FRAZEOLOGIK BIRLIKLARNI KORPUS UCHUN TEGLASH MASALASIDA ZAMONAVIY YONDASHUVLAR VA TADQIQOTLAR SHARHI

G‘anijonova Shaxlo Dilmurod qizi
Kompyuter lingvistikasi magistratura mutaxassisligi 1- kurs magistranti
O‘zbekiston Milliy universiteti, Toshkent, O‘zbekiston.
E-mail: ganijonovashaxlo2003@gmail.com
Tel: (998) 88 871 07 87
... taqrizi asosida

Annotatsiya:
Mazkur maqola o‘zbek tili milliy korpusida frazeologik birliklarni teglash va ularni avtomatik lingvistik annotatsiya qilishning nazariy-metodik asoslarini ishlab chiqishga bag‘ishlanadi. Frazeologik birliklarning semantik jihatdan yaxlitligi, shuningdek, tarkibining o‘zgaruvchanligi sabab NLP tizimlarida oddiy so‘z birikmalaridan ularni ajratish biroz murakkab. Tadqiqotda tizimli lisoniy tahlil, korpus metodlari va zamonaviy neyron tarmoq modellari imkoniyatlarini qiyoslash metodologiyasidan foydalanish orqali frazeologizmlarning elektron bazasini shakllantirishda xalqaro tajribalar va ularning o‘zbek tili korpusiga tatbiq etilish imkoniyatlari ko‘rib chiqiladi.
Kalit so‘zlar: *O‘zbek tili milliy korpusi, frazeologik birliklar, tabiiy tilni qayta ishlash, lingvistik annotatsiya, avtomatik teglash, ko‘p komponentli birikmalar, semantik yaxlitlik.*

MODERN APPROACHES AND RESEARCH REVIEW ON THE TAGGING OF PHRASEOLOGICAL UNITS

Abstract:
This article is devoted to developing the theoretical and methodological foundations for tagging phraseological units in the Uzbek national language corpus and their automatic linguistic annotation. Due to the semantic integrity of phraseological units, as well as the variability of their composition, distinguishing them from ordinary word combinations in NLP systems presents considerable complexity. The study examines international practices in the development of electronic databases of phraseological units through the application of systematic linguistic analysis, corpus-based methods, and a comparative methodological framework for evaluating the capabilities of modern neural network models, as well as the possibilities of their implementation within the Uzbek language corpus.

Keywords: *Uzbek national language corpus, phraseological units, natural language processing (NLP), linguistic annotation, automatic tagging, multi-word expressions (MWE), semantic integrity.*

СОВРЕМЕННЫЕ ПОДХОДЫ И ОБЗОР ИССЛЕДОВАНИЙ ПО ПРОБЛЕМЕ ТЕГИРОВАНИЯ ФРАЗЕОЛОГИЧЕСКИХ ЕДИНИЦ

Аннотация:

Данная статья посвящена разработке теоретико-методических основ тегирования фразеологических единиц в национальном корпусе узбекского языка и их автоматической лингвистической аннотации. Семантическая целостность фразеологических единиц, а также вариативность их состава обусловливают значительную сложность их разграничения с обычными словосочетаниями в системах обработки естественного языка (NLP). В исследовании рассматриваются международные практики формирования электронных баз данных фразеологических единиц посредством применения системного лингвистического анализа, корпусных методов и сравнительной методологии оценки возможностей современных моделей нейронных сетей, а также возможности их внедрения в корпус узбекского языка.

Ключевые слова: *национальный корпус узбекского языка, фразеологические единицы, обработка естественного языка, лингвистическая аннотация, автоматическое тегирование, многословные выражения, семантическая целостность.*

Kirish:

Frazeologik birliklar tilning semantik va madaniy jihatdan eng murakkab qatlamlaridan biri hisoblanadi va ular o‘zining ko‘chma ma’nosi va barqaror tarkibi bilan boshqa leksik birliklardan ajralib turadi. Frazeologizmlar tilning “oltin fondi” hisoblanib, ularning ma’nosi ko‘pincha komponentlarning oddiy yig‘indisidan kelib chiqmaydi. Shuning uchun frazeologik birliklarni korpus lingvistikasi va NLP tizimlarida avtomatik aniqlash ham qiziqarli, ham murakkab vazifalardan biri hisoblanadi. O‘zbek tili frazeologiyasini raqamlashtirish, xususan, frazeologik birliklarni milliy korpusda avtomatik identifikatsiyalash masalasi hozirda yechimini kutayotgan dolzarb ilmiy muammolardandir.

