ALISHER NAVOIY NOMIDAGI TOSHKENT DAVLAT O‘ZBEK TILI VA ADABIYOTI UNIVERSITETI HUZURIDAGI ILMIY DARAJALAR BERUVCHI DSc.03/30.12.2019.Fil.19.01 RAQAMLI ILMIY KENGASH ASOSIDAGI BIR MARTALIK ILMIY KENGASH

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

XUSAINOVA ZILOLA YULDASHEVNA

O‘ZBEK TILI BIRLIKLARINI TOKENLASH, STEMLASH, LEMMALASHNING LINGVISTIK ASOSLARI VA DASTURIY TA’MINOTI

10.00.11 – Til nazariyasi. Amaliy va kompyuter lingvistikasi

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

Toshkent – 2024

UO'K 811.512.133'32

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

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

Оглавление автореферата диссертации доктора философии (PhD) по филологическим наукам

ALISHER NAVOIY NOMIDAGI TOSHKENT DAVLAT O'ZBEK TILI VA ADABIYOTI UNIVERSITETI HUZURIDAGI ILMIY DARAJALAR BERUVCHI DSc.03/30.12.2019.Fil.19.01 RAQAMLI ILMIY KENGASH ASOSIDAGI BIR MARTALIK ILMIY KENGASH

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

XUSAINOVA ZILOLA YULDASHEVNA

O'ZBEK TILI BIRLIKLARINI TOKENLASH, STEMLASH, LEMMALASHNING LINGVISTIK ASOSLARI VA DASTURIY TA'MINOTI

10.00.11 – Til nazariyasi. Amaliy va kompyuter lingvistikasi

Filologiya fanlari bo'yicha falsafa doktori (PhD) dissertatsiyasi AVTOREFERATI

Toshkent – 2024

Falsafa doktori (PhD) dissertatsiyasi mavzusi O‘zbekiston Respublikasi Oliy attestatsiya komissiyasida B2023.2.PhD/Fil3688 raqam bilan ro‘yxatga olingan.

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

Dissertatsiya avtoreferati uch tilda (o‘zbek, ingliz, rus (rezyume) Ilmiy kengashning veb-sahifasi (www.tsuull.uz) va "Ziyonet" Axborot ta'lim portalida (www.ziyonet.uz) joylashtirilgan.

Ilmiy rahbar:	**Elov Botir Boltayevich** texnika fanlari bo‘yicha falsafa doktori, dotsent
Rasmiy opponentlar:	**Xolmonova Zulxumor Turdiyevna** filologiya fanlari doktori, professor
	Abdurahmonova Nilufar Zaynobiddin qizi filologiya fanlari doktori, professor
Yetakchi tashkilot:	**Samarqand davlat universiteti**

Dissertatsiya himoyasi Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti huzuridagi DSc.03/30.12.2019.Fil.19.01 raqamli Ilmiy kengash asosidagi bir martalik Ilmiy kengashning 2024-yil "14" 06 soat 10⁰⁰ dagi majlisida bo‘lib o‘tadi. (Manzil:100100, Toshkent, Yakkasaroy tumani, Yusuf Xos Hojib ko‘chasi, 103. Tel: (99871) 281-42-44; faks: (99871) 281-42-44; faks: (99871) 281-12-44. (www.tsuull.uz). e-mail: monitoring@navoiy-uni.uz).

Dissertatsiya bilan Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti Axborot-resurs markazida tanishish mumkin. (290-raqam bilan ro‘yxatga olingan) (Manzil: 100100, Toshkent, Yakkasaroy tumani, Yusuf Xos Hojib ko‘chasi, 103. Tel: (99871) 281-42-44; faks: (99871) 281-42-44; faks: (99871) 281-12-44 (www.tsuull.uz).

Dissertatsiya avtoreferati 2024-yil "30" 05 kuni tarqatildi.

(2024-yil "30" 05 dagi 1- raqamli reestr bayonnomasi)

H.A. Dadaboyev
Ilmiy darajalar beruvchi ilmiy kengash asosidagi bir martalik ilmiy kengash raisi, filol.f.d., professor

Q.U. Pardayev
Ilmiy darajalar beruvchi ilmiy kengash asosidagi bir martalik ilmiy kengash ilmiy kotibi, filol.f.d., professor

S.E. Normamatov
Ilmiy darajalar beruvchi ilmiy kengash asosidagi bir martalik ilmiy kengash qoshidagi ilmiy seminar raisi, filol.f.d., professor

Kirish (falsafa doktori (PhD) dissertatsiyasi annotatsiyasi)

Dissertatsiya mavzusining dolzarbligi va zarurati. Jahon kompyuter lingvistikasi amaliyotida tabiiy tilga ishlov berish usullaridan keng foydalanilmoqda. Tabiiy tilga ishlov berish (Natural language prossecing – NLP) loyihalarida doimiy bajariladigan vazifalar mavjud. Ular fundamental xarakterga ega bo‘lganligi sababli keng ko‘lamda o‘rganilgan. Ushbu vazifalarni yaxshi tushunish NLPning turli ilovalarini yaratishga imkon beradi. Tilni modellashtirish, matnni tasniflash, axborot olish, ma'lumot izlash, suhbat agenti, matnni umumlashtirish, savol-javob tizimlari, mashina tarjimasi, tematik modellashtirish NLPning asosiy sohalaridir. Tabiiy tilni qayta ishlash – sun'iy intellekt sohasining mashinalarni inson tilini tushunish va qayta ishlashga qaratilgan yo‘nalishi. Tabiiy tilni qayta ishlash mashina uchun oson ish emas, chunki mashina matn bilan emas, raqamlar bilan ishlaydi.

Dunyo tilshunosligida XX asr oxirida avtomatik tarjima sifatini yaxshilash, tilni lingvistik modellashtirish, muayyan tillar doirasida so‘zlarni lemmalash nazariyasi, algoritmini yaratish kabi masalalar kun tartibiga qo‘yildi. Natijada axborot-qidiruv tizimida faol qo‘llanuvchi tokenizator, stemmer, lemmatizator singari ilovalar ishlab chiqildi. Binobarin, til korpuslarini yaratish, ularni lingvistik annotatsiyalash, matnni avtomatik qayta ishlaydigan dasturiy vositalar – morfoanalizator, stemmer, lemmatizator, parser, orfokorrektorlarni ishlab chiqish kompyuter lingvistikasining ilmiy-nazariy muammolariga aylandi.

O‘zbek kompyuter lingvistikasi sohasida til korpusi, nutq sintezatori, avtomatik tarjima, sun'iy intellekt, o‘zbek tilini tushunish va qayta ishlash borasida ilmiy izlanishlarga e'tibor qaratilmoqda. Lekin tabiiy tilga ishlov berishning dastlabki bosqichlarini amalga oshiruvchi vositalar: tokenizator, stemmer, lemmatizatorlarni ishlab chiqish masalasi ilmiy-nazariy jihatdan o‘rganilmagan. Zero, "...davlat tilining sofligini saqlash, uni boyitib borish va aholining nutq madaniyatini oshirish; davlat tilining zamonaviy axborot texnologiyalari va kommunikatsiyalariga faol integratsiyalashuvini ta'minlash"[1] hozirgi kunda o‘zbek kompyuter lingvistikasi oldida turgan dolzarb vazifalardan biridir. Shunga ko‘ra, o‘zbek tiliga zamonaviy axborot texnologiyalari vositasida ishlov berilishiga erishish, buning uchun til korpuslari, elektron tarjimon, tezaurus, orfokorrektor, lingvistik analizatorlar kabi dastlabki avtomatik ishlov vositalari hamda ularning lingvistik ta'minotini yaratish singari masalalar muhim ahamiyat kasb etadi.

O‘zbekiston Respublikasi Prezidentining 2016-yil 13-maydagi PF-4997-son "Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universitetini tashkil etish to‘g‘risida", 2019-yil 21- oktyabrdagi PF-5850-son "O‘zbek tilining davlat tili sifatidagi nufuzi va mavqeini tubdan oshirish chora-tadbirlari to‘g‘risida", 2020-yil 20-oktyabrdagi PF-6084-son "Mamlakatimizda o‘zbek tilini yanada rivojlantirish va til siyosatini takomillashtirish chora-tadbirlari to‘g‘risida"gi Farmonlari, 2017-yil 17-fevraldagi PQ-2789-son "Fanlar akademiyasi faoliyati, ilmiy-tadqiqot ishlarini tashkil etish, boshqarish va moliyalashtirishni yanada

[1] "O‘zbek tilining davlat tili sifatidagi nufuzi va mavqeyini tubdan oshirish chora-tadbirlari to‘g‘risida"gi 2019-yil 21-oktyabrdagi PF-5850-son Prezident farmoni. https://lex.uz/docs/-4561730

takomillashtirish chora-tadbirlari to'g'risida", 2019-yil 4-oktyabrdagi PQ-4479-son "O'zbekiston Respublikasining "Davlat tili haqida"gi Qonuni qabul qilinganligining o'ttiz yilligini keng nishonlash to'g'risida"gi Qarorlari, 2020-yil 24-yanvardagi O'zbekiston Respublikasi Prezidentining Oliy Majlisga Murojaatnomasi hamda boshqa me'yoriy-huquqiy hujjatlarda belgilangan vazifalarni amalga oshirishda ushbu tadqiqot muayyan darajada xizmat qiladi.

Tadqiqotning respublika fan va texnologiyalari rivojlanishining ustuvor yo'nalishlariga mosligi. Dissertatsiya respublika fan va texnologiyalari rivojlanishining I."Axborotlashgan jamiyat va demokratik davlatni ijtimoiy, huquqiy, iqtisodiy, madaniy, ma'naviy-ma'rifiy rivojlantirish, innovatsion iqtisodiyotni rivojlantirish" ustuvor yo'nalishiga muvofiq bajarilgan.

Muammoning o'rganilganlik darajasi. Jahon tilshunosligida tabiiy tilga ishlov berish masalasi atroflicha o'rganilgan. Tabiiy tilga ishlov berishning umumiy masalalari qator tadqiqotlar obyekti bo'lgan[2]. Jumladan, J.J.Vebster, C.Kit, A.Rai, S.Borah, X.Song, A.Salchianiy, Y.Song, D.Dopson, D.Zhou va boshqalar[3] o'z tadqiqotlarida tokenlash muammolarini yoritganlar. Stemlash masalasi va uning yechimiga turli yondashuvlarni K.Savitha, Y.Liu, M.Haidar, K.Kreslins, S.Jonatan ishlaridan o'rganish mumkin[4]. Lemmalash muammosiga M.Avhavhudzani, D.MakKarti, A.Chakrabarti ishlarida yechimlar taklif qilingan[5].

O'zbek tiliga ishlov berish, xususan stemlash, morfologik tahlil va tahrir dasturiy vositalarining lingvistik modullari, o'zbek tili matn korpusining lingvistik xususiyatlarni tokenlash masalalari borasida Z.Xolmanova[6], B.Elov, Sh.Xamroyeva[7], M.Abjalova[8], N.Abduraxmonova, A.Ismailov[9], O.Abdullayeva[10], R.Alayev[11], I.Bakayev[12], M.Sharipov va O.Sobirov[13]larning ayrim qaydlari mavjud. O'zbek tilida so'zshakllari, ularning yasalishi va morfemalar[14] hamda o'zbek tiliga avtomatik ishlov berish masalasi[15] tadqiqi ham e'tiborga molik. Ammo o'zbek tili leksik birliklarini tokenlash, stemlash va lemmalashning lingvistik asoslari va dasturiy ta'minotini ishlab chiqish masalasi maxsus monografik tadqiqot sifatida o'rganilmagan.

Tadqiqotning dissertatsiya bajarilgan oliy ta'lim yoki ilmiy-tadqiqot muassasasining ilmiy-tadqiqot ishlari rejalari bilan bog'liqligi. Ushbu tadqiqot Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universitetida 2021-2024 yillarga bajarilishi mo'ljallangan "Axborot-qidiruv tizimlari (Google, Yandex, Google translate) uchun avtomatik ishlov berish vositasi – o'zbek tilining morfoleksikoni va morfologik analizatori dasturiy vositasini yaratish" mavzusidagi innovatsion loyiha doirasida bajarilgan.

Tadqiqotning maqsadi o'zbek tili leksik birliklarini tokenlash, stemlash va lemmalashning lingvistik asoslari va dasturiy ta'minotini ishlab chiqishdan iborat.

Tadqiqotning vazifalari:

tokenlashning umumiy xususiyatlari va o'zbek tiliga xos tomonlarini aniqlash;

stemlash va lemmalashning nazariy asoslarini aniqlash;

agglyutinativ tillar uchun stemlash jarayoni muammolarini aniqlash;

o'zbek tili leksik birliklarini stemlash va lemmalashning lingvistik asoslarini ishlab chiqish;

o'zbek tili milliy korpusi qidiruv tizimini optimallashtirish usullarini ishlab chiqish;

BPE algoritmi asosida tokenlash jarayonini amalga oshirish;

affikslarni ajratishga asoslangan uzb.stemming algoritmini ishlab chiqish;

[2] Vajjala S., Majumder B., Gupta A. Surana H. Practical Natural Language Processing. A Comprehensive Guide to Building Real-World NLP Systems. – USA, 2020. – 424 p.; Bölücü N., Can B. Unsupervised joint PoS tagging and stemming for agglutinative languages. ACM Transactions on Asian and Low-Resource Language Information Processing, New York, 2019, – №18. – P. 1-21.

[3] Webster J.J., Kit Ch. Tokenization as the initial phase in NLP. Hong Kong, 1992. – P. 1106-1110; Rai A., Borah S. Study of various methods for tokenization // In Lecture Notes in Networks and Systems. India, 2021, – № 137. – P. 193-199; Song X., Salcianu A., Song Y., Dopson D., Zhou D. Fast WordPiece Tokenization // Conference on Empirical Methods in Natural Language Processing, Proceedings. China, 2021. – P. 2089-2103.

[4] Savitha K. Study of stemming algorithms // UNLV Theses, Dissertations, Professional Papers and Capstones. – Las Vegas, 2010. – 48 p.; Liu Yi Yu. The advances of stemming algorithms in text analysis from 2013 to 2018 // MCom. Dissertation, University of Pretoria. – Pretoria, 2019. – 192 p.; Haidar M. "A comparison of root and stemming techniques for the retrieval of Arabic documents". Dissertation, McGill University. – Montreal, 2001. – 204 p; Kreslins K. "A stemming algorithm for Latvian". PhD Dissertation, Loughborough University. – Boston, 1996. – 264 p.; Jonathan S. Designing a stemming algorithm for Kambaata text: a rule based approach. st.mary's University. – Addis Ababa, Ethiopia, 2018. – 104 p.

[5] Avhavhudzani M.V. "The lemmatization of Tshivenda lexical items". Thesis, University of Limpopo. – Africa, 2012. – 105 p.; McCarthy D. Lexical acquisition at the syntax-semantics interface: Diathesis alternation, subcategorization frames and selectional preferences. Doctoral dissertation, University of Sussex. – Cambridge, 2001. – 189 p.; A. Chakrabarty. On Lemmatization and Morphological Tagging for Highly Inflected Indic Languages. Doctoral dissertation, Indian Statistical Institute. – India, 2018. – 116 p.

[6] Xolmonava Z.T. Kompyuter lingvistikasi. – Toshkent. 2020. – 198 b.

[7] Elov B.B., Xamroyeva Sh.M. The Problem of POS Tagging and Stemming for Agglutinative Languages // 8 th International Conference on Computer Science and Engineering UBMK 2023. Turkey, Burdur, 2023. – P. 57-62.

[8] Abjalova M. Tahrir va tahlil dasturlarining lingvistik modullari. (NLP) // Linguistic modules of the text editing and analysis program (NLP). – Monografiya. – Toshkent, 2020. – 165 b.

[9] Abdurakhmonova N., Ismailov A., Sayfullaeva R. Turkic language stemmer python package for Natural Language Processing // "Computer Linguistics challenges and sollutions" international scientific-practical conference. Tashkent, 2023. https://www.researchgate.net/publication/369503003

[10] Abdullayeva O. O'zbek tilining internet axborot matnlari korpusini shakllantirishning nazariy va amaliy asoslari. Filol. fan. bo'yicha falsafa doktori (PhD) diss. – Andijon, 2022. – 120 b.

[11] Elov B., Alayev R. Affikslarni ajratishga asoslangan uzb-temming algoritmi / O'zbekiston Informatika va energetika mummolari jurnali. O'zbekiston Jurnali. 2023, – № 1. – P. 14-27.

[12] Bakaev I. Linguistic features tokenization of text corpora of the Uzbek language // Bulletin of TUIT: Management and Communication Technologies. Tashkent, 2021. – P. 98-105.

[13] Sharipov M., Sobirov O. Development of a Rule-Based Lemmatization Algorithm Through Finite State Machine for Uzbek Language // The International Conference and Workshop on Agglutinative Language Technologies as a challenge of Natural Language Processing. Koper, Slovenia, 2022. – P. 1-6.

[14] Ҳожиев А. Ўзбек тили сўз ясалиши. –Тошкент, 1989. – Б. 10.; Миртожиев М.М. Ўзбек тили фонетикаси. – Тошкент, 2013.; Qo'ziboyeva G. Tilimizga kirib kelgan neologizmlar va ularning tahlili / International scientific journal. Toshkent, 2022, – № 3. – B. 78-83.

[15] Elov B., Hamroyeva Sh., Axmedova X. Methods for creating a morphological analyzer, 14th International Conference on Intellegent Human Computer Interaction. Tashkent, 2022. – P. 27-38.; Elov B., Hamroyeva Sh. Morfologik analizatorni yaratish usullari. / O'zbekiston: Til va Madaniyat (Amaliy filologiya). Toshkent, 2022, – № 5. – B. 40-64.

o‘zbek tili matnlari uchun uzb.tokenizator, uzb.stemming va uzb.lemmatizator dasturiy ta’minotini ishlab chiqish.

Tadqiqotning obyektini o‘zbek tili leksik birliklari tashkil qiladi.

Tadqiqotning predmetini o‘zbek tili leksik birliklarini tokenlash, stemlash va lemmalashning lingvistik asoslari va dasturiy ta’minoti tashkil qiladi.

Tadqiqotning usullari. Tadqiqot mavzusini yoritishda tasniflash, tavsiflash, qiyoslash, modellashtirish metodlaridan foydalanildi.

