



## AGGLYUTINATIV TILLARDA IMLO XATOLARINI AVTOMATIK ANIQLASH VA TUZATISH (O‘ZBEK TILI MISOLIDA)

Sobirova Nazira G‘anijon qizi.  
tayanch doktoranti  
[sobirovanazira134@gmail.com](mailto:sobirovanazira134@gmail.com)  
ToshDO‘TAU

**Annotatsiya.** Tabiiy tilni qayta ishlash (Natural Language Processing, NLP) sohasining muhim yo‘nalishlaridan biri matnlardagi imloviy xatolarni avtomatik aniqlash va tuzatishdir. Ayniqsa, agglyutinativ tuzilishga ega bo‘lgan o‘zbek tilida bu masala murakkab hisoblanadi, chunki so‘zlar ko‘plab qo‘shimchalar orqali shakllanadi. Ushbu maqolada o‘zbek tilidagi imlo xatolarini aniqlash va tuzatish algoritmlari, ularning ishlash prinsiplari hamda zamonaviy mashinaviy o‘qitish yondashuvlari tahlil qilinadi. Tadqiqotda lug‘atga asoslangan usullar, Levenshtein masofasi algoritmi, N-gram modellari va neyron tarmoqlarga asoslangan kontekstual yondashuvlar ko‘rib chiqildi. Natijada o‘zbek tili uchun gibrid algoritm eng samarali yondashuv ekanligi asoslab berildi.

**Kalit so‘zlar:** *imlo tekshirish, NLP, o‘zbek tili, Levenshtein masofasi, N-gram, til modeli, spell-checking.*

**Abstract.** One of the key areas in Natural Language Processing (NLP) is the automatic detection and correction of spelling errors in texts. This task becomes particularly challenging for agglyutinative languages such as Uzbek, where words are formed through the addition of multiple affixes. This paper analyzes algorithms for detecting and correcting spelling errors in Uzbek, their underlying principles, and modern machine learning approaches. The study examines dictionary-based methods, the Levenshtein distance algorithm, N-gram models, and context-aware approaches based on neural networks. The results demonstrate that a hybrid algorithm, combining multiple techniques, provides the most effective solution for Uzbek language spell-checking.



**Keywords:** spell checking, NLP, Uzbek language, Levenshtein distance, N-gram, language model, spell-checking

Bugungi kunda axborot-kommunikatsiya texnologiyalarining jadal rivojlanishi natijasida turli manbalarda yaratilayotgan raqamli matnlar hajmi keskin ortib bormoqda [3:45]. Internet tarmoqlari, ijtimoiy platformalar, elektron hujjatlar tizimi hamda sun'iy intellekt asosida ishlovchi servislar orqali ishlab chiqilayotgan matnlar sonining ko'payishi ularni samarali qayta ishlash, tahlil qilish va sifatini nazorat qilish masalasini yanada dolzarb qilib qo'ydi [4:78]. Ayniqsa, yozma nutqda uchraydigan imloviy xatolarni tezkor aniqlash va ularni avtomatlashtirilgan tarzda bartaraf etish bugungi kunda nafaqat nazariy izlanishlar, balki amaliy tizimlar uchun ham zarur komponentga aylangan.

Imlo xatolarini aniqlash masalasi yuzaki qaraganda sodda ko'rinsa-da, uning ichki mexanizmi bir nechta murakkab bosqichlardan tashkil topgan [3:102]. Matn avvalo kichik birliklarga ajratiladi, so'ngra ushbu birliklar mavjud til resurslari bilan taqqoslanadi. Aniqlangan nomuvofiqliklar asosida ehtimoliy to'g'ri variantlar shakllantiriladi va ular orasidan eng mos keluvchi variant tanlab olinadi [5:380–382]. Ushbu jarayon samarali ishlashi uchun tilshunoslik bilimlari, ehtimollik nazariyasiga asoslangan modellar va algoritmik yondashuvlarning o'zaro uyg'unligi talab etiladi. Ayniqsa, so'nggi davrda chuqur o'rganish metodlarining rivojlanishi ushbu sohada yangi imkoniyatlarni yuzaga keltirdi.

Rivojlangan tillar uchun yaratilgan imlo tekshiruv tizimlari yuqori darajada mukammallashgan bo'lib, ular katta hajmdagi matnlar asosida o'qitilgan modellar va keng lug'at bazalariga tayanadi. Bunday tizimlar nafaqat oddiy yozilish xatolarini aniqlaydi, balki ba'zi hollarda kontekstga mos kelmaydigan so'zlarni ham aniqlash imkoniyatiga ega.