Shavkat Rahmatullayevning tadqiqotlarida ta’kidlanganidek, frazeologik birliklar nutqimizning ko’rki bo’lish bilan birga, o’zining barqarorligi va yaxlitligi bilan ajralib turadi. Biroq bu xususiyatlar raqamli muhitda algoritmlar uchun jiddiy qiyinchiliklarni tug‘dirishi mumkin. An’anaviy tilshunoslikda iboralarning semantik va uslubiy xususiyatlari atroflicha

o‘rganilgan bo‘lsa ham, milliy korpusda avtomatik teglash va annotatsiya qilish uchun yangi texnologik yondashuvlar talab etiladi. So‘nggi yillarda ushbu muammo NLP tadqiqotlarida alohida e’tibor qaratilayotgan yo‘nalishlardan biriga aylandi.

Jahon miqyosida ko‘p so‘zli birikmalar (Multiword Expressions — MWE) va idiomalarni identifikatsiya qilish uchun zamonaviy transformator modellaridan foydalanish sezilarli darajada kengaydi. Jurafsky va Martin ta’kidlaganidek, til modellarining kontekstual semantikani anglash qobiliyati frazeologizmlarni matn tarkibida aniqlashda muhim o‘rin tutadi. O‘zbek tili agglyutinativ til sifatida o‘ziga xos murakkab morfologik tarkibga ega bo‘lib, frazeologik birliklarning tarkibiy qismlari turli grammatik shakllarda keladi yoki gap ichida bir-biridan uzoqda joylashadi, bu esa avtomatik tahlil jarayonini yanada murakkablashtiradi.

Maqolada frazeologik birliklarni avtomatik teglash va aniqlashga doir lingvistik hamda kompyuter lingvistikasi sohasidagi tadqiqotlar tahlil etiladi. Shuningdek, xalqaro tajribalar va ularning o‘zbek tili korpuslariga tatbiq etilish imkoniyatlari ko‘rib chiqiladi.

Mavzuga oid adabiyotlar tahlili:

Frazeologik birliklarni raqamli muhitda o‘rganishga oid tadqiqotlar dastlab an’anaviy leksikografiya doirasida shakllangan bo‘lsa, endilikda bu yo‘nalishda korpus lingvistikasi va NLP usullaridan foydalanish rivojlanib bormoqda.

O‘zbek frazeologiyasining dastlabki shakllanishida Shavkat Rahmatullayevning tadqiqotlari muhim o‘rin tutadi. Muallifning “O‘zbek tilining izohli frazeologik lug‘ati” va “Nutqimiz ko‘rki” asarlari bugungi kunda ham shu sohaning muhim manbalaridan biri hisoblanadi [15], [16]. Rahmatullayev frazeologizmlarni “ko‘chma ma’noli turg‘un bog‘lanmalar” deb ta’riflaydi va ularni erkin birikmalardan ajratuvchi asosiy mezon deb ma’no yaxlitligini ko'rsatadi. Mazkur qarash kompyuter lingvistikasidagi non-compositionality tushunchasiga yaqin turadi. Ammo klassik tadqiqotlarda frazeologizmlarning korpusdagi xususiyatlari yetarlicha o‘rganilmaganligi sabab zamonaviy korpus lingvistikasida bu masala yetarli darajada o‘zlashtirilmagan yo‘nalishlardan biri bo‘lib qolmoqda.

