AMEA Nizami Gəncəvi adına Ədəbiyyat İnstitutu
Poetika.izm, 2025, № 2 *ISSN: 2663-2926*
Səhifə: 161-171 *E-ISSN: 2709-2699*
https://doi.org/10.59849/2663-2926.2025.2.161

Azərbaycan dilinin bədii mətnlər korpusunun formalaşdırılmasında süni intellekt (Sİ) və korpus dilçiliyi metod və texnologiyaları

Həmidə Xəlilova
Moskva Dövlət Universitetinin Bakı filialı
AMEA Nəsimi adına Dilçilik İnstitutu
E-mail: hamida.khalilova.a@gmail.com
Orcid ID: 0000-0002-7423-8850

Annotasiya. Məqalə Azərbaycan dilinin bədii mətnlər korpusunun yaradılmasında süni intellekt (Sİ) texnologiyalarının tətbiqinə həsr olunmuşdur. Tədqiqat çərçivəsində Azərbaycan dilinin bədii mətn korpusu hazırlanmışdır. Tədqiqatın aktuallığı və yeniliyi, korpus dilçiliyi və linqvostatistika metodlarından, süni intellekt texnologiyalarından istifadə edərək mətn korpusları və tezlik lüğətlərinin yaradılmasında, həmçinin tədqiqat materialının düzgün seçilməsindən ibarətdir. Xüsusilə qeyd etmək lazımdır ki, Azərbaycan dili mətn korpusları ilə təmsil olunmamışdır. Bu tədqiqat mətn korpuslarının və Azərbaycan dilinin bədii mətn korpusunun və tezlik lüğətlərinin yaradılmasında pioner iş olmaqla yanaşı, həmçinin Azərbaycan dilinin Milli Korpusunun formalaşdırılması üçün mühüm əhəmiyyət daşıyır. Məqalədə avtomatlaşdırılmış işarələmə, tematik modelləşdirmə, lemmatizasiya və digər yanaşmalar da nəzərdən keçirilir. Bundan əlavə, korpus dilçiliyi metodları və Sİ texnologiyalarının vasitəsilə mətnləri təsnif etmək, müəyyən bir yazıçının üslubunu təhlil etmək, əsərin müəllifliyini müəyyənləşdirmək və digər aspektləri araşdırmaq mümkün olduğu vurğulanır. Müasir texnologiyaların tətbiqi ədəbiyyatşünaslar və dilçilərə müəlliflərin yaradıcılığını daha dərindən və hərtərəfli araşdırmağa, əsərlərdəki qanunauyğunluqları aşkar etməyə, eləcə də ədəbiyyatşünaslıq təhlilində korpus dilçiliyinin metodlarından istifadə etmək imkanı verəcəkdir. Həmçinin qeyd etmək lazımdır ki, müəllif korpuslarının yaradılması da innovativ xarakter daşıyır. Rusiyada Puşkinin mətnlər korpusu, Çexovun mətnlər korpusu, İngiltərədə isə Şekspirin mətnlər korpusu yaradılmışdır. Müəllif korpusları fərdi dil şəxsiyyətinin xüsusiyyətlərini araşdırmağa imkan verir. Bir şəxsiyyətin mətnlərində dil tam şəkildə vahid bir sistem kimi təzahür edir. Məqalədə M.Ə.Sabirin mətnlər korpusu, Nizamin Gəncəvinin mətnlər korpusu, M.Füzulinin poetik mətnlər korpusu, İ. Nəsiminin mətnlər korpusunun tərtibi və formalaşdırılması da yer almışdır. Həmçinin çoxdilli paralel mətnlərə əsaslanan korpusların yaradılmasında Sİ istifadəsinin üstünlükləri və faydaları barədə də danışılır, bu isə ədəbiyyat üçün olduqca vacibdir. Bu həm orijinal dildə, həm də tərcümələrdə bədii mətnlərin müqayisəsi üçün imkanlar açacaq və tərcümənin xüsusiyyətlərini daha dərindən təhlil etməyə, şərh fərqlərini müəyyən etməyə, həmçinin əsərlərin stilistik və linqvokulturoloji təsirlərini araşdırmağa şərait yaradacaqdır. Bu yanaşma təkcə leksikoloji və qrammatik təhlil üçün deyil, həm də müəllif üslubunun, mətnin emosional-ekspressiv rənginin və digər aspektləri öyrənilməsi üçün faydalı ola bilər. Beləliklə, tədqiqat göstərir ki, süni intellekt texnologiyaları və korpus dilçiliyi metodları humanitar elmlərin imkanlarını genişləndirir və Azərbaycan dili və ədəbiyyatını araşdırmaq üçün yeni perspektivlər açır.
Açar sözlər: korpus, süni intellekt, mətn korpusu, bədii ədəbiyyat, konkordans, Azərbaycan dili, lemmatizasiya, token, müəllif korpusu

Məqalə tarixçəsi: göndərilib – 10.11.2025; qəbul edilib – 25.11.2025

Artificial intelligence (AI) and corpus linguistics methods and technologies in the formation of a literary text corpus of the Azerbaijani language

Hamida Khalilova
Baku Branch of Moscow State University
Institute of Linguistics named after Nasimi ANAS
E-mail: hamida.khalilova.a@gmail.com

Abstract. The article is dedicated to the application of artificial intelligence (AI) technologies in the creation of a literary text corpus in the Azerbaijani language. Within the framework of the study, a literary text corpus of the Azerbaijani language was compiled, as well as corpora of texts by Azerbaijani authors. The relevance and novelty of the research lie in the use of corpus linguistics and linguostatistics methods, AI technologies in the creation of text corpora and frequency dictionaries, and in the proper selection of research material. It should be particularly noted that Azerbaijani language texts have been scarcely represented in corpora. This study is pioneering both in the creation of Azerbaijani text corpora and in the development of literary text corpora and frequency dictionaries in Azerbaijani, and it also holds significant importance for the formation of the National Corpus of the Azerbaijani language.
The article also addresses automated annotation, topic modeling, lemmatization, and other modern approaches. Furthermore, it emphasizes that corpus linguistics methods and AI technologies make it possible to classify texts, analyze the style of a specific author, determine authorship of works, and study other related aspects. The application of modern technologies allows literary scholars and linguists to investigate authors' works more deeply and comprehensively, identify patterns in their texts, and apply corpus linguistics methods in literary analysis. It should also be noted that the creation of author-specific corpora is innovative. In Russia, author corpora such as the "A.S.Pushkin Text Corpus" and the "A.P.Chekhov Text Corpus" have been created, while in the United Kingdom, the "Shakespeare Text Corpus" exists. Author corpora allow the study of individual linguistic personality features. In the texts of a single author, language manifests as an integral and unified system. Furthermore, the article discusses the advantages and benefits of applying AI in the creation of corpora based on multilingual parallel texts, which is of critical importance for literary studies. This approach enables the comparison of literary texts both in the original language and in translations, allows for a deeper analysis of translation features, reveals differences in interpretations, and facilitates the study of stylistic and cultural adaptations of works. Such an approach is useful not only for lexical and grammatical analysis but also for the investigation of authorial style, the emotional and expressive qualities of the text, and other aspects. Thus, the study demonstrates that artificial intelligence technologies and corpus linguistics methods expand the possibilities of the humanities and open new perspectives for the study of the Azerbaijani language and literature.
Keywords: corpus, artificial intelligence, text corpus, literary fiction, concordance, the Azerbaijani language, lemmatization, token, author-specific text corpus

