

Достоверность результатов исследования объясняется тем, что примененный подход, методы и теоретические данные были получены из надежных научных источников, представленный анализ основывался на рационально-типологических и комплексных методах анализа, теоретические идеи и выводы были реализованы на практике, а полученные результаты были одобрены уполномоченными структурами.

Научная и практическая значимость результатов исследования. Научная значимость результатов исследования определяется разработкой научно-теоретических основ лингвистического моделирования грамматических, семантических и прагматических функций знаков препинания в узбекском языке и их представления в корпусной лингвистике, совершенствованием лингвистических и алгоритмических подходов к автоматическому определению структуры текста, синтаксических связей и интонационных границ на основе пунктуационных единиц, развитием технологий обработки естественного языка (Natural Language Processing, NLP) для узбекского языка, а также расширением возможностей использования результатов исследования в качестве теоретической базы при

преподавании дисциплин “Теория языкознания”, “Компьютерная лингвистика”, “Лексикология” и “Пунктуация”, а также при создании современных лингвистических корпусов и интеллектуальных систем.

Практическая значимость результатов исследования определяется разработкой программного обеспечения для автоматической обработки текстов узбекского языка, выявления и исправления пунктуационных ошибок, внедрением его в практику посредством электронной платформы https://punct-restorer.uznlp.uz/, возможностью применения разработанной системы в образовательном процессе, при автоматическом редактировании текстов, в обучающих языковых программах и системах контроля качества онлайн-коммуникации, а также использованием результатов исследования при разработке учебно-методического обеспечения по дисциплинам “Пунктуационная лингвистика” и “Искусственный интеллект в языкознании”, что способствует повышению эффективности использования узбекского языка в современной цифровой информационно-коммуникационной среде.

Внедренность результатов исследования. На основе результатов, полученных в ходе разработки технологии представления знаков препинания в корпусе и создания компьютерной программы для обнаружения пунктуационных ошибок:

теоретические выводы, относящиеся к разработанной в результате применения принципа формализации к системе пунктуации в языкознании лингво-функциональной модели, в рамках которой на основе функционально-системного подхода были оценены грамматические (обособление однородных членов с помощью запятой), семантические (выражение пояснительных отношений посредством двоеточия и тире) и прагматические (выражение эмоционально-экспрессивного и вопросительного значений посредством восклицательного и вопросительного знаков) функции знаков препинания узбекского языка, а также их взаимосвязь в обеспечении структурной и смысловой целостности текста, были использованы в 2022-2024 годах в рамках научного проекта REP-25112021/113 “UzUDT: Универсальный корпус деревьев зависимостей узбекского языка для обработки естественного языка и его семантический анализ” (справка Национального университета Узбекистана имени Мирзо Улугбека № 04/11-20706 от 29 сентября 2025 года). В результате создана возможность нормализации знаков препинания узбекского языка для современных корпусов на основе их историко-сопоставительного анализа;

научные выводы, относящиеся к разработанной на основе методологии корпусной лингвистики лингвистической модели, в рамках которой закономерности нормативного употребления знаков препинания в узбекском языке, их вариативность в разрезе жанров, а также критерии автоматического выявления пунктуационных ошибок смоделированы на основе формальных правил и контекстуальных признаков, а пунктуационные единицы классифицированы по их грамматическим (запятая, точка с запятой, двоеточие) и коммуникативно-прагматическим (вопросительный знак,

восклицательный знак, многоточие) функциям и предложена их автоматическая разметка на основе системы POS-тегов и пунктуационных тегов, были использованы в 2024 году в рамках проекта AL-662204396 “Создание электронной платформы по научно-литературному наследию деятелей Бухарской литературной среды” (справка Бухарского государственного университета № 6202 от 10 октября 2025 года). В результате обоснован теоретический подход, обеспечивающий осознанное освоение пунктуации при формальной классификации языковых единиц;

научные результаты, связанные с разработкой на основе архитектуры Transformer, современных технологий NLP и искусственного интеллекта модели автоматического восстановления пунктуации, адаптированной к узбекскому языку, обоснованием лингвистических и алгоритмических принципов контекстно-зависимого определения и восстановления знаков препинания, а также подтверждением эффективности модели по результатам экспериментальных испытаний на основе показателей Accuracy, Precision, Recall и F1-score, были применены в рамках корпусного проекта AM-ФЗ-201908172 “Создание образовательного корпуса узбекского языка”, выполненного в 2020-2023 годах (справка Ташкентского государственного университета узбекского языка и литературы имени Алишера Навои № 01/4-3886 от 2 сентября 2025 года). В результате на основе корпусных данных определены статистические характеристики знаков препинания и уровни их контекстуальной активности;

научные выводы о выявленных на основе POS-разметки грамматических характеристиках знаков препинания, таких как определение последовательности частей речи и синтаксических связей, а также семантических характеристиках, выражающих предикативные, пояснительные отношения и отношения перечисления, были использованы в 2023-2024 годах в передачах “Islohotlar amalda”, “Kelajak yoshlari”, “Inson qadri”, “Diyor 24” телеканала “Navoiy” Навоийской телерадиокомпании, осуществляющей деятельность в сфере информационных технологий (справка Навоийской областной телерадиокомпании № 108 от 10 ноября 2025 года). В результате компоненты системы пунктуационной разметки, основанные на синтаксических и морфологических признаках, внедрены в практику в качестве самостоятельного программного модуля;

научно-практические подходы, связанные с разработкой программного обеспечения, обеспечивающего автоматическое выявление и исправление пунктуационных ошибок в узбекском языке, а также автоматическое восстановление знаков препинания в процессе преобразования устной речи в письменный текст, были использованы при подготовке учебников “Корпусная лингвистика” и “Компьютерная лингвистика” для магистрантов образовательного направления 70230401 – “Компьютерная лингвистика” (приказ Министерства высшего образования, науки и инноваций Республики Узбекистан № 109-У от 18 февраля 2026 года, разрешение на издание № 109-3; приказ № 220-У от 8 апреля 2025 года, разрешение на издание № 220-07). В результате обогащено научное содержание учебников и

усовершенствованы теоретические и практические материалы по корпусной лингвистике, обработке естественного языка и технологиям пунктуационного анализа.

Апробация результатов исследования. Результаты данного исследования обсуждались на 4 международных и 5 национальных научно-практических конференциях.

Опубликованность результатов исследования. В общей сложности по теме диссертации опубликовано 25 научных работ, в том числе 1 монография и 20 научных статей (10 в республиканских и 4 в зарубежных журналах) в научных изданиях, рекомендованных Высшей аттестационной комиссией Республики Узбекистан для публикации основных результатов докторских диссертаций.

Структура и объем диссертации. Диссертация состоит из введения, четырех глав, заключения и списка литературы, общим объемом 232 страницы.

ОСНОВНОЕ СОДЕРЖАНИЕ ДИССЕРТАЦИИ

Введение обосновывает актуальность и необходимость темы, описывает цели и задачи, объект и предмет исследования, указывает на его соответствие приоритетным направлениям развития науки и техники республики, описывает научную новизну и практические результаты исследования, раскрывает научно-практическое значение полученных результатов, предоставляет информацию о применении результатов исследования на практике, опубликованных работах и структуре диссертации.

Первая глава диссертации озаглавлена **“История знаков препинания и их современные проблемы”**. В этой главе **“Теоретические основы системы пунктуации в мировом языкознании”** дается всестороннее научное объяснение теоретических, исторических и проблемных аспектов знаков препинания. Анализируются теоретические основы системы пунктуации в мировой лингвистике, история знаков препинания в узбекской лингвистике и ее периодизация, а также современные проблемы знаков препинания. На основе научных источников обобщены взгляды на функционально-лингвистическую сущность знаков препинания, их грамматическую и стилистическую роль в письменной речи и их коммуникативные функции в языковой системе.

Правила и принципы использования знаков препинания изучаются в лингвистике в рамках самостоятельного раздела – пунктуация (лат. punctum – точка). Исторически этот раздел также назывался “пунктуацией” и теоретически регулировал функции знаков препинания в письменной речи, такие как уточнение синтаксической структуры, разделение предложений и частей предложений, а также определение различия между письменной и устной речью.

Корни системы пунктуации восходят к Древнему Риму, но первоначальные подходы к ней тесно связаны с деятельностью

древнегреческих грамматических школ. В этот период функция знаков препинания заключалась главным образом в адаптации устной речи к произношению, установлении норм использования точек и заглавных букв в конце предложений.

Греки использовали всего три различных формы знаков препинания того периода. Хотя Никанор использовал 8 знаков в развитии пунктуации, эти попытки привели к путанице синтаксических и логических структур. До XV века не существовало единых правил для знаков препинания. Только систематизированная норма знаков, использованная венецианским издателем Альдом Мануцием, стала поворотным моментом в истории пунктуации, и он признан одним из основателей системы пунктуации. Позже в европейской и русской лингвистике к системе пунктуации были добавлены такие знаки, как точка с запятой, двоеточие, вопросительный знак, восклицательный знак, тире, кавычки и скобки.

Вторая часть этой главы, озаглавленная **“История знаков препинания в узбекской лингвистике и её периодизация”,** анализирует процесс формирования знаков препинания в языке, их первоначальные проявления и способы выражения в хронологическом порядке. В частности, показано отсутствие знаков препинания в древних письменных источниках или их условное выражение символами, а также изменения и этапы развития системы пунктуации в узбекском языке под влиянием арабской, персидской и русской лингвистики в более поздние периоды. История узбекской пунктуации формировалась в разных этапах, и важную роль в этом процессе сыграли научные исследования таких учёных, как Ш. Шоабдурахмонов и К. Назаров. В I–VIII веках первоначальное появление знаков препинания в арамейских и рунических надписях было в виде интервалов, которые не выражали смысла, а лишь выполняли функцию разделения. В IX–XIII веках знаки арабской письменности начали выполнять семантические функции, однако развитие знаков препинания было ограничено из-за несовместимости арабской письменности с узбекским языком. Со второй половины XIII века до конца XX века формирование знаков препинания, разработка правил и теоретическое обоснование пунктуации происходили под влиянием книгопечатания и русского языка. С 90-х годов XX века, с возвышением узбекского языка до статуса государственного языка, требования к пунктуации возросли, а её образовательное и социальное значение ещё больше увеличилось. В работе Сайдризо Ализоды “Муросалот” классификация и использование знаков препинания анализируются с периодической точки зрения. Историческое развитие, теоретические основы и современные правила узбекской пунктуации научно анализируются в работах таких учёных, как А.Фитрат, Г.Гозиев, Г.Абдурахмонов, Р.А.Алимухаммедов, К.Назаров, Х.Гафуров, С.Каримов, Н.Махмудов, Б.Бахриддинова, Ф.Т.Каримов, Ў.Дж.Эсонов, Н.М.Гамбаров и Д.Тураева[6].

[6] Фитрат А. Танланган асарлар (дарслик ва ўқув қўлланмалар, илмий мақола ва тадқиқотлар). Нашрга тайёрловчи Ҳ.Болтабоев. – Тошкент: Маънавият, 2009. – 336 б.; Ғозиев Ҳ. Ўзбек пунктуациясининг тарихий тараққиёти (пунктуацион белгилар, уларнинг вазифалари ва қўлланиш асослари). – Тошкент: Фан. 1979. –

Формирование и практическое использование знаков препинания в узбекском языке стало результатом длительного исторического процесса. Они развивались главным образом под влиянием арабской, персидской, тюркской и русской письменных традиций. Одним из древнейших знаков является точка, которая впервые была использована в качестве письменного знака в арабских источниках, в частности, во время халифата Абдул Малика (I век хиджры). В узбекском языке она появилась как знак препинания во второй половине XIX века. А.Фитрат определяет точку как “знак остановки”. Вопросительный знак, с другой стороны, происходит от латинского слова quæstio и начал появляться в узбекских текстах после 1885 года. Фитрат определяет этот знак как “знак вопроса”. Восклицательный знак вошел в узбекский язык через русский язык и начал регулярно использоваться в печатных текстах начиная с 1901 года. Он происходит от латинского слова Io и используется для выражения эмоций и восклицаний.