O‘zbek olimlaridan N. Abduraxmonova korpus lingvistikasining zamonaviy yo‘nalishlari qatorida avtomatik annotatsiya va semantik teglash masalalariga alohida e’tibor qaratadi [1]. Mazkur yondashuvlar frazeologik birliklarni korpus tarkibida mustaqil semantik birlik sifatida ajratish masalasi bilan ham chambarchas bog‘liqdir. Agostini va hammualliflar tomonidan yaratilgan UZWORDNET tizimi esa o‘zbek tilining 28 000 dan ortiq synsetlarni qamrab olgan yirik leksik-semantik bazasi sifatida tavsiflanadi [6]. Bu tadqiqot o‘zbek tili frazeologizmlari uchun semantik annotatsiya va NLP texnologiyalarini rivojlantirishda muhim bosqich hisoblanadi.

O‘zbek tili tabiiy tilni qayta ishlash sohasida morfologik analizator va annotatsiyalangan datasetlar yaratishga oid tadqiqotlar ham olib borilgan. Xususan, Abduraxmonova, Mengliyev va Baraxninlar yaratgan morfologik analizator va keyinchalik nashr etilgan annotatsiyalangan morfologik dataset o‘zbek tili tabiatini kompyuter modellariga o‘rgatish sohasini rivojlantirishga xizmat qildi [3], [4]. Iboralarni avtomatik teglash jarayonida morfologik segmentatsiya muhim ahamiyatga ega. MorphUz analizatori o‘zbek tilidagi o‘zak va qo‘shimchalarni aniqlash asosida lingvistik birliklarni avtomatik tahlil qilish imkonini beradi, bu esa frazeologik konstruksiyalarni korpus muhitida aniqlash samaradorligini oshiradi [5]. Yodgorovning “Formation of a database of phraseological units based on the corpus of the Uzbek language” tadqiqoti esa bu ikki yo‘nalish: lingvistik tavsif va korpus texnologiyasini frazeologiya sohasida birlashtirish uchun frazeologik birliklar bazasini shakllantirish masalalariga e’tibor qaratadi [19]. Shunday bo‘lsa-da, mavjud o‘zbek tili korpuslarida asosiy e’tibor morfologik va sintaktik teglashga qaratilgan bo‘lib, frazeologik birliklarni alohida semantik segment sifatida ajratish hamon hal etilmagan masala bo‘lib qolmoqda.

Xalqaro adabiyotlarda frazeologiya masalasi, asosan, “Multiword Expressions” (MWE) atamasi bilan o‘rganiladi. Kompyuter lingvistikasida o‘ziga xos o‘ringa ega bo‘lgan Jurafsky va Martinning “Speech and Language Processing” nomli darsligida ko‘p so‘zli birikmalarni aniqlash NLP tizimlari uchun murakkab vazifalardan biri sifatida ko‘rsatiladi [13]. Ide va boshqalar tomonidan olib borilgan tadqiqot esa bu sohadagi metodologik xilma-xillikni tizimlashtiradi va iboralar identifikatsiyasining ikki yo‘nalishda rivojlanayotganini ko‘rsatadi. Bunda statistik qarash so‘zlarning birga kelish ehtimolligi, o‘zaro axborot koeffitsientlariga tayanadi. Chuqur o‘rganish uslubi esa kontekstual embeddinglar orqali matn ichida idiomalarni topishga qaratiladi [12].

V. Nedumpozhimana va boshqalarning “Shapley Idioms” tadqiqoti ham transformator modellar sohasida alohida ahamiyatga ega bo‘lib, mualliflar diqqat mexanizmlarining idiomalarni aniqlashdagi rolini tahlil qiladi [14]. Transformator modellari nafaqat ko‘chma ma’noni anglashda yaqin atrofdagi so‘zlarga, balki uzoq masofadagi kontekstual bog‘liqliklarga ham tayanadi. Bu xususiyat o‘zbek tilidagi komponentlari gap ichida bir-biridan uzoqlashib kelgan iboralar, ya’ni distant frazeologizmlar uchun, ayniqsa, muhimdir.