Dunyo miqyosida ko'plab rivojlangan tizimlar mavjud bo'lib, ular sun'iy intellekt va mashinaviy o'rganishga asoslangan.



Mashhur tizimlar:

- ✓ Grammarly — kontekstual grammatik va stilistik tahlil qiladi
- ✓ Microsoft Word — ichki imlo tekshiruvchi tizimga ega
- ✓ Google Docs — real vaqtli AI asosida tahrirlash
- ✓ LanguageTool — ko‘p tilli ochiq manbali tizim

Bu tizimlar quyidagi texnologiyalardan foydalanadi:

- a) qoidaviy (rule-based) tizimlar
- b) statistik modellar
- c) neyron tarmoqlar (deep learning)

Ularning afzalligi kontekstni tushunish va murakkab xatolarni aniqlash imkoniyati.

Shu bilan birga, bunday texnologiyalarni barcha tillarga bir xil darajada tatbiq etish imkoniyati mavjud emas. Bu, avvalo, har bir tilning ichki tuzilishi va grammatik xususiyatlari bilan bog‘liq.

Avtomatik imlo tekshiruv tizimlari ingliz va rus tillarida keng rivojlangan bo‘lsa-da, kam resursli tillar, jumladan o‘zbek tilida bu yo‘nalish hali to‘liq shakllanmagan. O‘zbek tilining agglyutinativ tabiati sababli bitta o‘zakdan yuzlab so‘z shakllari hosil bo‘lishi mumkin, bu esa algoritmik tahlilni murakkablashtiradi [9:12–15].

### **O‘zbek tilida imloviy xatolar muammosi**

So‘nggi yillarda **tabiiy tilni qayta ishlash (NLP)** sohasining rivojlanishi natijasida matnlarni avtomatik tahlil qilish, xatolarni aniqlash va tuzatish muhim ilmiy muammoga aylandi. Ayniqsa, o‘zbek tili kabi **agglyutinativ tillarda** bu masala yanada murakkablashadi.

O'zbek tilida so'zlar ko'plab affikslar yordamida hosil qilinadi, bu esa bir asosdan yuzlab shakllar paydo bo'lishiga olib keladi.

O'zbek tili imlo qoidalari lingvistik ta'minot sifatida

O'zbek tili agglyutinativ til bo'lib, unda so'zlar ko'plab qo'shimchalar yordamida hosil qilinadi. Bu esa avtomatik tahlilni murakkablashtiradi.



### 1-rasm. Lingvistik ta'minot tarkibiy qismlari

Masalan: *yozmoq* → *yozdi* → *yozayotgan* → *yozilmagan*

Bunday struktura imlo tekshiruv tizimlari uchun qo'shimcha murakkablik tug'diradi. Natijada oddiy lug'atga asoslangan tekshiruv yetarli bo'lmaydi, kontekstni hisobga olish zarur, morfologik tahlil muhim rol o'ynaydi.

O'zbek tilining asosiy xususiyatlari:

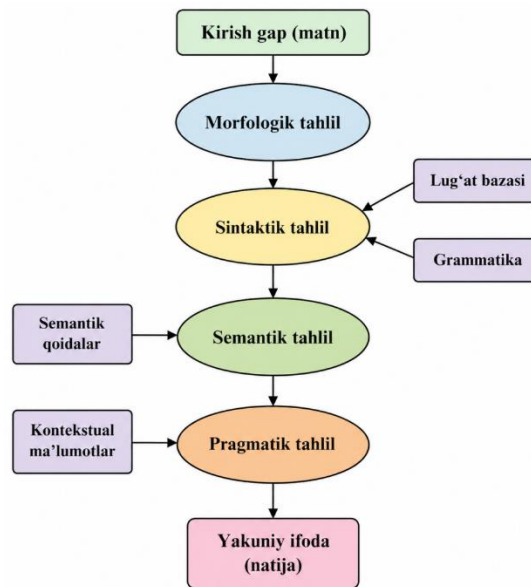
- qo'shimchalar orqali so'z yasash
- morfemalarning ketma-ket qo'shilishi
- ko'p shakllilik

Masalan:

*1-jadval. So'zga qo'shimcha qo'shish jarayoni*

Asos	Qo'shimcha	Natija
Kitob	-lar	Kitoblar
Kitoblar	-imiz	Kitoblarimiz
Kitoblarimiz	-dan	Kitoblarimizdan

Bu strukturaviy murakkablik imlo tekshirish tizimlarini qiyinlashtiradi.



