

N-GRAM MODELI VA TAHRIRLASH MASOFASI ALGORITMLARIGA ASOSLANGAN IMLO TUZATISH TIZIMI

Ahmedova Maftuna Kaxramon qizi,
II bosqich magistrant
maftunaahmedova1997@gmail.com
ToshDO‘TAU

Annotatsiya. Tabiiy tilni qayta ishlash (NLP) zamonaviy kompyuter fanlari va lingvistika sohasi kesishmasida jadal rivojlanayotgan yo‘nalishlaridan biri hisoblanadi. Ushbu sohada matnlarni avtomatik tahlil qilish, imloviy xatolarni aniqlash va tuzatish masalasi katta amaliy ahamiyatga ega. Ushbu maqolada n-gram til modeli va Damerau–Levenshteyin masofasiga asoslangan gibril yondashuv yordamida kontekstual imlo tuzatish tizimi ishlab chiqish masalasi ko‘rib chiqiladi. Taklif etilayotgan gibril yondashuv boshqa faqat masofaga asoslangan usullarga nisbatan yuqoriroq aniqlik ko‘rsatdi.

Kalit so‘zlar: *imlo tuzatish, N-gram, tahrir masofasi, Levenshteyin masofasi, korpus.*

Abstract. Natural Language Processing (NLP) is one of the rapidly developing fields at the intersection of modern computer science and linguistics. In this field, automatic text analysis, particularly the detection and correction of spelling errors, is of great practical importance. This article examines the development of a context-based spelling correction system using a hybrid approach based on the N-gram language model and the Damerau–Levenshtein distance. The proposed hybrid approach demonstrates higher accuracy compared to methods based solely on edit distance.

Keywords: *spell correction, n-gram, edit distance, Levenshtein distance, corpus.*

Kirish. Tabiiy tilni qayta ishlash sohalari, sun‘iy intellekt va kognitiv lingvistika sohalari so‘ngi yillarda eng tez rivojlanayotgan yo‘nalishlardan biriga aylanib bormoqda. NLP texnologiyalarining asosiy maqsadi inson tilini kompyuter

tomonidan tushunish, tahlil qilish va qayta ishlash imkonini yaratishdan iborat bo'lib, bu jarayonda matnlarni avtomatik tahrirlash va to'g'rilash muhim o'rin egallaydi. Ayniqsa, matnlardagi imlo xatolarini aniqlash va tuzatish masalasi amaliy jihatdan katta ahamiyatga ega bo'lib, u turli intellektual tizimlar samaradorligini bevosita belgilab beradi.

Hozirgi kunda matn muharrirlari, qidiruv tizimlari, mashina tarjimas platformalari va elektron kommunikatsiya vositalarida avtomatik imlo tekshiruv tizimlari keng qo'llanilib kelinmoqda. Bunday tizimlar juda qulay bo'lib, foydalanuvchilarning xatosiz matn yozishda, shu bilan birga, katta hajmdagi matnlarni avtomatik qayta ishlash jarayonida aniqlikni oshirishga yordam beradi.

Imlo xatolarini avtomatik aniqlash va tuzatish muammosi uzoq yillardan buyon ilmiy tadqiqotlarning markazida bo'lib kelmoqda. Ushbu muammo murakkabligi bilan ajralib turadi, chunki u nafaqat so'zning to'g'ri yozilishini aniqlash, balki uning kontekst ichidagi to'g'ri qo'llanilishini ham hisobga olishni talab qiladi. An'anaviy yondashuvlarda asosan faqat so'z darajasidagi tekshiruvga asoslangan bo'lib, ular kontekstual bog'lanishlarni yetarli darajada hisobga olmaydi. Natijada, bir xil yozilishdagi so'zlar turli kontekstlarda turlicha ma'no anglatishi mumkinligi sababli, noto'g'ri tanlov ehtimoli saqlanib qoladi.