Tadqiqotning ilmiy yangiligi quyidagilardan iborat:

jahon va o‘zbek tilshunosligida leksik birliklarni tokenlash, stemlash va lemmalashning nazariy asoslari tizimlashtirilib, belgi va so‘zga asoslangan tokenlash, lingvistik qoidalar va korpus vostasidagi gibrid yondashuvli stemlash, lug‘atga asoslangan lemmalash usullari aniqlangan;

agglyutinativ tillar uchun stemlash jarayonidagi o‘zak va qo‘shimchaning bitta o‘zak bilan omonim bo‘lishi, so‘zning tovush o‘zgarishiga uchrashi, neologizm va NERlarni stemlash kabi muammolar aniqlanib, o‘zbek tilidagi turli tuzilishli leksik birliklarni stemlash va lemmalashning lingvistik asoslari ishlab chiqilgan;

o‘zbek tili milliy korpusi matnlarini qayta ishlash asosida qidiruv tizimini optimallashtirish usullari ishlab chiqilib, leksik birliklarni stemlash va lemmalash jarayoni ochib berilgan;

tokenlash jarayoni BPE algoritmi asosida amalga oshirilib, affikslarni ajratishga asoslangan stemlash, lug‘at va morfologik tahlilga tayanilgan lemmalash algoritmi ishlab chiqilib, o‘zbek tili matnlari uchun uzb.tokenizator, uzb.stemming va uzb.lemmatizator dasturiy ta’minoti yaratilgan.

Tadqiqotning amaliy natijalari quyidagilardan iborat:

tadqiqot natijalari asosida ishlab chiqilgan algortimlar o‘zbek tili morfologik analizatori samaradorligini oshirishga xizmat qilgan;

dissertatsiyada keltirilgan tokenlash, stemlash va lemmalash amallari o‘zbek tili milliy korpusi qidiruv tizimi natijasini optimallashtirgan;

o‘zbek tili matnlarini tokenlash, stemlash va lemmalash dasturiy vositasi ishlab chiqilgan.

Tadqiqot natijalarining ishonchliligi o‘rganilgan materiallarning o‘zbek, turk, uyg‘ur va ingliz tili tabiatidan kelib chiqqan holda xulosalar qilishga yordam berganligi, ularning asosli ekanligi, metodologik mukammalligi, o‘zbek, turk, uyg‘ur va ingliz tili chog‘ishtirma tadqiqotlarning amalda isbotlangan manbalarga tayanilganligi bilan izohlanadi.

Tadqiqot natijalarining ilmiy va amaliy ahamiyati. Tadqiqot natijalarining ilmiy ahamiyati shundaki, mazkur tadqiqot materiallari o‘zbek tili matnlarini qayta ishlash, til korpusini tuzish, mashina tarjimasi tizimlarini ishlab chiqishga oid nazariy ishlarni muayyan ilmiy qarash va faktlar bilan boyitadi. Shuningdek, o‘zbek kompyuter lingvistikasi bo‘yicha amalga oshiriladigan tadqiqot, monografiya, darslik va o‘quv qo‘llanmalarning mukammallashishiga xizmat qiladi.

Tadqiqot natijalarining amaliy ahamiyati tabiiy tilni morfologik, sintaktik, semantik tahlil qilish vositalarini tuzishda tokenlash, stemlash va lemmalash jarayonini bajaruvchi https://uznatcorpara.uz/uz/POSTag dasturiy ta’minotining ishlab chiqilganligi bilan izohlanadi.

Tadqiqot natijalarining joriy qilinishi. O‘zbek tili birliklarini tokenlash, stemlash, lemmalashning lingvistik asoslari va dasturiy ta’minoti tadqiqi bo‘yicha olingan natijalar asosida:

o‘zbek tili birliklarini tokenlash, stemlash va lemmalashni amalga oshiruvchi dasturiy ta’minotdagi belgi va so‘zga asoslangan tokenlash, lingvistik qoidalar va korpus vositasidagi gibrid yondashuvli stemlash, lug‘atga asoslangan lemmalash haqidagi ilmiy xulosalardan Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti huzuridagi Davlat tilida ish yuritish asoslarini o‘qitish va malaka oshirish markazi davlat muassasasi markazida foydalanilgan (Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universitetining 2023-yil 12-oktyabrdagi 123-son dalolatnomasi). Natijada, o‘zbek tili birliklarini tokenlash, stemlash va lemmalash dasturiy ta’minotining sinovdan o‘tkazilishi natijasida o‘zbek tili gaplarining morfologik tahlilidagi 32000 dan ortiq sodda, 7500 dan ortiq qo‘shma leksemadan iborat ma’lumotlar bazasi shakllantirilgan;

o‘zbek tili birliklarini tokenlash, stemlash, lemmalash natijalaridan, xususan tabiiy tilga ishlov berishda o‘zbek, rus, ingliz tillarida stemlash va lemmalash sifatini oshirishga doir ilmiy natijalaridan Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universitetida 2021-2023-yillarda bajarilgan I-OT-2019-42 raqamli “O‘zbek adabiyotining ko‘p tilli (o‘zbek, rus, ingliz tillarida) elektron platformasini yaratish” mavzusidagi amaliy loyihada foydalanilgan (Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universitetining 2023-yil 19-oktyabrdagi 01/10-2182-raqamli ma’lumotnomasi). Natijada, tabiiy tilga ishlov berishda o‘zbek, rus, ingliz tillarida stemlash va lemmalash sifatini oshirish elektron platformada qidiruv sifatini yaxshilagan; o‘zbek tili birliklarini tokenlash, stemlash, lemmalash natijalaridan o‘zbek tiliga xos birliklar qidiruv natijasi samaradorligi ta’minlangan;

o‘zbek tili birliklarini tokenlash, stemlash, lemmalashning natijasini asoslovchi algoritmlar va ular asosida yaratilgan dasturiy ta’minotidan Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universitetida 2021-2023-yillarda bajarilgan PZ-2020042022 “Turkiy tillarning lingvodidaktik elektron platformasini yaratish” mavzusidagi amaliy loyihada foydalanilgan (Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universitetining 2023-yil 19-oktyabrdagi 01/10-2183-raqamli ma’lumotnomasi). Natijada, stemlash va lemmalash algoritmi asosida turkiy tillarning lingvodidaktik elektron platformasi qidiruv tizimi ko‘rsatkichi oshishiga erishilgan.

Tadqiqot natijalarining aprobatsiyasi. Tadqiqot natijalari 2 ta xalqaro, 2 ta respublika ilmiy-amaliy anjumanlarida muhokamadan o‘tkazilgan.

Tadqiqot natijalarining e’lon qilinganligi. Dissertatsiya mavzusi bo‘yicha 12 ta jumladan, O‘zbekiston Respublikasi Oliy attestatsiya komissiyasining doktorlik dissertasiyalari asosiy ilmiy natijalarini chop etish uchun tavsiya etilgan ilmiy nashrlarda 8 ta ilmiy maqola nashr ettirilgan, ulardan 3 tasi xorijiy jurnallarda chop qilingan.

Dissertasiyaning tuzilishi va hajmi. Dissertasiya kirish, uch bob, xulosa, foydalanilgan adabiyotlar ro‘yxati va ilovadan iborat bo‘lib, hajmi 129 sahifani tashkil etadi.

DISSERTATSIYANING ASOSIY MAZMUNI

Kirish qismida ilmiy tadqiqotning mavzusi va uning dolzarbligi asoslangan, tadqiqot ishining maqsadi va vazifalari, obyekti va predmeti aniqlangan, ilmiy ishning fan va texnologiyalarni rivojlantirishning muhim yo'nalishlariga mosligi ko'rsatib o'tilgan. Shu bilan birgalikda, dissertatsiyaning ilmiy yangiligi, amaliy natijalari va ularning ishonchliligi, ishning nazariy va amaliy ahamiyati, erishilgan natijalarning amaliyotga joriy etilishi, ilmiy nashrlarda e'lon qilinganligi, ishning tuzilishi haqidagi ma'lumotlar keltirilgan.

Dissertatsiyaning birinchi bobi **"Tokenlash, stemlash va lemmalashning nazariy asoslari"** deb nomlanadi. Bobning birinchi paragrafi *"Jahon va o'zbek tilshunosligida tokenlash jarayonining umumiy tavsifi"* deb ataladi. Matnni tahlil qilish, odatda, aniq belgilangan bosqichlarga muvofiq amalga oshiriladi. Matnga boshlang'ich ishlov berishning dastlabki bosqichi uni **tokenlar**ga aylantirishdir. Tokenlar matnni tahlil qilish uchun mazmunli birliklar sifatida aniqlangan matn segmentlaridan iborat bo'ladi. D.Manning, P.Raghavan va H.Schutzelarning ilmiy tadqiqotlarida tokenlarni alohida so'zlardan katta yoki kichikroq bo'laklar (*so'z ketma-ketligi, so'zostilar, paragraflar, gaplar* yoki *satrlar*)dan iborat bo'lishi mumkinligi ta'kidlangan[16].

Tokenlash NLPda matnli ma'lumot bilan ishlashning asosiy va majburiy bosqichi bo'lib, ushbu vazifani bajaruvchi dasturiy vosita tokenizatordir. Tokenizator NLP vazifasiga qarab turli yondashuv asosida ishlab chiqilishi mumkin.

Turli tizimli tillarning tokenlash dasturiy vositalari tavsifi. Dunyoning turli tillari uchun ko'plab tokenlash usullari mavjud bo'lib, qoidaga asoslangan[17], statistik[18], oshkor bo'lmagan[19], leksik[20] usullar farqlanadi. Velbers va V.Atteveldt o'zining "Text Analysis in R" nomli maqolasida: "Tokenlash jarayoni matnni kichikroq bo'laklarga ajratish jarayoni hisoblanib, u ko'pincha tinish belgilarini olib tashlash va barcha belgilarni kichik harflarga aylantirish uchun matnni boshlang'ich qayta ishlashni o'z ichiga oladi", – deb ta'rif beradilar[21].

Tokenlash jarayonida qabul qilingan qarorlar matnni qayta ishlashning keyingi jarayonlariga sezilarli ta'sir ko'rsatadi. M.Denni, A.Spirling[22] hamda D.Guzri[23] o'z

[16] Manning C.D., Raghavan P., Schütze H. Introduction to Information Retrieval. – Cambridge, 2008. – 544 p.
[17] Zhou L., Liu Q. A character-net based Chinese text segmentation method. 2002. https://aclanthology.org/W02-1117.pdf. (Saytga murojaat); Kaplan R.M. A Method for Tokenizing Text. Complexity and Education: Inquiries into Words, Constraints and Contexts. Cambridge, 2005. – P.55-64.
[18] Yang C.C., Li K.W. A heuristic method based on a statistical approach for Chinese text segmentation / Journal of the American Society for Information Science and Technology. America, 2005, – №56. – P. 1438-1447.
[19] Shahabi A.S., Kangaveri M.R. Intelligent processing system // IFIP International Federation of Information Processing Springer Boston. Pakistan, 2007. – P. 411-420.
[20] Wu D., Fung P. Improving Chinese tokenization with linguistic filters on statistical lexical acquisition // 4th Conference on Applied Natural Language Processing. Stuttgart, Germany, 1994. – P. 180-181; Labadié A., Prince V. Lexical and semantic methods in inner text topic segmentation: A comparison between C99 and transeg. // Conference: Proceedings of the 13th international conference on Natural Language and Information Systems: Applications of Natural Language to Information Systems. France, 2008. – P. 347-349.
[21] Welbers K., Van Atteveldt W., Benoit K. Text Analysis in R. Communication Methods and Measures. London, 2007, – №11. – P. 245-265.
[22] Denny M. J., Spirling A. Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It / Political Analysis. Cambridge, 2018, – №26. – P. 168-189.
[23] Guthrie D., Allison B., Liu W., Guthrie L., Wilks Y. A closer look at skip-gram modelling // Proceedings of the 5th International Conference on Language Resources and Evaluation. Chicago, 2006. – P. 1222-1225.

tadqiqotlarida katta hajmdagi til korpuslari uchun tokenlashni amalga oshirish jarayonida murakkab amallar bajarilishini, tokenlashning tabiiy til xususiyatlariga bog'liqligini ta'kidlashgan. A.Mullen, K.Benoit[24] va R.Core Team[25] dasturchilar guruhi tomonidan **R tilida** ishlab chiqilgan tokenizatorlar to'plami orqali tabiiy tildagi matn uchun tez, izchil tokenlash amalga oshirilgan hamda GitHub[26] va Zenododa[27] arxivlangan.

D.Eddelbettel va J.Balamutalar tomonidan ishlab chiqilgan tokenizator tezligini oshirish uchun **n-gram** va **skip n-gram tokenizer** kabi asosiy komponentlar **Rcpp** dasturiy vositasi yordamida ishlab chiqilgan[28].

Bugungi kunda **o'zbek tili** leksik birliklarini tokenlash bo'yicha amaliy ishlar uchramaydi. M.Abjalova "Tahrir va tahlil dasturlarining lingvistik modullari" nomli monografiyasida tokenlash, stemlash va lemmalash jarayonlariga ta'rif bergan[29].

I.Bakayev tomonidan o'zbek tili matn korpusini tokenlash amalga oshirilgan bo'lib, unda imlo xususiyatlarini hisobga olish lozimligi qayd etilgan[30]. So'zlarning turlari tahlil qilinib, *murakkab, juft, takroriy* va *qo'shma so'zlar* uchun so'z yasalishining matematik modeli taklif etilgan.

Tokenlashni keng ma'noda 3 turga – so'z, belgi va so'z tarkibi elementlari (n-gram belgilar)ga ajratish mumkin. Misol uchun, quyidagi gapni ko'rib chiqamiz: *"Hech qachon taslim bo'lmang"*. Tokenlarni shakllantirishning eng keng tarqalgan usuli bo'shliqqa (probel) asoslangan. Bo'sh joyni ajratuvchi sifatida qabul qilsak, gapning tokenlash natijasida 4 ta token hosil qilinadi – *Hech-qachon-taslim-bo'lmang*. Har bir token so'zdan iborat bo'lgani uchun u **word(so'z)ni tokenlash** deyiladi. Xuddi shunday, tokenlar sifatida belgilar va so'z tarkibi elementlari hosil qilinishi mumkin. Misol uchun, *"aqlliroq"* so'zini ko'rib chiqamiz:

belgilar tokeni: *a-q-l-l-i-r-o-q;*

so'z tarkibi elementlari tokeni: *aql-li-roq.*

Tokenlar tabiiy tilning qurilish bloklari bo'lganligi sababli strukturlanmagan matnni qayta ishlashning eng keng tarqalgan usuli token darajasida sodir bo'ladi.

Birinchi bobning ikkinchi paragrafi *"Stemlash jarayonining umumiy tavsifi"* deb nomlanadi. Stemlash NLPda matnga ishlov berishning dastlabki bosqichi bo'lib, mazkur jarayon NLPning ko'plab vazifalarida amalga oshirilishi talab etiladigan amaldir. Stemlash jarayonining asosiy maqsadi – so'zning flektiv shakllarini o'zak shakliga olib kelish. NLPning ushbu amali axborot qidiruv tizimlari (Information Retrieval systems, IRS)ning aksariyatida juda ahamiyatli. Stemming jarayoni

[24] Mullen A., Benoit L., Keyes K., Selivanov D., Arnold J. Fast. Consistent Tokenization of Natural Language Text / Journal of Open Source Software. America, 2018, – №3. – P.1-3.
[25] Team R.C. R: A Language and Environment for Statistical Computing / R Foundation for Statistical Computing. Vienna, 2021. https://www.R-project.org (Saytga murojaat).
[26] https://github.com/ropensci/tokenizers
[27] https://zenodo.org/record/1205017
[28] Eddelbuettel D. Seamless R and C++ integration with Rcpp. — USA, 2013. – 220 p.; Eddelbuettel D., Balamuta J.J. Extending R with C++: A Brief Introduction to Rcpp / The American Statistician, Taylor and Francis Journals. France, 2018, – № 72. – P. 28-36.
[29] Abjalova M. Tahrir va tahlil dasturlarining lingvistik modullari. (NLP). // Linguistic modules of the text editing and analysis program (NLP). – Monografiya. – Toshkent, 2020. – 165 b.
[30] Bakaev I. Linguistic features tokenization of text corpora of the Uzbek language // Bulletin of TUIT: Management and Communication Technologies, Tashkent, 2021. – P. 98-105.

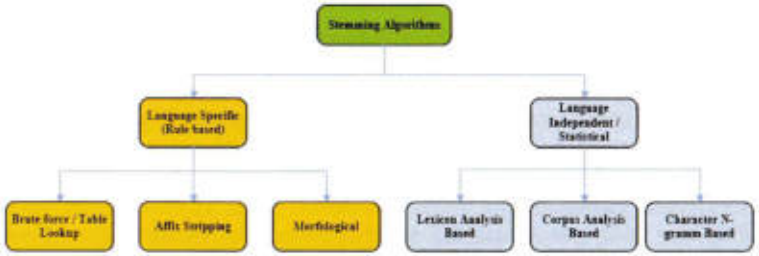
so‘zshakllardagi barcha qo‘shimchalarni kesib tashlash orqali amalga oshadi. Buning natijasi esa imlo tekshiruvi kabi NLP ilovalari samaradorligini oshirishga xizmat qiladi.

Stemlash jarayoni so‘zlarning yasalishi bilan bog‘liq: u *o‘zak* va *yasalmaning* o‘zgarishiga asoslanadi. Buning natijasida so‘zshakllar konflyatsiya[31]ga olib kelinadi. So‘zshakllarida prefiks va suffikslar ishtirok etishi mumkin.

1980-yilda M.Porter qo‘shimchalarni ajratish qoidalariga asoslangan stemmer algoritmini taklif qildi. Algoritm besh bosqich va 10000 so‘zli lug‘atdan iborat stemmer matn/korpus hajmini **33%**ga qisqartirishini aniqladi.

1990-yilda Lankaster universitetida Kris Peys grammatik qoidalar asosida qo‘shimchalarni iterativ ravishda olib tashlaydigan yangi stemmerni ishlab chiqdi. Mazkur stemmer algoritmi qoidalar istisnolari yuzaga kelmasligi uchun qo‘shimchalarni almashtirishga qaratilgan. Bu stemmer odatda Peys/Hask stemmer yoki Lankaster stemmer deb ataladi.

Stemlash algoritmlarining tasnifi. Hozirgi kunda NLP tadqiqotchilari tomonidan juda ko‘plab stemlash algoritmlari ishlab chiqilgan bo‘lib, ishning mazkur qismida keng miqyosda foydalanilayotgan algoritmlar xususida bahs yuritiladi. http://devopedia.org saytida zamonaviy stemmerlar quyidagicha tasniflangan:



1-rasm. http://devopedia.org ma’lumotlariga ko‘ra stemming algoritmlari tasnifi

Qo‘shimchalarni olib tashlash usullari. Mazkur usul so‘zdagi prefiks yoki affikslarni olib tashlash (**truncating**)ga asoslanadi. Bu turdagi eng oddiy stemmer **Tuncate**(n) stemmeri bo‘lib, u so‘zni **n**-belgisigacha qisqartiradi, ya’ni so‘zning birinchi **n** ta harfini saqlagan holda, qolganlarini olib tashlaydi[32]. Bu usulda uzunligi **n** dan qisqa so‘zlar o‘z shaklida saqlanadi. So‘z uzunligi kichik bo‘lgan holda ortiqcha amallar soni ortadi. Affiksni olib tashlashga asoslangan yana bir yondashuv – **S-stemmer** algoritmi ingliz tilidagi otlarning birlik va ko‘plik shakllarini birlashtiradi. Bu algoritm D.Xarman tomonidan taklif qilingan[33]. S-stemmer

[31] Konflatsiya (conflation) – so‘zshakllarining umumiy o‘zagini aniqlash.