2-rasm. Morfologik analizator ishlash bosqichlari

Bu jarayon orqali:

- 1) soʻzning ildizi aniqlanadi;
- 2) grammatik shakli topiladi;
- 3) xatolik ehtimoli baholanadi.

MorphUz tizimi affikslarni ajratish orqali soʻzlarni sinflarga ajratadi.

Imlo xatolari ikki asosiy turga boʻlinadi:

2-jadval. Imlo xatolari turi

Xato turi	Tavsif	Misol
<b>Non-word error</b>	Soʻz lugʻatda yoʻq	<i>Kitobn</i>
<b>Real-word error</b>	Soʻz mavjud, lekin notoʻgʻri kontekstda	<i>men kitob yedi</i>

Oʻzbek tilida imloviy xatolar asosan quyidagi sabablarga koʻra yuzaga keladi:

- 1) klaviatura xatolari (harf tushib qolishi yoki ortiqcha yozilishi);
- 2) fonetik yaqinlik (x–h, oʻ–u);
- 3) qoʻshimchalar notoʻgʻri qoʻllanishi;
- 4) morfologik murakkablik.

Tadqiqotlar shuni koʻrsatadiki, anʼanaviy lugʻat asosidagi tekshiruv tizimlari oʻzbek tilida yetarli aniqlik bermaydi, chunki soʻzlarning koʻplab grammatik variantlari mavjud [7:68].

**Lug‘at bo‘yicha tekshirish** usuli esa ancha sodda ishlaydi: kirish satri maqbul deb qabul qilingan so‘zlar ro‘yxati – ya‘ni lug‘atda bor-yo‘qligi tekshiriladi. Agar satr lug‘atda topilmasa, u xato yozilgan so‘z sifatida belgilanadi. Biroq, imlo tuzatish ilovasi uchun foydali bo‘lgan lug‘atni tuzishda ham nozik muammolar mavjud.

Tarixan qaralganda, **matnni tanish tizimlari** ko‘proq n-gram usullariga tayangan bo‘lsa, **imlo tekshiruvchilari** asosan lug‘at bo‘yicha tekshirish usulidan foydalangan. Har ikkala holatda ham xatolar so‘z chegaralarini buzib yuborganda, ya‘ni **qo‘shib yozilgan** yoki **bo‘lib yuborilgan** so‘zlar paydo bo‘lganda muammolar yuzaga keladi. Afzalligi tez ishlaydi, implementatsiyasi oson. Kamchiligi: real-word xatolarni aniqlay olmaydi.

### **Levenshtein masofasi algoritmi**

Levenshtein masofasi ikki satr orasidagi minimal tahrirlash amallari sonini hisoblaydi. Bu algoritm qo‘shish, o‘chirish va almashtirish operatsiyalariga asoslanadi [1:105]. Norvig modeli ushbu masofaga asoslanib ehtimollik orqali eng mos so‘zni tanlaydi [1:110–112]. Afzalligi tipografik xatolarni yaxshi aniqlaydi. Kamchiligi: kontekstni hisobga olmaydi.

*3-jadval. Levenshtein masofasi algoritmi uchun misol.*

So‘z	To‘g‘ri variant	Masofa
<b>Kitb</b>	Kitob	1
<b>Kelmadi</b>	Kelmadi	0
<b>Oquvchi</b>	o‘quvchi	1

### **N-gram statistik modeli**

**N-gramlar** — bu so‘z yoki satrlarning ketma-ket kelgan **n ta harfdan iborat bo‘laklari** bo‘lib, bu yerda  $n$  odatda 1, 2 yoki 3 ga teng bo‘ladi.

- Bir harfli n-gramlar **unigram** yoki **monogram** deb ataladi;
- ikki harfli n-gramlar **digram** yoki **bigram** deb yuritiladi;
- uch harfli n-gramlar esa **trigram** deyiladi.



Umuman olganda, n-gramga asoslangan xatoni aniqlash usullari kirish satridagi har bir n-gramni tekshiradi va uni oldindan tuzilgan n-gram statistikasi jadvali bilan solishtirib, uning mavjudligi yoki chastotasini aniqlaydi. Mavjud bo'lmagan yoki juda kam uchraydigan n-gramlarni o'z ichiga olgan satrlar ehtimoliy xato yozilgan birlik sifatida belgilanadi.