Hozirda axborot tizimlari va Internetning turli sohalarga tobora keng qo'llanilishi natijasida katta hajmdagi ma'lumotlar hosil bo'lmoqda. Shuning uchun O'zbek kontekstida ma'lumotlarni avtomatik qayta ishlash ehtiyojga aylanib bormoqda.[1:60] Shu sababli zamonaviy tadqiqotlarda kontekstga asoslangan imlo tuzatish yondashuvlariga alohida e'tibor qaratiladi. Bunday yondashuvlarda statistik til modellari va lingvistik usullar birgalikda qo'llanilib, matn ichidagi so'zlarning o'zaro bog'liqligi hisobga olinadi. N-gram til modellari so'zlar ketma-ketligi ehtimolligini hisoblash orqali kontekstni modellashtirish imkonini beradi, Levenshteyn va uning kengaytirilgan ko'rinishi bo'lgan Damerau–Levenshteyn

masofasi esa xato soʻz bilan toʻgʻri soʻz oʻrtasidagi oʻxshashlikni aniqlashda samarali vosita hisoblanadi.

Shu nuqtayi nazardan, ushbu maqolada N-gram modeli va Damerau–Levenshteyn algoritmini birlashtirgan gibridd yondashuv asosida kontekstual imlo tuzatish tizimini ishlab chiqish masalasi koʻrib chiqiladi. Taklif etilgan yondashuvning asosiy gʻoyasi shundan iboratki, avvalo xato soʻz uchun ehtimoliy tuzatish variantlari generatsiya qilinadi, soʻngra ular kontekstual ehtimollik asosida saralanadi va eng mos variant tanlab olinadi.

Ushbu tadqiqotning dolzarbligi shundaki, u nafaqat klassik masofaviy algoritmlarga, balki statistik til modellariga ham tayanib, imlo tuzatish jarayonining aniqligini oshirishga qaratilgan. Ayniqsa, morfologik jihatdan murakkab tillarda masalan, oʻzbek tilida kontekstni hisobga olish muhim ahamiyatga ega. Bu yondashuv morfologik jihatdan murakkab boʻlgan tillar uchun samarali yechimlardan biri sifatida qaraladi.

Imlo tuzatish jarayoni notoʻgʻri yozilgan soʻzga eng yaqin boʻlgan soʻzlarni taklif etish jarayonidan iboratdir. Ushbu muammoning yechimini izlash uzoq yillardan beri dolzarb ilmiy muammo boʻlib kelmoqda. Koʻplab tadqiqotchilar bu muammoni oʻrganib chiqqan va ularning saʼy-harakatlari natijasida turli xil usullar hamda koʻplab algoritmlar ishlab chiqilgan. Ular xatolarni ikki asosiy turga ajratadilar.[3:377-439]

Leksik xatolar: ular soʻzlarning oʻziga nisbatan yuzaga keladigan xatolar natijasida paydo boʻladi, yaʼni soʻzlarning gap ichidagi vazifasi va oʻrnidan qatʼi nazar, faqat soʻzning oʻzi notoʻgʻri yozilgan yoki notoʻgʻri ishlatilgan holatlarni anglatadi[7:121-131].

Grammatik xatolar: morfosintaktik xatolarni anglatadi. Ular grammatik imlo xatolari hamda sintaktik xatolarni oʻz ichiga oladi. Bu xatolar soʻzlarning oʻzaro birikishi yoki ularning grammatik oʻzgarishi bilan bogʻliq boʻlib, feʼllarning

tuslanishi yoki otlarning kelishik qo'shimchalari kabi shakllar bilan bog'liq holatlarni qamrab oladi.

Xatolar kelib chiqish sabablari juda ko'p bo'lib, ular kognitiv va tipografik xatolar o'rtasida farqlanadi[5:676-687]. Kognitiv xatolar so'zning to'g'ri yozilishni bilmaslik natijasida yuzaga keladi. Biroq tipografik xatolar klaviatura tugmalarining noto'g'ri tartibda bosilishi oqibatida paydo bo'ladi.

