

IMLO TEKSHIRISHDA MORFOANALIZ BOSQICI

Nilufar Abduraxmonova

filologiya fanlar doktori, professor O'zbekiston Milliy universiteti
Kompyuter lingvistikasi va amaliy tilshunoslik kafedrasini mudiri

Kenjayeva Sevinch Bahodir qizi

O'zbekiston Milliy universiteti Jurnalistika va o'zbek filologiyasi fakulteti,
Kompyuter lingvistikasi mutaxassisligi I bosqich magistranti

Annotatsiya. Ushbu maqolada tabiiy tilni qayta ishlash (NLP) tizimlarining ajralmas qismi bo'lgan imlo tekshirish va matnni avtomatik tahrirlash yondashuvlari tizimli tahlil qilinadi. Tadqiqot doirasida qoidaviy, statistik va zamonaviy neyron yondashuvlarning evolyutsion rivojlanishi hamda ularning o'ziga xos afzalliklari o'rganilgan. Ayniqsa, o'zbek tili kabi agglutinativ va kam resursli tillar uchun morfologik analizatorlarning o'rni va Transformer arxitekturasiga asoslangan (BERT, RoBERTa) modellarning samaradorligi qiyosiy tahlil etilgan. Tahlillar natijasida, yuqori aniqlikka erishish uchun qoidaviy morfologik tahlil va chuqur o'qitish modellarini birlashtirgan gibril tizimlardan foydalanish eng istiqbolli yo'nalish ekanligi asoslangan.

Kalit so'zlar: tabiiy tilni qayta ishlash (NLP), imlo tekshirish, qoidaviy yondashuv, statistik modellar, neyron tarmoqlar, o'zbek tili, morfologik analiz, Transformer, BERT.

Abstract. This article examines the main challenges of automatic named entity recognition (NER) in Uzbek language texts. Particular attention is given to the difficulties of identifying and classifying entities such as personal names, geographical locations, organizations, and other types of named units. The agglutinative nature of the Uzbek language, its relatively free word order, and rich morphological structure introduce additional complexity to the tagging process. Furthermore, the study analyzes rule-based, statistical, and neural approaches to named entity recognition, highlighting their advantages and limitations. Key challenges include the scarcity of linguistic resources, the lack of annotated corpora, and the absence of standardized lexicons and ontologies.

The findings emphasize the importance of hybrid approaches and the expansion of national linguistic resources to improve the performance of named entity recognition systems for the Uzbek language.

Keywords: natural language processing (NLP), spelling correction, rule-based approach, statistical models, neural networks, Uzbek language, morphological analysis, Transformer, BERT.

Аннотация. В данной статье проводится системный анализ подходов к проверке орфографии и автоматическому редактированию текста, которые являются неотъемлемой частью систем обработки естественного языка (NLP). В рамках исследования изучено эволюционное развитие правил, статистических и современных нейронных подходов, а также их специфические преимущества. Особое внимание уделено роли морфологических анализаторов для агглютинативных и малоресурсных языков, таких как узбекский, и проведен сравнительный анализ эффективности моделей на базе архитектуры Transformer (BERT, RoBERTa). Результаты анализа подтверждают, что использование гибридных систем, сочетающих правила морфологического анализа и модели глубокого обучения, является наиболее перспективным направлением для достижения высокой точности.

Ключевые слова: обработка естественного языка (NLP), проверка орфографии, правильный подход, статистические модели, нейронные сети, узбекский язык, морфологический анализ, Transformer, BERT.