В целом, формирование и закрепление знаков препинания в узбекском языке происходило в несколько этапов, каждый из которых методологически обосновывался такими просветителями, как Фитрат, Зохирий и Чолпон.

Третий раздел этой главы называется **“Актуальные проблемы использования знаков препинания”** и охватывает ряд теоретических и практических проблем, связанных с вопросами пунктуации в современной лингвистике. Сегодня неопределенность в использовании знаков препинания стоит на повестке дня как актуальная научная проблема не только в узбекском языке, но и во многих других языках. В частности, использование различных знаков препинания в предложениях с одинаковой структурой под влиянием контекстуальных факторов или, наоборот, проявление одного и того же знака в разных функциях требует серьезного анализа как для изучающих язык, так и для лингвистов. Эти проблемы возникают из-за специфической стилистической, синтаксической и семантической полисемии. В то же время в современной лингвистике существуют лингвистические и технические препятствия для обнаружения, анализа и правильного размещения знаков препинания в процессе автоматизированного анализа текста. Например, в таких процессах, как машинный перевод, автоматический синтаксический анализ, преобразование речи в текст, соблюдение правил пунктуации часто создает трудности. Особенно в процессе написания спонтанной речи до сих пор остается нерешенной проблемой выражение восклицаний, интонационных пауз, имитации голоса и единиц, выражающих эмоции, с помощью пунктуации,

---

Б.156-187.; Абдураҳмонов Ф.А. Пунктуация ўргатиш методикаси. – Тошкент: Ўқитувчи, 1968. – 157 б.; Алимуҳаммедов Р.А. Қадимги туркий битикларда пунктуация ва матнни англаш учун ишлатилган ишоратлар ҳақида // Хорижий филология. – 2016. – №1, – Б. 49-53.; Назаров К. Тиниш белгилари ва ёзма нутқ. –Тошкент: Фан, 1974. – 68 б.; Ғафуров Ҳ. Атов гапларда интонация ва пунктуация масалалари. – Тошкент, 1963. – 98 б.; Каримов С.А. Ўзбек тилида тиниш белгилари синонимияси // https://primedu.jspi.uz/index.php/primedu/.; Маҳмудов Н. “Ўзбек тили пунктуациясининг асосий қоидалари”. – Тошкент: Ўзбекистон, 2015. – 88 б.; Бахриддинова Б.М. Замонавий ўзбек пунктуацияси асослари. Монография. – Тошкент: Академнашр, 2015. – 125 б.; Каримов Ф.Т. Пунктуация белгиларининг лингвопсихологик хусусиятлари. Research and education, 1(3), 2022. – P. 163-167 https://researchedu.org/index.php/raye/article/viyesh/1179.

возникающей под влиянием звука. В разделе отдельно анализируются различия между письменной и устной речью, неполное выражение их просодических особенностей посредством пунктуации и ограниченное использование знаков препинания, особенно в экспрессивных, эмоциональных или функциональных стилистических ситуациях речи. В статье рассматривается нормативная и возможная структура отражения просодических элементов посредством знаков препинания. Это указывает на необходимость всестороннего охвата не только грамматических, но и стилистических и коммуникативных функций пунктуации. В данном разделе подчеркивается необходимость изучения пунктуации не только как средства традиционных грамматических норм, но и как методологического, прагматического и познавательного инструмента. На основе этого подхода рассматриваются возможности использования интерактивных технологий обучения, лингводидактических моделей, мультимедийных элементов, электронных программ и средств дистанционного обучения для преподавания знаков препинания и их эффективного применения в образовательном процессе. Также возрастает потребность в научно-практических разработках по внедрению упражнений по пунктуации в автоматизированную систему проверки на современных платформах электронного обучения.

В этом разделе также освещаются проблемы пунктуации, возникающие при транскрипции спонтанной речи в письменное выражение, роль знаков препинания в синтаксических функциях восклицаний и звукоимитирующих единиц. Акцент делается на необходимости пересмотра существующих норм для решения этих проблем, включая анализ пунктуации, основанный не только на формальных структурах, но и на таких факторах, как коммуникативная цель речи, намерение высказывания и отношение субъекта.

Вторая глава диссертации называется **“Формализация в лингвистике – как метод теоретического исследования”** и состоит из четырех разделов. Первый раздел называется **“Понятие формализации в лингвистике”**. В этом разделе анализируется содержание понятия формализации, его исторические корни в лингвистике, этапы формирования и его место в современной лингвистике. Подчеркивается неразрывная связь формальных подходов с семиотическими, математическими и информационными теориями, а также возможности моделирования языковых единиц с помощью этих подходов. В этом разделе раскрывается теоретическая основа лингвистических моделей, которые служат для выражения языковой системы в ясной, непротиворечивой и воспроизводимой форме. Вопрос формализации рассматривается не только с точки зрения философии лингвистики, но и как общая методологическая проблема. Он составляет методологическую основу современной лингвистики и непосредственно связан, в частности, с математической логикой, искусственными языками и информатикой[7].

[7] Ветров А.А. Методологические проблемы современной лингвистики. – Москва, 1973. – С. 4.; Шрейдер Ю.А. О понятии Математическая модель языка. – Москва: Знание. 1971. – 40 с.; Большая советская энциклопедия. в 30-ти т.. – 3-е изд. – Москва, 1977. – С. 540-541.; Новейший философский словарь / сост. и

Формализация позволяет моделировать языковые процессы в точных и согласованных формах, выражая их посредством системы искусственных символов. Этот метод обеспечивает согласованность, обработку и автоматизацию языковых знаний. Формальные системы основаны на точных символах (алфавите), правилах построения формул (синтаксисе) и системе придания значения (семантике). С их помощью моделируются единицы фонологического, грамматического и семантического уровней языка. Поэтому формализация является важным инструментом для уменьшения неопределенностей естественного языка, создания искусственных языков, которые широко используются в алгоритмическом анализе и процессах программирования.

В лингвистических источниках понятие формализации широко интерпретируется в рамках различных подходов — генеративной грамматики, структурализма, метаязыков, математической и вычислительной лингвистики[8]. Благодаря таким подходам формализация позволяет организовывать научные теории в конкретные структуры и систематизировать связи между языковыми объектами. Поэтому она считается неотъемлемой частью методологии научных исследований в современной лингвистике.

Слово “формализация” на узбекском языке означает “систематизировать”, “стандартизировать”. Это процесс уточнения и формализации теоретических систем путем формулирования теоретических концепций, теоретических моделей или математических правил, в котором большое значение придается форме их выражения, и в то же время эти системы классифицируются или обрабатываются в соответствии со строгими теоретическими требованиями. В то же время, поскольку математическая логика и кибернетика определили актуальность философско-эпистемологического вопроса, формализация важна в различных дисциплинах, включая лингвистику[9] .

На основе вышеуказанных источников и специальной литературы стало ясно, что понятие “формализация” изучалось в науке с философской, логической, математической и кибернетической точек зрения. Формализация – это процесс уточнения содержания, осуществляемый путем определения его формы и включающий описание определенного процесса, который служит основой для построения математического описания. Этот метод также используется в лингвистике. В лингвистике формализация определяется как понятие, которое “вытекает из формы, основывается на свойствах выражения и служит формой”. Формализация языка – это набор правил, допускающих их механическое применение и позволяющих применять результаты лингвистических знаний на практике. Формализация

гл. науч. ред. А.А.Грицанов. – Минск: Интерпрессервис : Кн. Дом, 2001. – С. 772.

[8] Рузавин Г.И. 1997. Логика и аргументация: Учебное пособие. – Москва, 1997. – С. 35-36; Рузавин Г.И. Математизация научного знания. – Москва, 1984. – 160 с.; Лингвистический энциклопедический словарь / Глав. ред. В.НЛрцева. – Москва: Советская энциклопедия, 1990. – С. 98-99.

[9] Бирюков Б.В.Кибернетика и методология науки. – Москва: Нука, 1974. – 170 с.

также служит основой для процессов алгоритмизации, программирования и компьютеризации языкового материала.

Второй раздел называется **“Новые тенденции в формализации”.** В нем анализируются новые научные взгляды и тенденции в лингвистике XXI века, в частности, новые подходы к формализации в когнитивной лингвистике, корпусной лингвистике и вычислительной лингвистике. Также описываются тенденции в моделировании динамики речевого процесса и отражении функционально-семантических структур в формальных рамках.

Формализация в компьютерных технологиях составляет теоретическую основу стремительно развивающегося в настоящее время компьютерного моделирования. Она объединяет достижения в математическом моделировании и системном программировании, открывая путь к значительному расширению возможностей формализации знаний в будущем[10]. Идея использования электронных компьютеров (ЭК) в анализе текста привела к развитию математических и прикладных лингвистических методов, основанных на формализации задач и материалов. Одной из важнейших особенностей автоматизации лингвистических исследований сегодня является создание автоматизированных систем обработки естественного языка[11].

Процесс автоматического редактирования текста–это один из видов автоматизированной обработки, при котором в текст, хранящийся в памяти компьютера, вносятся соответствующие изменения и дополнения. Текстовые редакторы используются как средство подготовки и ввода в компьютер программ, программных документов, научных отчетов и другой информации.

Неформализованные задачи, в отличие от формализованных, не подразумевают алгоритмического или систематического (точного) подхода и характеризуются одной или несколькими из следующих особенностей:

1. Алгоритм решения задачи неизвестен или не может быть использован из-за ограниченных вычислительных ресурсов.

2. Задача не может быть выражена численно.

3. Невозможно выразить цели и задачи в виде четко определенных целевых функций. [12].

Формализация обеспечивает процесс автоматической обработки текста. В этом процессе проявляется новый метод формализации — компьютерная формализация, являющаяся основой компьютерного моделирования. Существование автоматизированных систем в издательском деле обусловлено тем, что компьютеры обрабатывают текстовые данные в соответствии с заданными алгоритмами.

---

[10] Юрченко В.В. Функциональный подход к формализации знаний. // Искусственный интеллект и проблемы организации знаний. – Москва, 1991. – 132 с.

[11] Асиновский А.С., Кузнецова Е.Ж. и др. К вопросу об автоматизации лингвистических исследований // Вопросы языкознания. 1986. – № 4 , – 86 с.

[12] Корнеев В.В., Гареев А.Ф. и др. Базы данных. Интеллектуальная обработка информации. – Москва: Нолидж, 2000. 127 с; Попов Э.В. Экспертные системы: Решение неформализованных задач в диалоге с ЭВМ. – Москва: Наука, 1987. – С. 8.

Таким образом, было установлено, что формализация, моделирование, математическое обоснование и алгоритмизация языкового материала необходимы для автоматизации этого стандарта пунктуации.