Imlo tekshiruvchi (spell checker) tizimi mohiyatan ikki asosiy bosqichda ishlaydi: xatolarni aniqlash va ularni tuzatish. Xatolarni aniqlash va tuzatish usullari xato turlariga (leksik va grammatik) qarab farqlanadi va ilmiy adabiyotlarda ular odatda alohida ko'rib chiqiladi. Ushbu maqolada esa asosan leksik imlo xatolarini aniqlash va tuzatish uchun qo'llaniladigan yondashuvlar tahlil qilinadi.

Mumkin bo'lgan tuzatish variantlarini taklif etish uchun imlo tekshiruvchi tizim avvalo noto'g'ri yozilgan so'zlarni aniqlashi lozim. Bu esa xato modelini ishlab chiqishni talab etadi. Ushbu yo'nalishda ko'plab tadqiqotchilar tomonidan izlanishlar olib borilgan bo'lib, matndagi xatoli so'zlarni aniqlashda qo'llaniladigan asosiy usullar odatda ikki yondashuvga asoslanadi: lug'atga asoslangan tekshiruv va n-gram tahlili. Ushbu usullar imlo xatolarini aniqlash jarayonida turli yondashuvlarni taqdim etadi va har biri o'zining afzallik hamda cheklovlariga ega.

Imlo tuzatishda eng keng qo'llaniladigan usul lug'atga (to'g'ri so'zlar bazasiga) murojaat qilishga asoslanadi. Ushbu lug'at tilning barcha so'zlarini yoki hech bo'lmaganda ularning asosiy shakllarini o'z ichiga oladi. Bu usul matndagi so'zlarni lug'at bilan solishtirish orqali amalga oshiriladi va lug'atda mavjud bo'lmagan so'zlar dastur tomonida noto'g'ri deb belgilanadi.

Biroq, tilning barcha so'z shakllarini lug'atda saqlash qidiruv jarayonini murakkablashtirishi va sekinlashtirishi mumkin. Shu sababli, odatda lug'atda faqat asosiy (lemma) shakllar saqlanadi, qolgan shakllarni hosil qilish uchun esa morfologik tahlil vositalaridan foydalaniladi. Imlo tuzatishda morfologik tahlil

uchun eng ko'p qo'llaniladigan vositalar lemmatizatorlar, tuslovchilar va turlovchilardir. Ushbu jarayonlar odatda avtomatlar yordamida amalga oshiriladi.

N-gram tahlili

N-gram modelining tilni qayta ishlashda qo'llanilishi birinchi marta Shannon tomonidan tadqiq etilgan. Ushbu yondashuvdan so'ng N-gram modeli imlo tuzatish kabi ko'plab muammolarga yechim sifatida qo'llanila boshlandi. N-gram modelining asosiy afzalliklaridan biri uning tilga bog'liq emasligida.

N-gram tahlili matndagi noto'g'ri so'zlarni aniqlaydigan xatolarni aniqlash usulidir. U matndagi so'zlarni lug'at bilan solishtirmaydi, balki oldindan o'rgatilgan va eng ko'p uchraydigan n-gramlar saqlanadigan kvadrat matritsa asosida taqqoslashni amalga oshiradi. Har bir matn ketma-ketligi qo'shni n-gramlarga ajratiladi.[6: 358-368]

N-gram asosida tizim ishlab chiqishda avvalo har bir n-gramning xatolik ehtimolini alohida hisoblash talab etiladi. Bu ehtimollik n-gramning noto'g'ri so'zlarda uchrash sonining umumiy uchrash soniga nisbati orqali aniqlanadi. Shuningdek, bir xil n-gram bir kontekstda to'g'ri, boshqa kontekstda esa noto'g'ri bo'lishi mumkin. Shu sababli ushbu yondashuv o'qitish korpusini talab qiladi. Faqat noto'g'ri so'zlarda uchraydigan n-gramlar xato sifatida aniqlanadi. Agar n-gram ham to'g'ri, ham noto'g'ri shakllarda uchrasa, u xatoga hissa qo'shmaydi deb hisoblanadi.