Ko‘p tilli tahlil yo‘nalishida I. Zaitova va boshqalarning multilingual tahlili esa BERT asosidagi modellarning idiomalar va mikrosintaktik konstruksiyalarga nisbatan ishlash xususiyatini yanada chuqurroq ochib beradi [20]. Tadqiqot natijalariga ko‘ra, modelning idiomatik ma’noni to‘g‘ri talqin etishi ko‘p jihatdan ta’limiy korpusning hajmi va sifatiga bog‘liq. Kuzatuv o‘zbek tili uchun annotatsiyalangan frazeologik korpus yaratishning nafaqat lingvistik, balki texnologik jihatdan ham zarurligini yana bir bora tasdiqlaydi. Xuddi shu

model asosida Yayavaram va boshqalar ham tadqiqot olib borib, BERT modelini idiomalarni aniqlashga moslashtirish bo‘yicha yangicha yechim taklif etishdi. Bunda so‘zlarning o‘zaro yaqinligi (word cohesion) va tarjima usullari yordamida model ko‘chma ma’noni literal ma’nodan ajrata olish qobiliyati sezilarli yaxshilanganligi kuzatildi [18]. Agglyutinativ tuzilishga ega bo‘lgan o‘zbek tili uchun bu tadqiqot, ayniqsa, perspektiv bo‘lib, morfologik tomondan murakkab komponentlardan tashkil topgan frazeologizmlarni aniqlashda qo‘shimcha samaradorlik berishi kutiladi.

Frazeologik birliklarning nutqdagi faolligi va zamonaviy qo‘llanish darajasini aniqlash masalasi Budiltseva va Novikova tadqiqotida ham ko‘rib chiqilgan. Mualliflar somatik komponentli frazeologizmlar misolida ayrim idiomalar faol iste’molda saqlanib qolishini, ayrimlari asta-asta passiv qatlamga o‘tib qolishini qayd etadi. Bunda “aktual frazeologik korpus” modeli, ya’ni frazeologik birliklarni qo‘llanilish chastotasiga ko‘ra saralash usuli taklif etilgan [8]. Ularning ushbu yondashuvini o‘zbek tili frazeologizmlari uchun ham amaliy jihatdan qo‘llash mumkin, chunki milliy korpus asosida faol va passiv qatlamdagi iboralarni ajratish semantik teglash sifatini sezilarli miqdorda oshirishi mumkin.

Xorij tajribasi bilan bir qatorda bu sohada turkiy tillarning tadqiqotlarini ham yodga olish o‘rinli bo‘ladi. Sababi turkiy tillar doirasidagi so‘nggi ishlar ham muhim hisoblanadi. Xususan, G.Aslantash va T.Gungo‘rlarning “A Unified Turkic Idiom Understanding Benchmark” deb atalgan tadqiqoti turk, ozarbayjon, o‘zbek kabi beshta turkiy tillar uchun umumiy idiomalar bazasini yaratdi va o'zaro transfer qobiliyatini (cross-lingual transfer) sinovdan o‘tkazdi [7]. Sh.Hakimovning “TurkicNLP” loyihasi esa turkiy tillar uchun yagona API va CoNLL-U formatidagi annotatsiya standartlarini taklif etib, o‘zbek tili NLP tizimini xalqaro platformaga ulash uchun imkon yaratdi [11]. D.O.Dobrovolskiyning frazeologiyada korpus metodologiyasini qo‘llashga bag‘ishlangan ishlari esa qiyosiy perspektiva beradi: muallif parallel korpuslar orqali ayrim frazeologik birliklarning faqat ma’lum grammatik shaklda yoki ma’lum uslubda ishlatilishini ta’kidlaydi — bu kuzatuv o‘zbek tili uchun ham semantik teglash tizimini boyitish zarurligini ko‘rsatadi [9], [10].

Tadqiqot metodologiyasi:

Frazeologik birliklarni avtomatik aniqlash va teglashga oid tadqiqotlar metodologik jihatdan qoidalarga asoslangan, statistik va zamonaviy neyron tarmoq (transformator) modellari imkoniyatlarini qiyoslashga tayanadi. Tadqiqot doirasida quyidagi yondashuvlar tahlil qilindi:

1. Qoidalarga asoslangan yondashuv: Sh.Rahmatullayevning frazeologik lug‘atlaridagi birliklar asosida lug‘at va lingvistik qoidalarga tayanib ishlaydi [15], [16].

2. Statistik yondashuv: Korpusdagi so‘zlarning o‘zaro bog‘liqlik koeffitsientlari (Mutual Information, Log-likelihood) va birga qo‘llanish chastotasini tahlil qiladi [11], [18].