В главе **“Важность формализации в лингвистике и анализ знаков препинания”** подробно обсуждаются аналитические преимущества метода формализации при изучении различных уровней языковой системы– морфологических, синтаксических, семантических и прагматических структур. В частности, дается тщательный научный анализ на примере знаков препинания на основе методов формальной классификации, согласованности с грамматической структурой и контекстно-ориентированного моделирования. В лингвистике формализация как теоретический метод исследования играет важную роль в понимании и интерпретации изменяющихся и динамических свойств языка. Этот метод предполагает моделирование языка в соответствии с определенными нормами и правилами. Формализация четко определяет направление исследований в лингвистике и позволяет анализировать языковые единицы структурно и функционально. В то же время такой подход позволяет выявлять многофункциональные и прагматические особенности языка и организовывать их модельный анализ.

Знаки препинания в нашем языке, безусловно, являются частью графической структуры языка и средством ясного, понятного и беглого выражения мыслей в письменной речи. С течением времени сформировались и развились методы и функции использования этих знаков. Поэтому в лингвистике система знаков препинания, динамика их изменения и налагаемые на них правила могут быть определены с помощью формализации. Это важный фактор в правильной интерпретации текста, устранении ошибок и формировании высокоуровневой речевой культуры.

Таким образом, изучение знаков препинания тесно связано с новыми тенденциями формализации в лингвистике, а их интерпретация и глубокое понимание являются одной из приоритетных задач теоретических исследований. В этом процессе требуется совместное развитие грамматических, орфографических, синтаксических и пунктуационных норм языка и комплексное изучение всех их аспектов. Новые направления исследований системы пунктуации служат для выявления взаимосвязи правил, практических проблем знаков и возможностей их автоматизации.

Обозначение знаков препинания с помощью формальных тегов не только повышает эффективность процесса редактирования, но и создает возможность для совершенствования методологии, углубления знаний и автоматизации анализа в образовательном процессе. Такой подход предлагает многофункциональные решения для связи лингвистических знаний с практическими областями, расширяя возможности применения учебно-редакционной деятельности, особенно в рамках вычислительной лингвистики.

Формализация процесс преобразования языковых единиц в технические, вычислимые структуры – открывает большие возможности в

лингвистическом анализе, обработке естественного языка (NLP) и корпусных исследованиях. Однако этот процесс также сопряжен с определенными рисками. В частности, существует вероятность отрицания естественности, индивидуальных особенностей и стилистической свободы языка. В конце концов, любой язык –это не просто структура, основанная на правилах, но и живая система, включающая в себя личный стиль, творческие исследования и контекстно-зависимые значения.

Глава под названием **“Потенциал вычислительной лингвистики в моделировании знаков препинания”** дает широкий обзор методологических подходов к анализу и автоматизации пунктуации с использованием инструментов вычислительной лингвистики.

При создании системы автоматического распознавания знаков препинания на узбекском языке необходимо четко определить ряд задач. Каждая из этих задач направлена на обеспечение высокой лингвистической и технологической точности, и к ним относятся:

1. Обнаружение знаков препинания и определение их типа.

Прежде всего, автоматическая система должна уметь обнаруживать наличие знака препинания в тексте и правильно определять его тип (точка, запятая, точка с запятой, запятая и т. д.). Этот процесс осложняется тем, что он основан на порядке слов, синтаксической структуре и контекстной информации.

2. Оценка соответствия знаков цели речи.

Знаки препинания используются не только на основе грамматических критериев, но и в зависимости от цели речи. Поэтому автоматическая система обнаружения должна уметь анализировать соответствие указанного знака препинания контексту. Это особенно важно, когда знаки препинания используются по-разному в текстах разных стилей.

3. Различать факультативные, обычные и стандартные знаки препинания.

В то время как некоторые типы знаков препинания используются в соответствии с обычными или стандартными правилами, в некоторых случаях они используются добровольно, как стилистический выбор автора. Существуют также случаи, распространенные в практическом языке, но не полностью соответствующие грамматическим стандартам. Автоматическая система должна уметь различать эти три типа случаев и выполнять соответствующую маркировку.

В автоматическом определении знаков препинания в узбекском языке существует ряд лингвистических и технических проблем, которые сгруппированы следующим образом:

1. Свобода и фрагментация в структуре предложения.

Относительная свобода порядка слов в узбекском языке, вариативность расположения частей речи в предложении усложняют автоматический синтаксический анализ. Кроме того, во многих случаях авторы используют фрагментарные конструкции в структуре предложений (т.е., дополнительные

комментарии, склонения, не разделенные знаками препинания), что делает автоматический выбор знаков препинания неоднозначным.

2. Точки и сокращения в словах.

Например, в таких сокращениях, как “т.” (из), “ш.” (город), “д.” (минута), точка не является отдельным знаком препинания, а частью словоформы. В таких случаях программа проверки пунктуации может воспринимать её как точку, что может привести к неоднозначности. Для решения этой проблемы рекомендуется использовать специальные правила или регулярные выражения в процессе токенизации.

3. Различия между литературными, журналистскими и интернет-текстами в корпусе. Различия в стилях и жанрах также оказывают существенное влияние на автоматический анализ знаков препинания. Например, в интернет-текстах (форумах, социальных сетях) часто встречаются случаи неуместного или чрезмерного использования знаков препинания. В литературных текстах знаки препинания могут использоваться в ограниченном объеме в стилистических целях. Автоматическая система, не учитывающая эти случаи, может некорректно отражать особенности, присущие различным жанрам.

При оценке автоматических систем распознавания знаков препинания обычно используются стандартные метрики, такие как точность, полнота и F1-мера. Эти метрики позволяют оценить точность и полноту знаков препинания, обнаруженных автоматической системой:

1. Точность – показывает, сколько из автоматически обнаруженных знаков препинания являются действительно правильными.

2. Полнота – измеряет, сколько из всех реальных знаков препинания было обнаружено автоматически.

3. F1-мера – это сбалансированный показатель, сочетающий в себе точность и полноту.

Для автоматизированных систем размеченные корпуса должны быть не только большими по размеру, но и иметь стандартизированный формат. Поэтому крайне необходимо создать стандарты разметки XML/TEI для узбекского языка, работающие с такими тегами, как <пункт>, <запятая>, <точка>. Третья глава диссертации называется **“Формализация знаков препинания на основе корпуса и технология обнаружения ошибок”** и состоит из трех разделов. Первый раздел главы называется **“Формальная классификация нормативных и обычных функций знаков препинания в корпусе”**. В этом разделе рассматриваются функциональные задачи знаков препинания на основе синтаксических и грамматических норм, а также механизмы их автоматического обнаружения и разметки, основанные на научном подходе.

Грамматические функции знаков препинания проявляются прежде всего в синтаксической структуре, интонации и полном и ясном выражении мысли. Например, такие знаки, как точка, запятая, точка с запятой, кавычка, тире и двоеточие, выражают определенные грамматические нормы. Эти знаки позволяют логически связывать, разделять, организовывать и читать речь

четко и без ударения. Задача автоматического распознавания знаков препинания и их маркировки в корпусе тесно связана с компьютеризацией языка. В этом процессе в корпусной лингвистике знаки препинания классифицируются по форме и функции с использованием знаков препинания (например, <.>, <,>, <:>, <...>). Автоматическое распознавание учитывает положение знака, группы слов в нем и контекстную функцию. Это позволяет алгоритмизировать нормативную пунктуацию.

Нормативные функции могут также служить поэтическим целям. В частности, в литературном тексте использование точек, запятых, скобок, тире, вопросительных и восклицательных знаков раскрывает эстетические функции автора, такие как выражение эмоций, интонация, паузы и драматизм. Например:

Seni ko‘zg u qilib quyoshga tutgan
Mening xudoxohish yuragim uchun.
Har bir lahza uchun qon yutayotgan,
“Seni… Seni…” – degan tilagim uchun… (Z.Mirzo) →
<ellipsis_emotive>

Здесь герой не может говорить из-за чрезмерных эмоций. Это наглядный образ лирического возбуждения.

Siz – qimmat bahosiz, Men – arzon, arzon.
Sizga, balki, faqat yuragim arzir. (U.Azim)

В этом примере тире выполняет функцию сравнения и противопоставления притяжательного падежа и причастия. Такие утверждения, сделанные с логическим акцентом, могут быть отмечены в корпусе тегом <emph_eq_taqqos>.

Kim darg‘azab, Kim hayron,
Chiqib keldilar qator: Dante, Shiller va
Bayron, Firdavsiy, Balzak, Tagor.
(E.Vohidov)

В этом примере двоеточие ставится в конце предложения и служит для выделения и описания имен без использования обобщающих слов. В этом случае можно использовать тег <inverted_colon_list>.

Конструкции, приведенные в скобках, служат для усиления художественной силы текста, особенно в поэзии, за счет образности, символизма и эмоционального воздействия:

Sevgilim, sevgini sevmagan zotlar
Qaro aylasalar oq asotirni,
(Darg‘azab pishqirsa quturgan otlar)
Go‘rdan chiqarsalar Qaro botirni… (I.Mirzo)

В данном случае скобки используются для дополнения изображения, усиления воображения и страсти и могут быть помечены тегом <poetic_parenthesis>.

Второй раздел называется **“Корпусная идентификация знаков препинания в современной узбекской литературе”** и анализирует выявление устойчивых, хотя и распространенных, применений в письменных текстах, не отклоняющихся от нормативных границ, а также вариантов пунктуации, используемых в поэтическом и экспрессивном контекстах.

Для автоматической идентификации случайных функций необходим статистический, синтаксический и контекстуальный анализ данных, полученных из корпуса. Для этого корпусные алгоритмы обнаружения работают в следующих этапах:

1. Формирование условных шаблонов, определяющих нормативные функции: например, когда точка используется для паузы, незавершенной мысли или списка – это определяется как основная модель.

2. Выявление отклонений от шаблона: Если точка стоит в середине или конце предложения, не в синтаксическом или стандартном месте, а в ситуации, связанной с эмоциями, это помечается как случайный» случай.

3. NLP-теги, связанные с интонационными маркерами: Например, функции, идентифицирующие элементы пунктуации, связанные с такими тегами, как intensity_high, interruption, mental_state.

4. N-грамный контекстный анализ: Функция пунктуации определяется словами, расположенными рядом со знаком препинания (особенно эмоциональными словами: “vayy”, “ha”, “o-ox”, “havas”, “afsus” и т. д.).

5. Средства, представляющие пространство и ритм: Алгоритм оценивает уровень выразительности таких символов, как точки и тире, на основе частоты их повторения.

Третья глава называется **“Создание программы для обнаружения пунктуационных ошибок на основе идентифицированных функций и корпусных данных”**. В ней разрабатывается функциональная модель, основанная на автоматическом анализе, показаны механизмы в процессе распознавания нормативных, обычных и необычных символов как ошибок, а также обосновывается алгоритм программы, созданный на основе корпусных данных. Таким образом, третья глава охватывает не только теоретическую классификацию знаков препинания, но и возможности их автоматической обработки в практической системе. В автоматическом анализе знаков препинания важным этапом является их идентификация путем тегирования. В этом процессе становится возможным автоматическое определение их функций путем сопоставления их с грамматической и синтаксической информацией. Большинство систем обработки естественного языка используют POS-теги (части речи), синтаксические структуры и элементы синтаксического анализа зависимостей.

В процессе POS-тегирования знакам препинания присваивается отдельный тег (например, **“**PUNCT”). При синтаксическом анализе определяется их положение в предложении и точки соединения. С помощью инструментов синтаксического анализа зависимостей определяется, к какому слову относится знак препинания и какое синтаксическое отношение он выражает[13].