N-gram xatolik ehtimoli quyidagi formula orqali hisoblanadi:

$$P = \frac{E}{E + V}$$

Bu yerda:

- **E** – n-gramning noto'g'ri deb belgilangan holatlarda uchrash soni
- **V** – n-gramning to'g'ri deb belgilangan holatlarda uchrash soni

Imlo tuzatish jarayonida xatolar aniqlangandan so'ng navbatdagi bosqich noto'g'ri yozilgan so'zning to'g'ri shakliga yaqin bo'lishi mumkin bo'lgan

variantlarni hosil qilishdan iborat bo'ladi. Ushbu bosqichda qo'llaniladigan tuzatish usullari asosan so'zning o'ziga qaratilgan bo'lib, ularning ishlatilish konteksti inobatga olinmaydi. Tuzatish jarayoni ikki asosiy qadamdan tashkil topadi: birinchi bosqichda xato so'z uchun ehtimoliy to'g'ri yozilish variantlari shakllantiriladi, ikkinchi bosqichda esa ushbu variantlar baholanib, eng mos keluvchi variant tanlab olinadi.

Tahrir masofasi bir so'zni boshqa so'zga aylantirish uchun zarur bo'lgan minimal amallar sonini hisoblaydi. Bu amallar harflarni qo'shish, o'chirish, almashtirish va o'rin almashtirish (transpozitsiya)dan iborat bo'lib, ikki so'z orasidagi farqlar soni ortgan sari ularning masofasi ham ortadi.[4:1091-1095]

Eng mashhur “tahrirlash masofasi” (edit distance) sifatida ham tanilgan Levenshteyin masofasi uchta asosiy amalni hisobga oladi: almashtirish, qo'shish va o'chirish. Faraz qilaylik, $lev(i, j)$ – bu W_1 so'zining birinchi i ta harfi va W_2 so'zining birinchi j ta harfi asosida hisoblangan Levenshteyin masofasi. Uzunligi mos ravishda $|W_1|$ va $|W_2|$ bo'lgan ikkita so'z orasidagi Levenshteyin masofasi quyidagi rekursiv formula orqali aniqlanadi:

$$lev(i, j) = \min \begin{cases} lev(i-1, j) + 1 \\ lev(i, j-1) + 1 \\ lev(i-1, j-1) + I(w_{1i} \neq w_{2j}) \end{cases}$$

bu yerda $I(w_{1i} \neq w_{2j})$ – agar belgilar turlicha bo'lsa 1, aks holda 0 qiymatni oladi.

Levenshteyin masofasini rekursiya orqali hisoblash to'g'ridan-to'g'ri eksponensial vaqt talab qiladi, ammo dinamik dasturlash algoritmi [lev] matritsasini qatorma-qator hisoblash orqali uni $O(|W_1| \cdot |W_2|)$ vaqt ichida hisoblab chiqishi mumkin. Bu yerda $lev_{i,j}$ qiymati $W_1(w_{11}, \dots, w_{1i})$ va $W_2(w_{21}, \dots, w_{2j})$ ketma-ketliklari orasidagi Levenshteyin masofasini bildiradi. Shunday qilib, ushbu ikki ketma-ketlik orasidagi to'liq Levenshteyin masofasini olish uchun $lev_{|W_1|, |W_2|}$ ni

hisoblash zarur. Bu matritsa rekursiyani yuqorida keltirilganidek qo'llash imkonini beradi, lekin qiymatlarni bir necha marta hisoblashga hojat qolmaydi.