[32] Frakes W.B., Fox C.J. Strength and similarity of affix removal stemming algorithms / ACM SIGIR Forum. Virginia, 2003, – №37. – P. 26-30.; Sirsat S.R., Chavan V., Mahalle H. S. Strength and Accuracy Analysis of Affix Removal Stemming Algorithms / International Journal of Computer Science and Information Technologies. Indonesia, 2013, – №4. – P. 265-269.

[33] Harman D. How effective is suffixing? / Journal of the American Society for Information Science. America, 1991, – №42. – P. 7-15.

algoritmida ko‘plikdagi qo‘shimchalarni birlik shaklga aylantirish qoidalari ishlab chiqilgan.

O‘zbek tilida stemlash masalasiga oid ba’zi tadqiqotlar amalga oshirilgan, jumladan, N.Abduraxmonova, R.Sayfullayeva va A.Ismailovlar o‘zbek tili so‘zlari uchun stemlash dasturiy vositasini ishlab chiqish masalasini o‘rganishgan va axborotni qidirish jarayonida o‘zakni aniqlash muhim ahamiyat kasb etishini qayd etishgan[34]. Tabiiy tilni qayta ishlashda matnli ma’lumotlar to‘plamini qayta ishlashning boshlang‘ich bosqichi odatda ko‘p vaqt va resurs talab qiladi. N.Abduraxmonova va boshqalar o‘z tadqiqotida strukturalanmagan matnni qayta ishlashning birinchi usullaridan biri stemlash jarayoni ekanligini ta’kidlashgan[35]. Ularning qayd etishicha, stemlash odatda ma’lumot olish va mashinali o‘rganishda qo‘llaniladi. N.Abduraxmonova va A.Ismailov tomonidan TLstemmer nomli python stemmer paketi ishlab chiqilgan bo‘lib, u turkiy tillar guruhiga mansub matndagi stemlarni aniqlaydi.

Xulosa qilib aytganda, boshqa tabiiy tillar uchun ishlab chiqilgan stemlash algortimlarini o‘zbek tili leksik birliklariga bevosita qo‘llab bo‘lmaydi. O‘zbek tilining xususiyatlaridan kelib chiqqan holda UzbStemming algoritmini ishlab chiqish dolzarb vazifa hisoblanadi.

Bobning uchinchi paragrafi “*Lemmalash jarayoni: umumiy tavsif va tahlil*” deb nomlangan. **Lemmalash** – so‘zshakl yoki stemdan lemma (leksema)ning ma’noli shaklini aniqlash va bu birliklarni tabiiy tilda qayta ishlash usuli. Lemmalash ko‘plab NLP vazifalari uchun muhim qadam hisoblanadi. Lemmalash – so‘zning lug‘aviy jihatdan normallashgan shaklini topish jarayoni. Lemmalash atroflicha o‘rganilgan jarayon sanaladi. Jumladan, A.Bjo‘rkeland, D.Seddah va A.Fraserlar matnlarni sintaktik tahlil qilish va mashina tarjimasida lemmalashdan foydalanish masalasini tadqiq qilishgan[36]. Lemmani aniqlash so‘zni lug‘aviy manbalarga solishtirish va flektiv shakllari o‘rtasidagi munosabatni o‘rnatish uchun talab qilinadi. Ayniqsa, morfologik jihatdan boy tillar uchun lemmalash nihoyatda ahamiyatli. G.Chrupala tabiiy tilga bog‘liq bo‘lmagan, tokenlarga asoslangan lemmalash usulini taklif etgan, ammo ushbu usulning samadorligi pastligi tufayli amaliyotda mazkur algoritmdan juda kam foydalanilmoqda[37].

[34] Abdurakhmonova N., Ismailov A., Sayfullaeva R. Turkic language stemmer python package for Natural Language Processing // “Computer Linguistics challenges and sollutions” international scientific-practical conference. Tashkent, 2023. https://www.researchgate.net/publication/369503003. (Saytga murojaat).

[35] Abdurakhmonova N., Babajanov U., Ismailov A., Baysariyeva S. TLstemmer (Turkic language stemmer) python package for Natural Language Processing // Conference: International Conference on Information Science and Communications Technologies. Tashkent, 2022. https://www.researchgate.net/publication/349467555. (Saytga murojaat).

[36] Björkelund A., Bohnet B., Hafdell L., Nugues P. A high-performance syntactic and semantic dependency parser // 23rd International Conference on Computational Linguistics, Proceedings of the Conference. China, 2010, – №2. – P. 33-36.; Seddah D., Chrupała G., Çetinoğlu Ö., Van Genabith J., Candito M. Lemmatization and Lexicalized Statistical Parsing of Morphologically Rich Languages: the case of French // 1st Workshop on Statistical Parsing of Morphologically-Rich Languages. Los Angeles, 2010. – P. 85-93.; Fraser A., Weller M., Cahill A., Cap F. Modeling inflection and word-formation in SMT // Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. France, 2012. – P. 664–674.

[37] Chrupała G. Simple data-driven context-sensitive lemmatization // Procesamiento Del Lenguaje Natural. Ireland, 2006, – №37 – P. 121-127

O‘zbek tilidagi leksik birliklarni lemmalash algoritmlari bo‘yicha bir necha tadqiqotlar olib borilgan hamda kichik matnlarda sinovdan o‘tkazilgan. Xususan, I.Bakayev[38] *o‘zbek tili so‘zlariga lingvistik qoidalarga asoslangan yondashuv asosida stemlash algoritmini ishlab chiqish* masalalarini o‘rgangan. M.Sharipov, O.Sobirovlar **o‘zbek tili** uchun chekli holatlar mashinasi **(Finite State Machine)** orqali qoidaga asoslangan lemmalash algoritmini yaratgan[39]. Affikslarni olib tashlashda ularning ma’lumotlar bazasi va POS teglash qo‘llanilgan. Mashinali o‘rganish va neyron tarmog‘iga asoslangan lemmalashni turli tillarda qo‘llash bo‘yicha tadqiqotlar natijasi yuqori samaradorlikka ega bo‘lsa-da, amalda ular o‘zbek tiliga qo‘llanilmagan.

Dissertatsiyaning II bobi **“O‘zbek tili leksik birliklarini stemlash va lemmalashning lingvistik asoslari”** deb nomlangan bo‘lib, birinchi paragraf *“O‘zbek tili so‘zshakllarini stemlash: muammo va yechimlar”* deb ataladi. O‘zbek tilida so‘zning eng kichik ma’noli qismidan biri o‘zak **(root)** bo‘lsa, so‘zshaklning eng katta ma’noli qismi **stem**dir. Shuning uchun NLPda so‘z ikki qismdan iborat deb hisoblanadi: **ma’noni anglatuvchi stem** va **qo‘shimchalar**.

Morfologik tahlil sohasida hozirgacha bajarilgan ishlarni ikki sinfga ajratish mumkin: **stemga** va **affikslarni olib tashlashga asoslangan** yondashuv. *Stemga asoslangan morfologik tahlil*ni amalga oshirish uchun dastlab so‘zning stemi, so‘ngra affikslari aniqlanadi. *Affikslarni olib tashlashga asoslangan yondashuv*da dastlab so‘zdagi affikslar aniqlanadi.

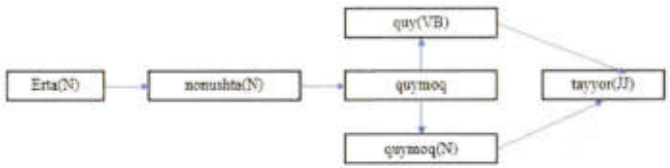
O‘zbek tilidagi so‘zlarning stemini aniqlash uchun o‘zak va unga birikadigan barcha turdagi qo‘shimchalar aniqlanadi. An’anaviy stemlash algoritmlari qo‘shimchalar va morfologik qoidalarga asoslangan bo‘lib, stemlash jarayoni natijasida quyidagi muammolar yuzaga kelishi mumkin:

o‘zak va qo‘shimchaning bitta o‘zak bilan omonim bo‘lishi;

so‘zning tovush o‘zgarishiga uchrashi;

neologizm va NER(Named entity recognition)larni stemlash.

O‘zbek tilidagi gaplarda stemdagi noaniqlikni quyidagi 2-rasmda ko‘rish mumkin:



2-rasm. O‘zbek tilidagi gaplarda stemdagi noaniqlik

[38] Bakayev I. Development of a stemming algorithm based on a linguistic approach for words of the uzbek language // International Conference on Scientific, Educational & Humanitarian Advancements. Bukhara, 2021. – P. 195-202.

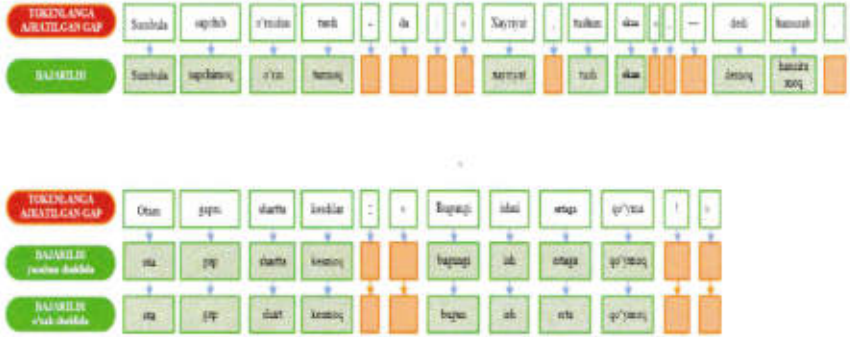
[39] Sharipov M., Sobirov O. Development of a Rule-Based Lemmatization Algorithm Through Finite State Machine for Uzbek Language // The International Conference and Workshop on Agglutinative Language Technologies as a challenge of Natural Language Processing. Koper, Slovenia, 2022. – P.1-6.

Stemni aniqlashdagi tovush o‘zgarishiga uchrashi muammosini hal qilish uchun birinchi bosqichda o‘zak va qo‘shimchalarning chegaralari aniqlanadi, ikkinchi bosqichda esa lemmalash amalga oshiriladi. Lemmalash natijasida xato hosil qilingan stemlar lug‘atda mavjud root (o‘zak)ga o‘zgartiriladi. NER va neologizmlarning stemini topish muammosi yuzaga kelganda shakl yasovchilar kesiladi, so‘z yasovchi shaklidagi qo‘shimcha yoki so‘zning qismi qoldiriladi va ushbu qism stemga teng keladi.

Bobning ikkinchi paragrafi *“O‘zbek tilida turli tuzilishli so‘zlarni lemmalash usullari”* deb ataladi. Lemmalash – so‘zning turli xil flektiv shakllarini bitta element sifatida tahlil qilish uchun guruhlash jarayoni. U so‘zning normallashgan shaklini yoki lemma deb ataladigan so‘zning lug‘at shaklini aniqlashda ishlatiladi. Leksema bir xil ma’noga ega bo‘lgan barcha shakllarning (lug‘at bosh so‘zlari deyiladi) birlashmasi bo‘lsa, lemma leksemani ifodalovchi tayanch shakl sifatida tanlangan so‘zdir. O‘zbek tilida so‘z tarkibining odatiy strukturasi quyidagi ko‘rinishda bo‘ladi:

“O‘zak + so‘z yasovchi + lug‘aviy shakl + sintaktik shakl”.

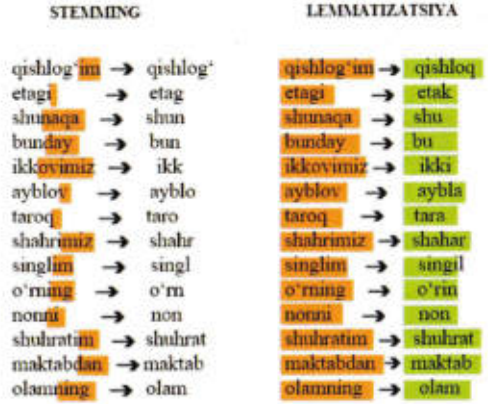
Quyida lemmalash jarayoni ko‘rsatiladi[40]:



3-rasm. Lemmatizatsiya jarayoni

O‘zbek tilida stemlash va lemmalash natijasi bir xildek ko‘rinadi, ammo ular batamom farqli jarayonlar sanaladi: stemlash o‘zakdan qo‘shimchani kesish, lemmalash esa o‘zakning lug‘atdagi variantini aniqlash jarayonidir (4-rasm).

[40] http://uznatcorpara.uz/uz/Lemmatizer



4-rasm. Stemlash va lemmalash o'rtasidagi farq

Sodda tub va sodda yasama so'zlar. Sodda tub va sodda yasama so'zlar odatda bitta lemmaga teng. Ularni lemmalash uchun sintaktik va lug'aviy shakl yasovchi qo'shimchalarni kesib tashlash kifoya.

Qo'shma so'zlar. Qo'shma so'zlarning lemmasini aniqlashda quyidagi qoidalarga rioya qilinadi:

1. Qo'shib yoziladigan qo'shma so'zlar lemmasi shakl yasovchi qo'shimchalarni *kesib tashlash orqali hal qilinadi.*
2. Ajratib yoziladigan qo'shma so'zlarni lemmalashda *lug'atga asoslaniladi.*
3. Tarkibli atoqli otlar ajratib yoziladi va NER hisoblanadi: *Yangi O'zbekiston, Birlashgan Millatlar Tashkiloti, Jahon Sog'liqni Saqlash Tashkiloti, Milliy tiklanish partiyasi.* Bunday birliklarni lemmalash alohida lingvistik va matematik modellarni talab qiladi.

Qo'shma fe'l bilan shaklan o'xshash birliklar mavjud, masalan, iboralar. Iboralar tarkibiga ko'ra quyidagi tuzilishga ega:

1) $W_1 + W_2$
2) $W_1 + W_2 + W_3$
3) $W_1 + W_2 + W_3 + W_4$

Bular orasidan $W_1 + W_2$ tarkibli iboralar qo'shma fe'l tuzilishiga o'xshab ketadi. Ot+fe'l shaklidagi iboralar (*yaxshi ko'rmoq, ko'zda tutmoq, quloq solmoq, jon bermoq, ko'zini uzmoq*) bitta lemmaga teng, ammo ular lemma emas, frazeologik birlik sifatida teglanadi.

Juft so'zlar. Juft so'zlarning lemmasini aniqlashda quyidagi qoidalarga rioya qilinadi:

1. **Ikkala qismi ma'no beruvchi:** *achchiq-chuchuk.* Bunday holda stem 2 ta, lemmasi bitta bo'lib, bunda ham lug'atga asoslaniladi.
2. **Biri ma'no beruvchi, biri ma'no bermaydigan:** *kalta-kul*ta, bunday holda stem faqat ma'no beruvchi qism, lemmasi bitta bo'lib, bunda ham lug'atga asoslaniladi.

3. **Ikkala qismi ham ma'no bermaydigan:** *g'idi-bidi, jiz-biz.* Juft so'zlarning ikkita qismi ham ma'noga ega bo'lmasa, shu juft so'zning o'zi stem, o'zak va lemma bo'ladi.

Takror so'zlar. Takror so'zlarning lemmasini aniqlashda quyidagi qoidalarga rioya qilinadi:

1. **Aynan takror:** *katta-katta, tez-tez, baland-baland.*
2. **Ikkinchi qismi birinchi qismining fonetik o'zgarishga uchragani:** *tuz-puz, non-pon.*

Takror so'zlarni lemmalashda ham lug'at ma'lumotiga asoslaniladi. Yuqoridagilardan ko'rinadiki, lemmalashda lug'at ma'lumoti – lingvistik ma'lumotlar bazasi muhim. Faqat stemni aniqlashda shakl yasovchi (sintaktik va lug'aviy)lar kesiladi. Takror so'zdan biri alohida kelganda POS tegi boshqa turkum, takror kelganda POS tegi boshqa turkumga mansub bo'ladi: *dedi* yakka holda qo'llanganda fe'l, *dedi-dedi* shaklida takror so'z bo'lib kelganda ot turkumiga mansub bo'ladi. Lekin bu tipik holat emas: *tez* ravish, *tez-tez* ravish; *katta* sifat, *katta-katta* sifat.

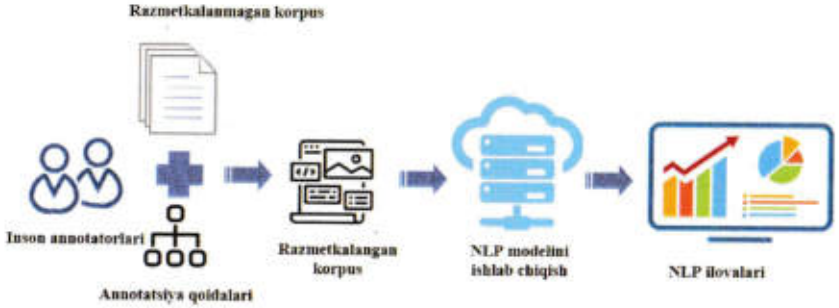
Bobning uchinchi paragrafi "*Til korpusi qidiruv tizimini optimallashtirishda lemmalashning ahamiyati*" deb nomlanadi. Bugungi kunda NLP algoritmlaridan *elektron pochtada spamni aniqlash* va *qidiruv tizimlarida foydalanuvchi so'rovlarini optimallashtirish* kabi masalalarni hal qilishda foydalanilmoqda. NLP – mashinalarga insonlar bilan *tabiiy tilda muloqot qilishda, matnni intellektual tahlil qilish, nutqni tushunish* va *uni to'g'ri formatda talqin qilish* kabi vazifalarni bajarishda yordam beradi[41]. Tizimlashtirilmagan ma'lumotlarni NLP vositalari orqali qayta ishlash murakkab va dolzarb amal bo'lib, NLP usullari vositasida matndan kerakli ma'lumotlarni yetarli aniqlikda olish nihoyatda ahamiyatli.

Lemmalash jarayoni qidiruv tizimini optimallashtirish (Search Engine Optimization, SEO)da muhim ahamiyatga ega bo'lib, lemmalashni amalga oshirish bosqichlari, metod va algoritmlarining qanday ishlashini, uni korpus matnlariga qo'llash usullarini ishlab chiqish lozim. Korpus matnlari (veb-sayt)dagi *kalit so'zlarni o'rganish, sahifani optimallashtirish* va *qidiruv reytinglariga ta'sir ko'rsatadigan boshqa omillarga e'tibor qaratish* ham muhim. Lemmalash eng yaxshi natijalarga erishish uchun keng qamrovli SEO strategiyasining bir qismi sifatida ishlatilishi kerak. Korpus matnlarini qo'lda lemmalash ko'proq vaqt talab qilsa-da, aniqroq natijalarni beradi. 5-rasmda korpus matnlarini razmetkalash bosqichlari keltirilgan.