Article history: received – 10.11.2025; accepted − 25.11.2025

Giriş / *Introduction*

Son illərdə süni intellekt (Sİ) texnologiyalarının sürətli inkişafı humanitar elmlərin müxtəlif sahələrində, eləcə də dilçilik və ədəbiyyatşünaslıqda yeni tədqiqat istiqamətlərinin formalaşmasına səbəb olmuşdur. Müasir texnologiyalar vasitəsilə ədəbi mətnlərin daha dərindən və sistemli şəkildə öyrənilməsi mümkün olmuş, bu isə tədqiqatçılara müxtəlif istiqamətlər üzrə zəngin analiz və genişmiqyaslı araşdırma aparmaq üçün şərait yaratmışdır. Bu, korpus dilçiliyi və onun metodlarının daha da inkişaf etməsində rol oynamışdır.

Korpus dilçiliyi kompüter dilçiliyi və linqvostatistikanın metodlarından istifadə edərək böyükhəcmli mətn korpusunun yaradılması ilə məşğul olan dilçilik sahələrindən biridir. Korpus dilçiliyinin terminoloji aparatına mətn korpusu, konkordans, korpus meneceri və s. daxildir.

Mətn korpusu – korpus dilçiliyinin əsas terminidir. Mətn(lər) korpusu kompüter texnologiyasının köməyi ilə tərtib edilmiş müəyyən mətnlərin toplusundan, hətta bu mətnlərin bir növ lüğətindən ibarətdir. Korpus müəyyən mətnlər üçün elektron axtarış sistemidir: Mətnlər korpusu (linqvistik) müəyyən qaydalara uyğun olaraq müəyyən edilmiş və tərtib edilmiş mətnlər toplusunun elektron informasiya sistemidir. *"Linqvistik mətnlər korpusu termini xüsusi linqvistik problemləri həll etmək üçün nəzərdə tutulmuş böyük, elektron şəkildə təqdim edilmiş, vahid, strukturlaşdırılmış, etiketlənmiş, filoloji cəhətdən səlahiyyətli dil məlumatlarına aiddir"* [2, s.6].

Korpus dilçiliyi metodlarının sayəsində yazıçıların üslubu, əsərlərin leksik və qrammatik xüsusiyyətləri, mövzu və janr müxtəlifliyi, həmçinin müəllifliyin müəyyənləşdirilməsi kimi məsələlər elmi əsaslarla araşdırıla bilir. Korpus dilçiliyi sayəsində yazıçıların fərdi mətn korpusları da yaradıla bilər. Dünya təcrübəsində bu istiqamətdə müxtəlif nümunələr mövcuddur: Rusiyada Puşkinin [5] və Çexovun [3], İngiltərədə isə Şekspirin mətn korpusları yaradılmış və bu korpuslar əsasında bir sıra mühüm tədqiqatlar aparılmışdır. Azərbaycan dili və ədəbiyyatında belə genişhəcmli və sistemli mətn korpusları bu günə qədər formalaşmamışdır.

Məqalə Azərbaycan yazıçılarının bədii mətn korpuslarının yaradılması və bu prosesdə süni intellekt texnologiyalarının tətbiqi məsələlərinə həsr olunmuşdur. Məqalənin məqsədi Azərbaycan ədəbiyyatının nümunələri əsasında elektron mətn korpuslarının formalaşdırılması və bu korpuslar üzərində müxtəlif avtomatlaşdırılmış təhlil metodlarının (lemmatizasiya, morfoloji işarələmə, tematik modelləşdirmə və s.) tətbiqi yolu ilə dil və ədəbiyyat materiallarının sistemli şəkildə araşdırılmasına yeni imkanlar açmaqdır.

Tədqiqatın aktuallığı bu sahədə ilk dəfə olaraq Azərbaycan yazıçılarının mətn korpuslarının yaradılması və strukturlaşdırılmasından ibarətdir. Xüsusilə qeyd etmək lazımdır ki, tədqiqat Azərbaycan dilinin Milli Korpusunun yaradılması istiqamətində atılmış mühüm addımdır. Belə ki, məhz bədii mətnlər milli-mədəni reallıq, presedent fenomen, kulturoloji və üslubi xüsusiyyətləri ən dolğun şəkildə əks etdirir. Bu baxımdan, bədii mətn korpuslarının yaradılması Milli Korpusun analitik imkanlarını genişləndirəcək və onun elmi-tədqiqat potensialını artıracaqdır.

Əsas hissə / *Main Part*

Ədəbiyyat icmali və metodoloji baza

Korpus dilçiliyi və süni intellekt (Sİ) texnologiyalarının dil və ədəbiyyat araşdırmalarında tətbiqi son illərdə dünya elmi ictimaiyyətinin diqqət mərkəzində olan aktual sahələrdən birinə çevrilmişdir. Bu yanaşma ilk dəfə XX əsrin ortalarından etibarən inkişaf etməyə başlamış və elektron mətn bazalarının yaranması ilə yeni mərhələyə qədəm qoymuşdur. Hazırda bu sahədə ən inkişaf etmiş tədqiqat mərkəzləri və nümunəvi layihələr müxtəlif ölkələrdə mövcuddur.

