TIL KORPUSI MATNLARIDA IMLO TUZATISH

Botir Elov Boltayevich,
texnika fanlari falsafa doktori, dotsent
elov@navoiy-uni.uz
ToshDO‘TAU

Ahmedova Maftuna Kaxramon qizi,
Kompyuter lingvistikasi mutaxasisligi magistranti
MaftunaAhmedova1997@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu maqola til korpusi matnlarida imlo xatolarini aniqlash va tuzatish muammosiga bag‘ishlangan. Tadqiqot kompyuter lingvistikasi va tabiiy tilni qayta ishlash sohasidagi zamonaviy yondashuvlarni tahlil qiladi. Korpus matnlaridagi imlo xatolarini avtomatik aniqlash va tuzatish algoritmlari, usullari va dasturiy vositalari ko‘rib chiqiladi. Tadqiqot natijasida imlo tuzatishning statistik, qoidaga asoslangan va mashina o‘rganishiga asoslangan usullarining samaradorligi taqqoslanadi. Xususan, til korpuslarida imlo xatolarini aniqlash va tuzatish jarayonida qo‘llaniladigan zamonaviy yondashuvlar va ularning o‘zbek tili korpuslariga tatbiq etish masalalari muhokama qilinadi.

Abstract. This article deals with the problem of detecting and correcting spelling errors in language Corpus texts. The study analyzes modern approaches in the field of computer linguistics and natural language processing. Algorithms, methods and software tools for automatic detection and correction of spelling errors in Corpus texts are considered. The study compares the effectiveness of statistical, rule-based, and machine learning-based methods of spelling correction. In particular, the language Corps discusses the modern approaches used in the process of detecting and correcting spelling errors and their application to the Corps of the Uzbek language.

Аннотация. Данная статья посвящена проблеме выявления и исправления орфографических ошибок в текстах языкового корпуса. В исследовании анализируются современные подходы в области компьютерной лингвистики и обработки естественного языка. Рассмотрены алгоритмы, методы и программные средства автоматического обнаружения и исправления орфографических ошибок в текстах корпуса. В ходе исследования сравнивается эффективность статистических методов коррекции орфографии, основанных на правилах, и методов машинного обучения. В частности, обсуждаются современные подходы к выявлению и исправлению орфографических ошибок в языковых корпусах и их применение в корпусах узбекского языка.

Kalit so‘zlar: *imlo tuzatish, til korpusi, tabiiy tilni qayta ishlash, mashina o‘rganishi, statistik modellar*.

KIRISH

Zamonaviy axborot texnologiyalari asrida til korpuslarining sifati va to‘g‘riligi juda muhim ahamiyatga ega. Til korpusi – bu muayyan tildagi matnlarning elektron to‘plami bo‘lib, tilshunoslik tadqiqotlari, kompyuter lingvistikasi, tabiiy tilni qayta ishlash, statistik tahlil va boshqa ko‘plab sohalarda qo‘llaniladi.

Til korpusi – bu tilning haqiqiy foydalanuvchilari tomonidan ishlab chiqarilgan va so‘zlar,iboralar va umuman, til qanday ishlatilishini tahlil qilish uchun ishlatiladigan juda katta matnlar to‘plami. U tilshunoslar, leksikograflar, ijtimoiy olimlar, gumanitar fanlar, tabiiy tillarni qayta ishlash bo‘yicha mutaxassislar va boshqa ko‘plab sohalarda qo‘llaniladi. Korpus shuningdek, dasturiy ta’minotni ishlab chiqishda ishlatiladigan turli til ma’lumotlar bazalarini yaratish uchun ishlatiladi, masalan, bashoratli klaviaturalar, imloni tekshirish va tuzatish, matn/nutqni tushunish tizimlari,matndan nutqqa modullar, mashina tarjimasi tizimlari va boshqalar[1].

Til korpusi – bu biror tilni to‘g‘ri va kompleks tarzda o‘rganish uchun zarur bo‘lgan ma’lumotlarni o‘z ichiga olgan matnlar to‘plamidir. Bunday matnlar, odatda, turli xil yozuvlar va nutq namunalarini o‘z ichiga oladi, masalan, kitoblar, maqolalar, she’rlar, onlayn matnlar, ijtimoiy tarmoq postlari va hokazo. Korpuslar tilshunoslar va kompyuter lingvistikasi mutaxassislari tomonidan tilni tahlil qilish, o‘rganish, semantik va sintaktik strukturalarni aniqlash hamda tilni qayta ishlashning turli texnologiyalarini ishlab chiqish uchun foydalaniladi[2].