3. Transformator modellar: BERT va mBERT kabi modellar yordamida frazeologizmlarning kontekstual semantikasini va diqqat mexanizmlari orqali idiomatiklik darajasini aniqlashga qaratiladi [7], [13], [17].

Bu yondashuvlarning o‘zbek va turkiy tillar korpuslari misolidagi qiyosiy samaradorlik ko‘rsatkichlarini quyidagi jadval umumlashtiradi:

1-jadval. NLP modellari samaradorligining qiyosiy tahlili

modellash usuli	aniqlik (F1-score)	tadqiqot yo‘nalishi
qoidalarga asoslangan (rule-based)	65 %	lug‘at va morfologik qoidalar
statistik metodlar (MI, log-likelihood)	72 %	So‘zlarning korpusda birga kelish ehtimoli
transformator modellari (BERT)	85 %	kontekstual semantika va diqqat mexanizmi
Ko‘p tilli modellar (mBERT/Turkic)	88 %	turkiy tillar uchun transfer qobiliyati

Ko‘rib chiqilgan tadqiqotlarda o‘zbek tilining agglyutinativ tabiati va komponentlari gap ichida uzoq joylashgan frazeologizmlarni aniqlashda transformator modellarining kontekstual imkoniyatlari yuqori samaradorlik berishi kuzatildi.

Tahlil va natijalar:

Mavjud adabiyotlar tahlili shuni ko‘rsatadiki, frazeologizmlarni teglash masalasi NLP tadqiqotlarida faol rivojlanayotgan yo‘nalishlardan biriga aylangan. Xalqaro tadqiqotlarda, ayniqsa, ingliz va yevropa tillari korpuslarida transformator modellardan foydalanish keng tarqalib, ayrim ishlarda frazeologik birliklarni aniqlashda yuqori samaradorlik ko‘rsatkichlari qayd etilgan [7], [13]. Qoidalarga asoslangan va statistik metodlar ham frazeologik birliklarni aniqlashda qo‘llaniladi, biroq komponentlari gap ichida uzoq joylashgan frazeologizmlarni aniqlashda ularning imkoniyatlari cheklanishi mumkin.

O‘zbek tili bilan bog‘liq tadqiqotlarda esa asosiy e’tibor ko‘proq morfologik va sintaktik annotatsiyaga qaratilgan. Abdurakhmonova va boshqalar tomonidan yaratilgan morfologik analizator va annotatsiyalangan datasetlar o‘zbek tilini avtomatik tahlil qilish yo‘nalishidagi muhim ishlardan hisoblanadi [3], [4]. Biroq frazeologik birliklarni alohida semantik birlik

sifatida avtomatik aniqlash va teglash masalasi hali yetarli darajada ishlab chiqilmagan. Turkiy tillar doirasidagi ayrim tadqiqotlar esa umumiy annotatsiya standartlari va ko‘p tilli modellarni qo‘llash imkoniyatlariga e’tibor qaratadi [2], [20].

Xulosa va takliflar:

Ko‘rib chiqilgan manbalar asosida shuni aytish mumkinki, xalqaro va mahalliy tadqiqotlarda transformator modellari, statistik metodlar va qoidalarga asoslangan usullardan keng foydalanilmoqda. O‘zbek tilida esa frazeologik birliklarni avtomatik tahlil qilish bo‘yicha tadqiqotlar hali yetarli darajada shakllanmagan bo‘lib, mavjud ishlarda, asosan, morfologik va sintaktik annotatsiyaga e’tibor qaratilgan. Shu bilan birga, turkiy tillar doirasidagi zamonaviy tadqiqotlar hamda ko‘p tilli NLP modellari o‘zbek tili korpuslarida frazeologik birliklarni semantik va kontekstual jihatdan avtomatik aniqlash bo‘yicha yangi imkoniyatlar mavjudligini ko‘rsatadi.