Məsələn, Rusiyada Rus dilinin Milli Korpusu, Puşkinin mətnlər korpusu və Çexovun mətnlər korpusu yaradılmışdır. Fərdi korpuslar yazıçıların dil üslubunun, leksik və sintaktik xüsusiyyətlərinin sistemli şəkildə öyrənilməsinə imkan verir, həmçinin onların vasitəsilə yazıçıların fərdi dil xüsusiyyətləri, dövrün dili ilə müqayisədə fərqlilikləri və inkişaf meyilləri aşkar edilir. Böyük Britaniyada isə Britaniya Milli Korpusu (British National Corpus) ilə yanaşı, Şekspirin mətnlər Korpusu da mövcuddur. Bu korpuslar ingilis dilinin tarixi inkişaf mərhələlərinin və yazıçı üslubunun təhlili baxımından böyük əhəmiyyət daşıyır.

Metodoloji baxımdan bu tədqiqatlar əsasən korpus dilçiliyi, linqvostatistika, lemmatizasiya və tematik modelləşdirmə kimi yanaşmalarla aparılır. Bu metodların tətbiqi nəticəsində mətnlərdəki

leksik tezliklər, sintaktik strukturlar, semantik sahələr və stilistik elementlər müəyyənləşdirilir və müqayisəli təhlil həyata keçirilir.

Süni intellekt texnologiyaları bu metodlara əlavə imkanlar gətirmişdir. Məlumatların avtomatik emalı, təbii dilin işlənməsi (NLP) alqoritmləri, dərin öyrənmə (deep learning) və maşın öyrənməsi (machine learning) kimi yanaşmalar sayəsində mətnlərin avtomatlaşdırılmış işarələnməsi, müəllifin tanınması, stilistik analiz və janr təsnifatı daha dəqiq və sürətli şəkildə həyata keçirilə bilir.

Bu məqalədə istifadə olunan metodoloji baza da məhz bu müasir yanaşmalara əsaslanır. Azərbaycan ədəbiyyatı nümunələri üzərində aparılan tədqiqatda korpusların formalaşdırılması, leksik-statistik analiz, avtomatik işarələmə və lemmatizasiya prosesləri icra olunmuşdur. Bununla yanaşı, tematik modelləşdirmə metodları əsasında yazıçı üslublarının təhlili və mətnlərin semantik sahələr üzrə təsnifatı həyata keçirilmişdir.

Beləliklə, tədqiqat həm dünya təcrübəsinə əsaslanır, həm də Azərbaycan dilində bənzərsiz bir nümunə olaraq bu sahədə ilk təşəbbüsdür.

Azərbaycan dilində mətn korpusunun vəziyyəti

Azərbaycan dili korpus dilçiliyi sahəsində sistemli tədqiqatlara hələ yeni başlayır. Hal-hazırda Azərbaycan dili mətn korpusları ilə təmsil edilməyib, lakin Azərbaycan dili elektron lüğət korpusları yaradılmışdır. Ənənəvi dilçilik araşdırmaları əsasən nəzəri müstəvidə aparıldığı üçün, rəqəmsal dil resurslarının və avtomatlaşdırılmış təhlil alətlərinin olmaması bu sahədə irəliləyişin zəif olmasına səbəb olmuşdur.

Qeyd etmək lazımdır ki, Azərbaycan dilində genişhəcmli və sistemli şəkildə formalaşdırılmış, leksik, morfoloji və sintaktik baxımdan işarələnmiş ədəbi mətn korpuslarının olmaması tədqiqatçılar üçün ciddi bir boşluq yaradır. Məhz bu səbəbdən, yazıçıların dil üslubunun, fərdi dil şəxsiyyətinin, bədii dilin inkişaf dinamikasının və digər bu kimi sahələrin təhlili çətinləşir və yalnız intuitiv yanaşmalarla məhdudlaşır.

Ədəbiyyatşünaslıqda və stilistik analizdə korpuslara əsaslanan metodların olmaması həm də obyektiv və müqayisəli təhlillərin aparılmasını çətinləşdirir. Bu yalnız Azərbaycan ədəbiyyatının deyil, eyni zamanda dilin ümumi inkişafını, normativliyi və müasir dil modelini öyrənmək imkanlarını da məhdudlaşdırır.

Digər tərəfdən, Azərbaycan ədəbiyyatının zənginliyi və janr müxtəlifliyi, klassik və müasir yazıçıların fərqli üslub və dil xüsusiyyətləri bu sahədə korpus əsaslı tədqiqatların aparılması üçün geniş imkanlar yaradır. Azərbaycan yazıçıları – Nizami, Nəsimi, Füzuli, Mirzə Ələkbər Sabir, Seyid Əzim Şirvani, Hüseyn Cavid, Cəfər Cabbarlı, Anar, Elçin və digər Azərbaycan yazıcı və mütəfəkkirlərin bədii irsi bu cür tədqiqatlar üçün zəngin material bazasıdır.

Araşdırmada təqdim edilən tədqiqat korpus dilçiliyi və Azərbaycan dilini mətn korpusunun, eləcə də Azərbaycan Milli Korpusunun yaradılması sahəsində ilk genişmiqyaslı təşəbbüsdür, elmi iş həm dilçilik, həm də ədəbiyyatşünaslıq üçün yeni imkanlar açmağı hədəfləyir. Süni intellekt texnologiyalarının və korpus dilçiliyi metodlarının vasitəsilə yaradılan mətn korpusları tədqiqatçılara daha dəqiq, sistemli və çoxaspektli analiz imkanı verəcəkdir.

Beləliklə, Azərbaycan dilində mətn korpuslarının yaradılması yalnız bir elmi nailiyyət deyil, eyni zamanda milli dilin və ədəbiyyatın rəqəmsal dövrdə inkişaf etdirilməsi baxımından da strateji əhəmiyyət daşıyır.

Tədqiqat materialı və korpusun yaradılması prinsipləri

Tədqiqat çərçivəsində yaradılan bədii mətn korpusunun əsasını Azərbaycan yazıçılarının müxtəlif dövrlərə və janrlara aid əsərləri təşkil edəcəkdir. Məqsəd həm klassik, həm də müasir ədəbiyyat nümunələrinin birgə təhlilinə imkan verən müxtəlif tipli və zəngin mətn bazası formalaşdırmaqdır. Bu məqsədlə aşağıdakı meyarlar nəzərə alınaraq mətn seçimi aparılmışdır:

Tarixi müxtəliflik: Korpusda XII əsrdən başlayaraq, həmçinin XX əsrin əvvəllərindən bu günə qədər yazılmış ədəbi əsərlər təmsil olunacaqdır. Bu, dilin tarixi inkişaf mərhələlərini izləməyi və dilin dəyişikliklərini müşahidə etməyi mümkün edir.