Veb-saytda lemmalashni amalga oshirish. Matn tasnifi, ma'lumotlarni qidirish va *mashina tarjimasi* kabi vazifalarini hal qilishda lemmalash yordam beradi. Veb-saytda lemmalashni amalga oshirish uchun turli yondashuvlar mavjud.

1. *Oldindan o'qitilgan lemmatizatsiya modelidan foydalanish.*
2. *Maxsus lemmatizatsiya modelini ishlab chiqish.*
3. *Onlayn lemmatizatsiya xizmatidan foydalanish.*

[41] Elov B.B., Hamroyeva Sh.M., Xusainova Z.Y. NLP (tabiiy tilga ishlov berish)ning vazifalari va zamonaviy yondashuvlar / Filologik tadqiqotlar: til, adabiyot, ta'lim. Termiz, 2022, – №5. – B. 19-26.



5-rasm. Korpus matnlarini razmetkalash bosqichlari

Tanlangan yondashuvdan qat'i nazar, veb-saytda lemmatizatsiyani amalga oshirish uchun bir necha qadamlarni bajarish kerak:

boshlang'ich ishlov berish;
tokenlash;
POS teglash;
lemmalash;
integratsiya.

Umuman olganda, lemmalash veb-saytning SEO faoliyatini o'stirishda zaruriy vosita hisoblanadi. *Tegishli kalit so'zlarni aniqlash, mavjud kontentni tahlil qilish, lemmalashni amalga oshirish, samaradorlikni kuzatish, natijalarni taqqoslash* veb-saytning SEOda lemmalash samaradorligini samarali sinab ko'rish va qidiruv tizimlari uchun kontentni qanday optimallashtirish haqida to'g'ri qaror qabul qilishga olib keladi.

Dissertatsiyaning **III bobi "O'zbek tili leksik birliklarini tokenlash, stemlash va lemmalashning dasturiy asoslari"** deb atalgan. Mazkur bobning birinchi paragrafida *"O'zbek tili birliklarini BoW va BPE algoritmlari asosida tokenlash"* masalasi ko'rib chiqiladi. BoW (so'zlar sumkasi) modeli – mashinali o'rganish algoritmlari tomonidan qayta ishlash lozim bo'lgan matnning raqamli ko'rinishi. Bag of Words (BoW) modellashtirish algoritmidan foydalanib, matnni raqamli matritsalarga aylantirish va qayta ishlash mumkin.

BoW algoritmi yordamida matnli ma'lumotlardan zarur *xususiyatlar ajratib olinadi*. Ushbu yondashuv hujjatlardan ma'lum xususiyatlarni ajratib olishning *oddiy va moslashuvchan* usulidir. BoW usulida faqat matndagi so'zlar soni (statistikasi) aniqlanadi. Biroq grammatik tafsilotlar va so'z tartibi e'tiborsiz qoldiriladi. Lug'at hajmi – BoW modeli oldida turgan katta muammo. Agar model hali ishlatilmagan yangi so'zga duch kelsa, BoW modelida bu so'z inobatga olinmaydi.

Byte Pair Encoding (BPE) – transformerga asoslangan modellar (*transformer-based models*) orasida keng qo'llaniladigan tokenlash usuli. BPE so'zlar va belgilar tokenizatorlari muammolarini hal qiladi[42]:

[42] Bostrom K., Durrett G. Byte pair encoding is suboptimal for language model pretraining // Findings of the Association for Computational Linguistics: EMNLP. Austin, 2020. – P. 4617-4624.; Heinzerling B., Strube M.

1. BPE OOV muammosini samarali hal qiladi. U OOVni so'z tarkibi elementlari sifatida ajratadi va qayta ishlaydi.
2. BPEdan keyin kirish va chiqish gaplarining uzunligi belgilarni tokenlashga nisbatan qisqaroq.

BPE – so'zlarni segmentlash algoritmi bo'lib, u eng ko'p uchraydigan belgilar yoki belgilar ketma-ketligini iterativ tarzda birlashtiradi. BPE algoritmini amalga oshirish qadamlari:

1. Korpusdagi so'zlarni </w> qo'shimchasidan keyin belgilarga ajratish.
2. Korpusdagi noyob so'zlar bilan lug'atni shakllantirish.
3. Korpusdagi bir juft belgilar yoki belgilar ketma-ketligi chastotasini hisoblash.
4. Korpusdagi eng ko'p uchraydigan juftlikni birlashtirish.
5. Eng yaxshi juftlikni lug'atga saqlash.
6. Muayyan miqdordagi iteratsiya uchun 3-5-bosqichlarni takrorlash.

Korpusni samarali tarzda qayta ishlash uchun BPEning har bir iteratsiyasida uning chastotasiga qarab birlashishning har qanday potentsial variantidan o'tadi va eng yaxshi yechimni optimallashtirish uchun o'ziga xos yondashuvga amal qiladi. Yuqorida keltirilgan usullar o'zbek tili korpusi(matnlar)ga qo'llanildi va https://uznatcorpara.uz/uz/Tokenizer dasturiy ta'minoti ishlab chiqildi.

Bobning ikkinchi paragrafi *"Affikslarni ajratishga asoslangan UzbStemming algoritmini ishlab chiqish"* deb atalgan. Affikslarni olib tashlashga asoslangan o'zbek tili stemmerini ishlab chiqish uchun *til korpusi, affikslar bazasi, o'zbek tili morfoleksikoni, root (o'zak)lar bazasidan* foydalaniladi[43]. So'zshaklning turkumi (POS tegi)ni inobatga olib, stemni aniqlash uchun quyidagi ikki bosqichdagi amallar bajarildi.

I-bosqich. Ma'lumotlar (dataset)ni tayyorlash

1. O'zakka qo'shilishi mumkin bo'lgan **prefiks va suffikslar** aniqlanadi (1-, 2-jadvallar).

1-jadval

O'zbek tilidagi prefikslar		
Prefix qo'shimchalar		
ba-	be-	ham-
bo-	no-	bar-
bad-	ser-	be-
bar-	g'ayri-	xush-

BPEMB: Tokenization-free pre-trained subword embeddings in 275 languages // 11th International Conference on Language Resources and Evaluation. Miyazaki, Japan, 2019. – P. 2989-2993.; Gutierrez-Vasques X., Bentz C., Sozinova O., Samardžić T. From characters to words: The turning point of BPE merges // Proceedings of the16th Conference of the European Chapter of the Association for Computational Linguistics. 2021.– P. 3454-3468. https://aclanthology.org/2021.eacl-main.302.pdf (Saytga murojaat)

[43] Elov B., Hamroyeva Sh., Abdullayeva O, Xusainova Z., Xudayberganov N. Agglyutinativ tillar uchun POS teglash va stemming masalasi (turk, uyg'ur, o'zbek tillari misolida) / O'zbekiston: Til va Madaniyat (Kompyuter lingvistikasi). Toshkent, 2023, – №1. – B. 35-50.

2-jadval

O‘zbek tilidagi suffikslar (POS bo‘yicha)

POS	so‘z yasovchi (derevatsion)	shakl yasovchi (flektiv)
Ot (N)	51	17
Son (Num)	1	5
Fe’l (Vb)	30	22+10 (nisbat)
Sifat (JJ)	97	9
Olmosh (P)	-	-
Ravish (R)	18	5

2. Har bir so‘z turkumiga oid bir nechtadan (faol) so‘zshakllar olinib, ularning stemini so‘zshakldan o‘chirib tashlash orqali berilgan so‘z turkumidagi stemlarga qo‘shilishi mumkin bo‘lgan **prefiks va suffikslar satrlari to‘plami** aniqlanadi.

3. Qo‘shimchalar satri bu so‘zshaklini hosil qilish uchun stemga qo‘shilgan o‘zbek tilidagi bir nechta **qo‘shimchalar birikmasi** hisoblanadi (3-jadval).

3-jadval

Til korpusidan aniqlangan o‘zbek tilidagi prefiks va suffiks birikmalari to‘plami

	Ot (N)	Son (Num)	Fe’l (Vb)	Sifat (JJ)	Olmosh (P)	Ravish (R)
1.	chiligimizning	Inchi	dılarki	almashtirgichda	nga	aydiki
2.	dagilarning	larcha	nganlaridan	bardoshligini	ngacha	aydimi
3.	doshlarimiznikidan	larda	yapsizlarmi	imizning	ngagina	chiligimizning
4.	korlarimizning	ovi	ilmayotganidan	lamaganidan	nikilarning	dangina
5.	laganlaridagina	ovimiz	ishimizdan	lashayotganligini	dek	gilarining
...	...	...	...	...	...	...
Jami:	1773	229	15870	725	509	1376

4. Stemga qo‘shilishi mumkin bo‘lgan barcha qo‘shimchalar satrini **til koprusidan olish uchun** maxsus dasturiy ilovadan[44] foydalaniladi.

5. Ushbu dasturiy vosita yordamida o‘zbek tili korpusidagi mavjud so‘zshakllaridan berilgan **regular ifodalar asosida kerakli so‘zshakllar bazasini shakllantirish** imkoniyati taqdim etiladi.

6. Qidiruv natijasida aniqlangan ba’zi so‘zshakllarning prefiks, stem va suffikslar satri kabi qismlarga ajratiladi.

7. $\boldsymbol{p + s' + q = w}$ shartni qanoatlantiradigan, barcha berilgan har bir so‘zshakli uchun $\boldsymbol{g = \langle p, s, q \rangle}$ **qiymatlar to‘plami** aniqlandi. Bunda,

$\boldsymbol{w}$ – so‘zshakli;

$\boldsymbol{s'}$ – stem;

$\boldsymbol{p \in P}$;

$\boldsymbol{P}$ – prefiks qo‘shimchalar satri to‘plami;

$\boldsymbol{s \in S, S}$ – stem so‘z turkumlari to‘plami;

$\boldsymbol{q \in Q, Q}$ – suffiks qo‘shimchalar satri to‘plami;

$\boldsymbol{g \in G, G}$ – har bir elementi $\{\boldsymbol{p, s, q}\}$ uchta parametrdan iborat to‘plam.

8. Aniqlangan $\{\boldsymbol{p, s, q}\}$ qiymatlariga misol jadvalda keltiriladi.

Qiymatlar aniqlanganidan so‘ng **II bosqichda UzbStemming algoritmi** yordamida so‘zshaklning stemi aniqlanadi.

[44] http://uznatcorpara.uz/uz/Dictionary/Search

Affikslarni olib tashlashga asoslangan yuqorida keltirilgan UzbStemmerini o‘zbek tili ta’limiy korpusi (http://uzschoolcorpara.uz/uz/Search)dagi 100000 dan ortiq gaplarga qo‘llash natijasida **97,5%** lik aniqlikdagi natija olindi.

Uchinchi paragraf *“O‘zbek tilidagi so‘zlarni lemmalash muammosi va usullari”* deb nomlangan. Barcha tillarda lemmani aniqlash uchun turli qoidalar talab qilinadi, chunki tillardagi so‘zlarning tuzilishi har xil. O‘zbek tili agglyutinativ tillar oilasiga mansub bo‘lganligi sababli gaplarda nechta token bo‘lsa, shuncha lemma bo‘lishi mumkin. Lekin har doim ham bu fikr to‘g‘ri bo‘lmaydi. Ba’zi gaplarda lemma soni tokenlarga soniga teng kelmasligi mumkin. Gapni so‘zlarning mantiqan birikuvidan iborat deb hisoblasak, quyidagicha bo‘lishi mumkin:

1. $W_1, W_2, \ldots W_n$ *(tokenlar)*

2. $L_1, L_2, \ldots L_m$ *(lemmalar),* $(m \leq n)$ shaklida bo‘lishi kuzatiladi.

Ko‘p hollarda bitta token bitta lemmaga mos keladi. O‘zbek tilida so‘zshakllar tuzilishi va hosil bo‘lish usuliga ko‘ra turlarga ajratiladi. So‘zlarning tuzilishiga ko‘ra turli shakllarini lemmalash qoidalari quyidagicha belgilanadi:

1. Juft so‘zga so‘z yasovchi qo‘shimcha qo‘shilganda chiziqcha saqlanib qoladi: *baxt-saodatli.* O‘zbek tilida juft so‘zlarning lemmasini aniqlashda lug‘atga asoslanamiz, ya’ni ikkala qismi ham mustaqil ma’noli bo‘lgan juft so‘zlarda ikkita lemma mavjud deb hisoblanadi.

2. Takroriy so‘zlar yasama so‘zlar hisoblanmagani uchun tarkibida ikkita token va ikkita lemma mavjud deb hisoblaymiz. Ammo takroriy so‘zning ikkinchi qismi ma’noga ega bo‘lmagan istisnoli holat mavjudki, bu lemmalar sonini aniqlashda muammo tug‘diradi. Bu holatda lug‘atga asoslanib, ma’noli qismi olinadi.

1) (ma’noli) o‘zak +(ma’noli) o‘zak: tez-tez, qator-qator, qop-qop, kunma-kun, oyma-oy, shu-shu, tez-tez, yilma-yil, baland-baland, ko‘p-ko‘p;

2) (ma’noli) o‘zak + (ma’nosiz) o‘zak: tuz-puz, don-dun, osh-posh, choy-poy, mol-hol, irim-sirim, qon-pon, pul-mul, meva-cheva.

3. Qo‘shma so‘zlar turli o‘zak so‘zlarning birikib, yangi so‘z hosil qilishi, yasama so‘z vujudga kelishi sababli tarkibida nechta negiz bo‘lsa ham bitta lemma deb olinadi.

4. Birdan ortiq token bitta lemmaga teng keladi, chunki qo‘shma fe’llar ikki va undan ortiq asosning birikuvidan hosil bo‘ladi, ulardan yaxlit bir ma’no anglashiladi. Mazkur tokenlar turli turkumlarga mansub bo‘lishi, qo‘shma fe’lning tarkibi quyidagicha bo‘lishi mumkin:

1. “Ot + fe’l” tuzilishli (fe’l bo‘lmagan so‘z bilan fe’ldan tarkib topgan) qo‘shma fe’llar: *ta’zim qilmoq, g‘ayrat qilmoq, fikr qilmoq* kabi;

2. “Fe’l + fe’l” tuzilishli (ikki fe’ldan tarkib topgan) qo‘shma fe’llar: *olib bormoq; sotib olmoq, olib qochmoq, chiqib ketmoq* kabi.

“Fe’l + fe’l” tuzilishli qo‘shma fe’llar. Qo‘shma fe’lning bu turi ikki fe’lning birikuvidan hosil bo‘ladi. Birinchi qism odatda fe’lning -a, -y hamda -b (-ib) ravishdoshi shaklida bo‘ladi. Qo‘shma fe’l tarkibidagi qismlar o‘zaro birikib, ularning ma’nosidan farq qilgan yangi ma’no anglatadi: *olib bormoq, olib kelmoq, sotib olmoq, tashlab ketmoq.*

Ko‘makchi fe’lli so‘z qo‘shilmalari tarkibidagi tokenlar soniga ko‘ra quyidagicha bo‘lishi kuzatiladi:

1. Ikki tokenli: *aytib bermoq, yozib yubormoq.*

2. Uch tokenli: *yozib berib turmoq.*

3. To‘rt tokenli: *yozib berib turgan edi/emish/ekan.*

4. Yetakchi fe’li qo‘shma fe’ldan iborat KFSQ: *javob berib turmoq, imzo qo‘yib bermoq va h.*

Qo‘shma fe’l va KFSQ tarkibida nechta token bo‘lsa ham, bitta lemma hisoblanadi.

Qo‘shma ravish. Qo‘shma ravish tarkibida nechta token bo‘lsa ham, lemmasi bitta hisoblanadi. Qo‘shma ravishlar ikki so‘z asosining birikishi orqali tuziladi: *har gal, bir vaqt, tez fursatda, shu yoqqa, bir mahal, bir zum, bir nafas, har dam, har on* kabi.

Qo‘shma son. Qo‘shma son qismlarida ham nechta token bo‘lsa-da, bitta lemma deb olinadi. Qo‘shma son tarkibi ham ajratib yoziladi. Masalan, *o‘n besh, ikki ming yigirma uch, besh yuz olti, yigirma bir* va h.

Qo‘shma olmosh. Qo‘shma olmoshlarda ikkita token, bitta lemma bo‘ladi. *Hech kim, hech qachon, har bir, har kim, mana shu, ana o‘sha* kabi olmoshlar qo‘shma olmoshlardir. Sodda olmosh qismlari qo‘shib, qo‘shma olmosh qismlari esa ajratib yoziladi.

Ibora. Iboralar ba’zan ikki tokenli, uch tokenli, hatto to‘rt va undan ko‘p tokenli bo‘lishi mumkin, ammo bitta lemma deb olinadi. O‘zbek tilida frazeologik iboralar ikki, uch va undan ortiq so‘zlardan tashkli topadi. Ibora – tilimizning lug‘at tarkibini tashkil etuvchi birliklardan biridir. Ular ikki va undan ortiq so‘zning ko‘chgan ma’nolari asosida tarkib topgan lug‘aviy birlik bo‘lib, xuddi so‘z kabi lug‘aviy ma’noni anglatadi. Masalan, *xamirdan qil sug‘urganday* iborasi, “osonlik bilan”, “qiyinchiliksiz ” ma’nosini, *ko‘ngil bermoq* iborasi “sevmoq” ma’nosini, *qo‘y og‘zidan cho‘p olmagan* iborasi “yuvosh” ma’nosini bildiradi.

Shuningdek, tilimizda nomni bildiruvchi so‘zlar (NERlar), hali lug‘at boyligiga o‘zlashib ulgurmagan yangi so‘zlar (neologizmlar) va hujjatlarda belgilab qo‘yilgan qisqartmalar (abbreviaturalar) mavjudki, ularning lemmasini aniqlashda ham bir tartibga kelish talab etiladi. Mazkur so‘zshakllarda nechta token bo‘lsa ham, stemi kabi lemmasi bitta deb belgilanadi va quyidagicha xulosa qilinadi.

1. NERlarning lemmasi va stemi bir xil. (Atoqli ot tarkibida uchragan so‘z yasovchi qo‘shimcha inobatga olinmaydi).

2. Neologizmlarning lemmasi va stemi bir xil.

3. Qisqartmalar (abbreviatura (BMT), shartli qisqartmalar (sh.k., va b., sm, mlrd.) ning lemmasi va stemi bir xil.

Demak, tilda so‘zshaklning lemmasini aniqlashda kontekstda vazifa bajaradigan mustaqil ma’noga ega qismi olinadi. So‘zshakllardagi barcha shakl hosil qiluvchi affikslar olib tashlanadi va so‘z yasovchi qo‘shimchalar qoldiriladi.