Biroq korpus matnlari turli manbalardan olingan bo‘lib, ko‘pincha kiritish xatolari, noto‘g‘ri so‘zlar, texnik xatolar va boshqa nuqsonlarni o‘z ichiga oladi. Bu xatolar korpus asosida ishlaydigan tizimlar ishini sezilarli darajada yomonlashtirishi mumkin. Shuning uchun korpus matnlaridagi imlo xatolarini aniqlash va tuzatish – kompyuter lingvistikasining dolzarb muammolaridan biridir [3].

Turli tillar uchun imlo tuzatish tizimlari mavjud bo‘lsa-da, o‘zbek tili kabi resurslar cheklangan tillar uchun bu masala hali ham to‘liq hal qilinmagan. O‘zbek tilining o‘ziga xos xususiyatlari – lotin va kirill alifbolaridan foydalanish, morfologik boylik, so‘z yasalishining murakkabligi kabi omillar imlo tuzatish tizimlarini yaratishda qo‘shimcha qiyinchiliklar tug‘diradi [4].

Ushbu maqolada til korpusi matnlaridagi imlo xatolarini aniqlash va tuzatish bo‘yicha mavjud yondashuvlar tahlil qilinadi, ularning afzalliklari va kamchiliklari aniqlanadi, hamda o‘zbek tili korpuslarida imlo tuzatish tizimlarini takomillashtirish bo‘yicha takliflar beriladi.

METODOLOGIYA VA ADABIYOTLAR TAHLILI

Tadqiqot metodologiyasi tizimli adabiyotlar tahlili (systematic literature review) usulini qo‘llashga asoslangan. Imlo tuzatish tizimlarini yaratishda asosan uch xil yondashuv qo‘llaniladi: qoidaga asoslangan, statistik va mashina o‘rganishiga asoslangan usullar[5].

Qoidaga asoslangan yondashuvlar til grammatikasi va imlo qoidalariga tayangan holda xatolarni aniqlaydi va tuzatadi. Qoidaga asoslangan tizimlar aniq va tushunarli natijalar bersa-da, barcha hollarda ham yaxshi natija bermaydi. Chunki tilning barcha grammatik qoidalarini ifodalash murakkab. Bundan tashqari, yangi so‘zlar, chet tilidan kirgan so‘zlar va adabiy tilda qo‘llanilmaydigan so‘zlar va eskirgan so‘zlar bilan ishlashda qiyinchiliklar yuzaga keladi.

Statistik yondashuvlar katta hajmdagi korpuslar asosida so‘zlarning chastotasi va ularning kontekstdagi ishlatilishini tahlil qiladi. Bu usul “N-gram” modellari, Noisy Channel modellarni o‘z ichiga oladi. Statistik usullar faqat katta hajmdagi ma’lumotlar mavjud bo‘lgandagina yaxshi natija bergani uchun kam resursli tillar uchun statistik yondashuvni qo‘llash qiyin.

So‘nggi yillarda mashina o‘rganishi va chuqur o‘rganish (deep learning) usullari korpus matnlaridagi imlo xatolarini aniqlash va tuzatishda keng qo‘llanilmoqda. Bu yondashuvda neyron tarmoqlar, xususan rekurrent neyron tarmoqlar (RNN), uzun-qisqa muddatli xotira (LSTM) va transformerlar kabi modellar ishlatiladi. Masalan, I. Petrov, Smirnov va V.Ivanov “Transformer-based spelling correction for Russian text corpora” nomli maqolasida transformer arxitekturasiga asoslangan modelni rus tili korpuslaridagi imlo xatolarini tuzatish uchun taklif etganlar va bu model 92% aniqlikka erishgan[6]. Mashina o‘rganishi usullari yuqori aniqlikka erishishi mumkin, lekin bu usullarni qo‘llash uchun katta hajmdagi annotatsiyalangan ma’lumotlar to‘plami (dataset) talab etiladi.

Bundan tashqari, gibrid (hibrid) yondashuvlar ham mavjud bo‘lib, ular yuqoridagi usullarning kombinatsiyasidan foydalanadi.