Axborot texnologiyalarining jadal rivojlanishi va global miqyosda raqamli ma'lumotlar hajmining keskin ortishi tabiiy tillarni kompyuter yordamida qayta ishlash (NLP) sohasini zamonaviy ilm-fanning eng ustuvor yo'nalishlaridan biriga aylantirdi. Matnli ma'lumotlar bilan ishlashda sifat va aniqlik darajasi bevosita imlo tekshirish (spelling correction) va avtomatik tahrirlash tizimlarining samaradorligiga bog'liq. Imlo xatolari nafaqat matnning tushunilishini qiyinlashtiradi, balki qidiruv tizimlari, mashina tarjimai va nutqni aniqlash kabi yuqori bosqichli NLP vazifalarida xatoliklarning ko'payishiga sabab bo'ladi. Imlo tekshirish tizimlarida morfoanaliz bosqichi til birliklarining ichki tuzilishini aniqlash va ularni lingvistik jihatdan to'g'ri talqin qilishda muhim o'rin tutadi. Ushbu bosqichda so'zlarning morfemik tarkibi, ya'ni ildiz va qo'shimchalar ajratilib, ularning grammatik kategoriyalari (so'z turkumi, son, kelishik, zamon va boshqalar) aniqlanadi. Natijada, nafaqat yuzaki yozuv xatolarini, balki morfologik jihatdan noto'g'ri shakllangan birliklarni ham aniqlash imkoniyati kengayadi.

Morfoanaliz, ayniqsa, agglutinatив tillarda, jumladan o'zbek tilida, imlo tekshirish jarayonining muhim bosqichidir. Bunday tillarda so'zlar ko'plab qo'shimchalar orqali yasalishi sababli, oddiy lug'atga asoslangan tekshiruv yetarli bo'lmaydi. Shu bois, morfologik tahlil yordamida so'zlarning to'g'ri yasalgan-yasalmaganligi, qo'shimchalarning mosligi va ularning ketma-ketligi aniqlanadi.

Umuman olganda, morfoanaliz bosqichi imlo tekshirish tizimining aniqligi va ishonchliligini oshiruvchi asosiy mexanizmlardan biri bo'lib, u keyingi bosqichlarda kontekstual va sintaktik tahlilni samarali amalga oshirish uchun zarur lingvistik asosni yaratadi.

O'zbek tili o'zining boy morfologik tuzilishi va agglutinatив tabiati bilan ajralib turadi. Bu xususiyat so'z ildiziga ko'plab affikslarning ketma-ket qo'shilishi orqali bitta so'zdan minglab shakllar hosil bo'lishini anglatadi. Bunday murakkablik an'anaviy lug'atga asoslangan oddiy tizimlar uchun katta qiyinchilik tug'diradi. Dunyo tajribasida imlo tekshirish tizimlari oddiy qoidaviy yondashuvlardan murakkab neyron tarmoqlar va kontekstni tahlil qiluvchi Transformer modellari qadar evolyutsion yo'lni bosib o'tdi. Biroq, o'zbek tili uchun bu texnologiyalarni tatbiq etishda hali ham hal qilinishi lozim bo'lgan qator muammolar mavjud. Ushbu maqolaning maqsadi imlo tekshirish sohasidagi mavjud xalqaro va mahalliy yondashuvlarni tizimli ravishda tahlil qilish, ularning o'zbek tili kabi agglutinatив tillar uchun samaradorligini baholash va kelajakdagi tadqiqotlar uchun istiqbolli yo'nalishlarni belgilashdan iborat. Tadqiqot doirasida qoidaviy morfologik analizatorlarning aniqligi va neyron tarmoqlarning kontekstual moslashuvchanligini birlashtiruvchi gibridd modellarning ahamiyati asoslab beriladi.

Tabiiy tilni qayta ishlashda (NLP) imlo tekshirish va matnni avtomatik tahrirlash tizimlari tilshunoslik va kompyuter fanlari o'rtasidagi murakkab vazifalardan biri bo'lib qolmoqda. Ushbu sohadagi tadqiqotlar evolyutsiyasini shartli ravishda uchta asosiy bosqichga: qoidaviy, statistik va neyron yondashuvlarga ajratish mumkin.

Qoidaviy yondashuvlar imlo tekshirishning ilk fundamental asosi bo'lib, ular qat'iy til qoidalari va leksik lug'atlarga tayanadi. Bunday tizimlar kiritilgan so'zni lug'at bazasidagi birliklar bilan solishtirish va morfologik qoliplar asosida tahlil qilish orqali ishlaydi. O'zbek tili kabi agglutinatив tillar uchun bu yondashuv o'ta murakkab morfologik analizatorlarni