N-граммовые модели эффективны для предсказания того, какой знак, скорее всего, появится в данном контексте (или какова функция результирующего знака), путем вычисления вероятностей комбинаций слов или знаков. В этой модели знаки препинания, такие как точка (“.”),

[13] Chen D., Manning C. D. A Fast and Accurate Dependency Parser using Neural Networks // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). – Doha: ACL, 2014. – P. 740-750. https://doi.org/10.3115/v1/D14-1082

восклицательный знак (“!”), двоеточие (“...”) и вопросительный знак (“?”), имеют одинаковую формальную структуру, но их контекстно-зависимые функции могут быть различными. Например:

Точка – конец предложения, описывающий ситуацию, завершающий мысль;

Вопросительный знак – просьба о предоставлении информации, риторический вопрос, сомнение;

Восклицательный знак – сильная эмоция, удивление, эмоциональная форма приказа;

Несколько точек – пауза в речи, оставляющая мысль открытой, драматическая пауза.

В практическом анализе во многих случаях функции определяются анализом того, какое слово или группа слов стоит до и после определенного знака. Например, модель [существительное/местоимение + глагол + .] означает конец предложения, модель [междометие + !] – эмоциональную реакцию, а модель [незаконченная фраза + ...] – паузу в речи и драматическую паузу.

Возможности языка Python для автоматического анализа пунктуации весьма обширны. Он широко использует различные библиотеки обработки естественного языка (spaCy, NLTK), pandas для анализа данных и регулярные выражения (regex) для обнаружения знаков препинания.

1. Оценка функции символа с помощью блоков if–elif. Условные операторы используются для оценки функции знака препинания в предложении. Например:

```
def evaluate_punctuation(punct, context):
    if punct == '?':
        if 'ми' in context or context.strip().endswith('ми'):
            return 'Savol belgisi - to‘g‘ri'
        else:
            return 'Savol belgisi - noto‘g‘ri yoki noadatiy'
    elif punct == '...':
        return 'Ko‘p nuqta - ekspressiv yoki okkazional holat'
    elif punct == ',':
        if 'agar' in context.lower():
            return 'Vergul - shartli bog‘lovchi oldida - to‘g‘ri'
        else:
            return 'Vergul - ehtimoliy xato'
    else:
        return 'Aniqlanmagan belgi'
```

“Рисунок 3.3.1” Архитектура алгоритма

2. Распознавание конструкций с помощью регулярных выражений. Распознавание определенных структур возможно с помощью регулярных выражений (модуль `re`). Например, чтобы распознать конструкцию в форме

вопросительного предложения:

```
import re

text = "U keladimi?"
pattern = r'\b\w+ми\?'

if re.search(pattern, text):
    print("Savol konstruktsiyasi to‘g‘ri")
else:
    print("Ehtimoliy xato")
```

“Рисунок 3.3.2” Архитектура алгоритма.

3. Сравнение данных корпуса с данными из pandas или CSV-файлов. Программа должна сравнивать входной текст с нормативным приложением в корпусе. Например:

```
import pandas as pd

# Корпус маълумотини юклаш
df = pd.read_csv("corp_norms.csv")  # устунлар: context, punctuation, status

# Фойдаланувчи матни
user_text = "U ketdi..."

# Таққослаш
matches = df[df['context'].str.contains("ketdi", case=False)]
if not matches.empty:
    print("Ushbu kontekstda '...' ekspressiv deb tan olingan")
else:
    print("Korpusda mos holat topilmadi – ehtimoliy xato")
```

“Рисунок 3.3.3” Архитектура алгоритма.

Оценка результатов программы (точность, полнота, F1-мера)

Классическими метриками НЛП для оценки эффективности программы автоматического распознавания знаков препинания являются точность, полнота и F1-мера.

Четвертая глава называется **“Современные методы восстановления знаков препинания в узбекском языке на основе языковых моделей”** и состоит из четырех разделов. Первый раздел главы называется **“Модели естественного языка”.** Языковые модели играют важную роль в процессе автоматической обработки естественного языка (обработка естественного языка, НЛП). Языковые модели позволяют анализировать текст и генерировать новый текст, присваивая вероятности последовательностям символов или слов. В последние годы разработка моделей на основе искусственных нейронных сетей, наряду со статистическими подходами, повысила эффективность различных задач НЛП. В данной статье анализируются теоретические основы статистических языковых моделей (N-граммы, марковская модель), критерии их оценки и результаты практических исследований с использованием пакета UzbNlp на основе национального корпуса узбекского языка.

Языковая модель (ЯМ) описывает вероятностную структуру текста. Языковые модели можно разделить на два основных типа:

1. Статистические языковые модели (СЯМ) – основаны на N-граммовых моделях.

2. Нейронные языковые модели (НЯМ) – созданы с использованием методов глубокого обучения.

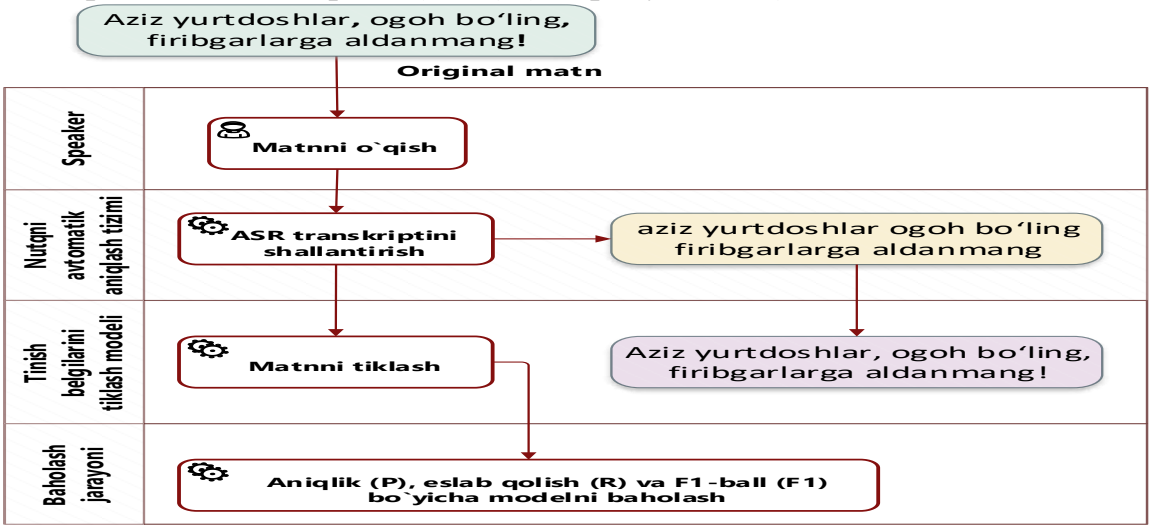
N-граммовые модели вычисляют вероятность следующего слова в последовательности на основе предыдущего слова (слов). Марковские модели учитывают только контекст определенной длины (биграмма, триграмма и т. д.).

Эффективность языковых моделей определяется с помощью методов внутренней и внешней оценки. Перплексия является основным критерием внутренней оценки. Например, в корпусе узбекского языка индекс перплексии биграммной модели значительно ниже, чем у униграммной модели, что указывает на высокую точность модели.

Проблемы языковых моделей: Обобщение и чувствительность – модель сильно зависит от обучающего корпуса; Неизвестные слова (OOV) – единицы, не встречающиеся в корпусе; Контекстная чувствительность – требует применения методов сглаживания и дисконтирования.

Второй раздел главы называется **“Методы восстановления знаков препинания в тексте в системах распознавания речи”**. Системы автоматического распознавания речи (ASR) в последние годы широко используются в различных областях. В частности, ASR приобретает все большее значение в автоматической транскрипции текста, мультимодальном анализе, подготовке субтитров, синтезе речи и интерактивных приложениях. Однако транскрипции, полученные с помощью систем ASR, грамматически неполны: они не сохраняют регистр букв, теряется структура абзацев, и, что наиболее важно, отсутствие знаков препинания затрудняет чтение текста и уточнение смысла.

Цель задачи обработки естественного языка (NLP) по поиску знаков препинания в тексте состоит в том, чтобы восстановить знаки препинания в системе распознавания речи (ASR) (см. рисунок 4.1).



“Рисунок 4.1” Архитектура задачи обработки естественного языка по расстановке знаков препинания в тексте.

Очевидно, что восстановление знаков препинания не только улучшает

читабельность и содержание текста, но и положительно влияет на эффективность последующих задач обработки естественного языка (синтаксический анализ, машинный перевод, системы вопросов и ответов и т. д.).

UzbPunct – это набор текстовых и аудиоданных, созданный сотрудниками кафедры вычислительной лингвистики и цифровых технологий Тошдотауского университета. Набор данных разделен на две части: **разговоры (UzbTalks)** и **новости (UzbNews).** В создании аудиокомпонента участвовало более ста сотрудников, исследователей, аспирантов и студентов. Общая длина (объем) аудиоданных, включая тестовый набор, достигает почти **тридцати шести часов.** Были приняты меры для обеспечения сбалансированного соотношения мужчин и женщин при формировании аудиоданных. База данных UzbPunct содержит более 32 000 текстовых файлов и 1200 аудиофайлов, из которых 1000 входят в обучающий набор и 200–в тестовый. Для каждого текста имеется **расшифровка** автоматически распознанной речи и текст, выровненный вручную. Информация о формате данных и показателях оценки представлена в следующих разделах абзаца.

Статистика набора данных UzbPunct следующая:

- **Текст:** 32 000 текстов из корпуса узбекского языка;

- Аудио:

Процедура отбора:

случайно отобранные UzbNews (80%, т.е. 800 записей для обучающего набора) с количеством слов в диапазоне [150-300];

случайно отобранные UzbTalks (20%) с количеством слов в диапазоне [150-300] и предложением, содержащим хотя бы один вопросительный знак;

Разделение набора данных:

Обучающие данные: 1000 записей;

Тестовые данные: 274 записи;

Спикеры:

Мужчины: 51 лекторов, 16,7 часов речи;

Женщины: 54 лектора, 19 часов речи;

Представлены статистические данные по пунктуации для неструктурированного (необработанного) текста (см. Таблицу 4.1).

“Таблица - 4.1”

Статистика пунктуации для неструктурированного текста.

Tinish belgisi	yozilishi	mean	median	Max	Sum	Included
Fullstop	**.**	12.44	7.0	1129.0	404378	Yes
Comma	**,**	10.97	5.0	1283.0	356678	Yes
question mark	**?**	0.83	0.0	130.0	26879	Yes
exclamation mark	**!**	0.22	0.0	55.0	7164	Yes
Hyphen	**–**	2.64	1.0	363.0	81190	Yes
Colon	**:**	1.49	0.0	202.0	44995	Yes
Ellipsis	**...**	0.27	0.0	60.0	8882	Yes
Semicolon	**;**	0.13	0.0	51.0	4270	No

Quote	"	3.64	0.0	346.0	116874	No
Jami so‘zlar soni		**169.50**	**89.0**	**17252.0**	**5452032**	-

Как упоминалось выше, набор данных состоит из двух частей: **разговоры (UzbTalks)** и **новости (UzbNews).**

В ходе задачи распознавания речи (см. Таблицу 4.2) оцениваются следующие знаки препинания:

“Таблица - 4.2”

Знаки препинания и их обозначение.

№	**Tinish belgisi**	**Punctuation mark**	**Symbol**
1.	nuqta	Fullstop	**.**
2.	vergul	Comma	**,**
3.	so‘roq belgisi	question mark	**?**
4.	undov belgisi	exclamation mark	**!**
5.	tire	Hyphen	**-**
6.	ikki nuqta	Colon	**:**
7.	uch nuqta	Ellipsis	**...**
8.	nuqtali vergul	semi-colon	**;**
9.	bo‘sh joy	blank(no punctuation)	

Результатом анализа является текст, содержащий только знаки препинания.