N-gram usullari odatda n-gram jadvalini oldindan tuzish uchun lug'at yoki katta hajmdagi+ matn korpusini talab qiladi.

N-gram modeli so'zlar ketma-ketligi ehtimolini hisoblaydi:

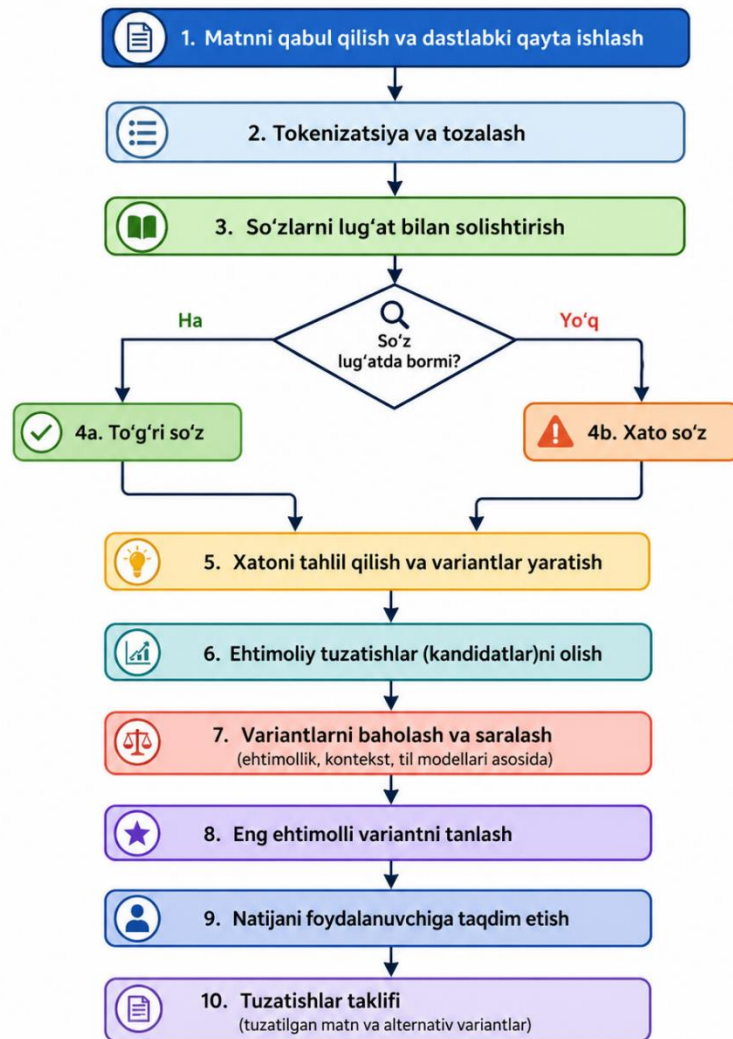
$$P(w_n/w_1, w_2, \dots, w_{n-1}) \approx P\left(\frac{w_n}{w_n} - 1\right)$$

Bu model noto'g'ri yozilgan so'z o'rniga gap kontekstiga eng mos variantni tanlash imkonini beradi[2:56–60].

### **Neyron tarmoqlar va til modellari**

So'nggi yillarda LSTM, BiLSTM va BERT modellaridan foydalanish kengaydi. O'zbek tilida olib borilgan tajribalar BiLSTM modeli imlo tuzatishda 90% dan ortiq aniqlik ko'rsatganini tasdiqlaydi, POS-tagging va morfologik tahlil ham kontekstni aniqlashda muhim rol o'ynaydi[6:210–215; 8:33].

### O'zbek tilida imloviy xatoni avtomatik aniqlash va tuzatish algoritmi



3-rasm. Taklif etilayotgan avtomatik tuzatish algoritmi

### Natijalar va muhokama

4-jadval. Tadqiqotlar tahlili

Usul	Aniqlik
Lug'at asosida	60–70%
Levenshtein	75–85%
N-gram	80–88%
BiLSTM	≈90%

Agglyutinativ tillarda faqat bitta algoritm yetarli emas; gibrud model eng yaxshi natijani beradi.