Kompyuter lingvistikasi mutaxassislari yaxshi biladilarki, imloviy sifati past bo‘lgan korpus uning asosida yaratiladigan dasturiy mahsulotlar sifatiga salbiy ta’sir qiladi. Masalan, neyron tarmoqlar asosida ishlaydigan sun’iy intellektni xatolari ko‘p matn bilan mashq qildirilsa, u ana shu matndagi xatolarni tilning ajralmas qismi deb qabul qilishga va shundan kelib chiqib ishlashga o‘rganadi. Oddiy misol: sun’iy intellektga berilgan korpusda “tatbiq” so‘zi 40% hollarda “tadbiq” deb yozilgan bo‘lsa, u shu so‘z o‘zbek tilida ikki xil usulda yozilar ekan deb xulosa chiqaradi va shu xulosa asosida ishlaydi. Bu juda soddalashtirilgan misol bo‘lgani bilan, uning o‘zi lingvistik dasturni “*o*‘qitish” uchun ishlatilgan korpus sifati uning yakuniy ishlash sifatiga qanday ta’sir qilishini tushunib yetish uchun yetarli[7].

O‘zbek tili korpuslaridagi imlo xatolarini tuzatish masalasi bo‘yicha maxsus tadqiqotlar juda kam. Mavjud tadqiqotlar ko‘proq rus, ingliz,ispan va boshqa keng tarqalgan tillarga qaratilgan. Bu esa o‘zbek tili korpuslari matnlarini takomillashtirish uchun yangi yondashuvlarni ishlab chiqish kerakligini ko‘rsatadi.

Adabiyotlar tahlili natijasida til korpusi matnlarida imlo tuzatishning turli usullari samaradorligi aniqlandi va qiyosiy tahlil qilindi. 1-jadvalda imlo tuzatish usullarining qiyosiy tahlili keltirilgan.

1-jadval. Imlo tuzatish usullarining qiyosiy tahlili

Usul	**Afzalliklari**	**Kamchiliklari**	**O‘rtacha aniqlik (%)**	**O‘zbek tiliga moslashish imkoniyati**
Qoidaga asoslangan	Tushunarliligi, izohlanishi oson,kichik korpus uchun qo‘llanilishi mumkin.	Yangi so‘zlarni qamrab ololmasligi.	75-80%	Yuqori, lekin morfologik qoidalarni to‘liq kiritish talab etiladi.
Statistik	Kontekstga asoslanganligi, yangi so‘zlarni ham qamray olishi.	Katta hajmdagi ma’lumotlar talab etilishi.	80-85%	O‘rta, katta hajmli korpus talab etiladi.
Mashina o‘rganishi	Yuqori aniqlik, kontekstni yaxshi tushunishi, yangilanish imkoniyati.	Katta annotatsiyalangan ma’lumotlar bazasi talab etilishi, hisoblash resurslarining ko‘pligi	85-92%	Past-o‘rta, annotatsiyalangan ma’lumotlar yetishmasligi.
Gibrid	Turli usullarning afzalliklarini birlashtirishi.	Tizimning murakkabligi, sozlash qiyinligi.	88-94%	O‘rta-yuqori, moslashtirilgan holda yaxshi natija beradi.

Jadvaldan ko‘rinib turibdiki, zamonaviy yondashuv sifatida gibrid tizimlar eng yuqori aniqlikka erishadi, ammo ularni yaratish va sozlash murakkab jarayondir. O‘zbek tili uchun qoidaga asoslangan tizimlar nisbatan oson moslashtirish imkoniyatiga ega bo‘lsa-da, ularning aniqligi pastroq. Mashina o‘rganishiga asoslangan usullar yuqori aniqlikka erishishi mumkin, lekin o‘zbek tili uchun annotatsiyalangan ma’lumotlar yetishmasligi bu usulni qo‘llashni qiyinlashtiradi.

2-jadvalda esa turli tillarda imlo tuzatish tizimlari uchun erishilgan aniqlik ko‘rsatkichlari taqqoslangan.

2-jadval. Turli tillarda imlo tuzatish tizimlarining aniqligi

Til	**Qoidaga asoslangan (%)**	**Statistik (%)**	**Mashina o‘rganishi (%)**	**Gibrid (%)**
Ingliz	82%	88%	94%	96%

Rus	80%	85%	92%	94%
Turkcha	78%	83%	90%	93%
O‘zbek	75%	80%	83%	88%

Jadval ma’lumotlaridan ko‘rinib turibdiki, resurslari cheklangan tillar, xususan turkiy tillarda imlo tuzatish tizimlarining aniqligi ingliz va rus tillariga nisbatan pastroq. Bu ushbu tillar uchun korpuslar hajmining kichikligi, annotatsiyalangan ma’lumotlarning yetishmasligi va tilning o‘ziga xos murakkabliklari bilan bog‘liq.