talab etadi. Bu yo'nalishda ushbu mavzuga doir olib borgan tadqiqotlarimiz (2022) o'zbek tili NLP muhitini yaratishda muhim poydevor bo'ldi. Ushbu izlanishlarda tilning agglyutinativ tabiatini hisobga olgan holda, affiksalar birikish qonuniyatlariga asoslangan morfologik analiz modellari va o'zbek tili korpusi uchun zaruriy lingvistik resurslar ishlab chiqilgan. Shu bilan birga, zamonaviy tadqiqotlarimiz natijasida O'zbek tili milliy korpusi (uzbekcorpus.uz) kabi yirik va balanslashgan ma'lumotlar bazasining yaratilishi o'zbek tili NLP tizimlari uchun yangi imkoniyatlar ochmoqda. Ushbu korpus nafaqat statistik ma'lumotlarni yig'ish, balki neyron tarmoq modellari uchun zaruriy kontekstual muhitni ta'minlovchi asosiy resurs sifatida baholanadi. N. Murzintcev va Sh. Yuldasheva (2026) tomonidan Hunspell kutubxonasi asosida ishlab chiqilgan o'zbek tili morfologik tahlilchisi ham aynan qoidaviy va lug'atli yondashuvning zamonaviy amaliy namunasidir. Ushbu tizimlarning ishlash mexanizmini shu yo'nalishda amalga oshirgan tadqiqotlarimiz (2022) natijasida ishlab chiqilgan MorphUz analizatori misolida ko'rish mumkin. Tizim kiritilgan so'zni (masalan, "bolalarga") ildiz va affikslarga ajratishda chekli holatli avtomatlar va leksik bazalardan foydalanadi (1-jadval).

Kiritilgan so'z	So'z o'zagi	So'z turkumi	Morfologik analiz
Bolalarga	Bola	NOUN (Ot)	Stemma: bolalarga -> bola Qo'shimchasi: larga Model: NOUN+Plur+Dat Izohi: lar - Ko'plik son; ga - Jo'nalish kelishigi
Bolalarga	Bolala	VERB (Fe'l)	Stemma: bolalarga -> bolala Qo'shimchasi: rga Model: VERB+FUT+Conv Izohi: a - Kelasi zamon; a - Ravishdosh qo'shimchasi (harakat holati, payti, sababini bildiradi)

1-jadval. Tadqiqotlarimiz doirasida ishlab chiqilgan MorphUz analizatorining tahlil natijalari namunasi (2022)

Tahlillar shuni ko'rsatadiki, qoidaviy tizimlar bitta so'z shakli uchun bir nechta qonuniy variantlarni taqdim etishi mumkin. Masalan, "dasturlash" so'zi ham ot, ham fe'l sifatida tahlil qilinishi yoki egalik qo'shimchalarining turlicha talqin etilishi tizimning aniqlik ko'rsatkichlariga ta'sir etadi (2-jadval). Qoidaviy usullar tizimning shaffofligi va tezkorligi bilan ajralib tursa-da, yangi so'zlarga moslashuvchanlikning pastligi hamda yuqoridagi jadvalda ko'ringanidek, kontekstsiz to'g'ri variantni tanlay olmasligi ularning imkoniyatlarini cheklaydi. Aynan shu nuqtada, morfologik tahlil natijalari ichidan kontekstga mosini tanlash uchun neyron tarmoqli modellarga ehtiyoj tug'iladi.

Kiritilgan so'z	So'z o'zagi (Stem)	So'z turkumi (POS)	Morfologik analiz (Morfemalar va modellari)
Dasturlash	Dastur	NOUN (Ot)	Qo'shimcha: -lash Izoh: NOUN so'z turkumiga oid -lash qo'shimchasi tizimdan topilmadi.
Dasturlash	Dasturla	VERB (Fe'l)	Qo'shimcha: -sh Model 1: VERB+Rcp (Birgalik nisbati) Model 2: VERB+Vnoun (Harakat nomi) Model 3: VERB+Rfl (O'zlik nisbati)

2-jadval. Tadqiqotlarimiz doirasida ishlab chiqilgan MorphUz analizatorining tahlil natijalari namunasi (2022)