XULOSA

1. Tokenlash – berilgan matndagi gaplarni eng kichik o‘lchov birligi – tokenga ajratish jarayoni. Gapda tinish belgi, so‘z va raqamlar token sifatida belgilanishi mumkin. Tokenlash jarayonida tabiiy tilda matnga boshlang‘ich ishlov berish orqali tokenlarga ajratiladi; tokenlar orasidagi chegaralar belgiladi. Tokenlash so‘z, belgi yoki so‘z tarkibi elementlari darajasida amalga oshirilishi mumkin.

2. NLPning ko‘p vazifalarni yechishda so‘zshakllarini o‘zakkacha qisqartirish (stemlash)ga to‘g‘ri keladi. Stemlash – NLPda matnga ishlov berishning dastlabki bosqichi, bu amal NLPning imlo tekshiruvi kabi ilovalari samaradorligini oshirishga xizmat qiladi. Bugungi kunda Porter stemmeri algoritmi modifikatsiya qilinib, keng miqyosda foydalanilmoqda. Stemlashda korpusga asoslangan (1); gibrid yondashuv (2); statistik usul (3); n-gramm stemmer (4); kontekstga bog‘liq stemmer (5) kabi usul va yondashuvlarga asoslaniladi.

3. Stemmer algoritmlarining afzallik va kamchiliklari mavjud. Boshqa tabiiy tillar uchun ishlab chiqilgan Lovins, Porter, Peys/Hask, Davson, Snovball, Krovetz, Lankaster, YASS va HMM stemmerlarning birortasi 100 foiz natija bermaydi. Mazkur stemlash algortimlarini o‘zbek tili leksik birliklariga bevosita qo‘llab bo‘lmaydi, shu sababli o‘zbek tilining xususiyatlaridan kelib chiqqan holda UzbStemming algoritmini ishlab chiqish dolzarblik kasb etadi.

4. Lemmalash NLPning qator vazifalarida muhim qadam bo‘lib, so‘zning lug‘atdagi (asosiy) shaklini topishdan iborat. Stemlash va lemmalash – tabiiy tilda so‘zning o‘zagini aniqlashda keng tarqalgan usullardir. Lemmalash turli xususiyatli tillarda turlicha amalga oshiriladi. Lemmalash amali POS teglash, imloni tekshirish va hujjatlarni klasterlash kabi NLP ilovalarida qo‘llaniladi. Bugugi kunda o‘zbek tilidagi kichik hajmli matnlarda lemmalash algoritmlari sinovdan o‘tkazilgan. Turli tillarda mashinali o‘rganish va neyron tarmoqqa asoslangan lemmalash algoritmlari ishlab chiqilgan bo‘lsa-da, hozirgacha bu usullarning birortasi o‘zbek tiliga joriy etilmagan.

5. O‘zbek tili agglyutinativ xususiyatli bo‘lganligi sababli lingvistik qoidalarga asoslangan stemlash algoritmlarini ishlab chiqish maqsadga muvofiq. Stem so‘zshaklning qo‘shimchalarini kesib tashlashdan hosil bo‘luvchi qism bo‘lib, ba’zi hollarda ma’no anglatmasligi, stem so‘zning morfologik o‘zagi bilan aynan mos bo‘lmasligi yoki mos tushishi mumkin. O‘zbek tilida stemlash so‘zga qo‘shilgan so‘z yasovchi va shakl yasovchi qo‘shimchalarni olib tashlash orqali o‘zakkacha qisqartirishdir.

6. O‘zbek tilidagi so‘zlarning stemini aniqlashda tovush o‘zgarishiga uchrash muammosini hal qilishda birinchi bosqichda o‘zak va qo‘shimchalarning chegarasi aniqlanadi, ikkinchi bosqichda esa lemmalash amalga oshiriladi. Lemmalash natijasida xato hosil qilingan stemlar lug‘atda mavjud o‘zakka o‘zgartiriladi. NER va neologizmlarning stemini topish muammosi yuzaga kelganda shakl yasovchilar kesiladi, so‘z yasovchi shaklidagi qo‘shimcha yoki so‘zning qismi qoldiriladi va ushbu qism stemga teng keladi.

7. Lemmalash amali til korpusi matnlari mazmunini tahlil qilish aniqligini oshiradi. Lemmalash qidiruv tizimi natijalarining aniqligi va korpus matni kontentini tushunishda amaliy ahamiyat kasb etadi. Lemmalashning yana bir afzalligi veb-sahifalar tarkibining umumiy tuzilishi va kalit so‘zlarni aniqlashga yordam berishidir. Tanlangan yondashuvdan qat’i nazar, veb-sahifada lemmalashni amalga oshirishda boshlang‘ich ishlov berish (1); tokenlash (2); POS teglash (3); lemmalash (4); integratsiya (5) kabi qadamlar bajariladi.

8. Bugungi kunda tokenlash jarayonini amalga oshirishda BoW (Bag of Words) va BPE (Byte Pair Encoding) algoritmlaridan keng foydalanilmoqda. BoW (so‘zlar sumkasi) – bu tabiiy tilni qayta ishlashda matnni modellashtirish (raqamlashtirish) usuli, axborot tizimlaridagi katta hajmdagi matnli ma’lumotlarni qayta ishlash imkonini beradi. BoW algoritmi yordamida matnli ma’lumotlardan zarur xususiyatlar ajratib olinadi va matndagi so‘zlar soni (statistikasi) aniqlanadi.

9. Byte Pair Encoding (BPE) – transformerga asoslangan modellar (transformer-based models) orasida keng qo‘llaniladigan tokenlash usuli, so‘z va belgilar tokenizatorlari muammolarini yecha oladi. BPE so‘zlarni segmentlash algoritmi bo‘lib, u eng ko‘p uchraydigan belgi yoki belgilar ketma-ketligini birlashtiradi. O‘zbek tilidagi leksik birliklari BoW va BPE usullari orqali tokenlash amalga oshirildi va dasturiy ta’minoti (https://uznatcorpara.uz/uz/Tokenizer) ishlab chiqildi.

10. Affikslarni olib tashlashga asoslangan UzbStemmeri dasturiy ta’minotini ishlab chiqishda *til korpusi, affikslar bazasi, o‘zbek tili morfoleksikoni, root (o‘zak)lar bazasi, POS teglashdan* foydalanilib, ikki bosqichli algoritmni o‘zbek tili ta’limiy korpusi (http://uzschoolcorpara.uz/uz/Search)dagi 100000 dan ortiq gaplarga qo‘llash natijasida 97.5% lik aniqlikdagi samaradorlikka erishildi.

11. Lemmalash algoritmi – bu so‘z o‘zagini topishning murakkab usuli, algoritmni qo‘llashda tabiiy til va uning grammatikasini tushunish lozim. O‘zbek tili agglyutinativ tillar oilasiga mansub bo‘lganligi sababli so‘zshakllar so‘z+affiks yoki o‘zak+o‘zak shaklida hosil qilinadi. Gapdagi tokenlar soni lemmalar soniga teng bo‘lishi mumkin. Ba’zi gaplarda lemmalar soni tokenlar soniga teng kelmaydi. Mazkur tadqiqotda sodda so‘z, takror so‘z, juft so‘z, qo‘shma so‘z (qo‘shma fe’l, qo‘shma ravish, qo‘shma son), KFSQ va ibora kabi leksik birliklarni lemmalash usullari keltirildi va https://uznatcorpara.uz/uz/Lemmatizer dasturiy ta’minoti ishlab chiqildi.

ONE-TIME SCIENTIFIC COUNCIL ON THE BASIS OF SCIENTIFIC COUNCIL OF GRANTING SCIENTIFIC DEGREE DSc.03/30.12.2019. Phil.19.01 UNDER TASHKENT STATE UNIVERSITY OF UZBEK LANGUAGE AND LITERATURE NAMED AFTER ALISHER NAVOI

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

XUSAINOVA ZILOLA YULDASHEVNA

LINGUISTIC BASES AND SOFTWARE OF TOKENIZATION, STEMMING, LEMMING OF LEXICAL UNITS OF THE UZBEK LANGUAGE

10.00.11 – Language theory. Applied and computer linguistics

DISSERTATION ABSTRACT OF DOCTOR OF PHILOSOPHY (PhD) IN PHILOLOGICAL SCIENCES OREFERATI

Tashkent – 2024

The theme of the dissertation of Doctor of Philosophy (PhD) was registered at the Supreme Attestation Commission of the Republic of Uzbekistan under the number B2023.2.PhD/Fil3688.

The dissertation has been carried out at Tashkent State University of Uzbek Language and Literature named after Alisher Navoi.

The abstract of dissertation is posted in three languages (Uzbek, English and Russian (resume)) on the website of Scientific Council (www.tsuull.uz) and on the website "ZiyoNet" information-educational portal (www.ziyonet.uz).

Scientific adviser:	**Elov Botir Boltayevich** doctor of philosophy of technical sciences, associate professor
Official opponents:	**Xolmonova Zulxumor Turdiyevna** doctor of philological sciences, professor
	Abdurahmonova Nilufar Zaynobiddin qizi doctor of philological sciences, professor
Leading organization:	**Samarqand state university**

The defense of the dissertation will be held on "14" 06, 2024 at 10:00 at the meeting of Scientific Council awarding scientific degrees DSc. 03/30.12.2019.Phil.19.01, under Tashkent State University of Uzbek language and literature named after Alisher Navoi. (Address: 100100, Tashkent city, Yusuf Khos Hojib street, 103. Phone: (99871) 281-42-44; faks: (99871) 281-42-44; faks: (99871) 281-12-44. (www.tsuull.uz). e-mail: monitoring@navoiy-uni.uz).

The dissertation is available at the Information Resource Center of Tashkent State University of Uzbek language and literature named after Alisher Navoi. (Registered with No 290). (Address: 100100, Tashkent city, Yusuf Khos Hojib street, 103. Phone: (99871) 281-42-44; faks: (99871) 281-42-44; faks: (99871) 281-12-44 (www.tsuull.uz).

The abstract of the dissertation was distributed on 2024 "30" 05.

(Register Protocol No. 1 2024 "30" 05.

H.A. Dadaboyev
Chairman of the one-time scientific council based on the scientific council awarding scientific degrees, philol.ph.d., professor

Q.U. Pardayev
The Scientific secretary of one-time Scientific Council on awarding scientific, philol.ph.d., professor

S.E. Normamatov
The chairman of one-time scientific seminar at the scientific council awarding scientific degrees, philol.ph.d., professor

Introduction (annotation of the Doctor of Philosophy (PhD) dissertation)

Relevance and necessity of the dissertation topic. Natural language processing methods are widely used in the practice of world computational linguistics. Natural language processing (NLP) projects have ongoing tasks. They are widely studied because of their fundamental nature. A good understanding of these tasks allows for various applications of NLP. Language modeling, text classification, information retrieval, conversational agents, text summarization, question-and-answer systems, machine translation, thematic modeling are the main areas of NLP. Natural language processing is a branch of artificial intelligence aimed at understanding and processing human language by machines. Natural language processing is not an easy task for a machine, because a machine works with numbers, not text.

In world linguistics at the end of the 20th century, issues such as improving the quality of automatic translation, linguistic modeling of the language, the theory of lemming words within certain languages, and the creation of an algorithm were put on the agenda. As a result, applications such as tokenizer, stemmer, lemmatizer, which are actively used in the information search system, were developed. Consequently, the creation of language corpora, their linguistic annotation, the development of automatic text processing software tools - morphoanalyzer, stemmer, lemmatizer, parser, orthocorrector - have become scientific-theoretical problems of computer linguistics.

In the field of Uzbek computer linguistics, attention is paid to scientific research on language corpus, speech synthesizer, automatic translation, artificial intelligence, understanding and processing of the Uzbek language. However, the development of tools that implement the initial stages of natural language processing: tokenizer, stemmer, lemmatizer has not been scientifically and theoretically studied. Maintaining the purity of the state language, enriching it and improving the speech culture of the population; ensuring the active integration of the state language into modern information technologies and communications are one of the urgent tasks facing Uzbek computer linguistics today[45]. Accordingly, it is important to achieve processing of the Uzbek language by means of modern information technologies, for this, issues such as language corpora, electronic translator, thesaurus, orthocorrector, linguistic analyzers, and the creation of their linguistic support are important.

This study serves to a certain extent in the implementation of the tasks specified in No. PF-4997 of the President of the Republic of Uzbekistan dated May 13, 2016 "On the establishment of the Tashkent State University of Uzbek Language and Literature named after Alisher Navoi", No. PF-5850 dated October 21, 2019 "On measures to fundamentally increase the prestige and status of the Uzbek language as a state language", No. PO-6084 dated October 20, 2020 "On measures of further development of the Uzbek language in our country and improve language policy",

[45] Presidential decree No. PF-5850 of October 21, 2019 "On measures to radically increase the prestige and status of the Uzbek language as a state language". https://lex.uz/docs/-4561730

PF-2789 dated February 17, 2017 "On measures for further improving the activities of the Academy of Sciences, organizing, managing and financing of research and development activities", Resolution No. PQ-4479 dated October 4, 2019 "On the wide celebration of the thirtieth anniversary of the adoption of the Law of the Republic of Uzbekistan "On the State Language", the Address of the President of the Republic of Uzbekistan to the Oliy Majlis on January 24, 2020 and other regulatory and legal documents.

Relevant research priority areas of science and developing technology of the Republic. The dissertation was completed in accordance with the priority direction of the republic's science and technology development I."Social, legal, economic, cultural, spiritual and educational development of the information society and democratic state, development of innovative economy".

The level of study of the problem. The issue of natural language processing has been thoroughly studied in world linguistics. General issues of natural language processing have been the object of many studies[46]. In particular, J. J. Webster, C. Keith, A. Rai, S. Borah, X. Song, A. Salchiani, Y. Song, D. Dopson, D. Zhou and other[47] covered the issues of tokenization in their research. The issue of stemming and various approaches to its solution can be learned from the works of K. Savitha, Y. Liu, M. Haidar, K. Kreslins, S. Jonathan[48]. Solutions to the problem of lemming were proposed in the works of Avhavhudzani, D. McCarthy, A. Chakrabarti[49].

[46] Vajjala S., Majumder B., Gupta A. Surana H. Practical Natural Language Processing. A Comprehensive Guide to Building Real-World NLP Systems. – USA, 2020. – 424 p.; Bölücü N., Can B. Unsupervised joint PoS tagging and stemming for agglutinative languages. ACM Transactions on Asian and Low-Resource Language Information Processing, New York, 2019, – №18. – P. 1-21.

[47]Webster J.J., Kit Ch. Tokenization as the initial phase in NLP. Hong Kong, 1992. – P. 1106-1110; Rai A., Borah S. Study of various methods for tokenization // In Lecture Notes in Networks and Systems. India, 2021, – № 137. – P. 193-199; Song X., Salcianu A., Song Y., Dopson D., Zhou D. Fast WordPiece Tokenization // Conference on Empirical Methods in Natural Language Processing, Proceedings. China, 2021. – P. 2089-2103.

[48]Savitha K. Study of stemming algorithms // UNLV Theses, Dissertations, Professional Papers and Capstones. – Las Vegas, 2010. – 48 p.; Liu Yi Yu. The advances of stemming algorithms in text analysis from 2013 to 2018 // MCom. Dissertation, University of Pretoria. – Pretoria, 2019. – 192 p.; Haidar M. "A comparison of root and stemming techniques for the retrieval of Arabic documents". Dissertation, McGill University. – Montreal, 2001. – 204 p. Kreslins K. "A stemming algorithm for Latvian". PhD Dissertation, Loughborough University. – Boston, 1996. – 264 p.; Jonathan S. Designing a stemming algorithm for Kambaata text: a rule based approach. st.mary's University. – Addis Ababa, Ethiopia, 2018. – 104 p.

[49]Avhavhudzani M.V. "The lemmatization of Tshivenda lexical items". Thesis, University of Limpopo. – Africa, 2012. – 105 p.; McCarthy D. Lexical acquisition at the syntax-semantics interface: Diathesis alternation, subcategorization frames and selectional preferences. Doctoral dissertation, University of Sussex. – Cambridge, 2001. –189p.; A. Chakrabarty. On Lemmatization and Morphological Tagging for Highly Inflected Indic Languages. Doctoral dissertation, Indian Statistical Institute. – India, 2018. – 116 p.

There are some records of Z. Kholmanova[50], B. Elov, Sh. Khamroyeva[51] M. Abjalova[52], N. Abdurakhmonova, A. Ismailov[53], O. Abdullayeva[54], R. Alayev[55], I. Bakayev[56], M. Sharipov and O. Sobirov[57] regarding the processing of the Uzbek language, in particular, linguistic modules of the software tools for stemming, morphological analysis, editing and analysis, tokenization of the linguistic features of the Uzbek language text corpus. The study of Uzbek word forms, their formation and morphemes, as well as the automatic processing of the Uzbek language[58] is also noteworthy. However, the issue of linguistic foundations and software development of tokenization, stemming and lemming of Uzbek lexical units has not been studied as a special monographic study.

The connection of the research with the research plans of the higher education or scientific-research institution where the dissertation was completed. This research is performed within an innovative project on the topic of "Creating an automatic processing tool for information-search systems (Google, Yandex, Google translate) - a morpholexicon and morphological analyzer software tool of the Uzbek language" which is being carried out at the Tashkent State University of Uzbek Language and Literature named after Alisher Navoi for 2021-2024.

The purpose of the research: is to develop linguistic bases and software for tokenization, stemming and lemming of lexical units of the Uzbek language.

Tasks of the research:

to determine the general characteristics of tokenization and specific aspects of the Uzbek language;

to determine the theoretical basis of stemming and lemming;

to identify stemming process problems for agglutinative languages;

to develop linguistic bases of stemming and lemming of lexical units of the Uzbek language;

development of optimization methods for the search system of the national corpus of the Uzbek language;

implementation of the tokenization process based on the BPE algorithm;

[50] Xolmonava Z.T. Computing linguistics – Tashkent. 2020. – 198 p.

[51]Elov B.B., Xamroyeva Sh.M. The Problem of POS Tagging and Stemming for Agglutinative Languages // 8 th International Conference on Computer Science and Engineering UBMK 2023. Turkey, Burdur, 2023. – P. 57-62;

[52]Abjalova M. Tahrir va tahlil dasturlarining lingvistik modullari. (NLP) // Linguistic modules of the text editing and analysis program (NLP). – Monografiya. – Toshkent, 2020. – 165p.

[53]Abdurakhmonova N., Ismailov A., Sayfullaeva R. Turkic language stemmer python package for Natural Language Processing // "Computer Linguistics challenges and sollutions" international scientific-practical conference. Tashkent, 2023. https://www.researchgate.net/publication/369503003.