Окончательные результаты оцениваются с использованием показателей точности, полноты и F1-меры для прогнозирования каждого знака препинания отдельно. Представленные значения сравниваются со средневзвешенным значением F1-меры для каждого знака препинания.

Оценка документа:

$$\text{Точность } (\mathbf{precision}) = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FP}};$$

$$\text{Запомнить } (\mathbf{recall}) = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FN}};$$

$$\mathbf{F1} = \mathbf{2} * \frac{\mathbf{Precision} * \mathbf{Recall}}{\mathbf{Precision} + \mathbf{Recall}};$$

Общий (p) балл для каждого знака препинания:

$$\mathbf{Precision}_{\boldsymbol{P}} = \mathbf{avg}_{\boldsymbol{micro}}\mathbf{Precision}_{\boldsymbol{P}} = \frac{\sum_{\boldsymbol{d \in Documents}} \boldsymbol{TP}}{\sum_{\boldsymbol{d \in Documents}} \boldsymbol{TP + FP}};$$

$$\mathbf{Recall}_{\boldsymbol{P}} = \mathbf{avg}_{\boldsymbol{micro}}\mathbf{Recall}_{\boldsymbol{P}} = \frac{\sum_{\boldsymbol{d \in Documents}} \boldsymbol{TP}}{\sum_{\boldsymbol{d \in Documents}} \boldsymbol{TP + FN}};$$

Итоговая оценка модели рассчитывается как взвешенное среднее глобальных оценок.

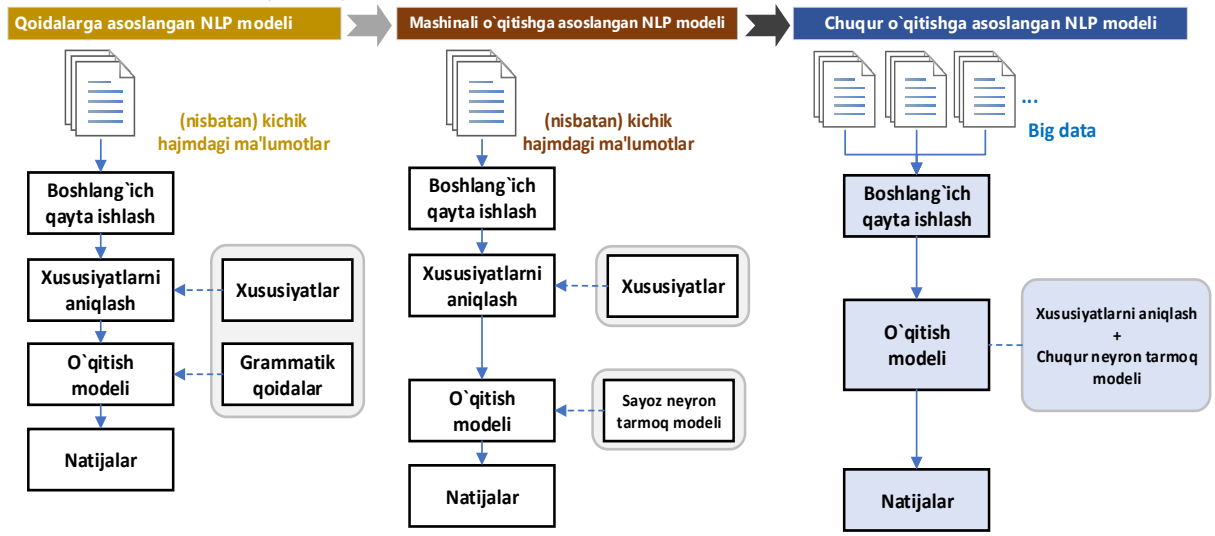
$$\frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{p \in Punctuation}} \boldsymbol{support(p)} * \mathbf{avg}_{\boldsymbol{micro}}\mathbf{F_1}(\boldsymbol{p})$$

Третий раздел главы называется **“Восстановление знаков препинания в узбекском тексте с использованием трансформерных моделей”**. В нем рассматривается проблема восстановления знаков препинания в узбекских

текстах в системах автоматического распознавания речи (ASR). Показана важность знаков препинания для определения читабельности и смысла текста. Также анализируется опыт использования трансформерных моделей в английском языке (BERT, RoBERTa и др.) и указывается, что для узбекского языка был создан специальный набор данных, и получены первые успешные результаты на основе модели UzbBERT. То есть, этот раздел посвящен проблеме автоматического восстановления знаков препинания для узбекского языка, ее решениям и предварительным научным результатам.

Ученые из отдела вычислительной лингвистики и цифровых технологий ТошДОТАУ разработали новые модели для восстановления знаков препинания в текстах на языках с ограниченными ресурсами, таких как узбекский. На основе ресурсов языкового корпуса UzbCorpus, разработанного ими, модели были обучены на различных вариантах модели BERT, а затем протестированы на оценочных текстах. В этом разделе рассматриваются различные модели на основе трансформеров и представлены методы решения этой задачи в текстах как на языках с большими ресурсами (английский), так и на языках с малыми ресурсами (узбекский). Результаты различных моделей BERT были проанализированы с точки зрения точности (P), полноты (R) и F1-меры (F1), и для восстановления знаков препинания в узбекских текстах была достигнута F1-мера 86,7%. Данное исследование может послужить основой для будущих работ по узбекскому языку.

Восстановление пунктуации с помощью машинного обучения включает в себя использование алгоритмов и статистических моделей для прогнозирования правильной пунктуации в заданном тексте. Этот процесс обычно включает три основных этапа: **подготовку данных, извлечение признаков и обучение модели.** На рисунке 4.2 ниже показаны модели восстановления пунктуации:



“Рисунок 4.2” Модели восстановления знаков препинания.

В задачах обработки естественного языка, таких как автоматическое распознавание речи (ASR), сгенерированный текст обычно не содержит

знаков препинания, что затрудняет его распознавание или **анализ.** Поскольку входной текст можно рассматривать как последовательность, где контекст важен для характеристик каждого слова, характеристики глубоких нейронных сетей являются хорошим подходом.

Для восстановления знаков препинания в узбекском тексте с помощью моделей Transformer мы отобрали статьи для подготовки обучающего (4000), оценочного (500) и тестового (200) наборов данных. Количество слов (токенов) в этих наборах данных составляет 1 380 000, 180 000 и 88 000 слов соответственно. Мы также подготовили два тестовых набора данных, состоящих из расшифрованных вручную и транскрипций ASR, для оценки модели. Для этого мы собрали 65 минут речевых отрывков из четырех узбекских рассказов (т.е. монологов, прочитанных вслух). Знаки препинания в них были расшифрованы вручную. Мы сгенерировали транскрипции ASR для аудиозаписей, используя **API speech Google Cloud.** API speech Google Cloud не предоставляет знаки препинания для узбекского языка. Следовательно, необходимо вручную разметить транскрипции, полученные с помощью API Google Cloud Speech. Показатель частоты ошибок распознавания слов (WER) полученных транскрипций ASR был рассчитан путем сравнения их с рукописными транскрипциями, и составил **14,8%.** Количество слов в транскрипциях, размеченных вручную, и в транскрипциях ASR составило 6821 и 6417 слов соответственно. В этом абзаце мы рассмотрим следующие четыре знака препинания в узбекском языке: **точка, запятая, вопросительный знак и Ø.**

В таблице 4.3 ниже представлено распределение тегов для английского и узбекского языков. Процент знаков препинания указан в скобках в этой таблице.

“Таблица 4.3”

Распределение наборов данных на английском и узбекском языках.

Dataset	Jami	Nutqa(.)	Vergul(,)	So‘roq(?)	Boshqa(Ø)
Ingliz tili					
O‘quv	2102417	132393(6.3%)	158392(7.53%)	9905(0.47%)	1801727(85.7%)
Baholash	295800	18910(6.39%)	22451(7.59%)	1517(0.51%)	252922(85.5%)
Test (Ref.)	12626	807(6.39%)	830(6.57%)	46(0.36%)	10943(86.67%)
Test (ASR)	12822	809(6.31%)	798(6.22%)	35(0.27%)	11180(87.19%)
O‘zbek tili					
O‘quv	1379986	98791(7.16%)	65235(4.73%)	4555(0.33%)	1211405(87.78%)
Baholash	179371	13161(7.34%)	7544(4.21%)	534(0.3%)	158132(88.16%)
Test (news)	87721	6263(7.14%)	4102(4.68%)	305(0.35%)	77051(87.84%)
Test (Ref.)	6821	996(17.6%)	279(4.09%)	170(2.49%)	5376(78.82%)
Test (ASR)	6417	887(13.82%)	253(3.94%)	125(1.95%)	5152(80.29%)

Распределение вопросительных предложений в приведенном выше наборе данных очень низкое (менее 1%), что объясняется тем, что набор данных в основном взят из новостного корпуса. Однако эти показатели имеют гораздо более высокие значения в узбекских и транскрипциях ACP. Это связано с тем, что эти тексты отобраны из коротких рассказов, в которых люди ведут беседу и задают друг другу вопросы. Поскольку литературный

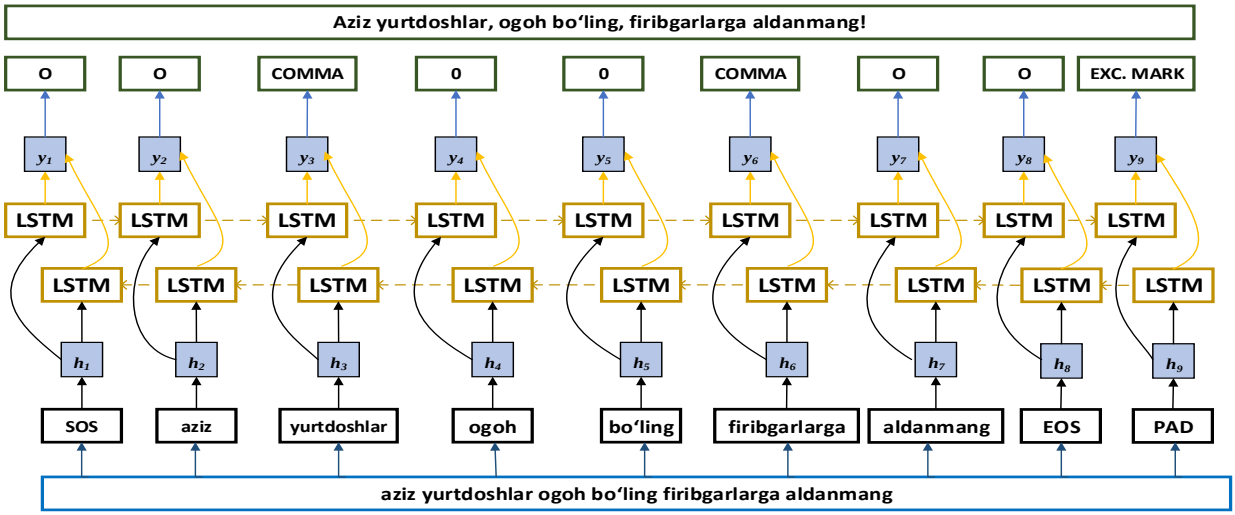
стиль рассказов отличается от новостного, распределение знаков препинания также выше в узбекских и транскрипциях ACP. Это, в свою очередь, приводит к гораздо меньшей средней длине предложения в наборе данных (см. Таблицу 4.4).

“Таблица 4.4”

Средняя длина предложения в наборе данных.