Janr müxtəlifliyi: Verilənlər bazasına roman, povest, hekayə, dram və s. kimi müxtəlif ədəbi janrlara aid əsərlər seçilmişdir. Bu, dilin fərqli üslub və funksional sahələrdə necə işlədiyini araşdırmağa imkan verir.

Müəllif fərqliliyi: Korpusda klassik şərq poeziyasının böyük mütəfəkkirləri – Nizami Gəncəvi, İmadəddin Nəsimi, Məhəmməd Füzuli; maarifçilik dövrünün tanınmış satirik şairi – Mirzə Ələkbər Sabir; XX əsrin əvvəlləri və sovet dövrünün görkəmli yazıçı və dramaturqları – Cəlil Məmmədquluzadə, Hüseyn Cavid, Səməd Vurğun, Mirzə İbrahimov; eləcə də çağdaş dövrün tanınmış qələm sahibləri – Anar, Elçin, Kamal Abdulla və digərləri daxil olacaqdır.

Mətnlərin rəqəmsallaşdırılması zamanı orfoqrafik düzgünlük, orijinal strukturun qorunması və linqvistik annotasiya üçün əlverişli formatda təqdim olunması əsas meyarlar kimi götürülmüşdür. Mənbə kimi etibarlı elektron kitab platformalarından, rəsmi nəşrlərdən və müəllif hüquqları qorunan açıq resurslardan istifadə olunmuşdur.

Qeyd etmək lazımdır ki, ilk olaraq korpusa Nizami Gəncəvinin, Mirzə Ələkbər Sabirin, Məhəmməd Füzulinin və İmadəddin Nəsiminin mətnləri daxil edilmişdir. Xüsusilə qeyd etmək istərdim ki, bu mərhələdə həm konkordansların, həm də qeyd olunan dörd korpusun yaradılmasında süni intellekt metodlarından istifadə olunmamışdır. Bu mərhələdə əsas diqqət korpus və kompüter metodlarına yönəldilmişdir. Bundan əlavə, kontekstlərin əksəriyyəti sonradan yenidən əl ilə yoxlanılmış və düzəliş edilmişdir. Belə ki, Azərbaycan dili aqqlutinativ dil olduğundan, korpusun formalaşdırılması zamanı çoxmənalılıq və leksik vahidlərin avtomatik müəyyənləşdirilməsi ilə bağlı problemlər yaranır. Aqqlutinativ dillərdə sözlər əsasən şəkilçilər vasitəsilə törədildiyindən eyni kökdən çoxsaylı formalar əmələ gəlir; bu isə bazada axtarış, lemmatizasiya və morfoloji analiz proseslərini çətinləşdirir. Nəticədə avtomatik sistemlərin çıxışları tez-tez natamam və ya səhv olur və onların dəqiqliyinin təmin edilməsi üçün əlavə əl ilə yoxlama və korreksiya tələb olunur.

İlkin materialların seçimi təsadüfi deyildir. Nizami Gəncəvi, M.Füzuli və M.Ə.Sabirin şəxsiyyətinə və yaradıcılıq irsinə maraq bu gün də öz aktuallığını itirmir [7]. Qeyd etmək lazımdır ki, ilkin materialların sırasına İmadəddin Nəsiminin də əsərləri əlavə edilmişdir.

Məşhur şair və filosof Nizami Gəncəvinin "Xəmsə"si həm Azərbaycanda, həm də xaricdə hərtərəfli tədqiqat obyektinə çevrilmişdir. Nizami irsinin öyrənilməsinin əhəmiyyətini Nizami Gəncəvi Mərkəzi ilə Oksford Universiteti arasında, AMEA-nın vitse-prezidenti, MDU-nun Bakı filialının rektoru, akademik Nərgiz Paşayevanın iştirakı və təşəbbüsü ilə imzalanmış müqavilə bir daha təsdiqləyir [7].

Azərbaycan satirik poeziyasının banisi, görkəmli satirik şair M.Ə.Sabirin satiraları bu gün də aktuallığını qoruyur. Sabir qabaqcıl bir şəxsiyyət idi. Satiralarında qaldırdığı mövzuların mürəkkəbliyi və aktuallığı ilə böyük satirik öz dövrünü xeyli qabaqlamışdır. M.Ə. Sabirin yazı üslubu və dili də tədqiqatımız üçün böyük əhəmiyyət kəsb edir. O, sadə və aydın yazaraq, canlı xalq danışığı ilə ədəbi dili birləşdirərək poeziyada yeni üslub yaratmışdır. Məhz bu xüsusiyyətinə görə Sabirin poetik dili xalqa asanlıqla başa düşülən idi və onun yaradıcılığının geniş kütlələr arasında yayılmasına şərait yaratmışdır [7].

Azərbaycan ədəbiyyatının digər önəmli nümayəndəsi – Məhəmməd Füzulini. M.Füzuli – görkəmli şair, filosof, mütəfəkkir və tərcüməçidir. Bütün bunlarla yanaşı, M. Füzuli ədəbi banisi kimi araşdırma üçün böyük əhəmiyyət kəsb edir. Onun yaradıcılığı bədii dilin leksik, sintaktik və üslubi xüsusiyyətlərinin formalaşmasına zəmin yaradır [7].

İlk yaradılan müəllif korpuslarından birinin İmadəddin Nəsimin mətn korpusu olması da təsadüfi deyildir. Anadilli poeziyamızın inkişafında müstəsna rol oynayan mütəfəkkirlərdən biri məhz İmadəddin Nəsimidir. O, Azərbaycan ədəbiyyatı tarixində fəlsəfi şeirin əsasını qoymuş, bədii

sözü forma və məzmunca zənginləşdirmiş, Azərbaycan dilini ədəbi-bədii dil səviyyəsinə yüksəldən ustad ədibdir.