O‘zbek tili korpuslarida imlo tuzatish tizimlarini takomillashtirish uchun quyidagi yondashuvlar taklif etiladi:

1. Morfologik tahlilga asoslangan qoidalar tizimini takomillashtirish. O‘zbek tilining morfologik boyligi va so‘z yasalishining o‘ziga xos xususiyatlarini hisobga olgan holda qoidalar tizimini taklif etish lozim.

2. “O‘zbek tili milliy korpusi” ma’lumotlaridan foydalangan holda keng qamrovli statistik modellarni yaratish. Bu korpus hajmini kengaytirish, sifatini oshirish orqali amalga oshirilishi mumkin.

3. Transfer learning usulidan foydalanib, rus, turk kabi yaqin tillarda o‘rgatilgan modellarni o‘zbek tiliga moslashtirish. Bu usul orqali ma’lumotlar yetishmasligi muammosini qisman hal qilish mumkin.

4. O‘zbek tili xususiyatlarini hisobga olgan gibrid tizimlarni yaratish: morfologik tahlil, statistik modellar va neyron tarmoqlarni birlashtiradigan yondashuvlarni ishlab chiqish.

Oldimizda turgan yana bir masalalardan biri:o‘zbek tili imlo qoidalarini bir standartga keltirish, alifbo masalasini uzil-kesil hal qilishdan iboratdir.

O‘zbek tili korpuslarida uchraydigan imlo xatolarini tahlil qilganda, xatolarning quyidagi asosiy turlari ko‘p uchraydi:

Lotin va kirill alifbolari orasidagi transpozitsiya xatolari (masalan, “sh” o‘rniga “ш” yoki “w” aksincha)

1. H va X harflari bilan bog‘liq xatolar (masalan, “h” va “x” harflarining almashtirilishi).

2. Qo‘shimchalarning noto‘g‘ri qo‘shilishi (“mening kitobim” emas “meni kitobim”).

3. Qo‘shma so‘zlarning alohida yozilishi bilan bog‘liq xatolar (“oqqush” emas “oq qush” tarzida).

4. Olinma so‘zlarni yozishda yuzaga keladigan xatolar (“multimedia” shaklida emas, “multimediya” shalida yozilish holatlari) ko‘p uchraydi.

Til korpuslarida imlo tuzatish jarayonini avtomatlashtirish uchun kontekstual ma’lumotlardan foydalanish muhim ahamiyatga ega. Chunki ko‘p so‘zlar kontekstga qarab to‘g‘ri yoki noto‘g‘ri yozilgan bo‘lishi mumkin. Masalan, “boshog‘riq” va “bosh og‘riq” so‘zlari kontekstga qarab bir-birining o‘rniga ishlatilishi mumkin, lekin ular turli ma’noga ega.Birinchisida “muammo” ma’nosida kelsa,ikkinchisida boshdagi og‘riq ma’nosida keladi.Bunday holatlarda kontekstual ma’lumotlardan foydalanish imlo tuzatish tizimining aniqligini sezilarli darajada oshiradi.

Zamonaviy transformer asosli modellar (BERT, GPT, T5) kontekstual ma’lumotlarni yaxshi tushunishi va saqlashi bilan ajralib turadi. Ushbu modellarni o‘zbek tili korpuslariga moslashtirish va imlo tuzatish tizimlarida qo‘llash istiqbolli yo‘nalish hisoblanadi.

Korpus matnlarida imlo tuzatish jarayoni faqat xatolarni aniqlash va tuzatishdan iborat emas. Bu jarayon matn sifatini oshirish, uni standartlashtirish va keyingi ishlov berish uchun tayyorlashning muhim bosqichi hisoblanadi. Sifatli tuzatilgan korpus matnlari tilshunoslik tadqiqotlari, mashinali tarjima, avtomatik annotatsiyalash va boshqa ko‘plab vazifalar uchun muhim ahamiyatga ega.

O‘zbek tili korpuslarida imlo tuzatishning hozirgi holati va istiqbollari tahlil qilinganda, quyidagi muammolar va yechimlar aniqlanadi:

1. Annotatsiyalangan ma’lumotlar yetishmasligi – bu muammoni kraudsorsing, mavjud korpuslar asosida sun’iy xatolar yaratish va qo‘shni tillar ma’lumotlaridan foydalanish orqali hal qilish mumkin.