Foydalanilgan adabiyotlar

1. Abdurakhmonova N. Korpus lingvistikasi. – Toshkent, 2024. – 320 b.
2. Abdurakhmonova N., Ismailov A., Mengliev D. Developing NLP Tool for Linguistic Analysis of Turkic Languages // 2022 IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON). – 2022. – Pp. 1–6.
3. Abdurakhmonova N., Mengliev D., Barakhnin V. Development of Intellectual Web System for Morph Analyzing of Uzbek Words // Applied Sciences. – 2021. – Vol. 11. – No. 19. – Pp. 9117.
4. Abdurakhmonova N., Shirinova R., Sayfullayeva R., Mengliev D., Ibragimov B., Ernazarova M. An annotated morphological dataset for Uzbek word forms: Towards rule-based and machine learning approaches // Data in Brief. – 2025. – Vol. 61. – Pp. 111702.
5. Abdurakhmonova N., Alisher I., Sayfulleyeva R. Morphological analyzer for the Uzbek language // 2022 7th International Conference on Computer Science and Engineering (UBMK). – 2022. – Pp. 61–66.
6. Agostini A., Usmanov T., Khamdamov U., Abdurakhmonova N., Mamasaidov M. UzWordNet: A lexical-semantic database for the Uzbek language // Proceedings of the 11th Global Wordnet Conference. – 2021. – Pp. 8–19.
7. Aslantaş G., Güngör T. A Unified Turkic Idiom Understanding Benchmark: Idiom Detection and Semantic Retrieval Across Five Turkic Languages // Proceedings of SIGTURK. – 2026. – Pp. 12–24.

8. Budiltseva M.B., Novikova N.S. Life of an idiom – defining the current corpus of phraseological units: experimental research experience // RUDN Journal of Language Studies, Semiotics and Semantics. – 2023. – Vol. 14. – No. 3. – Pp. 931–945.

9. Dobrovol’skij D.O. Korpusniy podkhod k issledovaniyu frazeologii: novie rezultati po dannim parallelnikh korpusov // Vestnik SPbGU. Yazik i literatura. – 2020. – T. 17. – Vip. 3. – Pp. 398–411.

10. Dobrovol’skij D.O. Kriterii semanticheskoy chlenimosti idiom // RUDN Journal of Language Studies, Semiotics and Semantics. – 2025. – Vol. 16. – No. 3. – Pp. 638–655.

11. Hakimov Sh. TurkicNLP: An NLP Toolkit for Turkic Languages // arXiv preprint arXiv:2602.19174. – 2026.

12. Ide Y., Tanner J., Nohejl A., Hoffman J., Vasselli J., Kamigaito H., Watanabe T. COAM: Corpus of All-Type Multiword Expressions // arXiv preprint arXiv:2412.18151. – 2024.

13. Jurafsky D., Martin J.H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. – Stanford University, 2025. – 640 p.

14. Nedumpozhimana V., Klubička F., Kelleher J.D. Shapley Idioms: Analysing BERT Sentence Embeddings for General Idiom Token Identification // Frontiers in Artificial Intelligence. – 2022. – Vol. 5. – Pp. 813967.

15. Rahmatullayev Sh. Nutqimiz ko‘rki. – Toshkent: Fan, 1970. – 60 b.

16. Rahmatullayev Sh. O‘zbek tilining izohli frazeologik lug‘ati. – Toshkent: O‘qituvchi, 1978. – 408 b.

17. Tanner J., Hoffman J. MWE as WSD: Solving Multiword Expression Identification with Word Sense Disambiguation // arXiv preprint arXiv:2303.06623. – 2023.

18. Yayavaram A., Yayavaram S., Upadhyay P., Das A. BERT-based Idiom Identification using Language Translation and Word Cohesion // Proceedings of MWE-2024. – 2024. – Pp. 110–118.

19. Yodgorov U.S. Formation of a database of phraseological units based on the corpus of the Uzbek language // American Journal of Philological Sciences. – 2025. – Vol. 5. – No. 3. – Pp. 148–151.

20. Zaitova I., Hirak V., Abdullah B.M., Klakow D., Möbius B., Avgustinova T. Attention on Multiword Expressions: A Multilingual Study of BERT-based Models with Regard to Idiomaticity and Microsyntax // arXiv preprint arXiv:2505.06062. – 2025.