Dataset	Ingliz tili		O‘zbek tili	
	Avg.	Std.	Avg.	Std.
O‘quv	13.8	10.8	12.4	7.6
Baholash	13.5	10.7	12.1	7.2
Test (news)	-	-	12.4	7.2
Test (Ref.)	13.8	9.6	4.8	3.2
Test (ASR)	14.2	9.7	5.3	3.6

В обучающих данных, состоящих из устной речи, доля неизвестных токенов составляла 79,1%, тогда как в тестовых данных этот показатель достигал 78,82%. На основании этих результатов распределение наших транскрибированных данных больше похоже на естественные разговоры. Архитектура модели трансформера, использованной в экспериментах, представлена (см. рисунок 4.3).



“Рисунок 4.3” Общая архитектура модели трансформера для наших экспериментов.

Три типа заполнения, соответствующие трем возможным ошибкам, которые могут возникнуть в результате ASR, следующие:

1. **Первый** (т.е. замена), мы заменяем один токен другим токеном. В наших экспериментах мы случайным образом заменяем токен специальным неизвестным токеном.

2. **Второй** (т.е. удаление), мы случайным образом удаляем некоторые токены из обработанной входной последовательности.

3. **Наконец,** мы добавляем неизвестный токен в некоторую случайную позицию во входных данных (т.е. вставка).

Поскольку три упомянутые выше ошибки не одинаково распространены

в текстах корпуса, различные операции заполнения по-разному влияют на производительность модели. Поэтому для обработки заданного текста мы используем три регулируемых параметра:

1. **α** – параметр, определяющий вероятность изменения токена;
2. **αsub** – параметр, контролирующий вероятность замены;
3. **αdel** – параметр, контролирующий вероятность удаления.

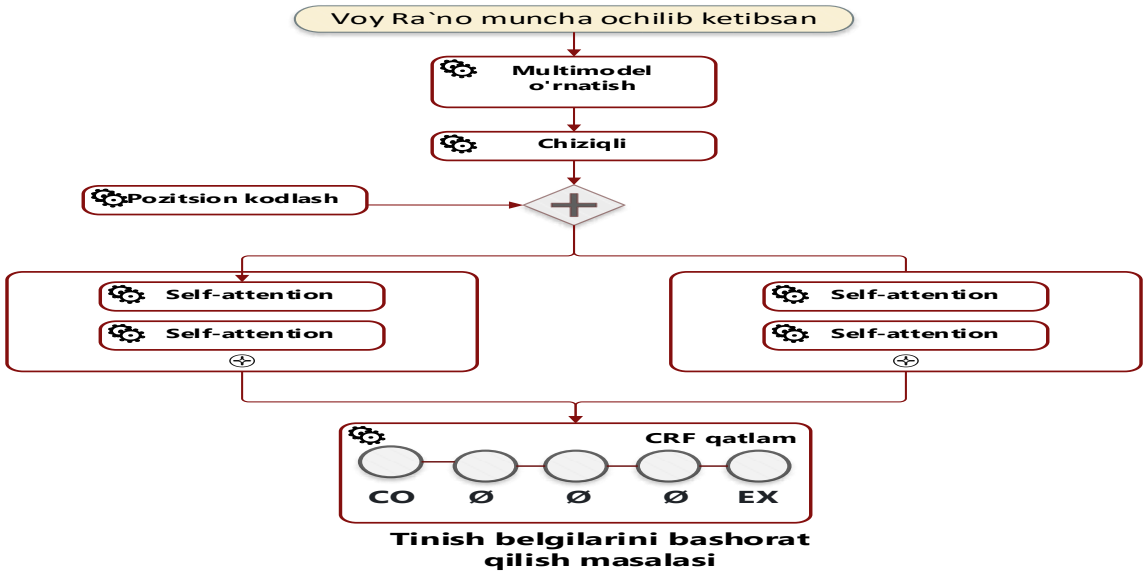
Вероятность вставки определяется выражением **1-(αsub+αdel).**

При применении *операции модификации текста* мы заменили токен в этой позиции неизвестным токеном, оставив целевой знак препинания без изменений.

Четвертый раздел главы называется **“Знаки препинания в узбекских текстах с использованием POS-тегирования”.** Вопрос восстановления знаков препинания является одной из широко изучаемых проблем обработки естественного языка в рамках систем автоматического распознавания речи (ASR). Технология ASR – это программный инструмент, позволяющий компьютерам распознавать человеческую речь и преобразовывать ее в текст. Актуальность этой проблемы заключается в том, что устные тексты формируются без знаков препинания, что создает определенные трудности на последующих этапах – синтаксическом анализе, переводе или семантическом моделировании.

Этап декодирования и показатели оценки

На этапе декодирования задача разметки частей речи и задача классификатора исключаются. Однако на этом этапе используется задача прогнозирования пунктуации (см. рис. 4.4). Таким образом, модель может использовать инвариантные синтаксические особенности задачи восстановления пунктуации в задаче разметки частей речи без выполнения дополнительных вычислений.

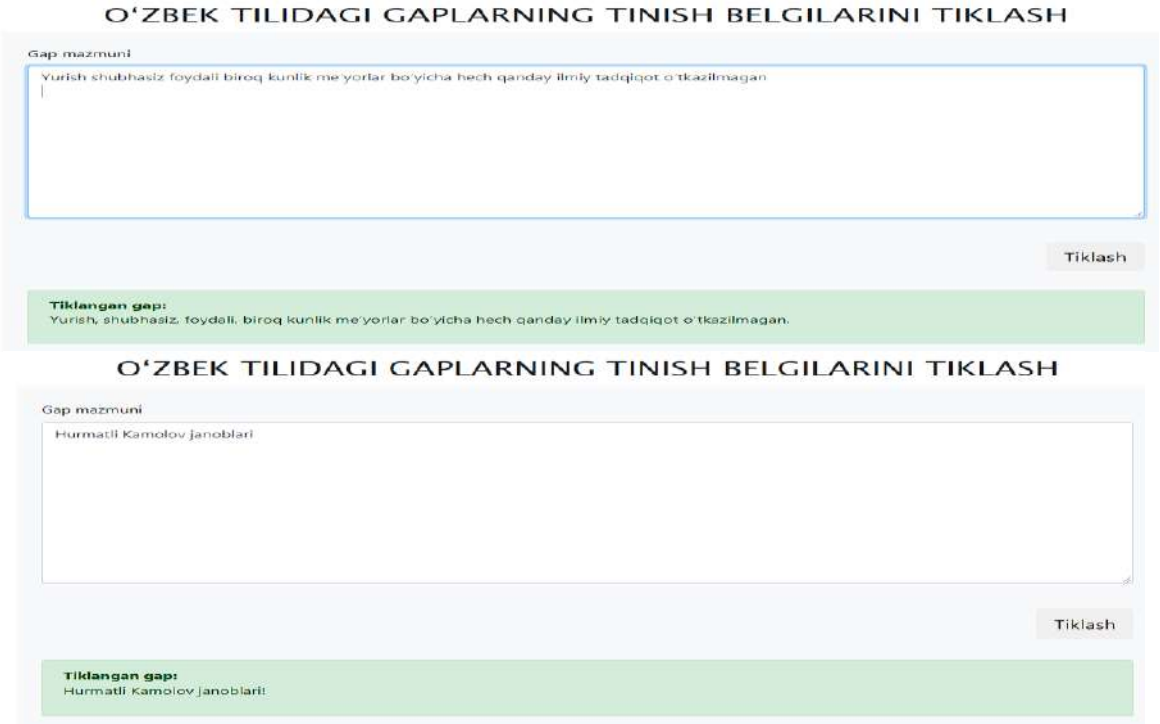


“Рисунок 4.4” Архитектура предлагаемой модели на этапе декодирования.

Алгоритм Витерби (Sammut и Webb, 2010) может быть использован для

обнаружения последовательностей знаков препинания с высокой вероятностью. Модели во всех экспериментах оцениваются по показателю **F1-score (F1)**. Показатель F1-score рассчитывается на основе значений **точности (P)** и **полноты (R)**. Мы фокусируемся на правильном использовании знаков препинания в тексте. Таким образом, правильно предсказанный знак препинания “Ø” игнорируется. Мы оцениваем производительность только трех знаков препинания: **запятой (CO), точкой (PE) и вопросительным знаком (QU).** Все эксперименты были протестированы на двух тестовых наборах данных: **Ref**. и **ASR.** Для задач предсказания знаков препинания и определения частей речи используется набор данных из Национального корпуса узбекского языка (UzCorpus).

Скриншоты разработанного программного обеспечения (https://punct-restorer.uznlp.uz/) (см. рисунок 4.5)



“Рисунок 4.5” Программное обеспечение.

ВЫВОДЫ

1. Постепенная классификация исторического развития пунктуации в узбекском языке и выявление различий по периодам способствуют созданию нового теоретического ландшафта в лингвистике. Этот процесс позволяет изучать влияние внешних факторов – арабской, персидско-таджикской и русской лингвистики – наряду с внутренними законами формирования национальной системы пунктуации. Сравнительный анализ функций знаков препинания в разные периоды позволяет теоретически выявить, как формировались грамматические и семантические нормы письменной речи и какие изменения они претерпевали. Одновременно эволюция системы пунктуации открывает широкий спектр возможностей для сравнительных исследований с точки зрения общей тюркской лингвистики, а также для сравнения с общими законами и нормами мировой лингвистики. Таким

образом, изучение системы пунктуации на историческом уровне демонстрирует необходимость изучения системы пунктуации национального языка не только как практического правила, но и как самостоятельного и методологически значимого направления лингвистической теории. Такой подход позволяет проводить систематический анализ роли пунктуации в языковой структуре, ее функций и исторической эволюции.

2.Пунктуация является важным инструментом обеспечения грамматического порядка письменной речи, определения смысловых и стилистических аспектов, а также уточнения коммуникативной цели. Поскольку система пунктуации в узбекском языке формировалась под влиянием исторического развития и других языков, актуальной задачей является совершенствование её нормативных основ и глубокое изучение её функциональных свойств. Кроме того, вопрос автоматического размещения знаков препинания является синтаксически и семантически сложным и требует дальнейшего совершенствования для достижения высокой эффективности в компьютерном анализе и технологиях обработки естественного языка.

3. Формализация имеет большое значение в лингвистике как методологическая основа для моделирования и анализа языковых процессов. Она служит для понимания внутренних законов языка, достижения синтаксической и семантической точности в письменной речи и совершенствования процессов автоматического анализа, машинного перевода и создания корпусов в вычислительной лингвистике.

4. Изучение знаков препинания посредством формализации, наряду с выявлением их грамматических, семантических и коммуникативных функций, позволяет эффективно решать такие задачи, как восстановление пунктуации в технологиях обработки естественного языка. Несмотря на морфологическую сложность и свободный порядок слов в узбекском языке, модели, такие как BERT и BiLSTM, показали высокие результаты в автоматическом размещении знаков препинания, послужив научной основой для создания национального корпуса и развития интеллектуальных систем.

5. Знаки препинания, наряду с их нормативными и формальными аспектами, также проявляются в корпусе через свои поэтические и выразительные функции, обеспечивая грамматическую точность, коммуникативную цель и художественный эффект письменной речи. Через них в тексте выражаются индивидуальные стилистические особенности, и очевидно, что пунктуацию следует изучать не только как формальный инструмент, но и как многогранный семиотический элемент.

6. На основе корпусной лингвистики и технологий обработки естественного языка становится возможным автоматический анализ формального и выразительного использования знаков препинания, эффективно решающий такие задачи, как выявление ошибок, определение стилистических различий и обогащение национального корпуса. Этот процесс демонстрирует необходимость изучения пунктуации на основе новых методологических подходов, сочетающих лингвистику и вычислительную лингвистику.