Tahrirlash masofasining turli variantlari satrlar ustida turlicha amallarni qo'llash imkonini beradi. Shulardan biri – Damerau–Levenshteyin masofasi – klassik Levenshteyin masofasiga qo'shimcha kiritadi. U qo'shish, o'chirish va almashtirish amallaridan tashqari, ikki qo'shni belgining o'rin almashishi amalini ham hisobga oladi. Bu yondashuvni birinchi bo'lib, Damerau [2:171-176] taklif qilgan bo'lib, u satrlarni tahrirlash jarayonida yanada realistik va amaliy natijalar olish imkonini beradi.

Damerau–Levenshteyin masofasi

Levenshteyin masofasiga o'xshash holda, Damerau – Levenshteyin masofasi ham rekursiv tarzda hisoblanishi mumkin. Faraz qilaylik, $dl(i, j)$ – bu W_1 so'zining birinchi i ta harfi va W_2 so'zining birinchi j ta harfi asosida hisoblangan Damerau–Levenshtein masofasi bo'lsin. U holda:

$$dl(i, j) = \min \begin{cases} dl(i-1, j) + 1 \\ dl(i, j-1) + 1 \\ dl(i-1, j-1) + I(w_{1i} \neq w_{2j}) \\ dl(i-2, j-2) + 1 (\text{transpozitsiya sharti}) \end{cases}$$

agar quyidagi shartlar bajarilsa:

- $i, j > 1$
- $w_{1i} = w_{2(j-1)}$
- $w_{1(i-1)} = w_{2j}$

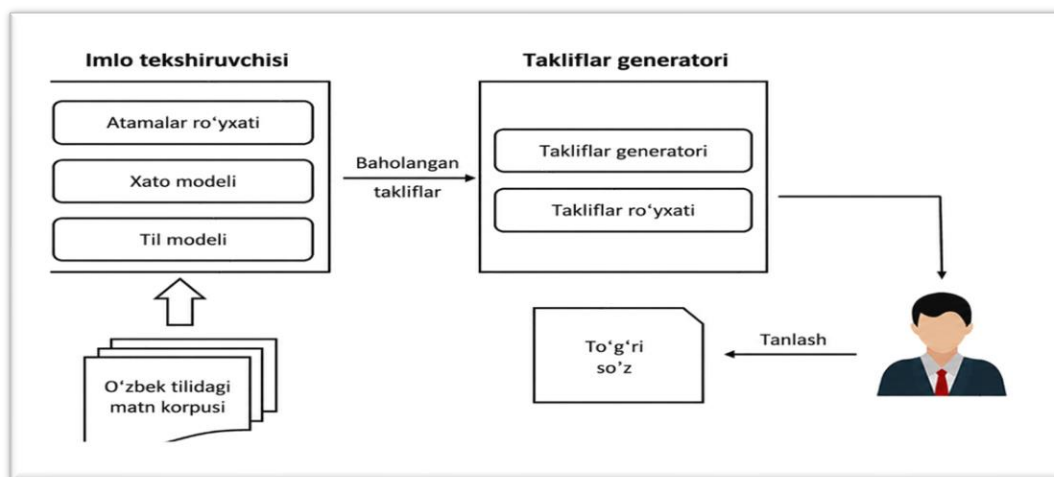
ya'ni, yonma-yon belgilar o'rin almashgan holatda transpozitsiya amali ham hisobga olinadi. Levenshteyin masofasi va uning kengaytirilgan variantlari imlo tuzatish tizimlarida xato so'zga eng yaqin variantlarni aniqlash va turli algoritmlarni ishlab chiqishda muhim ahamiyatga ega.

Taklif etilayotgan tizim

O'zbek tilida imlo tuzatuvchi tizimni yaratish jarayonida dastlabki bosqichda alohida so'zlardagi imlo xatolarini aniqlash va tuzatishga e'tibor qaratildi. Shu

asosda N-gram modeliga asoslangan tizim taklif etildi. Ushbu tizim samaradorligini yanada oshirish maqsadida Damerau–Levenshteyin masofasi algoritmi qo‘shildi.