[54]Abdullayeva O. Theoretical and practical foundations of the formation of the corpus of Internet information texts of the Uzbek language. Philol. science. Doctor of Philosophy (PhD) diss. - Andijan, 2022. - 120 p.

[55]Elov B., Alayev R. Algorithm of Uzb-temming based on separation of affixes / Uzbekistan Journal of Informatics and Energy. Journal of Uzbekistan. 2023, - No. 1. - P. 14-27.

[56]Bakaev I. Linguistic features tokenization of text corpora of the Uzbek language // Bulletin of TUIT: Management and Communication Technologies. Tashkent, 2021. – P. 98-105.

[57] Sharipov M., Sobirov O. Development of a Rule-Based Lemmatization Algorithm Through Finite State Machine for Uzbek Language // The International Conference and Workshop on Agglutinative Language Technologies as a challenge of Natural Language Processing. Koper, Slovenia, 2022. – P. 1-6.

[58] Elov B., Hamroyeva Sh., Axmedova X. Methods for creating a morphological analyzer, 14th International Conference on Intellegent Human Computer Interaction. Tashkent, 2022. – P. 27-38.; Elov B., Hamroyeva Sh. Methods of creating a morphological analyzer. / Uzbekistan: Language and Culture (Applied Philology). Tashkent, 2022, - No. 5. - P. 40-64

development of the uzb.stemming algorithm based on the separation of affixes.

The object of the research is the lexical units of the Uzbek language.

The subject of the research is the linguistic foundations and software of tokenization, stemming and lemming of lexical units of the Uzbek language.

Research methods. Systematic, comparative, component analysis, description and classification methods were used in the research.

The scientific novelty of the research is as follows:

the theoretical foundations of tokenization, stemming and lemming of lexical units in world and Uzbek linguistics were systematized, tokenization based on signs and words, stemming with a hybrid approach based on linguistic rules and corpus, and methods of lemming based on the dictionary were determined;

for agglutinative languages, problems such as homonymy of stem and suffix in the process of stemming with one stem, meeting the sound change of the word, stemming of neologisms and NERs were identified, and the linguistic bases of stemming and lemming of lexical units of different structures in the Uzbek language were developed;

based on the processing of the texts of the national corpus of the Uzbek language, searching engine optimization methods were developed, and the process of stemming and lemming of lexical units was revealed;

the tokenization process is carried out on the basis of the BPE algorithm, stemming based on the separation of affixes, a lemming algorithm based on dictionary and morphological analysis has been developed, Uzbek.tokenizer, Uzbek.stemming and Uzbek.lemmatizer software have been created for Uzbek language texts.

The practical results of the research are as follows:

Algorithms developed on the basis of research results served to increase the efficiency of the morphological analyzer of the Uzbek language;

Tokenization, stemming and lemming actions presented in the dissertation optimized the result of the search system of the national corpus of the Uzbek language;

A software tool for tokenization, stemming and lemming of Uzbek language texts has been developed.

The reliability of the research results is explained by the fact that the studied materials helped to draw conclusions based on the nature of the Uzbek, Turkish, Uyghur and English languages, their reasonableness, methodological excellence, and the fact that the Uzbek, Turkish, Uyghur and English cross-language studies are based on practically proven sources.

Scientific and practical significance of research results. The scientific significance of the research results is that these research materials enrich theoretical works related to the processing of Uzbek language texts, the creation of a language corpus, and the development of machine translation systems with certain scientific views and facts. It also serves to improve research, monographs, textbooks and training manuals on Uzbek computer linguistics.

The practical significance of the research results is explained by the development of the software https://uznatcorpara.uz/uz/POSTag, which performs the process of tokenization, stemming and lemma in the creation of morphological, syntactic, semantic analysis tools of natural language.

Implementation of research results. Based on the results obtained in the study of linguistic fundamentals and software of tokenization, stemming, lemming units of the Uzbek language:

scientific conclusions about the character and word-based tokenization in software that implements token, stemming, and lemming of Uzbek language units, linguistic rules, and hybrid approach stemming in Corpus tool, dictionary-based lemming were used in the center for teaching and professional development of the basics of conducting business in the state language under the Tashkent State University of Uzbek language and Literature named after Alisher Navoi (certificate of the Tashkent State University of Uzbek language and literature dated October 12, 2023). As a result, a database of more than 32,000 simple and more than 7,500 combined lexemes in the morphological analysis of Uzbek language sentences was formed as a result of the testing of the software for tokenization, stemming and lemming of Uzbek language units;

the results of tokenization, stemming and lemmatization of Uzbek units were used in the practical project numbered I-OT-2019-42 "Creating a multilingual electronic platform of Uzbek literature (in Uzbek, Russian, English)" carried out in 2021-2023 at the Tashkent State University of Uzbek Language and Literature named after Alisher Navoi for improving the quality of stemming and lemming in Uzbek, Russian, and English languages in natural language processing (reference No. 01/10-2182 dated October 19, 2023 of Alisher Navoi Tashkent State University of Uzbek Language and Literature). As a result, improving the quality of stemming and lemming in Uzbek, Russian, and English languages in natural language processing improved the quality of search on the electronic platform; from the results of tokenization, stemming, and lemming of units of the Uzbek language, the effectiveness of the search result for units specific to the Uzbek language is ensured;

Algorithms justifying the result of tokenization, stemming, lemming of Uzbek language units and the software created based on them were used in the practical project PZ-2020042022 "Creating a language-didactic electronic platform of Turkish languages" carried out in 2021-2023 at the Alisher Navoi Tashkent State University of Uzbek Language and Literature (reference No. 01/10-2183 dated October 19, 2023 of Alisher Navoi Tashkent State University of Uzbek Language and Literature). As a result, based on the stemming and lemming algorithm, the index of the search system of the linguodidactic electronic platform of Turkish languages has been increased.

Aprobation of research results. The results of the study were discussed at 2 international, 2 republican scientific and practical conferences.

The publication of the research results. On the topic of the dissertation, 12 scientific articles were published, 3 of which were published in foreign journals and 8 of them The publication of the research results. The publication of the research results. in scientific publications recommended for the publication of the main

scientific results of doctoral dissertations of the Higher Attestation Commission of the Republic of Uzbekistan.

Structure and size of the dissertation. The dissertation consists of an introduction, three chapters, a conclusion, a list of used literature and an appendix, the volume of which is 129 pages.

THE MAIN CONTENT OF THE DISSERTATION

In the introduction, the topic of scientific research and its relevance are covered, the purpose and objectives, object and subject of research work are identified, and the suitability of scientific work for important areas of development of science and technology is indicated. At the same time, the scientific novelty of the dissertation, the practical results and their reliability, the theoretical and practical significance of the work, the introduction of the achieved results into practice, publication in scientific publications, information about the structure of the work are presented.

The first chapter of the dissertation is called **"Theoretical foundations of tokenization, stemming, and lemming"**. The first section of the chapter is called "A general description of the tokenization process in world and Uzbek linguistics". Text analysis is usually carried out according to clearly defined steps. The first step in the initial processing of text is to convert it into tokens. Tokens will consist of text segments identified as meaningful units for text analysis. D.Manning, P.Raghavan and H.Scientific research by the Schutze states that tokens can be made up of larger or smaller pieces (word sequences, vocabularies, paragraphs, sentences, or lines)from individual words[59].

Tokenization is the main and mandatory step in working with text information in NLP, and the software tool that performs this task is the tokenizer. The tokenizer can be designed based on a different approach, depending on the NLP task.

Description of tokenization software tools of various systematic languages. There are many tokenization methods for different languages of the world, and rule-based[60], statistical[61], non-disclosure[62], lexical[63] methods differ. Welbers and W. Atteveldt in their paper "Text Analysis in R", describes: "the tokenization process is considered a process of breaking text into smaller pieces, which often involves initial processing of text to remove punctuation and convert all characters into lowercase"[64].

[59] Manning C.D., Raghavan P., Schütze H. Introduction to Information Retrieval. – Cambridge, 2008. – 544 p.
[60] Zhou L., Liu Q. A character-net based Chinese text segmentation method. 2002. https://aclanthology.org/W02-1117.pdf. (Saytga murojaat); Kaplan R.M. A Method for Tokenizing Text. Complexity and Education: Inquiries into Words, Constraints and Contexts. Cambridge, 2005. – P.55-64.
[61] Yang C.C., Li K.W. A heuristic method based on a statistical approach for Chinese text segmentation / Journal of the American Society for Information Science and Technology. America, 2005, – №56. – P. 1438-1447.
[62] Shahabi A.S., Kangaveri M.R. Intelligent processing system // IFIP International Federation of Information Processing Springer Boston. Pakistan, 2007. – P. 411-420.
[63] Wu D., Fung P. Improving Chinese tokenization with linguistic filters on statistical lexical acquisition // 4th Conference on Applied Natural Language Processing. Stuttgart, Germany, 1994. – P. 180-181; Labadié A., Prince V. Lexical and semantic methods in inner text topic segmentation: A comparison between C99 and transeg. // Conference: Proceedings of the 13th international conference on Natural Language and Information Systems: Applications of Natural Language to Information Systems. France, 2008. – P. 347-349.
[64] Welbers K., Van Atteveldt W., Benoit K. Text Analysis in R. Communication Methods and Measures. London, 2007, – №11. – P. 245-265.

Decisions made during the tokenization process have a significant impact on further text processing processes. M. Denny, A. Spirling[65] and D. Guzr[66] in their research argued that in the process of implementing tokenization for large-scale language Corpus, complex actions have been performed, depending on the natural language properties of tokenization. Consistent tokenization for natural language text has been implemented through a set of tokenizers developed in the **R language** by A.Mullen, K.Benoit[67] and the Core Team[68] group of programmers, and archived on GitHub[69] and Zenodo[70].

To increase the speed of the tokenizer developed by D. Eddelbettel and J. Balamuta, the main components such as **n-gram** and **skip n-gram** tokenizer were developed using the Rcpp software tool[71].

Today, there are no practical works on tokenization of lexical units of the **Uzbek language**. M. Abjalova described tokenization, stemming and lemming processes in her monograph "Linguistic modules of editing and analysis programs"[72].

Tokenization of the Uzbek text corpus was carried out by I. Bakayev, in which it was noted that spelling features should be taken into account[73]. The types of words are analyzed and a mathematical model of word formation is proposed for complex, double, repeated and compound words.

Tokenization can be broadly divided into 3 types - words, symbols and word structure elements (n-gram symbols). For example, consider the following sentence: *"Hech qachon taslim bo'lmang"*. The most common way to generate tokens is space-based. If we take a space as a delimiter, tokenization of the sentence produces 4 tokens – *Hech-qachon-taslim-bo'lmang*. Since each token consists of a word, it is called word tokenization. Similarly, characters and word structure elements can be generated as tokens. For example, consider the word *"aqlliroq"*:

character token: *a-q-l-l-i-r-o-q*;

word composition element token: *aql-li-roq*.

As tokens are the building blocks of natural language, the most common way to process unstructured text occurs at the token level.

The second section of the first chapter is titled "Overview of the Stemming Process". Stemming is the first step in text processing in NLP, and this process is

[65] Denny M. J., Spirling A. Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It / Political Analysis. Cambridge, 2018, – №26. – P. 168-189.
[66] Guthrie D., Allison B., Liu W., Guthrie L., Wilks Y. A closer look at skip-gram modelling // Proceedings of the 5th International Conference on Language Resources and Evaluation. Chicago, 2006. – P. 1222-1225.
[67] Mullen A., Benoit L., Keyes K., Selivanov D., Arnold J. Fast. Consistent Tokenization of Natural Language Text / Journal of Open Source Software. America, 2018, – №3. – P.1-3.
[68] Team R.C. R: A Language and Environment for Statistical Computing / R Foundation for Statistical Computing. Vienna, 2021. https://www.R-project.org
[69] https://github.com/ropensci/tokenizers
[70] https://zenodo.org/record/1205017
[71] Eddelbuettel D. Seamless R and C++ integration with Rcpp. – USA, 2013. – 220 p.; Eddelbuettel D., Balamuta J.J. Extending R with C++: A Brief Introduction to Rcpp / The American Statistician, Taylor and Francis Journals. France, 2018, – №. 72. – P. 28-36.
[72] Abjalova M. Linguistic modules of editing and analysis programs. (NLP). // Linguistic modules of the text editing and analysis program (NLP). – Monograph. – Tashkent, 2020. – 165 p.
[73] Bakaev I. Linguistic features tokenization of text corpora of the Uzbek language // Bulletin of TUIT: Management and Communication Technologies, Tashkent, 2021. – P. 98-105.

required for many NLP tasks. The main goal of the stemming process is to bring the inflected forms of the word to the root form. This application of NLP is very important in most information retrieval systems (IRS). The process of stemming is done by removing all suffixes from word forms. The result of this is to improve the performance of NLP applications such as spell checking.

The process of stemming is related to the formation of words: it is based on the change of stem and formation[74]. As a result, word forms are brought to conflation. Word forms can include prefixes and suffixes.

In 1980, M. Porter proposed a stemmer algorithm based on the rules for separating the suffixes. The algorithm found that a stemmer with five stages and a 10,000-word dictionary reduces text/corpus size by 33%.

In 1990, Chris Pace at Lancaster University developed a new stemmer that iteratively removes suffixes based on grammatical rules. This stemmer algorithm aims to replace suffixes so that rule exceptions do not occur. This stemmer is commonly called a Pace/Hask stemmer or a Lancaster stemmer.

Classification of stemming algorithms. Currently, many stemming algorithms have been developed by NLP researchers, and this part of the work discusses the algorithms that are widely used. http://devopedia.org classifies modern stemmers as follows:

Figure 1. Classification of stemming algorithms according to http://devopedia.org

Ways of removing suffixes. This method is based on removing (truncating) prefixes or affixes from the word. The simplest stemmer of this type is the Tuncate(n) stemmer, which truncates a word to the n^{th} character, that is, keeping the first n letters of the word and removing the rest[75]. In this method, words shorter than n are stored in their original form. The number of redundant operations increases as the word length becomes smaller. Another approach based on affix removal, the S-stemmer algorithm combines the singular and plural forms of English nouns. This

[74] Conflation - identifying the common root of word forms.

[75] Frakes W.B., Fox C.J. Strength and similarity of affix removal stemming algorithms / ACM SIGIR Forum. Virginia, 2003, – №37. – P. 26-30.; Sirsat S.R., Chavan V., Mahalle H. S. Strength and Accuracy Analysis of Affix Removal Stemming Algorithms / International Journal of Computer Science and Information Technologies. Indonesia, 2013, – №4. – P. 265-269.

algorithm was proposed by D. Harman[76]. In the S-stemmer algorithm, the rules for converting plural suffixes into singular forms have been developed.

Some studies on the issue of stemming in the Uzbek language have been carried out, including N. Abdurakhmonova, R. Saifullayeva and A. Ismailov studied the issue of developing a stemmer software tool for Uzbek words and noted that determining the stem is important in the process of searching for information[77]. In natural language processing, the initial stage of processing a text dataset is usually time and resource intensive. In their research, N. Abdurakhmonova and others noted that one of the first methods of unstructured text processing is the stemming process[78]. They note that stemming is commonly used in data mining and machine learning. A python stemmer package called TLstemmer was developed by N. Abdurakhmonova and A. Ismailov, which identifies stems in text belonging to the Turkic language group.

In short, the stemming algorithms developed for other natural languages cannot be directly applied to the lexical units of the Uzbek language. The development of the UzbStemming algorithm based on the characteristics of the Uzbek language is an urgent task.

The third section of the chapter is titled "*The Lemmalization Process: Overview and Analysis*". Lemmalization is a method of determining the meaningful form of a lemma (lexeme) from a word form or stem and processing these units in natural language. Lemmalization is an important step for many NLP tasks. Lemmalization is the process of finding the lexically normalized form of a word. Lemmas is a well-studied process. In particular, A.Bjorkeland, D.Seddah and A.Fraser studied the issue of syntactic analysis of texts and the use of lemming in machine translation[79]. Lemma identification is required to compare words to lexical sources and establish relationships between inflectional forms. Lemmalization is especially important for morphologically rich languages. G. Chrupala proposed a lemma method based on tokens, which does not depend on natural language, but due to the low efficiency of this method, this algorithm is rarely used in practice[80].

Several studies have been conducted and tested in small texts on algorithms for **lemmalization of lexical units in the Uzbek language**. In particular,

[76] Harman D. How effective is suffixing? / Journal of the American Society for Information Science. America, 1991, – №42. – P. 7-15.

[77] Abdurakhmonova N., Ismailov A., Sayfullaeva R. Turkic language stemmer python package for Natural Language Processing // "Computer Linguistics challenges and sollutions" international scientific-practical conference. Tashkent, 2023. https://www.researchgate.net/publication/369503003

[78] Abdurakhmonova N., Babajanov U., Ismailov A., Baysariyeva S. TLstemmer (Turkic language stemmer) python package for Natural Language Processing // Conference: International Conference on Information Science and Communications Technologies. Tashkent, 2022. https://www.researchgate.net/publication/349467555.

[79] Björkelund A., Bohnet B., Hafdell L., Nugues P. A high-performance syntactic and semantic dependency parser // 23rd International Conference on Computational Linguistics, Proceedings of the Conference. China, 2010, – №2. – P. 33-36.; Seddah D., Chrupala G., Çetinoğlu Ö., Van Genabith J., Candito M. Lemmatization and Lexicalized Statistical Parsing of Morphologically Rich Languages: the case of French // 1st Workshop on Statistical Parsing of Morphologically-Rich Languages. Los Angeles, 2010. – P. 85-93.; Fraser A., Weller M., Cahill A., Cap F. Modeling inflection and word-formation in SMT // Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. France, 2012. – P. 664–674.

[80] Chrupala G. Simple data-driven context-sensitive lemmatization // Procesamiento Del Lenguaje Natural. Ireland, 2006, – №37 – P. 121-127

I. Bakayev[81] studied the issues of developing a stemming algorithm based on an approach based on linguistic rules to Uzbek words. M. Sharipov, O. Sobirov created a rule-based lemmalization algorithm for the **Uzbek language** using **Finite State Machine**[82]. Their database and POS tagging were used to remove affixes. Although the results of research on the application of machine learning and neural network-based lemma in different languages have high efficiency, in practice they have not been applied to the Uzbek language.

Chapter II of the dissertation is entitled **"Linguistic bases of stemming and lemmalization of Uzbek language lexical units"**, and the first part is called "*Stemming of Uzbek language lexical units: problems and solutions*". In the Uzbek language, one of the smallest meaningful parts of a word is **the root**, and the largest meaningful part of the word form is **the stem**. Therefore, in NLP, a word is considered to be made up of two parts: **the stem, which carries the meaning**, and **the suffixes.**

The work done so far in the field of morphological analysis can be divided into two classes: **approaches based on stemming** and **affix removal**. To perform *stem-based morphological analysis*, first the stem of the word is determined, and then its affixes. In *the affix removal approach*, the affixes in the word are first identified.