7. Систематическое изучение знаков препинания на основе корпусных материалов и применение к ним формального подхода позволяет глубже

понять грамматические и семантические законы языка. Такой подход создает основу для научного анализа роли пунктуации в языковой структуре, ее функций и эффективности в контексте текста. Пунктуационные средства теоретически обосновываются не только как нормативный элемент, обеспечивающий нормативный порядок письменной речи, но и как важный семиотический инструмент, формирующий духовную целостность, стилистические аспекты и коммуникативную эффективность текста. Поэтому в процессе создания компьютерной программы, выявляющей пунктуационные ошибки, формализация языковых законов и правил, их сочетание со статистическими данными, извлеченными из корпуса, и моделирование с использованием алгоритмов искусственного интеллекта теоретически формируются как самостоятельное научное направление.

8. Важным шагом для национальной лингвистики является тот факт, что технологии автоматического обнаружения пунктуационных ошибок и преобразования устной речи в письменную форму были впервые протестированы на примере узбекского языка. Этот процесс служит для обеспечения синтаксической и стилистической точности письменной речи, упрощения процессов редактирования и эффективного освоения правил языка в образовательном процессе. Кроме того, эти технологии обогатят национальный корпус, улучшат электронные словари и системы автоматического перевода, а также предоставят возможность высокоточной обработки узбекских текстов на медиа- и информационных платформах.

9. Языковое моделирование и системы автоматического распознавания речи (ASR) являются одним из основных направлений в области обработки естественного языка (NLP), целью которого является изучение законов и правил естественного языка на основе языковых корпусов и их применение в практических процессах. Путем выявления возможных связей между языковыми элементами на основе статистических и нейронных моделей эффективно решаются различные задачи, такие как преобразование аудио в текст, автоматический перевод, анализ настроений, системы вопросов и ответов и классификация текста. В частности, современные языковые модели, такие как GPT, BERT, Claude, Falcon, обучаются на больших корпусах и позволяют создавать осмысленные представления в различных контекстах с помощью метода маскированного языкового моделирования.

10. Одной из наиболее актуальных проблем в системах ASR является задача пунктуации, обеспечивающая читаемость и грамматическую точность письменного текста. В современных исследованиях для автоматического восстановления пунктуации используются такие модели, как CRF, RNN, CNN и трансформеры, а модели на основе BERT демонстрируют высокие результаты в идентификации последовательностей. В то же время, для повышения эффективности автоматического распознавания речи необходимо учитывать типы ошибок, принимать во внимание несоответствия между интерпретаторами, а также сдвиги пунктуации и ошибки принудительного отслеживания. Поэтому исследования в области языкового моделирования и автоматического распознавания речи имеют большое значение не только в теоретическом, но и в практическом плане, и играют решающую роль в совершенствовании интеллектуальных программ и различных приложений в повседневной жизни.

**SCIENTIFIC COUNCIL AWARDING
SCIENTIFIC DEGREES DSc. DSc.03/2025.27.12 Fil.08.07
AT BUKHARA STATE UNIVERSITY**

BUKHARA STATE UNIVERSITY

JUMAEVA DILNOZA BAKHSHULLOYEVNA

TECHNOLOGY FOR DEVELOPING A COMPUTER PROGRAM FOR IDENTIFYING PUNCTUATION ERRORS AND REPRESENTING PUNCTUATION MARKS IN A CORPUS

10.00.11 – Applied and computational linguistics.

**DISSERTATION ABSTRACT
for a Doctor of Philological sciences (DSc)**

Bukhara – 2026

The topic of the Doctor of Science (DSc) dissertation has been registered with the Higher Attestation Commission under the number B2024.4.DSc/Fil905

The dissertation has been prepared at Bukhara State University.

The abstract of the dissertation is posted in three languages (Uzbek, Russian, English (resume)) on the webpage (www.buxdu.uz) of Bukhara State University and on the "Ziyonet" information educational portal (www.ziyonet.uz).

Scientific supervisor:	**Toirova Guli Ibragimovna** Doctor of Science in Philology (DSc), professor
Official opponents:	**Gulyamova Shaxnoza Kaxramonovna** Doctor of Science in Philology (DSc), professor
	Elov Botir Boltayevich Doctor of Science in Technical (DSc), dotsent
	Xolmuxamedov Baxtiyor Farxodovich Doctor of Science in Philology (DSc), dotsent
Leading organization:	**Berdakh Karakalpak State university**

The dissertation defense will be held on "5" september 2026 at 9.00 at the meeting of the Scientific Council awarding Scientific degrees DSc.03/2025.27.12 Fil.08.07 at Bukhara State University. (Address:Bukhara city, 200118, M.Ikbol str., 11. Phone: +99865 221-29-14; Fax: +99865 221-27-07; E-mail: buxdu_rektor@buxdu.uz.)

The dissertation is available at the Information-resource center of Bukhara State University (registration No.998). (Address:Bukhara city, 200118, M.Ikbol str., 11. Phone: +99865 221-29-14.)

Dissertation abstract was distributed on "20" august 2026.

(Mailing protocol- register No. 27 of "20" august 2026.)

B.E. Kilichev
Chairperson of the Scientific Council awarding scientific degrees, Doctor of Science in Philology (DSc), professor

M.Y. Xojiyeva
Scientific Secretary of the Scientific Council awarding scientific degrees, Doctor of Philological Sciences in Philology (PhD), dotsent

G.I. Toirova
Scientific Council providing Scientific degrees the chairman of the Scientific seminar, Doctor of Science in Philology (DSc), professor

INTRODUCTION (DSc thesis annotation)

The aim of the research is to develop a methodology for creating a computer program that identifies punctuation errors based on the analysis of the use of punctuation marks in the Uzbek language corpus.

The object of the research work is the use of 10 punctuation marks in texts from the National Uzbek Corpus, developed at Tashkent State University of Languages and Literature, as well as the punctuation errors associated with them.

The scientific novelty of the study is as follows:

the principle of formalization has been applied to the punctuation system in linguistics. The grammatical (separation of homogeneous sentence members by commas), semantic (expression of explanatory relations by means of colons and dashes), and pragmatic (expression of emotional-expressive and interrogative meanings through exclamation and question marks) functions of punctuation marks in the Uzbek language have been evaluated based on a functional-systemic approach, and a linguistic-functional model reflecting their interrelationship in ensuring the structural and semantic integrity of the text has been developed;

based on the methodology of corpus linguistics, the patterns of normative punctuation usage in the Uzbek language, their variability across genres, and the criteria for the automatic detection of punctuation errors have been modeled using formal rules and contextual features. A linguistic model has been proposed for classifying punctuation units according to their grammatical (comma, semicolon, colon) and communicative-pragmatic (question mark, exclamation mark, ellipsis) functions, as well as for their automatic tagging based on a system of POS and punctuation tags;

based on the Transformer architecture and modern NLP and artificial intelligence technologies, a punctuation restoration model adapted to the Uzbek language has been developed. The linguistic and algorithmic principles for the context-dependent identification and restoration of punctuation marks have been substantiated, and the effectiveness of the model has been demonstrated through experimental testing using the Accuracy, Precision, Recall, and F1-score metrics;

based on POS tagging, the grammatical characteristics of punctuation marks, including the identification of part-of-speech sequences and syntactic relations, as well as their semantic characteristics, including the expression of predicative, explanatory, and enumerative relations, have been identified;

Software has been developed for the automatic detection and correction of punctuation errors in the Uzbek language and for the automatic restoration of punctuation marks during the conversion of speech into written text. The software automatically restores punctuation in punctuation-free texts, for example, converting “Bugun havo issiq ertaga yom‘g‘ir yog‘ishi mumkin” into “Bugun havo issiq, ertaga yom‘g‘ir yog‘ishi mumkin,” thereby ensuring the grammatical correctness and readability of the text.

Implementation of the research results. Based on the results obtained during the development of the technology for representing punctuation marks in a corpus and the creation of a computer program for detecting punctuation errors:

theoretical conclusions concerning the linguistic-functional model developed through the application of the principle of formalization to the punctuation system in linguistics, within which the grammatical (separation of homogeneous sentence members by commas), semantic (expression of explanatory relations by means of colons and dashes), and pragmatic (expression of emotional-expressive and interrogative meanings through exclamation and question marks) functions of punctuation marks in the Uzbek language were evaluated on the basis of a functional-systemic approach, as well as their interrelationship in ensuring the structural and semantic integrity of the text, were used in the implementation of the research project REP-25112021/113 entitled "UzUDT: Universal Dependency Treebank Corpus of the Uzbek Language for Natural Language Processing and Its Semantic Analysis" in 2022-2024 (Reference No. 04/11-20706 of September 29, 2025, issued by the National University of Uzbekistan named after Mirzo Ulugbek). As a result, the possibility of normalizing Uzbek punctuation marks for modern corpora based on their historical-comparative analysis was established;

scientific conclusions concerning the linguistic model developed on the basis of corpus linguistics methodology, within which the patterns of normative use of punctuation marks in the Uzbek language, their variability across genres, and the criteria for the automatic detection of punctuation errors were modeled based on formal rules and contextual features, while punctuation units were classified according to their grammatical (comma, semicolon, colon) and communicative-pragmatic (question mark, exclamation mark, ellipsis) functions, and a linguistic model for their automatic tagging based on a system of POS tags and punctuation tags was proposed, were utilized in the project AL-662204396 entitled "Creation of an Electronic Platform for the Scholarly and Literary Heritage of the Figures of the Bukhara Literary Environment," implemented in 2024 (Reference No. 6202 dated October 10, 2025, issued by Bukhara State University). As a result, a theoretical approach enabling the systematic acquisition of punctuation through the formal classification of linguistic units was substantiated;

scientific results concerning the development of a punctuation restoration model adapted to the Uzbek language based on the Transformer architecture, modern NLP and artificial intelligence technologies, the substantiation of linguistic and algorithmic principles for the context-dependent identification and restoration of punctuation marks, and the demonstrated effectiveness of the model based on the results of experimental testing according to the Accuracy, Precision, Recall, and F1-score metrics were applied within the framework of the corpus project AM-FZ-201908172 entitled "Creation of an Educational Corpus of the Uzbek Language," implemented in 2020-2023 (Reference No. 01/4-3886 dated September 2, 2025, issued by the Tashkent State University of Uzbek Language and Literature named after Alisher Navoi). As a result, corpus-based statistical characteristics of punctuation marks and their contextual activity levels were identified through linguistic analysis;

scientific conclusions concerning the identification, through POS tagging, of the grammatical characteristics of punctuation marks, including the determination of part-of-speech sequences and syntactic relations, as well as their semantic

characteristics, such as the expression of predicative, explanatory, and enumerative relations, were utilized in 2023-2024 in the programs “Islohotlar amalda,” “Kelajak yoshlari,” “Inson qadri,” and “Diyor 24” broadcast by the “Navoiy” TV channel of the Navoi Television and Radio Company, operating in the field of information technologies (Reference No. 108 dated November 10, 2025, issued by the Navoi Regional Television and Radio Company). As a result, the components of a punctuation annotation system based on syntactic and morphological features were implemented in practice as an independent software module;

scientific and practical approaches related to the development of software enabling the automatic detection and correction of punctuation errors in the Uzbek language, as well as the automatic restoration of punctuation marks in the process of converting speech into written text, were applied in the preparation of the textbooks entitled Corpus Linguistics and Computational Linguistics for master’s students enrolled in the 70230401 – Computational Linguistics educational program (Order No. 109-U of the Ministry of Higher Education, Science and Innovation of the Republic of Uzbekistan dated February 18, 2026, Publication Permit No. 109-3; Order No. 220-U dated April 8, 2025, Publication Permit No. 220-07). As a result, the scientific content of the textbooks was enriched and the teaching materials on corpus linguistics, natural language processing, and punctuation analysis were enhanced.