2. O‘zbek tilining o‘ziga xos xususiyatlari – morfologik tahlil, qo‘shimcha qo‘shish qoidalari va so‘z yasalish qonuniyatlarini hisobga olgan algoritmlar yaratish.

3. Lotin va kirill alifbolari orasidagi transliteratsiya muammolari – ikki alifboda ham ishlay oladigan universal tizimlar yaratish.

4. Tilning doimiy rivojlanishi va yangi so‘zlarning paydo bo‘lishi – tizimni doimiy yangilab turish mexanizmlarini ishlab chiqish.

Ushbu muammolar va yechimlarni hisobga olgan holda, o‘zbek tili korpusi matnlarida imlo tuzatish tizimlarini takomillashtirish uchun tizimli yondashuv talab etiladi. Bu yondashuv tilshunoslik, kompyuter dasturlash, statistika va mashina o‘rganishi sohalarini birlashtirgan holda amalga oshirilishi lozim.

XULOSA. Til korpusi matnlarida imlo tuzatish tizimlari zamonaviy kompyuter lingvistikasi va tabiiy tilni qayta ishlash sohalarining muhim qismi hisoblanadi. Ushbu tadqiqot natijasida quyidagi xulosalar shakllantirildi:

birinchidan, o‘zbek tili korpuslarida imlo tuzatish tizimlari hozirgi kunda rivojlanish bosqichida bo‘lib, qoidaga asoslangan, statistik va mashina o‘rganishiga asoslangan usullarning kombinatsiyasi eng samarali natija beradi;

ikkinchidan, resurslari cheklangan tillar, xususan o‘zbek tili uchun imlo tuzatish tizimlarini yaratishda tilning o‘ziga xos xususiyatlarini hisobga olish, morfologik tahlil va kontekstual ma’lumotlardan foydalanish lozim;

uchinchidan, zamonaviy transformer-asosli modellarni o‘zbek tiliga moslashtirish va imlo tuzatish tizimlarida qo‘llash samarali natija beradi;

to‘rtinchidan, imlo tuzatish tizimlarining samaradorligini oshirish uchun annotatsiyalangan ma’lumotlar bazasini kengaytirish zarur.

Ushbu tadqiqot natijalari o‘zbek tili korpusi matnlarida imlo tuzatish tizimlarini takomillashtirish uchun asos bo‘lib xizmat qiladi va bu sohada yangi tadqiqotlar uchun yo‘nalishlar belgilab beradi.

Kelajakda bu sohada qilinadigan tadqiqotlar o‘zbek tili korpuslarining sifatini yanada oshirish, mashina tarjimasi, avtomatik annotatsiyalash va boshqa kompyuter lingvistikasi sohalari uchun muhim ahamiyatga ega bo‘ladi.

Foydalanilgan adabiyotlar:

1. Elov B.B., Yuldashev A.U., Yodgorov U.S. Til korpusi turlari va umumiy xususiyatlari //Xorazm Ma’mun akademiyasi, – 2024 – №12 . B.118-124.
2. Elov B., Samadboyeva M. Language corpora and their importance// Xorijiy lingvistika va lingvodidaktika. – 2025.
3. Kukich Karen. Techniques for automatically correcting words in text // ACM Computing Surveys. – 2019. – T. 24. – № 4. – P. 377-439.
4. Bekmurodov A. va Rixsiyeva G. O‘zbek tilida imlo xatolarini avtomatik aniqlash va tuzatish muammolari // O‘zbek tilshunosligi masalalari. – 2020. – T. 4. – № 2. – B. 67-82.
5. Flor, Michael va Futagi Yoko. Approaches to Automated Spelling Correction: A Comparative Study // Natural Language Engineering. – 2021. – T. 27. – № 1. – P. 1-33.
6. Petrov S., Smirnov I va Ivanov V. Transformer-based spelling correction for Russian text corpora // Computational Linguistics and Intellectual Technologies. – 2023. – T. 22. – № 1. – P. 123-138.
7. Jo‘rayev J. O‘zbek tili korpusini yaratish: Muammolar halqasi va yechimlar. // Kompyuter lingvistikasi:Muammolar,yechim,istiqbollar – 2021.– №. 01. – B. 23-26.