Statistik yondashuvlar tilni ehtimollik modeli sifatida tahlil qiladi. C. Manning (1999) tomonidan asos solingan n-gram modellari va Bayes yondashuvlari imlo tekshirishda soʻzlarning matndagi ketma-ketlik ehtimolligini hisoblashga imkon berdi. D. Jurafsky va J. Martin (2024) taʼriflaganidek, imlo xatolarini aniqlashda "noisy channel" (shovqinli kanal) modeli markaziy oʻrinni egallaydi. Bunda xato deb topilgan soʻz uchun eng kam "tahrirlash masofasiga" (Minimum Edit Distance) ega boʻlgan va kontekstda uchrash ehtimoli eng yuqori boʻlgan nomzod variant tanlab olinadi. Statistik metodlar qoidaviy tizimlarga nisbatan yuqori moslashuvchanlikka ega boʻlsa-da, ularning aniqligi oʻqitish uchun foydalaniladigan matnlar korpusining hajmiga bevosita bogʻliq.

Neyron yondashuvlar chuqur oʻqitish (Deep Learning) texnologiyalari orqali imlo tekshirishni yangi sifat bosqichiga olib chiqdi. A. Vaswani va boshqalar (2017) tomonidan taklif etilgan Transformer arxitekturasini va undagi "attention" (diqqat) mexanizmi matnni global kontekstda tahlil qilish imkonini berdi. J. Devlin va hamkorlari (2019) tomonidan ishlab chiqilgan BERT modeli ikki yoʻnalishli tahlil yordamida nafaqat imlo xatolarini, balki kontekstual xatolarni ham yuqori aniqlikda tuzatishga xizmat qilmoqda. Oʻzbek tili kabi kam resursli tillar uchun bu yoʻnalishda G. Yiming va boshqalar (2025) tomonidan taklif etilgan BiLSTM va Attention mexanizmlari asosidagi modellar morfologik segmentatsiyani avtomatlashtirish va ildiz ajratish jarayonini yaxshilash orqali imlo tekshirgichlar samaradorligini oshirmoqda. Shuningdek, S. Kasmaiee (2023) tadqiqotlari FastText embeddinglari va Encoder-Decoder neyron tarmoqlarining gibrid qoʻllanilishi imlo tuzatishda 87% dan yuqori aniqlik berishini isbotladi.

Yondashuvlarning qiyosiy tahlili shuni koʻrsatadiki, qoidaviy tizimlar lingvistik aniqlikni, statistik modellar ehtimollik moslashuvchanligini, neyron modellar esa chuqur kontekstual tahlilni taʼminlaydi. Oʻzbek tili uchun olib borilayotgan zamonaviy tadqiqotlarning asosiy tendensiyasi — bu qoidaviy morfologik analizatorlarni zamonaviy transformer modellari bilan birlashtirgan gibrid tizimlarni yaratishdir. Bu, ayniqsa, oʻzbek tili kabi resursi cheklangan tillar uchun eng samarali yoʻl hisoblanadi.

Tahlillarga koʻra oʻzbek tili uchun imlo xatolarini aniqlash va tuzatish tizimlarining samaradorligini baholash maqsadida tizimli va qiyosiy tahlil metodlaridan foydalanildi. Tadqiqot jarayoni quyidagi bosqichlarni qamrab oldi:

1. Maʼlumotlarni yigʻish va filtrlash: Tadqiqot uchun Scopus, Web of Science va Google Scholar kabi xalqaro ilmiy bazalardan 1999–2026-yillar oraligʻida chop etilgan 30 dan ortiq ilmiy manbalar dastlabki tahlildan oʻtkazildi. Ulardan imlo tekshirishning uchta asosiy yoʻnalishi (qoidaviy, statistik va neyron) boʻyicha eng yuqori iqtiboslik darajasiga ega boʻlgan 10 ta tayanch tadqiqot tanlab olindi.

2. Qiyosiy tahlil mezonlari: Tanlab olingan metodlar va algoritmlarni baholashda quyidagi asosiy koʻrsatkichlar mezon qilib olindi:

a) morfologik qamrov: agglyutinativ tillardagi murakkab affikslar bilan ishlash imkoniyati.

b) kontekstual aniqlik: soʻzning matndagi maʼnosiga qarab xatolarni tuzata olish qobiliyati.

c) resurs talabi: modelni oʻqitish uchun zarur boʻlgan matnlar korpusi.