To determine the stem of words in the Uzbek language, the stem and all types of affixes attached to it are determined. Traditional stemming algorithms are based on addition and morphological rules, and the stemming process can cause the following problems:

the root and suffix are homonymous with one root;

meeting of a word with a sound change;

stemming of neologisms and NER (Named entity recognition).

The stem ambiguity in Uzbek language sentences can be seen in Figure 2 below:

Ertalabdan nonushtaga quymoq tayyorlandi.

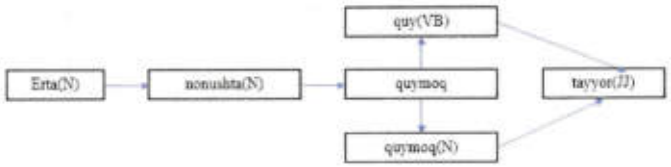


Figure 2. Stem uncertainty in Uzbek language sentences

In order to solve the problem of sound change in stem detection, the boundaries of stems and suffixes are determined in the first step, and lemma is performed in the second step. As a result of lemming, the resulting stems are changed to the root in the dictionary. When the problem of finding the stem of NER and neologisms arises,

[81] Bakayev I. Development of a stemming algorithm based on a linguistic approach for words of the uzbek language // International Conference on Scientific, Educational & Humanitarian Advancements. Bukhara, 2021. – P. 195-202.
[82] Sharipov M., Sobirov O. Development of a Rule-Based Lemmatization Algorithm Through Finite State Machine for Uzbek Language // The International Conference and Workshop on Agglutinative Language Technologies as a challenge of Natural Language Processing. Koper, Slovenia, 2022. – P.1-6.

the form-formers are cut, the suffix or part of the word in the form of the word-former is left, and this part is equal to the stem.

The second part of the chapter is called "*Methods of lemming words with different structures in the Uzbek language*". Lemmalization is the process of grouping different inflectional forms of a word for analysis as a single element. It is used to determine the normalized form of a word or the dictionary form of a word called a lemma. A lexeme is a combination of all forms (called vocabulary heads) with the same meaning, while a lemma is a word chosen as the base form representing the lexeme. The usual structure of word structure in Uzbek is as follows:

"Root + word-former + lexical form + syntactic form".

The lemmalization process is shown below[83]:

Figure 3. Lemmatization process

In the Uzbek language, the result of stemming and lemming seems to be the same, but they are completely different processes: stemming is the process of cutting the suffix from the root, and lemming is the process of determining the variant of the root in the dictionary (Fig. 4).

Figure 4. Difference between stemming and lemming

[83] http://uznatcorpara.uz/uz/Lemmatizer

Simple and simple formed words. A simple stem and a simple denominator are usually equal to one lemma. To lemme them, it is enough to cut syntactic and lexical form-forming additions.

Compound words. When determining the lemma of compound words, the following rules are followed:

1. The lemma of compound words is solved by cutting off form-forming suffixes.

2. Lemming of compound words to be written separately is based on the dictionary.

3. Compound nouns are written separately and are considered NER: *New Uzbekistan, United Nations, World Health Organization, Party of National Recovery*. The lemming of such units requires separate linguistic and mathematical models.

There are units that are similar in form to compound verbs, such as phrases. According to the content of the phrases, it has the following structure:

1) W1 + W2
2) W1 + W2 + W3
3) W1 + W2 + W3 + W4

Among them, phrases with W1 + W2 structure are similar to compound verb structure. Noun+verb phrases (see well, keep in mind, listen to, give life to, turn away from) are equivalent to one lemma, but they are tagged as a phraseological unit, not a lemma.

Pair of words. When determining the lemma of a pair of words, the following rules are followed:

1. **Both parts are meaningful**: bitter-sweet. In this case, there are two stems and one lemma, which is also based on the dictionary.

2. **One part gives meaning, one does not**: kalta-kulta, in this case the stem is only the meaning-giving part, the lemma is one, and it is also based on the dictionary.

3. **Neither part makes sense**: "*g'idi-bidi, jiz-biz*". If both parts of a word pair have no meaning, then the word pair itself is a stem, root, and lemma.

Repeated words. When determining the lemma of repeated words, the following rules are followed:

1. **Exact repetition:** *katta-katta, tez-tez, baland-baland.*

2. **The second part is the phonetic change of the first part**: tuz-puz, non-pon.

Lemming of repeated words is also based on dictionary information. It can be seen from the above that vocabulary information - linguistic database is important in lemma. Only when determining the stem, form-forming elements (syntactic and lexical) are cut. When one of the repeated words occurs separately, the POS tag belongs to another category, and when it occurs again, the POS tag belongs to another category: *dedi* is a verb when it is used alone, and when it is a repeated word in the form of *dedi-dedi*, it belongs to the noun category. But this is not a typical situation: *tez* is adverb, *tez-tez* is adverb; *katta* is adjective, *katta-katta* is adjective.

The third section of the chapter is entitled "The Importance of Lemmas in Linguistic Corpus Search Engine Optimization". Today, NLP algorithms are used to

solve problems such as detecting spam in e-mail and optimizing user queries in searching engines. NLP helps machines communicate with humans in natural language, intelligently analyze text, understand speech and interpret it in the correct format[84]. Processing unsystematized data using NLP tools is a complex and urgent process, and it is extremely important to obtain the necessary information from the text using NLP methods with sufficient accuracy.

The process of lemming is important in searching engine optimization (SEO), and it is necessary to develop how the stages, methods and algorithms of lemming work, and how to apply it to corpus texts. It's also important to focus on keyword research, on-page optimization, and other factors that affect search rankings. Lemmas should be used as part of a comprehensive SEO strategy for best results. Manual lemming of corpus texts takes more time, but gives more accurate results. 5-figure below shows the steps for parsing corpus texts.

Implementation of lemma on the website. Lemmas help solve tasks such as text classification, data retrieval, and machine translation. There are different approaches to implement lemma on a website.

1. *Using a pre-trained lemmatization model.*
2. *Development of a special lemmatization model.*
3. *Using the online lemmatization service.*

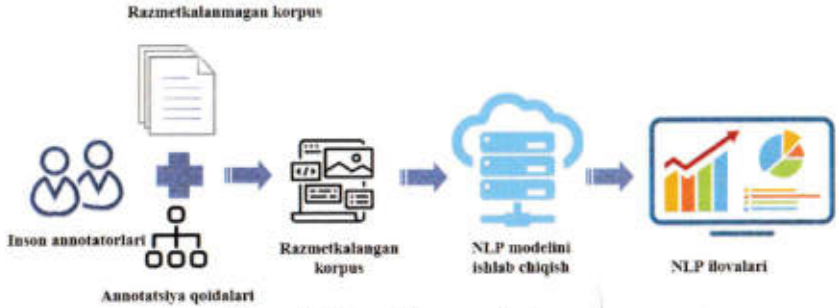


Figure 5. Steps of corpus text parsing.

Regardless of the chosen approach, there are a few steps to implement lemmatization on a website:

initial processing;
tokenization;
POS tagging;
lemming;
integration.

In general, lemma is an essential tool in increasing the SEO performance of a website. *Identifying relevant keywords, analyzing existing content, performing lemming, tracking performance, comparing results* will effectively test the

[84] Elov B.B., Hamroyeva Sh.M., Xusainova Z.Y. NLP (tabiiy tilga ishlov berish)ning vazifalari va zamonaviy yondashuvlar / Filologik tadqiqotlar: til, adabiyot, ta'lim. Termiz, 2022, – №5. – B. 19-26.

effectiveness of lemming in SEO of a website and make the right decision on how to optimize content for searching engines.

Chapter III of the dissertation is called **"Programmatic bases of tokenization, stemming and lemming of lexical units of the Uzbek language"**. In the first part of this chapter, the issue of "Tokenization of Uzbek language units based on BoW and BPE algorithms" is considered. A BoW (bag of words) model is a digital representation of text to be processed by machine learning algorithms. Using the Bag of Words (BoW) modeling algorithm, text can be converted and processed into digital matrices.

The necessary features are extracted from the textual data using the **BoW algorithm**. This approach is a simple and flexible way to extract certain features from documents. In the BoW method, only the number of words in the text (statistics) is determined. However, grammatical details and word order are neglected. Vocabulary size is a major challenge facing the BoW model. If the model encounters a new word that has not yet been used, the BoW model ignores that word.

Byte Pair Encoding (BPE) is a widely used tokenization method among transformer-based models. BPE solves the problems of word and character tokenizers[85]:

1. BPE effectively solves the problem of OOV. It parses and processes OOV as word structure elements.

2. Length of input and output statements after BPE is shorter compared to character tokenization.

BPE is a word segmentation algorithm that iteratively combines the most frequent characters or sequences of characters. Steps to implement the BPE algorithm:

1. Split the words in the corpus into characters after the suffix </w>.

2. Vocabulary formation with unique words from the corpus.

3. Calculation of the frequency of a pair of characters or a sequence of characters in the corpus.

4. Merge the most frequent pair in the corpus.

5. Save the best pair to the dictionary.

6. Repeat steps 3-5 for a certain number of iterations.

In order to process the corpus efficiently, in each iteration of BPE, it goes through every potential combination option based on its frequency and follows a unique approach to optimize the best solution. The above methods were applied to the Uzbek language corpus (texts) and the https://uznatcorpara.uz/uz/Tokenizer software was developed.

The second part of the chapter is called "Development of UzbStemming Algorithm Based on Affix Separation". To develop the Uzbek language stemmer

[85] Bostrom K., Durrett G. Byte pair encoding is suboptimal for language model pretraining // Findings of the Association for Computational Linguistics: EMNLP. Austin, 2020. – P. 4617-4624.; Heinzerling B., Strube M. BPEMB: Tokenization-free pre-trained subword embeddings in 275 languages // 11th International Conference on Language Resources and Evaluation. Miyazaki, Japan, 2019. – P. 2989-2993.; Gutierrez-Vasques X., Bentz C., Sozinova O., Samardžić T. From characters to words: The turning point of BPE merges // Proceedings of the16th Conference of the European Chapter of the Association for Computational Linguistics. 2021.– P. 3454-3468. https://aclanthology.org/2021.eacl-main.302.pdf (Saytga murojaat)

based on removing affixes, the language corpus, affix database, Uzbek morpholexicon, and root database are used[86]. The following two-step process was performed to determine the stem, taking into account the word form group (POS tag).

Stage I. Preparation of data (dataset)

1. Prefixes and suffixes that can be added to the root are determined (Tables 1, 2).

Table 1

Prefixes in the Uzbek language

Prefixes		
ba-	be-	ham-
bo-	no-	bar-
bad-	ser-	be-
bar-	g'ayri-	xush-

Table 2

Suffixes in the Uzbek language (by POS)

POS	Derevative	formative (inflective)
Noun (N)	51	17
Number (Num)	1	5
Verb (Vb)	30	22+10 (nisbat)
Adjective (JJ)	97	9
Pronoun (P)	-	-
Adverb (R)	18	5

2. A set of strings of prefixes and suffixes that can be added to the stems of the given word group is determined by taking several (active) word forms for each word group and deleting their stem from the word form.

3. A string of suffixes is a combination of several Uzbek suffixes added to the stem to form this word form (Table 3).

Table 3

A set of Uzbek prefix and suffix combinations identified from the language corpus

	Noun (N)	Number (Num)	Verb (Vb)	Adjective (JJ)	Pronoun (P)	Adverb (R)
1.	chiligimizning	Inchi	dilarki	almashtirgichda	nga	aydiki
2.	dagilarning	larcha	nganlaridan	bardoshligini	ngacha	aydimi
3.	doshlarimiznikidan	larda	yapsizlarmi	imizning	ngagina	chiligimizning
4.	korlarimizning	ovi	ilmayotganidan	lamaganidan	nikilarning	dangina
5.	laganlaridagina	ovimiz	ishimizdan	lashayotganligini	dek	gilarining
...	...	...	...	...	...	...
Total:	**1773**	**229**	**15870**	**725**	**509**	**1376**

4. A special software application is used to retrieve the list of all suffixes that can be added to the stem from the language corpus[87].

[86] Elov B., Hamroyeva Sh., Abdullayeva O, Khusainova Z., Khudaiberganov N. The issue of POS tagging and stemming for agglutinative languages (as examples of Turkish, Uyghur, Uzbek languages) / Uzbekistan: Language and Culture (Computer Linguistics). Tashkent, 2023, – No. 1. P- 35-50.

[87] http://uznatcorpara.uz/uz/Dictionary/Search

5. With the help of this software tool, it is possible to create a base of necessary word forms on the basis of regular expressions from the existing word forms of the Uzbek language corpus.

6. As a result of the search, certain word forms are divided into parts such as prefix, stem and suffix string.

7. A set of values $g = \langle p, s, q \rangle$ is determined for all given expressions satisfying the condition $p + s' + q = w$ In this,

w – word form;

s'– stem;

$p \in P$;

P is a set of prefix strings;

$s \in S$, S is a set of stem word groups;

$q \in Q$, Q is a set of suffix suffix strings;

$g \in G$, G is a set with three parameters, each element $\{p, s, q\}$.

8. An example of defined $\{p, s, q\}$values is given in the table.

After determining the values, the stem of the word form is determined using **the UzbStemming algorithm at the II stage**.

As a result of applying the above-mentioned UzbStemmer based on removal of affixes to more than 100000 sentences in the educational corpus of the Uzbek language (http://uzschoolcorpara.uz/uz/Search), a result of 97.5% accuracy was obtained.

The third section is called "*Problems and methods of lemming words in Uzbek*". All languages require different rules for determining a lemma, because languages have different word structures. Since the Uzbek language belongs to the family of agglutinative languages, there can be as many lemmas as there are tokens in sentences. But this opinion is not always correct. In some sentences, the number of lemmas may not be equal to the number of tokens. If we consider a sentence as a logical combination of words, it can be as follows:

1. $W_1, W_2, \ldots W_n$ *(tokens)*
2. $L_1, L_2, \ldots L_m$ *(lemmas), (m≤n) is observed*

In most cases, one token corresponds to one lemma. In the Uzbek language, word forms are divided into types according to their structure and formation method. According to the structure of words, the rules for lemming different forms are determined as follows:

1. When a word-forming suffix is added to a pair of words, the hyphen is preserved: *baxt-saodatli*. When determining the lemma of pairs of words in the Uzbek language, we rely on the dictionary, that is, pairs of words with independent meanings are considered to have two lemmas.

2. Since repeated words are not considered pseudowords, we assume that they contain two tokens and two lemmas. But there is an exceptional case where the second part of the repeated word has no meaning, which creates a problem in determining the number of lemmas. In this case, based on the dictionary, the meaningful part is taken.

1) *(meaningful) root + (meaningful) root*: *tez-tez, qator-qator, qop-qop, kunma-kun, oyma-oy, shu-shu, tez-tez, yilma-yil, baland-baland, ko'p-ko'p;*

2) *(meaningful) root + (meaningless) root*: *tuz-puz, don-dun, osh-posh, choy-poy, mol-hol, irim-sirim, qon-pon, pul-mul, meva-cheva.*

3. Compound words are considered to be one lemma even if they contain several bases due to the fact that different root words are combined to form a new word, a synthetic word is created.

4. More than one token is equal to one lemma, because compound verbs are formed from the combination of two or more bases, from which a single meaning is understood. These tokens can belong to different categories, the composition of the compound verb can be as follows:

1. Compound verbs with "noun + verb" structure (consisting of a non-verb word and a verb): *ta'zim qilmoq, g'ayrat qilmoq, fikr qilmoq*;

2. Compound verbs with "verb + verb" structure (consisting of two verbs): *olib bormoq; sotib olmoq, olib qochmoq, chiqib ketmoq.*

Compound verbs with the structure "verb + verb". This type of compound verb is formed by combining two verbs. The first part is usually in the form of -a, -y and -b (-ib) of the verb. The parts of the compound verb are combined and mean a new meaning different from their meaning: *olib bormoq, olib kelmoq, sotib olmoq, tashlab ketmoq.*

According to the number of tokens in the auxiliary verb phrases, it is observed that:

1. Two tokens: *aytib bermoq, yozib yubormoq.*
2. Three-token: *yozib berib turmoq.*
3. Four-token: *yozib berib turgan edi/emish/ekan.*
4. The leading verb consists of a compound verb KFSQ: *javob berib turmoq, imzo qo'yib bermoq and* etc.

No matter how many tokens a compound verb and KFSQ contain, it is a single lemma.

Compound adverbs. Even if there are several tokens in a compound expression, its lemma is one. Compound adverbs are formed by combining two word bases:

har gal, bir vaqt, tez fursatda, shu yoqqa, bir mahal, bir zum, bir nafas, har dam, har on.

Compound number. Even if the parts of a composite number have several tokens, they are treated as one lemma. The composition of the compound number is also written separately. For example, *o'n besh, ikki ming yigirma uch, besh yuz olti, yigirma bir*, etc.

Compound pronoun. Compound pronouns have two tokens, one lemma. Pronouns such as *Hech kim, hech qachon, har bir, har kim, mana shu, ana o'sha* are compound pronouns. Simple pronoun parts are added, and compound pronoun parts are written separately.

The phrase. Phrases can sometimes be two-token, three-token, or even four-token or more, but are considered a single lemma. In the Uzbek language, phraseological expressions consist of two, three or more words. A phrase is one of the units that make up the vocabulary of our language. They are a lexical unit formed on the basis of the transferred meanings of two or more words and have the same

lexical meaning as a word. For example, the phrase "*xamirdan qil sug'urganday*" means "easily", "without difficulty", the phrase "*ko'ngil bermoq*" means "to love", and the phrase "*qo'y og'zidan cho'p olmagan*" means "friendly".

Also, our language contains words denoting the name (NERs), new words (neologisms) that have not yet been assimilated into the vocabulary, and abbreviations defined in the documents, which need to be agreed upon in determining their lemma. Even if there are several tokens in these word forms, the lemma like stem is defined as one and it is concluded as follows.

1. Lemma and stem of NERs are the same. (A word-former found in a noun phrase is not taken into account).

2. Neologisms have the same lemma and stem.

3. The lemma and stem of abbreviations (BMT), conditional abbreviations (*sh.k., va b., sm, mlrd.*) are the same.

Therefore, in the language, a part with an independent meaning is taken, which performs a function in the context in determining the lemma of the word form. All form-forming affixes in word forms are removed and word-forming affixes are left.

CONCLUSION

1. Tokenization is the process of dividing sentences in a given text into the smallest unit of measurement - a token. Punctuation, words, and numbers can be tokenized in a sentence. In the tokenization process, natural language text is divided into tokens by initial processing; set boundaries between tokens. Tokenization can be done at the word, character, or word content element level.