Yaradılmış müəllif korpusları haqqında qısa məlumat [7]:

Nizami Gəncəvinin mətnlər korpusu. Hal-hazırda bu korpusu "Xəmsə"sənin tərkibinə daxil olan aşağıdakı poemaların Azərbaycan dilinə tərcümələri təşkil edir: "Sirlər xəzinəsi", "Xosrov və Şirin", "Leyli və Məcnun", "Yeddi gözəl", "İsgəndərnamə"nin "Şərəfnamə" və "İqbalnamə" hissələri. Bununla yanaşı, tədqiqata Nizami Gəncəvinin "Leyli və Məcnun" poemasının rus dilinə tərcüməsi də daxil edilmişdir. Tədqiqat nəticəsində:

- Çoxmənalılıq problemi qismən aradan qaldırılmışdır;
- Nizami Gəncəvinin araşdırılan mətnlərinin leksik vahidlərini əhatə edən verilənlər bazası (VB) tərtib edilmişdir;
- Konkordanslar düzəldilmişdir;
- Kontekstlərlə birlikdə təqdim olunan tezlik lüğətləri tərtib edilmiş və hal-hazırda genişləndirilir;
- Müəllifin dil xüsusiyyətlərindən bəziləri müəyyən edilmişdir;
- "Leyli və Məcnun" poemasının Azərbaycan dilindəki mətninə əsasən etnospesifik leksikanın kontekstli tezlik lüğəti və ona uyğun konkordans hazırlanmışdır;
- "Xəmsə"yə daxil olan mətnlər üzrə tökenlərin (word tokens – söz istifadələrinin) və word type-ların (söz formalarının) ümumi sayı müəyyən edilmişdir. Ümumi token sayı: 290 476 Ümumi word type sayı: 45 453;
- Bütün tokenlərin tezlik siyahısı tərtib edilmişdir;
- Bütün söz istifadələri üzrə konkordans əldə edilmişdir;
- Ən çox tezliyi olan söz formaları müəyyənləşdirilmişdir.

M.Ə.Sabirin mətnlər korpusu. M.Ə.Sabirin ikicildlik "Hophopnamə"sində toplanmış əsərləri və satiraları əsasında hazırlanmışdır. M.Ə.Sabirin mətnlər korpusuna daxil edilmişdir: Satiralar – 274; Məktublar – 12; Hekayə və felyetonlar – 4; Tərcümələr – 5; Uşaq şeirləri və hekayələri – 22; Müxtəlif şeirlər – 11; Qəzəllər – 11; Məqalə və müxbir qeydləri – 38; Təxmislər – 37. Tədqiqat nəticəsində:

- M.Ə.Sabirin mətnlər korpusunun leksik vahidlərini əhatə edən verilənlər bazası (VB) tərtib edilmişdir. Bu baza həm hər bir ilin satiraları üzrə ayrıca, həm də ümumi şəkildə hazırlanmışdır;
- Hər təqvim ili üzrə ayrı-ayrı konkordanslar, həmçinin 1906–1911-ci illəri əhatə edən ümumi konkordans tərtib edilmişdir;
- Kontekstlərlə birlikdə təqdim olunan tezlik lüğətləri tərtib edilmişdir.
- M.Ə. Sabirin bütün əsərləri üzrə söz istifadələrinin ümumi sayı (word tokens) və söz formalarının sayı (word types) müəyyənləşdirilmişdir: Word tokens – 67 138 Word types – 21 361;
- Ən çox tezliyi olan söz formaları müəyyən edilmişdir;
- M.Ə. Sabirin dilinin bəzi xüsusiyyətləri aşkar edilmişdir;
- Ən az tezliyi olan leksik vahidlər çox zaman M.Ə. Sabirə məxsus fərdi söz istifadələrini təşkil etdiyi müəyyənləşdirilmişdir. Məsələn: uçitel, uşkola, uçeniklər və s. Sabir rus mənşəli sözləri ədəbi dilə daxil edərək onları fonetik, morfoloji və semantik baxımdan Azərbaycan dilinə uyğunlaşdırmışdır. Bu sözlərin fərdi istifadəsi yazıçının dil üslubu çərçivəsində baş vermiş və nəticədə xarici leksikanın milli kontekstdə assimilyasiyası təmin olunmuşdur.

Füzulinin poetik mətnlər korpusu. Hal-hazırda Məhəmməd Füzulinin poetik mətnlər korpusu onun Azərbaycan dilindəki "Divan"ı və "Leyli və Məcnun", "Tiryək və Şərab", "Meyvələrin söhbəti", "Şikayətnamə" poemalarından ibarətdir. Tədqiqat nəticəsində:

- Ümumi tokenlərin (word tokens), yəni söz istifadələrin və word type-ların (söz formaların) sayı müəyyən edilmişdir. Ümumi token sayı: 39000; ümumi word type sayı: 11872;

- Bütün söz istifadələr üzrə tezlik siyahısı tərtib edilmişdir;
- Bütün söz istifadələrinin konkordansı tərtib olunmuşdur;
- Ən çox tezliyi olan söz formaları müəyyən edilmişdir;
- M.Füzulinin qəzəllərindəki token (word tokens) və word type (sözformalar) sayları müəyyən edilmişdir. Token sayı: 31372; word type sayı: 9289;
- Füzulinin dilinin bəzi xüsusiyyətləri müəyyən olunmuşdur. Məsələn, Füzuli poeziyasında köməkçi feillərin (etmək, eyləmək, olmaq) geniş istifadəsi, əvəzliklərin bağlayıcı funksiyası və ərəb-fars mənşəli sözlərin Azərbaycan dilinin qrammatik qanunlarına uyğun birləşdirilməsi yolu ilə yeni mürəkkəb sözlərin yaradılması müşahidə olunur. Bu yanaşma Füzulinin xarici leksikanı milli dil kontekstinə uyğunlaşdırmaqla Azərbaycan bədii dilinin zənginləşməsinə töhfə verdiyini göstərir.

İmadəddin Nəsimin mətnlər korpusu. Korpusa İmadəddin Nəsiminin ikicildlik nəşrdə təqdim olunmuş doğma Azərbaycan dilində yazdığı qəzəlləri, ictimai-fəlsəfi şeirləri, müstəzadları, məsnəviləri və digər poetik nümunələri daxil edilmişdir. İ. Nəsiminin mətnlər korpusuna daxildir: Hazırkı vəziyyət: qəzəllər – 287, ictimai və fəlsəfi şeirlər – 155, müstəzadlar – 4, əliflam və tərs əlifba – 5, tərcibəndlər – 3, məsnəvilər – 2, müləmmə – 1, tuyuğlar – 1, əlavələr (əlavə əsərlər) – 15. Tədqiqat nəticəsində:

Ümumi tokenlərin (word tokens), yəni söz istifadələrin və word type-ların (söz formaların) sayı müəyyən edilmişdir;

Bütün söz istifadələr üzrə tezlik siyahısı tərtib edilmişdir;

Bütün söz istifadələrinin konkordansı tərtib olunmuşdur;

Ən çox tezliyi olan söz formaları müəyyən edilmişdir.