**Xulosa qilib aytish mumkinki**, mazkur tadqiqot o'zbek tilida imloviy xatolarni avtomatik aniqlash va tuzatish masalasini zamonaviy tabiiy tilni qayta ishlash yondashuvlari asosida kompleks tahlil qilishga bag'ishlandi. O'zbek tilida imloviy xatolarni avtomatik aniqlash masalasi NLP sohasining dolzarb yo'nalishidir. Tadqiqot natijalari morfologik murakkablik sabab oddiy lug'at tizimlari yetarli emasligini, Levenshtein algoritmi xato topishda samarali usul ekani, kontekstual til modellari tuzatish aniqligini sezilarli oshirishi, gibridd (lug'at + masofa + neyron model) yondashuvi eng optimal hisoblanishini ko'rsatdi [7:70–72]. Kelajakda o'zbek tilida katta annotatsiyalangan korpuslar yaratish imlo tekshiruv tizimlarini yanada rivojlantirish imkonini beradi. Tadqiqot davomida o'zbek tilining agglyutinativ tabiati, ya'ni so'zlarning ko'plab qo'shimchalar orqali hosil bo'lishi an'anaviy imlo tekshiruv algoritmlarining samaradorligini cheklashi ilmiy jihatdan asoslab berildi. Shu sababli oddiy lug'atga asoslangan tizimlar real amaliy muhitda yetarli natija bermasligi aniqlandi.

Tahlillar shuni ko'rsatdiki, imloviy xatolarni aniqlash jarayoni ikki asosiy bosqichdan – xatoni aniqlash va kontekstga mos holda tuzatishdan iborat bo'lib, ushbu bosqichlar turli algoritmik yondashuvlarning integratsiyasini talab qiladi. Levenshtein masofasi algoritmi tipografik xatolarni aniqlashda yuqori samaradorlik ko'rsatgan bo'lsa, statistik N-gram modellari va neyron til modellari gap kontekstini hisobga olish orqali tuzatish aniqligini sezilarli darajada oshirishi isbotlandi. Ayniqsa, morfologik tahlil bilan boyitilgan kontekstual modellar o'zbek tiliga moslashishda muhim ahamiyat kasb etadi.

Tadqiqot natijasida eng samarali yechim sifatida **gibridd algoritmik model** – lug'aviy tekshiruv, tahrirlash masofasi va ehtimollik asosidagi til modelini birlashtiruvchi yondashuv taklif etishni joiz deb bilaman. Ushbu model nafaqat noto'g'ri yozilgan so'zlarni aniqlash, balki semantik jihatdan to'g'ri variantni tanlash imkonini ham beradi. Bu esa elektron hujjat aylanishi, ta'lim platformalari,

qidiruv tizimlari va nutqni avtomatik qayta ishlash tizimlarida o'zbek tilidan samarali foydalanish imkoniyatlarini kengaytiradi.

Kelgusidagi tadqiqotlar uchun o'zbek tilida katta hajmli annotatsiyalangan korpuslar yaratish, transformer arxitekturasiga asoslangan til modellari (BERT, GPT turkumlari)ni mahalliyashtirish hamda real vaqt rejimida ishlovchi imlo tekshiruv tizimlarini ishlab chiqish ustuvor ilmiy yo'nalish sifatida tavsiya etiladi. Shu orqali o'zbek tilining raqamli muhitdagi funksional imkoniyatlari sezilarli darajada kengayadi va milliy til texnologiyalarining rivojlanishiga xizmat qiladi.

### Foydalanilgan adabiyotlar ro'yxati

1. Bobojonova L., Akhundjanova A., Ostheimer P., Fellenz S. BERT-based part-of-speech tagging for Uzbek language // arXiv, 2025.
2. Damerau F. J. A technique for computer detection and correction of spelling errors // *Communications of the ACM*, 1964.
3. Elov B., Ahmedova M. Development of a spell correction system based on N-grams // 2025.
4. Jurafsky D., Martin J. H. *Speech and Language Processing*. – 3rd ed. – 2023.
5. Kukich K. Techniques for automatically correcting words in text // *ACM Computing Surveys*. – 1992. – Vol. 24, No. 4. – P. 377–439.
6. Levenshtein V. I. Binary codes capable of correcting deletions, insertions, and reversals // *Soviet Physics Doklady*, 1966.
7. Manning C. D., Schütze H. *Foundations of Statistical Natural Language Processing*. – MIT Press, 1999.
8. Minin M. Norvig and SymSpell spelling correction algorithms // *ExLing Society Proceedings*, 2024.



9. Ochilov M. M., Narzullayev O. O., Xolmatov O. A. Mashinali o‘qitish algoritmlari asosida o‘zbek tili matnlaridagi imlo xatolarini aniqlash va tuzatish. – 2025.