2. When solving many tasks of NLP, it is necessary to reduce (stemming) the word forms to the root. Stemming is the first stage of text processing in NLP, which is used to improve the performance of NLP applications such as spell checking. Today, the Porter stemmer algorithm has been modified and is widely used. In stemming methods and approaches such as corpus-based approach(1); hybrid approach (2); statistical method (3); n-gram stemmer (4); context-dependent stemmer (5) are used.

3. Stemmer algorithms have advantages and disadvantages. None of the Lovins, Porter, Pace/Hask, Dawson, Snowball, Krovetz, Lancaster, YASS, and HMM stemmers developed for other natural languages are 100 percent accurate. These stemming algorithms cannot be directly applied to the lexical units of the Uzbek language, therefore, the development of the UzbStemming algorithm based on the characteristics of the Uzbek language becomes relevant.

4. Lemmas is an important step in a series of NLP tasks, which consists in finding the (basic) form of a word in the dictionary. Stemming and lemming are common methods for determining the stem of a word in natural language. Lemmas are implemented differently in different feature languages. Lemmaing is used in NLP applications such as POS tagging, spell checking, and document clustering. Today, lemma algorithms have been tested on small texts in the Uzbek language. Although machine learning and neural network-based lemma algorithms have been developed in various languages, none of these methods have been introduced into the Uzbek language so far.

5. Since the Uzbek language is agglutinative, it is appropriate to develop stemming algorithms based on linguistic rules. Stem is a part formed by cutting off the suffixes of the word form, which in some cases does not have meaning; stem may or may not exactly match the morphological stem of the word. In the Uzbek language, stemming is the reduction to the root by removing the word-forming and form-forming suffixes added to the word.

6. When solving the problem of sound changes in determining the stem of words in the Uzbek language, at the first stage, the border of the stem and suffixes is determined, and at the second stage, lemma is performed. As a result of lemming, the resulting stems are changed to the existing stem in the dictionary. When the problem of finding the stem of NER and neologisms arises, the form-formers are cut, the suffix or part of the word in the form of the word-former is left, and this part is equal to the stem.

7. The action of lemming increases the accuracy of analysis of the content of language corpus texts. Lemmas are of practical importance in the accuracy of search engine results and the understanding of corpus text content. Another advantage of lemming is that it helps to identify the overall structure and keywords of web page content. Regardless of the chosen approach, initial processing steps such as (1) tokenization (2); POS tagging (3); lemming (4); integration (5) are performed on a web page in performing lemma;.

8. Today, BoW (Bag of Words) and BPE (Byte Pair Encoding) algorithms are widely used in the tokenization process. BoW (bag of words) is a method of text modeling (digitization) in natural language processing, which allows processing large volumes of text data in information systems. Using the BoW algorithm, necessary features are extracted from textual data and the number of words (statistics) in the text is determined.

9. Byte Pair Encoding (BPE) is a widely used tokenization method among transformer-based models, which can solve the problems of word and character tokenizers. BPE is a word segmentation algorithm that combines the most frequent characters or sequences of characters. Uzbek lexical units were tokenized using BoW and BPE methods, and software (https://uznatcorpara.uz/uz/Tokenizer) was developed.

10. In the development of the UzbStemmer software based on removing affixes, a language corpus, a base of affixes, a morpholexicon of the Uzbek language, a base of roots, POS tagging were used, and a two-stage algorithm was used in the educational corpus of the Uzbek language (http://uzschoolcorpara.uz/uz/Search), as a result of application 97.5% accuracy efficiency was achieved in more than 100,000 sentences.

11. The lemming algorithm is a complex method of finding the root of a word. When using the algorithm, it is necessary to understand the natural language and its grammar. As the Uzbek language belongs to the family of agglutinative languages, word forms are formed in the form of word+affix or stem+stem. The number of tokens in a sentence can be equal to the number of lemmas. In some sentences, the

number of lemmas is not equal to the number of tokens. In this research, the methods of lemming lexical units such as simple word, repeated word, double word, compound word (compound verb, compound adverb, compound number), KFSQ and phrase are presented and https://uznatcorpara.uz/uz/Lemmatizer software supply was developed.

РАЗОВЫЙ НАУЧНЫЙ СОВЕТ НА БАЗЕ НАУЧНОГО СОВЕТА ПО ПРИСУЖДЕНИЮ УЧЁНЫХ СТЕПЕНЕЙ ЗА НОМЕРОМ DSc.03/30.12.2019.Fil.19.01 ПРИ ТАШКЕНТСКОМ ГОСУДАРСТВЕННОМ УНИВЕРСИТЕТЕ УЗБЕКСКОГО ЯЗЫКА И ЛИТЕРАТУРЫ ИМЕНИ АЛИШЕРА НАВОИ

ТАШКЕНТСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ УЗБЕКСКОГО ЯЗЫКА И ЛИТЕРАТУРЫ ИМЕНИ АЛИШЕРА НАВОИ

ХУСАИНОВА ЗИЛОЛА ЮЛДАШЕВНА

ЛИНГВИСТИЧЕСКИЕ ОСНОВЫ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ТОКЕНИЗАЦИИ, СТЕММИНГА, ЛЕММИНГА ЕДИНИЦ УЗБЕКСКОГО ЯЗЫКА

10.00.11 - Теория языка. Прикладная и компьютерная лингвистика

АВТОРЕФЕРАТ ДИССЕРТАЦИИ ДОКТОРА ФИЛОСОФИИ (PhD) ПОФИЛОЛОГИЧЕСКИМ НАУКАМ

Ташкент – 2024

Тема диссертации доктора философии (PhD) зарегистрирована в Высшей аттестационной комиссии Республики Узбекистан под номером B2023.2.PhD/Fil3688.

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

Автореферат диссертации размещен на трех языках (узбекский, английский, русский (резюме)) на веб-странице Научного совета (www.tsuull.uz) и на Информационно-образовательном портале «ZiyoNet» (www.ziyonet.uz).

Научный руководитель:	**Элов Ботир Болтаевич** доктор философии по техническим наукам, доцент
Официальные оппоненты:	**Холмонова Зулхумор Турдиевна** доктор филологических наук, профессор
	Адбурахмонова Нилуфар Зайнобиддин кизи доктор филологических наук, профессор
Ведущая организация:	**Самаркандский государственный университет**

Защита диссертации состоится «14» 06 2024 г. в 10:00 часов на заседании Разового научного совета на базе Научного совета по присуждению ученых степеней DSc.03/30.12.2019.Fil.19.01 при Ташкентском государственном университете узбекского языка и литературы. (Адрес: 100100, г. Ташкент, Яккасарайский район, ул. Юсуф Хос Хожиб, дом 103. Тел.: (99871) 281-42-44; факс: (99871) 281-42-44; факс: (99871) 281-12-44. (www.tsuull.uz). e-mail: monitoring@navoiy-uni.uz)

С диссертацией можно ознакомиться в информационно-ресурсном центре Ташкентского государственного университета узбекского языка и литературы имени Алишера Навои. (зарегистрирована за № 290). (Адрес: 100000, г. Ташкент, Яккасарайский район, ул. Юсуф Хос Хожиб, дом 103. Тел.: (99871) 281-42-44; факс: (99871) 281-42-44; факс: (99871) 281-12-44. (www.tsuull.uz)).

Автореферат диссертации разослан «30» 05 2024 г.

(Реестр протокола рассылки за № 1 от «30» 05 2024 г.)

Х.А. Дадабоев
Председатель Разового научного совета на базе Научного совета по присуждению ученых степеней, доктор филологических наук, профессор

К.У.Пардаев
Ученый секретарь Разового научного совета на базе Научного совета по присуждению ученых степеней, доктор филологических наук, доцент

С.Э. Нормаматов
Председатель Разового научного семинара на базе Разового научного совета при Научном совете по присуждению ученых степеней, доктор филологических наук, профессор

ВВЕДЕНИЕ (аннотация диссертации доктора философии (PhD))

Актуальность и необходимость темы диссертации. Методы обработки естественного языка широко используются в практике мировой компьютерной лингвистики. У проектов обработки естественного языка (Natural language prossecing – NLP) есть постоянные задачи. Они широко изучаются из-за своей фундаментальности. Хорошее понимание этих задач позволяет создать разные приложения НЛП (NLP). Моделирование языка, классификация текста, получение информации, поиск информации, агент общения, обобщение текста, системы вопросов и ответов, машинный перевод, тематическое моделирование – основные направления НЛП. Обработка естественного языка – это направление области искусственного интеллекта, целью которого является заставить машины понимать и обрабатывать человеческий язык. Обработка естественного языка — непростая задача для машины, поскольку машина работает с числами, а не с текстом.

В мировой лингвистике в конце XX века на повестку дня были поставлены такие вопросы, как повышение качества автоматического перевода, лингвистическое моделирование языка, теория лемминговых слов внутри определённых языков, создание алгоритма. В результате были разработаны такие приложения, как токенизатор, стеммер, лемматизатор, которые активно используются в системе поиска информации. Следовательно, создание языковых корпусов, их лингвистическая аннотация, разработка программных средств автоматической обработки текста как морфоанализатор, стеммер, лемматизатор, синтаксический анализатор, орфокорректор стали научно-теоретическими задачами компьютерной лингвистики.

В области узбекской компьютерной лингвистики уделяется внимание научным исследованиям в области языкового корпуса, синтезатора речи, автоматического перевода, искусственного интеллекта, понимания и обработки узбекского языка. Однако научно-теоретически не изучены вопросы разработки средств как токенизатор, стеммер, лемматизатор, которые реализуют начальные этапы обработки естественного языка. Ибо, «... сохранение чистоты государственного языка и его обогащение, повышение культуры речи населения; обеспечение занятия государственным языком достойного места в области информационных и коммуникационных технологий»[88] является одной из актуальных задач, стоящих сегодня перед узбекской компьютерной лингвистики. Соответственно, важно добиться обработки узбекского языка средствами современных информационных технологий, для этого важны такие вопросы, как создание первичных средств автоматической обработки, таких как языковые корпуса, электронный переводчик, тезаурус, проверка орфографии, лингвистические анализаторы и их лингвистическое обеспечение.

[88] Указ Президента Республики Узбекистан от 21 октября 2019 года «О мерах по кардинальному повышению роли и авторитета узбекского языка в качестве государственного языка» // УП-5850-сон 21.10.2019. О мерах по кардинальному повышению роли и авторитета узбекского языка в качестве государственного языка (lex.uz)

Данная диссертационная работа, в определенной мере служит реализации задач, определенных в Указах Президента Республики Узбекистан № УП-4797 от 13 мая 2016 года «О создании Ташкентского государственного университета узбекского языка и литературы имени Алишера Навои», № УП-5850 от 21 октября 2019 года «О мерах по кардинальному повышению роли и авторитета узбекского языка в качестве государственного языка», № УП-6084 от 20 октября 2020 года «О мерах по дальнейшему развитию узбекского языка и совершенствованию языковой политики в стране», в Постановлениях Президента Республики Узбекистан № ПП-2789 от 17 февраля 2017 года «О мерах по дальнейшему совершенствованию деятельности Академии наук, организации, управления и финансирования научно-исследовательской деятельности», № ПП-4479 от 4 октября 2019 года «О широком праздновании тридцатилетия принятия закона Республики Узбекистан «О государственном языке»», в Послании Президента Республики Узбекистан Олий Мажлису от 24 января 2020 года и других нормативных правовых документах, связанных с этой деятельностью.

Цель исследования. Целью исследования является разработка лингвистических основ и программного обеспечения для токенизации, стемминга и лемминга лексических единиц узбекского языка.

Объект исследования. Объектом исследования являются лексические единицы узбекского языка.

Научная новизна исследования:

систематизированы теоретические основы токенизации, стемминга и лемминга лексических единиц в мировом и узбекском языкознании, определены способы токенизации на основе знаков и слов, стемминга с гибридным подходом на основе лингвистических правил и корпуса, а также методы лемминга на основе словаря;

для агглютинативных языков выявлены такие проблемы, как омонимия корня и суффикса с одним корнем, возникновение звуковых изменений слов в процессе стеммирования, стемминг неологизмов и НЭР (NER), а также разработаны лингвистические основы стеммирования и леммирования лексических единиц разной структуры на узбекском языке;

на основе обработки текстов национального корпуса узбекского языка разработаны методы поисковой оптимизации, выявлен процесс стеммирования и лемминга лексических единиц;

процесс токенизации осуществляется на основе алгоритма BPE, разработан алгоритм стемминга на основе разделения аффиксов, лемминга на основе словарного и морфологического анализа, создано программное обеспечение «узб.токенезатор (uzb.tokenizator)», «узб.стемминг (uzb.stemming)» и «узб.лемматизатор (uzb.lemmatizator)» для текстов на узбекском языке.

Внедрение результатов исследований. На основе результатов, полученных по исследованию токенизации, стемминга, лемматизации лексических единиц узбекского языка:

научные выводы о токенизации, основанной на символах и словах; о стемминге с гибридном подходом, осуществляемый посредством лингвистических правил и корпуса; о лемминге, основанном на словаре, составляющие программное обеспечение, которое реализует токенизацию, стемминг и лемминг единиц узбекского языка, использованы в Центре обучения и повышения квалификации основам работы на государственном языке при Ташкентском государственном университете узбекского языка и литературы имени Алишера Навои (Акт Ташкентского государственного университета узбекского языка и литературы имени Алишера Навои от 12 октября 2023 года № 123). В результате, в следствие тестирования программного обеспечения для токенизации, стемминга и лемминга единиц узбекского языка, была сформирована база данных из более чем 32 000 простых и более 7500 сложных лексем при морфологическом анализе предложений узбекского языка;

результаты токенизации, стемминга и лемминга единиц узбекского языка, в частности научные результаты по повышению качества стемминга и лемминга в узбекском, русском и английском языках при обработке естественного языка использованы в практическом проекте №И-ОТ-2019-42 «Создание многоязычной электронной платформы узбекской литературы (на узбекском, русском, английском языках)», выполненном в 2021-2023 годах в Ташкентском государственном университете узбекского языка и литературы имени Алишера Навои (справка Ташкентского государственного университета узбекского языка и литературы имени Алишера Навои № 01/10-2182 от 19 октября 2023 года). В результате, повышение качества стемминга и лемминга в узбекском, русском и английском языках при обработке естественного языка улучшило качество поиска на электронной платформе; по результатам токенизации, стемминга и лемминга единиц узбекского языка обеспечивается эффективность результата поиска единиц, специфичных для узбекского языка;

алгоритмы, обосновывающие результаты токенизации, стемминга и лемминга единиц узбекского языка и программное обеспечение, созданное на их основе, использованы в практическом проекте ПЗ-2020042022 «Создание лингводидактической электронной платформы тюркских языков», реализуемом в 2021-2023 годах в Ташкентском государственном университете узбекского языка и литературы имени Алишера Навои (справка Ташкентского государственного университета узбекского языка и литературы имени Алишера Навои № 01/10-2183 от 19 октября 2023 года). В результате на основе алгоритма стемминга и лемминга увеличен индекс поисковой системы лингводидактической электронной платформы тюркских языков.

Структура и объем диссертации. Диссертация состоит из введения, трёх глав, заключения, списка использованной литературы и приложения, ее объём составляет 129 страниц.

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

I bo'lim (Часть I; Part I)

1. Khusainova Z.Y. An Approach to Distinguishing Affixes in the Development of the Uzbstemming Algorithm //Pindus Journal of Culture, Literature, and ELT. 2023, 3(12) – B. 8-18. ISSN: 2792 – 1883.

2. Xusainova Z.Y. Tokenizatsiya algoritmlari // Filologik tadqiqotlar: til, adabiyot, ta'lim. Termiz, 2022. №5. – B.73-76.

3. Xusainova Z.Y. BPE algoritmi asosida tokenizatsiya jarayonini amalga oshirish // O'zbekiston milliy universiteti xabarlari jurnali. Toshkent, 2023. №1/3/1. – B. 296-298 (10.00.00. № 15).

4. Xusainova Z.Y. O'zbek tili milliy korpusi qidiruv tizimini optimallashtirishda lemmatizatsiyadan foydalanish / O'zbekiston: til va madaniyat (Kompyuter lingvistikasi). Toshkent, 2023, №2. – B. 20-37.

5. Xusainova Z.Y. NLPning zamonaviy vazifalari // "O'zbek tilining milliy korpusi: muammolar va vazifalar" mavzusidagi xalqaro ilmiy-amaliy anjuman. Samarqand, 2023. – B. 140-148.

6. Xusainova Z.Y O'zbek tilida stemmingni amalga oshirishning gibrid statistik yondashuvi // "Kompyuter lingvistikasi: muammolar, yechim, istiqbollar" mavzusidagi an'anaviy xalqaro ilmiy-amaliy konferensiya. Toshkent, 2023. – B. 70-76.

7. Xusainova Z.Y. NLP: tokenizatsiya, stemming, lemmatizatsiya va nutq qismlarini teglash // "O'zbek amaliy filologiyasi istiqbollari" mavzusidagi respublika ilmiy-amaliy konferensiyasi – Toshkent, 2022. №.1. – B. 154-163.

8. Xusainova Z.Y. O'zbek tili ta'limiy korpusida lemmalash algoritmlari // "O'zbek tili milliy va ta'limiy korpusining nazariy va amaliy masalalari" mavzusidagi respublika ilmiy-amaliy konferensiyasi materiallari. Toshkent, 2023. – B. 43-50.

II bo'lim (Часть II; Part II)

9. Xusainova Z.Y. Methods of Lemmatizing Various Structured Words in Uzbek Language / Central Asian Journal of Literature, Philosophy and Culture. 2023, 4(12). – B. 67-72. eISSN: 2660-6828.

10. Elov B.B., Hamroyeva Sh.M., Xusainova Z.Y. The pipeline processing of NLP // E3S Web of Conferencesç INTERAGROMASH. Rossiya, Rostov, 2023. https://doi.org/10.1051/e3sconf/202341303011

11. Elov.B.B., Xusainova Z.Y., Xudayberganov N.U. Tabiiy tilni qayta ishlashda bag of words algoritmidan foydalanish / O'zbekiston: til va madaniyat. (Amaliy filologiya masalalari). 2022. №2. – B. 35-50.

12. Elov B.B., Hamroyeva Sh.M., Xusainova Z.Y. NLP (tabiiy tilga ishlov berish)ning vazifalari va zamonaviy yondashuvlar / Filologik tadqiqotlar: til, adabiyot, ta'lim. Termiz, 2022. №5. – B. 19-26.

Avtoreferat «O'zbekiston: til va madaniyat» jurnali tahririyatida tahrirdan o'tkazildi.

Adadi 100 nusxa. Bichimi 60x84 $^{1}/_{16}$
Bosma tabog'i 3,4. "Times New Roman" garniturasi.
"BOOKMANY PRINT" MCHJ bosmaxonasida chop etildi.
Toshkent shahri, Uchtepa tumani, 22-mavze, 17-b uy.