The outline of the thesis. The thesis consists of an introduction, four chapters, conclusion, reference with the total volume of dissertation 232 pages

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

I bo‘lim (I часть; part I)

1. Jumayeva D.B. Korpus asosida tinish belgilarini avtomatik tahlil va xatolarni aniqlash texnologiyasi. Monografiya. Buxoro: “Bukhara hamd print” nashriyoti. 2026. – 180 b. https://drive.google.com/file/d/1JzjCaVKBWsiSejZw4jB0zLLk4DHZl4ji/view?usp=sharing

2. Jumayeva D.B. Methods of recovering punctuation mark in the text in speech recognition system and its specific aspects// American journal of language, literacy and learning in STEM Education. Volume 3, Issue 2, 2025 ISSN (E): 2993-2769. 200-201 (Impact factor 10.23) https://drive.google.com/file/d/1WG3hpDBxtoJmjwOLpHqC2QCqjmRKV3ST/view?usp=sharing

3. Jumayeva D.B. The mutual importance of the punctuation department and the national corpus//American journal of Pedagogical and Educational Research Volume35, April-2025. ISSN (E):2832-9791. 19-21 (Impact factor 10.36) https://drive.google.com/file/d/186-n_W8tST5NWA8llPWKbLOD4ZgfKFRD/view?usp=sharing

4. Jumayeva D.B. National Corpus and Punctuation// Miasto Przyszłości Kielce. VOL. 57, 2025 ISSN-L: 2544-980X:196-198(Impact factor 9.9.) https://drive.google.com/file/d/1fhpT-XiLvPwUiqUkO9WGqZkEs4OufVr9/view?usp=sharing

5. Jumayeva D.B. The purpose and Meaning of Punctuation Marks in the National corpus// American journal of open University Education Vol.2, No.5 May 2025, ISSN 2997-3899 25-27 (Impact factor 12.76) https://drive.google.com/file/d/1LOnyzxVQnUzjWmoJsT0o1rnSiWUdO8N2/view?usp=sharing

6. Jumayeva D.B. Punktuatsion xatolarning kompyuter dasturini yaratish texnologiyasi va uning o‘ziga xosliklari // Buxoro davlat universiteti ilmiy axboroti. 2024. – № 12. – B. 197-200 (10.00.00; №1) https://drive.google.com/file/d/1HDGUBdc_aR5_RoizRjecV9GSgu2q6tXf/view?usp=sharing

7. Jumayeva D.B. Tinish belgilarining takror va qator ishlatilishi va ularning korpusda berilishi //Qo‘qon DPI. Ilmiy xabarlari Qo‘qon, 2024. – № 4. – B. 1912-1918. (2023-yil 5-maydagi №337/6 sonli Rayosat qarori) https://drive.google.com/file/d/1sV4xPQtQWSxrW19jiZyy4VbHtWrCE-8H/view?usp=sharing

8. Jumayeva D.B. O‘zbek tilida tinish belgilarining morfologik va sintaktik xususiyatlari// Til va adabiyot.uz. Elektron metodik jurnal 2025. – № 4. – B. 85-88. (10.00.00; №9) https://drive.google.com/file/d/1iBjXDvIGBRxOr7A9jfSubX2fShqwiWAw/view?usp=sharing

9. Jumayeva D.B. Jahon tilshunosligida punktuatsiya tizimining nazariy asoslari va ahamiyati// Ayniy vorislari. Xalqaro elektron jurna**l**, 2025. – № 2(9). – B. 132-134. **(**OAK Rayosatining 2024-yil 7- iyundagi 355-son qarori**)** https://drive.google.com/file/d/1Plcb05dTQ5pQw1X5ylZaHlB1-7dxlryK/view?usp=sharing

10. Jumayeva D.B. Morphological and syntactic characteristics of punctuation in uzbeki //Mug‘allim. Ilmiy-metodik jurnal, 2025.-2/1-son 105-110. (2007-yil 25-oktabr №138 sonli Rayosat qarori) https://drive.google.com/file/d/1CdWVISAwQWXBHsgDI75SZlEbcGIUj4Vw/view?usp=sharing

11. Jumayeva D.B. O‘zbek tilshunosligida tinish belgilarining tarixi va uni davrlashtirish// Buxoro davlat universiteti ilmiy axboroti. 2025. – № 6. – B. 130-134 (10.00.00; №1) https://drive.google.com/file/d/1UAuFkOtY0ZvGsawStqkTIXwo2T0RgwOq/view?usp=sharing

12. Jumayeva D.B. Og‘zaki nutqni tanib olishda matndagi tinish belgilarini tiklash usullari// Til va adabiyot.uz. Elektron metodik jurnal 2025. – № 14. – B. 24-26. (10.00.00; №9) https://drive.google.com/file/d/1UUea_AoZBXF9JfFMsn2nc3baferfk9NK/view?usp=sharing

13. Jumayeva D.B Punktuatsiyalarni modellash va kompyuter lingvistikasining imkoniyatlari// Namangan davlat universiteti ilmiy axborotnomasi. 2025. – № 8.- B. 494-498. (10.00.00; №26) https://drive.google.com/file/d/1I0e3OynZc8D1DzHNkKJxduesnwPqyDQd/view?usp=sharing

14. Jumayeva D.B Aniqlangan funksiyalar va korpus ma’lumotlari asosida punktuatsion xatolarni aniqlovchi dastur yaratish//Guliston davlat universiteti axborotnomasi. 2025. – № 3.- B. 112-114. (2022-yil 28-fevraldagi 312/6-sonli qarori) https://drive.google.com/file/d/1xGyjkUye0KEAbOkXyDqt3z64XwOqvyQ1/view?usp=sharing

15. Jumayeva D.B., Nasirov Sh.F. Tinish belgilarini modellashtirishda kompyuter lingvistikasi imkoniyatlari va usullari// Qo‘qon DPI. Ilmiy xabarlari Qo‘qon, 2025. **– № 7. –** B. 608-614. (2023-yil 5-maydagi №337/6 sonli Rayosat qarori) https://drive.google.com/file/d/1XkSqJs_wwk_lt90kl8qlMvtqWvnIaQFZ/view?usp=sharing

16. Jumayeva D.B. Punktuatsion xatolarni aniqlovchi kompyuter dasturini yaratish texnologiyasi // “Zamonaviy tilshunoslik yo‘nalishlari va neyropsixolingvistika muammolari” mavzusidagi respublika ilmiy-amaliy anjuman materiallari. – Ilmiy nashr. – Farg‘ona: FarDU, 2024. – B. 314-318. https://drive.google.com/file/d/1B1OkMOwv2vgCH8l6XHQ_Ou4UJIHidAhb/view?usp=sharing

17. Jumayeva D.B. Punktuatsion qoidalar va kompyuter lingvistikasi// “Xalq maqollari nazariy tadqiqining dolzarb muammolari” mavzusidagi xalqaro ilmiy-amaliy anjuman. BuxDU. 2025-yil,19-sentabr. - B. 241-243. https://drive.google.com/file/d/1s3UPMieX7-hfvBxUE8JZj7HpVYr0D_oZ/view?usp=sharing

18. Жумаева Д.Б. Технология создания компьютерной программы для обнаружения пунктуационных ошибок // “Modern trends of teaching in the context of innovative and digital technologies in higher education: prospects, problems and solutions” International scientific and practical conference, November 29, -P.319-321. https://drive.google.com/file/d/1bxbuZPlkxwkfucm918OSN37AWcoUYNoB/view?usp=sharing

19. Jumayeva D.B. Matndagi tinish belgilarini milliy korpusdagi ahamiyati va o‘rni // VI Traditional International Scientific and Practical Conference “Professional development of the personality of the XXI century in the system of continuous education: problems and ways of their solution”, December 3-4, 2024. -B.68-69. https://drive.google.com/file/d/1-k9dqn0LbFQYrRS-0wyNFChihKz8exQH/view?usp=sharing

II bo‘lim (II part; II часть)

20. Jumayeva D.B. “Tinish belgilarining korpusda berilishi va punktuatsion xatolarni aniqlovchi kompyuter dasturini yaratish texnologiyasi”. DGU.2025, 55254. 15.10.2025 https://drive.google.com/file/d/1-F0kck2FJDAO8yHIKpt_dMxaYieBOrtT/view?usp=sharing

21. Jumayeva D.B. Technology for creating a computer program for detecting punctuation errors // Problems and solutions of scientific and innovative research Volume 01, Issue 05, 2024. – B. 340-344. https://drive.google.com/file/d/1FZAMdiBjQ9duwhfcnI0F9CDjJIInKLy5/view?usp=sharing

22. Jumayeva D.B., Usmonova S.M. Boshlang‘ich sinf o‘quvchilarini orfoepik me’yorlarga o‘rgatish va uning amaliy ahamiyati // Oliy ta’limda raqamli texnologiyalar va innovatsiya: Fan ta’lim,tarbiya mavzusidagi xalqaro miqyosidagi ilmiy-texnik anjuman. – Navoiy, 2024-yil, 18-aprel. – B. 108-111. https://drive.google.com/file/d/1nztCssYBYJbtLq8GUEbuWhTgeq4L6CKC/view?usp=sharing

23. Jumaveva D.B., Toirova G.I. Построение модели восстановления пунктуации в текстах на узбекском языке с использованием POS-тегирования// “Russhunoslik va turkshunoslik:adabiy an’analar, madaniy transfer va tarjima” xalqaro forumi. 2025-yil, 2-3-oktyabr. -B-363-367. https://drive.google.com/file/d/1theMWNLt7nkPe2p7WvF7CVd6QgyecwfK/view?usp=sharing

24. Jumayeva D.B., Norova N. O. Boshlang‘ich sinflarda imlo qoidalari asosida to‘g‘ri yozishni shakllantirishga o‘rgatish // International scientific-practical conference: “Contemporary technologies of computational linguistics - 2024”. Vol.2 / No. 22. 2024. March. -B-582-585. https://drive.google.com/file/d/1-n3CGHZo2wNQ26y5GITX5cCbKK5yHXJS/view?usp=sharing

25. Jumayeva D.B. Tabiiy til modellari va ularning o‘ziga xosligi// “Ta’limni rivojlantirishda innovatsion texnologiyalarning o‘rni va ahamiyati” nomli ilmiy-amaliy onlayn konferensiya. – 2025-yil, 8-avgust – B. 26-29. https://drive.google.com/file/d/1gKfGbi6I03a5cXRkfCYSMeCtcfFfRToK/view?usp=sharing

Avtoreferat “Durdona” nashriyotida tahrirdan o‘tkazildi
hamda o‘zbek, rus va ingliz tillaridagi matnlarini mosligi tekshirildi.

Bosishga ruxsat etildi: 3.08.2026 Bichimi: 60x84 1/16.
“Times New Roman” garniturada raqamli bosma usulda bosildi.
Shartli bosma tabog‘i: 5. Adadi 100. Buyurtma № 166.
Guvohnoma AI №178.08.12.2010

“Sadriddin Salim Buxoriy” MCHJ bosmaxonasida chop etilgan.
Buxoro shahri, M. Iqbol ko‘chasi, 11-uy. Tel.: 65 221-26-45