Taklif etilgan imlo tuzatish tizimi bir nechta bosqichda ishlaydi: xatolarni aniqlash, mumkin bo‘lgan tuzatish variantlarini tanlash, ushbu variantlarni filtrlash va foydalanuvchiga tuzatishlar ro‘yxati ko‘rsatiladi va u mos variantni tanlashi mumkin.



1-rasm. Imlo tuzatish tizimi

Tizim Damerau–Levenshteyin masofasi va N-gram modellarining kombinatsiyasiga asoslangan.

Tizimning leksik bazasi va korpus manbalari

To‘g‘ri so‘zlar lug‘ati bizning yondashuvimizning markaziy qismi hisoblanadi, chunki u imlo tuzatuvchi tizim uchun doimiy ravishda saqlanadigan va boyitib boriladigan asosiy resurs vazifasini bajaradi. Shu sababli, biz imlo tekshiruvchi lug‘atini ishga tushirish uchun turli matnli resurslarni to‘pladik. Foydalanilgan korpus turli mavzularni qamrab olgan. Badiiy, rasmiy, ilmiy kabi turli janrlardagi materiallardan yig‘ilgan. Ushbu matnlar imlo qoidalariga muvofiq mutaxassislar tomonidan tekshirilgan va lingvistik jihatdan tahrir qilingan.

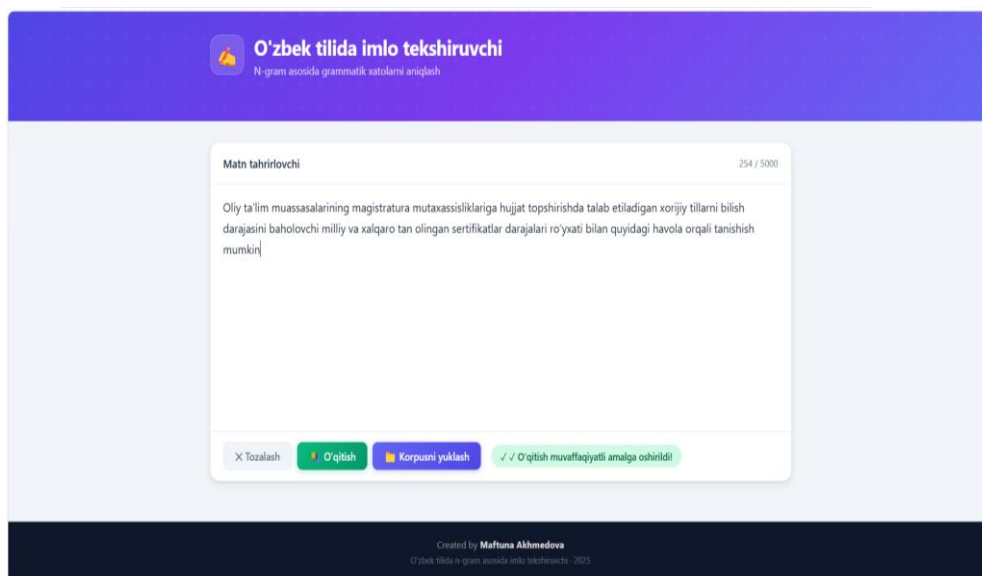
Tizim interfeysi

Taklif etilgan tizim avvalo matnni tokenlarga ajratadi. Har bir token uchun imlo xatolarini aniqlovchi modul ishlatiladi va u tokenning lug‘atda mavjud yoki

mavjud emasligini tekshiradi. Tuzatish moduli esa noto'g'ri deb topilgan so'z uchun ehtimoliy to'g'ri variantlar ro'yxatini qaytaradi. Agar ushbu ro'yxat bo'sh bo'lmasa, u tahrirlash masofasi (edit distance) va korpusdagi chastota asosida tartiblanadi. So'ngra nomzodlar ro'yxati foydalanuvchiga taqdim etiladi. Foydalanuvchi:

- to'g'ri variantni tanlashi,
- agar to'g'ri variant bo'lmasa, o'zi kiritishi,
- yoki so'z to'g'ri ekanini belgilashi mumkin.