Azərbaycan yazıçılarının bədii mətn korpusunun texniki baxımdan formalaşdırılması aşağıdakı mərhələlər üzrə həyata keçirilmişdir:

Mətnlərin toplanması və təmizlənməsi. Əsərlərdən əlavə informasiyanın (nəşr tarixi, redaktə qeydləri və s.) silinməsi və yalnız ədəbi mətndən ibarət "saf" tekstlərin əldə olunması [7].

Formatlaşdırma və bölmələmə. Mətnlərin cümlə və abzas strukturlarının qorunması, hər bir sənədin metaetiketlərlə (müəllif, əsərin adı, janr, yazılma tarixi və s.) təmin edilməsi.

Avtomatlaşdırılmış linqvistik işarələmə – Stanza və digər NLP alətlərindən istifadə etməklə sözlərin morfoloji analizinin, lemmatizasiyasının və sintaktik strukturunun işarələnməsi.

Tezlik lüğətlərinin və statistik cədvəllərin yaradılması. Hər bir müəllif və əsər üzrə istifadə olunan leksik vahidlərin tezlik sıralaması və bu göstəricilərin müqayisəli təhlili.

Formalaşdırılmış korpus tədqiqatçılar üçün həm ümumi dil statistikası, həm də müəllif üslubu, üslub qrupları və janr strukturlarının müqayisəli şəkildə araşdırılması üçün zəngin bir resurs təqdim etmiş olur. Eyni zamanda, bu mənbə Azərbaycan dilinin rəqəmsal humanitar tədqiqatlar üçün istifadəsinə imkan verən ilk sistemli addımlardan biri kimi çıxış edəcəkdir.

Süni intellekt texnologiyalarının tətbiqi: alətlər və yanaşmalar

Tədqiqatda korpusların yaradılması və mətnlərin avtomatik təhlili üçün bir sıra süni intellekt və təbii dilin işlənməsi (Natural Language Processing – NLP) [11] texnologiyalarından istifadə olunmuşdur. Bu texnologiyalar mətnin struktur və məna səviyyəsində avtomatlaşdırılmış işlənməsinə, statistik və semantik məlumatların çıxarılmasına, üslubi xüsusiyyətlərin müəyyən edilməsinə imkan yaratmışdır [8].

İstifadə olunan əsas texnologiyalar aşağıdakılardır:

Stanza (Stanford NLP). Stanford Universiteti tərəfindən hazırlanmış Stanza aləti Azərbaycan dili üçün uyğunlaşdırılaraq istifadə edilmişdir. Bu alət vasitəsilə aşağıdakı mərhələlər həyata keçirilmişdir:

- Tokenləşdirmə (sözlərin ayrılması) [13];
- Cümlələrin seqmentasiyası;
- Lemmatizasiya (sözlərin əsas forma üzrə tanınması) [1];

- Morfoloji analiz (söz növü, hal, zaman və s.);
- Sintaktik analiz (cümlədəki sözlərarası əlaqələrin müəyyənləşdirilməsi).

Stanza-nın istifadəsi nəticəsində korpusa daxil olan mətnlər çoxsəviyyəli linqvistik işarələmələrlə təmin olunmuşdur.

Python proqramlaşdırma dili və Natural Language Toolkit (NLTK). Python dili bu tədqiqatın texniki bazasını təşkil etmişdir. NLTK, Pandas və digər Python kitabxanaları ilə birlikdə aşağıdakı funksiyalar yerinə yetirilmişdir [9, 10]:

- Mətnlərin emalı və təmizlənməsi;
- Tezlik lüğətlərinin yaradılması;
- Qrafik təsvirlər və statistik analizlər.

Tematik modelləşdirmə (Topic Modeling). Tematik modelləşdirmə üçün Latent Dirichlet Allocation (LDA) kimi modellərdən istifadə edilmişdir. Bu modellər vasitəsilə korpusdakı mətnlərdə əsas semantik mövzuların avtomatik şəkildə ayrılması mümkün olmuşdur [12]. Mövzuların müəyyənləşdirilməsi əsərlərin janr və məzmun baxımından təsnifatına, müəllifin fərdi maraq dairəsinin və ideya spektrinin aşkar edilməsinə xidmət edir.

Stilometriya və müəllif tanıma modelləri. Stilometriya – yazıçı üslubunun statistik xüsusiyyətlərinin öyrənilməsinə yönəlmiş metoddur. Cümlə uzunluqları, sintaktik quruluşlar, söz növlərinin tezliyi, unikal söz ehtiyatı kimi göstəricilər əsasında müəlliflərin yazı üslubları müqayisə olunmuş və fərqləndirici xüsusiyyətlər də müəyyən edilmişdir.

Əldə olunan nəticələr əsasında müəyyən yazıçıların fərdi dil üslubunu xarakterizə edən spesifik xüsusiyyətlər (məsələn, poetik sintaksis, emosional leksika, dialoq quruluşu və s.) təyin olunmuşdur.

Bu texnologiyaların tətbiqi nəticəsində Azərbaycan ədəbiyyatının rəqəmsal təhlili mümkün olmuş və ənənəvi ədəbiyyatşünas yanaşmalara əlavə olaraq, daha obyektiv və çoxaspektli təhlil imkanı yaranmışdır.

Tədqiqatın nəticələri göstərdi ki, süni intellekt (Sİ) texnologiyaları və korpus dilçiliyi metodları Azərbaycan yazıçılarının dil və üslub xüsusiyyətlərini sistemli və obyektiv şəkildə öyrənməyə imkan verir. Aparılan analizlər bir neçə istiqamət üzrə həyata keçirilmişdir: leksik və morfoloji xüsusiyyətlər, sintaksis və üslub, tematik modelləşdirmə və stilometriya.

Leksik və morfoloji xüsusiyyətlər. Tezlik lüğətləri göstərmişdir ki, klassik yazıçıların əsərlərində ərəb-fars mənşəli sözlərin payı yüksəkdir, müasir yazıçılarda isə milli leksika və texnoloji, sosial terminlər üstünlük təşkil edir. Feil və sifətlərin istifadəsi müəlliflər arasında fərqlilik göstərir; bəzi yazıçılar deskriptiv, digərləri isə hərəkət yönümlü ifadələrə üstünlük verirlər. Bu fərqliliklər Azərbaycan ədəbiyyatında klassik və müasir dil üslublarının müqayisəli öyrənilməsinə imkan verir.