3. Oʻzbek tili korpusi va resurslarini tahlil qilish: Metodologiyaning yakuniy qismida neyron tarmoqlarni oʻqitishda asosiy manba sifatida oʻzbek tili elektron korpusidan (uzbekcorpus.uz, 2024) foydalanish imkoniyatlari tahlil qilindi va bizning tadqiqotlarimiz

(2022) natijasida yaratilgan mavjud morfologik analizatorlar va N.Murzintcev ishlanmalarining ushbu texnologiyalar bilan integratsiyalashuv imkoniyatlari nazariy jihatdan modellashtirildi.

O'tkazilgan tahlillar shuni ko'rsatadiki, o'zbek tili kabi agglyutinativ va resursi cheklangan tillar uchun imlo xatolarini tekshirish va tuzatish tizimlarini yaratishda faqatgina bir yo'nalishdagi modellardan foydalanish kutilgan yuqori aniqlikni ta'minlamaydi. Tadqiqot natijasida quyidagi muhim xulosalarga kelindi. Qoidaviy tizimlar (MorphUz, Hunspell) o'zbek tili morfologiyasini tahlil qilishda muhim fundamental baza bo'lib xizmat qilsa-da, ularning asosiy cheklovi matndagi kontekstual xatolarni aniqlay olmasligidir. Statistik va neyron tarmoq modellari, xususan, Transformer (BERT) arxitekturasi matnni kontekstda tahlil qilish orqali xatolarni tuzatish sifatini yangi bosqichga olib chiqdi. Biroq, bu modellar o'zbek tili uchun hali ham yetarlicha keng qamrovli va sifatli o'qitish ma'lumotlariga ehtiyoj sezadi.

Bizningcha, eng samarali yechimgibrid yondashuv hisoblanadi. Yuqori aniqlikka erishish uchun qoidaviy morfologik analizatorlarning lingvistik aniqligini neyron tarmoqlarning kontekstual tahlil imkoniyatlari bilan integratsiya qilish zarur. Bunday gibrid tizimda morfologik analizator so'zning barcha variantlarini hosil qilsa, BERT modeli gapdagi ma'nodan kelib chiqib eng mos variantni tanlab beradi.

Kelgusidagi tadqiqotlar O'zbek tili milliy korpusi resurslaridan foydalangan holda, ushbu nazariy gibrid modelni amaliy dasturiy vosita sifatida tatbiq etish va uning samaradorligini turli janrdagi matnlar ustida sinovdan o'tkazishga qaratiladi. Bu esa nafaqat imlo tekshirgichlarni, balki o'zbek tili uchun mashina tarjimai va sun'iy intellekt tizimlarini rivojlantirishda muhim qadam bo'ladi.

FOYDALANILGAN ADABIYOTLAR:

1. Abdurakhmonova, N., Alisher, S., & Mengliev, B. (2022). Uzbek Morphological Analysis and NLP Systems. Tashkent: National University of Uzbekistan.
2. Abdurakhmonova, N., Alisher, I. & Sayfullayeva, R (2022). Morphological Analyzer for the Uzbek Language. International Conference on Computer Science and Engineering (UBMK)
3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT, 4171–4186.
4. Hládek, D., Staš, J., & Pleva, M. (2020). Survey of Automatic Spelling Correction. Electronics, 9(10), 1670.
5. Jurafsky, D., & Martin, J. H. (2024). Speech and Language Processing (3rd ed. draft).
6. Kasmaiee, Sa., Kasmaiee, Si., & Homayounpour, M. (2023). Correcting spelling mistakes in Persian texts with rules and deep learning methods. Scientific Reports, 13, 19457.
7. Manning, C.D., & Schütze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press.
8. Murzintcev, N., & Yuldasheva, S.S. (2026). Uzbek language morphology analyser. Natural Language Processing Journal, 14, 100195.
9. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems (NeurIPS), 5998–6008.

10. Yiming, G., Ni, P., Kong, H., Zhang, X., & Wang, L. (2025). Research on morphological analysis of derived low-resource languages. Proceedings of the 2025 International Conference on Natural Language Processing.

11. <https://uzbekcorpus.uz>