Oxirgi holatda tizim so'z lug'atda yo'qligi sababli uni tanimagan bo'ladi. Foydalanuvchi tasdiqlagan to'g'ri so'z natija matniga kiritiladi va jarayon matn oxirigacha davom etadi. Agar so'z tanilmasa, foydalanuvchi uni shaxsiy lug'atga qo'shishi mumkin. Shu tariqa imlo tekshiruvchi tizim ushbu so'zni eslab qoladi va keyingi safar uni hisobga oladi.



2-rasm. Tizim interfeysi

Xulosa. Ushbu maqolada tabiiy tilni qayta ishlash sohasida imlo xatolarini aniqlash va tuzatish muammosi o'rganildi hamda n-gram til modeli va Damerau–Levenshteyin masofasiga asoslangan gibrid yondashuv taklif etildi. Tadqiqot davomida leksik xatolarni aniqlash va tuzatish jarayonida statistik til modellari

hamda tahrirlash masofasi algoritmlarning birgalikda qo'llanilishi tizim samaradorligini sezilarli darajada oshirishi aniqlandi.

Taklif etilgan yondashuvda avvalo xato so'zlar uchun ehtimoliy tuzatish variantlari generatsiya qilinadi, so'ngra n-gram modeli yordamida kontekstual moslik asosida eng to'g'ri variant tanlanadi. Bu jarayon an'anaviy faqat tahrirlash masofasiga asoslangan usullarga nisbatan yuqoriroq aniqlik va barqarorlikni ta'minlashi bilan ajralib turadi. Shuningdek, tadqiqot natijalari shuni ko'rsatadiki, gibrid yondashuv morfologik jihatdan murakkab tillarda imlo tuzatish muammosini hal etishda samarali yechim bo'lib xizmat qiladi. Ayniqsa, kontekstni hisobga olish imkoniyati noto'g'ri tanlovlarni kamaytirishga va tuzatish sifatini oshirishga yordam beradi. Kelgusida ushbu tizimni yanada takomillashtirish maqsadida morfo-sintaktik va semantik tahlil elementlarini qo'shish, shuningdek, so'z darajasidan gap darajasidagi imlo va grammatik xatolarni tuzatishga o'tish rejalashtirilmoqda. Bu esa tizimning amaliy qo'llanish doirasini kengaytirib, uning aniqligini yanada oshirish imkonini beradi.

Foydalanilgan adabiyotlar ro'yxati

1. Elov B N-gramm til modellari vositasida o'zbek tilida matn generatsiya qilish // *Kompyuter lingvistikasi: muammolar, yechim, istiqbollar*: xalqaro ilmiy-amaliy konferensiya materiallari. – 2022.
2. Damerau, F. J. A technique for computer detection and correction of spelling errors // *Communications of the ACM*. – 1964.
3. Kukich, K. Techniques for automatically correcting words in text // *ACM Computing Surveys (CSUR)*. – 1992.
4. Li, Y., Bo, L. A normalized Levenshtein distance metric // *IEEE Transactions on Pattern Analysis and Machine Intelligence*. – 2007.
5. Peterson, J. L. Computer programs for detecting and correcting spelling errors // *Communications of the ACM*. – 1980.



6. Pollock J. J., Zamora A. Automatic spelling correction in scientific and scholarly text // *Communications of the ACM*. – 1984.
7. Ten Hacken, P., Tschichold, C. Word manager and CALL: structured access to the lexicon as a tool for enriching learners' vocabulary // *ReCALL*. – 2001.