Sintaksis və üslub. Klassik ədəbiyyatda uzun, mürəkkəb cümlələr və poetik sintaksis – inversiya, paralelizm və ritmik quruluşlar geniş yayılmışdır. Müasir yazıçılarda isə cümlələr daha qısa, sadə və danışıq dilinə yaxın olmaqla, təbii dialoqlar və realist nitq formaları ilə zəngindir. Bu fərqlər müəlliflərin ifadə vasitələrinin zaman və janr kontekstində necə dəyişdiyini göstərir.

Tematik modelləşdirmə. LDA modelləşdirilməsi nəticəsində müəyyən edilmişdir ki, klassik əsərlərdə milli azadlıq, dini-əxlaqi dəyərlər, sevgi və ictimai ədalət mövzuları üstünlük təşkil edir. Müasir əsərlərdə isə insan psixologiyası, şəhər həyatı, texnoloji dəyişikliklər və qloballaşma əsas mövzular kimi çıxış edir. Müəlliflər arasında mövzu spektri fərqlidir: Anarın əsərləri geniş tematik diapazonla seçilirsə, Mirzə İbrahimovun yaradıcılığı daha spesifik sahələrə – məsələn, kənd həyatının təsvirinə – fokuslanır.

Stilometriya və müəllif üslubu. Statistik üsullar müəlliflərin üslublarını fərqləndirməyə imkan vermişdir. Əsas göstəricilər unikal söz ehtiyatının həcmi, cümlə uzunluğunun orta göstəricisi, modal ifadələrin tezliyi və emosional-ekspressiv leksikanın istifadəsidir. Bu göstəricilər əsasında yazıçıların üslub "profilləri" hazırlanmış, əsərlər üzrə üslub sabitliyi və dəyişkənliyi müəyyən edilmişdir.

Milli Korpusun yaradılması perspektivləri və gələcək istiqamətlər

Azərbaycan dili üçün genişmiqyaslı Milli Korpusun yaradılması bu tədqiqatın ən mühüm strateji nəticələrindən biri kimi qiymətləndirilə bilər. Mövcud tədqiqat yalnız bədii mətnlərlə məhdudlaşsa da, aparılan işlərin metodoloji bazası və əldə olunan nəticələr ümummilli səviyyədə tətbiq edilə biləcək bir model təqdim edir. Milli Korpusun yaradılması üçün aşağıdakı komponentlər nəzərdə tutulmalıdır:

- Bədii mətnlər – klassik və müasir ədəbiyyat nümunələri;
- Publisistik mətnlər – qəzet və jurnal yazıları;
- Elmi mətnlər – humanitar və texniki sahələr üzrə;
- Danışıq dili nümunələri – şifahi çıxışlar, müsahibələr;
- Sosial media mətnləri – yeni kommunikasiya formalarının təhlili üçün.

Bu komponentlərin hər biri Azərbaycan dilinin müxtəlif üslub və funksional sahələrdəki istifadəsini əks etdirəcək və dilin dinamik inkişafını müşahidə etməyə imkan verəcəkdir.

Dilin müxtəlif qatlarının əhatə olunması. Milli Korpusun yaradılmasında həm ədəbi dil, həm də dialekt və şivələrin təmsil olunması vacibdir. Bu isə həm dialektologiya, həm də dil tarixi sahələrində mühüm tədqiqat imkanları açacaqdır.

Paralel və çoxdilli korpusların yaradılması. Tərcümə olunmuş ədəbiyyat və çoxdilli paralel korpuslar, Azərbaycan ədəbiyyatının beynəlxalq əlaqələrini və tərcümə proseslərini öyrənmək üçün əvəzolunmaz resursdur. Süni intellekt vasitəsilə bu cür korpuslarda: tərcümə strategiyaları; stilistik uyğunluq, mədəni elementlərin adaptasiyası kimi məsələlərin avtomatlaşdırılmış şəkildə təhlili mümkün olur.

Açıq mənbə platformalarının yaradılması. Gələcəkdə yaradılacaq Milli Korpusun tədqiqatçılar və müəllimlər üçün əlçatan olması məqsədilə açıq mənbəli onlayn platforma kimi fəaliyyət göstərməsi tövsiyə olunur. Bu təkcə ədəbiyyat və dilçilik tədqiqatçıları üçün deyil, həm də dil tədrisi və tərcümə sahəsində çalışanlar üçün mühüm mənbəyə çevrilə bilər.

Gənc tədqiqatçılar üçün tədqiqat imkanlarının genişləndirilməsi. Yeni texnologiyaların tətbiqi gənc tədqiqatçılar üçün də motivasiyaedici olacaq, onları interdisiplinar yanaşmalara – dilçilik, informatika və ədəbiyyatşünaslığın sintezində araşdırmalara yönəldəcəkdir. Süni intellektin humanitar sahələrə inteqrasiyası bu sahələrin innovativ inkişafına təkan verir.

Nəticə */ Conclusion*

Məqalədə Azərbaycan yazıçılarının bədii mətnlərinə əsaslanan korpusların yaradılması və bu prosesdə süni intellekt (Sİ) texnologiyalarının tətbiqi məsələləri araşdırılmışdır. Tədqiqat nəticəsində sübut edilmişdir ki, Sİ alətləri və korpus dilçiliyi metodları Azərbaycan dilinin və ədəbiyyatının daha dərindən və sistemli şəkildə öyrənilməsi üçün yeni imkanlar yaradır. Aparılan işlər nəticəsində:

- Azərbaycan yazıçılarının mətnlərinə əsaslanan ilk tematik, leksik və üslubi göstəricilərlə zəngin bədii korpus hazırlanmışdır;
- Fərdi müəllif korpusları tərtib edilmişdir: M.Ə. Sabirin mətnlər korpusu, Füzulinin poetik mətnlər korpusu, Nizami Gəncəvinin mətnlər korpusu, Füzulinin poetik mətnlər korpusu, İmadəddin Nəsimin mətnlər korpusu;
- Leksik tezlik lüğətləri və stilometrik təhlillər (qismən) vasitəsilə yazıçıların fərdi dil və üslub xüsusiyyətləri müəyyən edilmişdir;
- Tematik modelləşdirmə metodları ilə mətnlərin əsas mövzu sahələri müəyyən olunmuş, ədəbi inkişaf dinamikası təhlil edilmişdir;
- Çoxdilli korpus modelləri əsasında orijinal və tərcümə mətnlərin müqayisəli analizi həyata keçirilmişdir;

- Bu tədqiqatın nəticələri Azərbaycan dilinin Milli Korpusunun yaradılması üçün elmi və texnoloji baza formalaşdırmışdır.

Tədqiqat göstərmişdir ki, müasir texnologiyaların humanitar elmlərə inteqrasiyası – xüsusilə dil və ədəbiyyat sahəsində – həm elmi yeniliklərə yol açır, həm də ənənəvi yanaşmaların effektivliyini artırır. Azərbaycan dilinin rəqəmsal məkanlarda daha geniş təmsil olunması, dilin və ədəbiyyatın beynəlxalq elmi müstəvidə araşdırılması üçün belə korpusların rolu əvəzsizdir.

Gələcəkdə bu cür tədqiqatların davam etdirilməsi, korpusların daha da genişləndirilməsi, açıq platformaların yaradılması, bədii və tərcümə mətnləri ilə yanaşı, dialekt materiallarının da daxil edilməsi Azərbaycan dilçiliyinin və ədəbiyyatşünaslığının yeni mərhələyə keçməsinə imkan verəcəkdir. Bu tədqiqat mətn korpuslarının və tezlik lüğətlərinin yaradılmasında pioner iş olmaqla, Azərbaycan dilinin Milli Korpusunun formalaşdırılması üçün mühüm əhəmiyyət daşıyır.

Ədəbiyyat / *References*

1. Баранов А.А., Скуратовская Л.И. Автоматическая лемматизация и морфологическая разметка текстов на русском языке. Вопросы языкознания, 2015, № 6.
2. Захаров В.П. Корпусная лингвистика: Учебно-метод. пособие. – СПб.,1 2005, 48 с.
3. Корпус текстов А.П. Чехова [Электронный ресурс] URL: https://lex.philol.msu.ru/proekty/istok-korpus-chehova/
4. Кунин А.В. Стилистика и корпусный анализ художественного текста. Вестник МГУ, 2005
5. Лаборатория Общей и Компьютерной лексикологии и лексикографии [Электронный ресурс] URL: http://www.philol.msu.ru/~lex/main.htm
6. Национальный корпус русского языка [Электронный ресурс] URL: https://ruscorpora.ru/new/instruction-main.pdf
7. Халилова Г.А. «Развитие азербайджанской корпусной лингвистики. Создание корпуса текстов азербайджанских авторов (поэтов и писателей) и национального корпуса азербайджанского языка». Вестник Дагестанского государственного университета. Серия2: Гуманитарные науки 2024 / 2. [Электронный ресурс] /URL: https://vestnik.dgu.ru/pol.aspx?razdel=2&artId=4670
8. Шмелёв А.Д. Корпусная лингвистика и лексикография. – Москва: Языки славянской культуры, 2003.
9. Bird S., Klein E. & Loper E. Natural Language Processing with Python. O'Reilly Media, 2009
10. Evert S. Corpora and collocations. In: Lüdeling A., & Kytö M. (eds.), Corpus Linguistics. – Berlin: Mouton de Gruyter, 2008.
11. Jurafsky D., & Martin J.H. Speech and Language Processing. 3rd Edition Draft. Stanford University, 2021.
12. Kilgarriff A., & Grefenstette G. Introduction to the Special Issue on the Web as Corpus. Computational Linguistics, 2003, 29(3).
13. McEnery T., & Hardie A. Corpus Linguistics: Method, Theory and Practice. – Cambridge: Cambridge University Press, 2012

Методы и технологии искусственного интеллекта (ИИ) и корпусной лингвистики при формировании корпуса художественных текстов азербайджанского языка

Гамида Халилова
Филиал Московского государственного университета в г. Баку
Институт языкознания им. Насими НАНА
E-mail: hamida.khalilova.a@gmail.com
Orcid ID: 0000-0002-7423-8850

Резюме. Статья посвящена применению технологий искусственного интеллекта (ИИ) при создании корпуса художественных текстов азербайджанского языка. В рамках исследования составлен корпус художественных текстов азербайджанского языка, а также корпуса текстов азербайджанских писателей. Актуальность и новизна исследования заключаются в использовании методов корпусной лингвистики и лингвостатистики, технологий искусственного интеллекта при создании корпуса текстов и частотных словарей, а также в правильном отборе исследовательского материала. Следует особо отметить, что азербайджанский язык практически не представлен корпусами текстов. Данное исследование является пионерским как в создании корпусов текстов азербайджанского языка, так и корпусов художественных текстов и частотных словарей на азербайджанском языке, а также имеет важное значение для формирования Национального корпуса азербайджанского языка. В статье также рассматриваются автоматизированная разметка, тематическое моделирование, лемматизация и другие современные подходы. Кроме того, подчёркивается, что методы корпусной лингвистики и технологии ИИ позволяют классифицировать тексты, анализировать стиль конкретного писателя, определять авторство произведений и изучать другие связанные аспекты. Применение современных технологий предоставляет литературоведам и лингвистам возможность более глубоко и всесторонне исследовать творчество авторов, выявлять закономерности в их произведениях и использовать методы корпусной лингвистики в литературоведческом анализе. Следует отметить и то, что создание авторских корпусов также носит новаторский характер. В России созданы корпуса текстов: «Корпус текстов А.С.Пушкина», «Корпус текстов А.П.Чехова», в Великобритании – «Корпус текстов Шекспира». Авторские корпуса позволяют изучать особенности индивидуальной языковой личности. В текстах одного автора язык проявляется как целостная и единая система. В статье представлены составленные корпуса текстов азербайджанский авторов: Корпус текстов М.А. Сабира, Корпус текстов Н. Гянджеви, М. Физули и И. Насими. Кроме того, в статье говорится о преимуществах и пользе применения ИИ при создании корпусов, основанных на многоязычных параллельных текстах, что чрезвычайно важно для литературоведения. Это откроет возможности для сравнения художественных текстов как на языке оригинала, так и в переводах, позволит глубже анализировать особенности переводов, выявлять различия интерпретаций, а также изучать стилистические и культурные адаптации произведений. Такой подход полезен не только для лексического и грамматического анализа, но и для исследования авторского стиля, эмоционально-экспрессивной окраски текста и других аспектов. Таким образом, исследование показывает, что технологии искусственного интеллекта и методы корпусной лингвистики расширяют возможности гуманитарных наук и открывают новые перспективы для изучения азербайджанского языка и литературы.

Ключевые слова: корпус, искусственный интеллект, текстовый корпус, художественная литература, конкорданс, азербайджанский язык, лемматизация, токен, авторский корпус