AMEA Nəsimi adına Dilçilik İnstitutu,
Dilçilik araşdırmaları, 2025, №2 *ISSN: 2664-5432*
Səhifə: 3-15
https://doi.org/10.59849/2664-5432.2025.2.3

Nadir MƏMMƏDLİ*
Məsud MAHMUDOV*
İlham TAHİROV*

MİLLİ KORPUS VƏ LEKSİK İNFRASTRUKTUR: AZƏRBAYCAN DİLİNİN SÖZ BAZASININ YARADILMASI TƏCRÜBƏSİ

XÜLASƏ

Məqalədə Azərbaycan dilinin söz bazasının hazırlanması ilə bağlı həyata keçirilən nəzəri və praktiki mərhələlər sistemli şəkildə təhlil edilmiş, dil resurslarının formalaşdırılması üçün zəruri olan texnoloji və linqvistik yanaşmalar əsasında söz bazasının qurulması prinsipləri ərsəyə gətirilmişdir. Söz bazası: a) Azərbaycan dilinin milli korpusunun əsas struktur komponentlərindən biri kimi nəzərdən keçirilmişdir; b) leksik vahidlərin mənbə əsasında avtomatik və yarıavtomatik emalı təcrübəsi əsasında yaradılmışdır; c) fərqli funksional üslubları, çoxsaylı mənbələri və böyükhəcmli materialları əhatə etməsi baxımından, mövcud korpustəməlli təşəbbüslərlə müqayisədə daha geniş miqyaslıdır. Müəlliflər hesab edirlər ki, strukturlaşdırılmış leksik ehtiyatları təmin edən belə bir verilənlər bazası həm akademik tədqiqatlar, həm də texnoloji tətbiqlər üçün zəruridir. Araşdırma zamanı müxtəlif funksional üslublara məxsus 520 milyon söz-forma bazası toplanmış, xüsusi proqram təminatı ilə təmizlənmiş və strukturlaşdırılmışdır. Avtomatik emal nəticəsində ilkin mərhələdə 2.918.910 söz-forma müəyyən olunmuş, sonrakı bir neçə mərhələdə texniki və leksik filtrləmə nəticəsində baza təmizlənmişdir. Son nəticədə 175.521 leksik vahiddən ibarət tezlik və əlifba sözlükləri tərtib olunmuşdur. Eyni zamanda, orfoqrafiya lüğətində yer almayan 93.287 söz müəyyənləşdirilərək ayrıca siyahı şəklində təqdim olunmuşdur. Məqalədə milli korpusun strukturu, onun bölmələri (bədii, publisistik, elmi, rəsmi, şifahi və tədris mətnləri), konkordanslar və linqvistik analizatorlar geniş şəkildə şərh olunur. Söz bazası həm nəzəri tədqiqatlar, həm də təbii dilin emalı, avtomatik tərcümə, səsləndirmə sistemləri və süni intellekt modelləri üçün strateji resurs kimi təqdim olunur. Məqalədə Azərbaycan dilinin rəqəmsal leksik resurslarının formalaşması istiqamətində mühüm elmi və tətbiqi nəticələr ortaya qoyulur.

Açar sözlər: *süni intellekt, korpus dilçiliyi, milli korpus, söz bazası, təbii dilin emalı, statistik leksikoqrafiya, konkordans.*

* AMEA Nəsimi adına Dilçilik İnstitutunun direktoru, professor. Email: nurlan1959@gmail.com
* AMEA Nəsimi adına Dilçilik İnstitutu, Hind-Avropa dilləri şöbəsinin müdiri, professor. Email: ilham_tahir@rambler.ru
* AMEA Nəsimi adına Dilçilik İnstitutu, Süni intellekt və kompüter dilçiliyi şöbəsinin müdiri, professor. Email: mmasud@bk.ru

Əsaslandırma və Kontekst

Azərbaycan Milli Elmlər Akademiyasının 21 oktyabr 2024-cü il tarixli qərarı ilə "AMEA-nın 2025–2030-cu illər üzrə İnkişaf Konsepsiyası və Yol Xəritəsi" təsdiq edilmişdir. Bu sənədin əsas məqsədi ölkə elminin müasir çağırışlara uyğun inkişafını təmin etməkdir. Konsepsiyada süni intellekt, rəqəmsal və ağıllı texnologiyaların humanitar və ictimai elmlərlə inteqrasiyası prioritet istiqamətlərdən biri kimi müəyyən olunmuşdur.

Bu kontekstdə AMEA-nın müxtəlif institutları, o cümlədən Nəsimi adına Dilçilik İnstitutu qarşısında bir sıra mühüm vəzifələr qoyulmuşdur. Bu vəzifələr arasında Azərbaycan dilinin rəqəmsal mühitdə təmsil olunması, texnoloji tətbiqlərə uyğunlaşdırılması, müxtəlif lüğət və korpusların hazırlanması, süni intellekt əsaslı dilçilik tədqiqatlarının aparılması xüsusi əhəmiyyət daşıyır.

Nəsimi adına Dilçilik İnstitutu bu çağırışlara cavab olaraq aşağıdakı istiqamətlər üzrə fəaliyyət göstərir:

1. Rəqəmsallaşdırma və texnoloji uyğunlaşma:
- Azərbaycan dilinin internet məkanında təmsil olunması, yerləşdirilmə və istifadə imkanlarının genişləndirilməsi;
- Azərbaycan dilinin elektron lüğətlərinin hazırlanması və mövcud lüğətlərin rəqəmsal versiyalarının təkmilləşdirilməsi;
- Azərbaycan dilinin tədrisinə dair onlayn dərslik və texnologiyaların hazırlanması.

2. Elmi-tədqiqat resurslarının inkişafı:
- Azərbaycan yazıçıları və şairlərinin əsərləri, dilimizin yazılı abidələri əsasında konkordansların hazırlanması;
- Azərbaycan dilinin milli korpusunun müxtəlif bölmələrinin tərtibi və ilkin versiyasının internetdə istifadəyə verilməsi;
- süni intellekt kontekstində dilin öyrənilməsi və bu istiqamətdə tədqiqatların genişləndirilməsi.

3. Kadr hazırlığı və beynəlxalq əməkdaşlıq:
- dilçiliyin yeni sahələri üzrə elmi kadrların hazırlanması;
- xarici, ölkə alimləri, eləcə də yerli ali təhsil müəssisələri ilə birgə elmi layihələrin icrası və beynəlxalq əməkdaşlığın gücləndirilməsi.

Bu kompleks tədbirlər çərçivəsində Azərbaycan dilinin söz bazasının yaradılması xüsusi əhəmiyyət kəsb edir. Məqsəd dilin leksik ehtiyatını sistemli şəkildə toplamaq və bu materialı həm elmi-tədqiqat, həm də texnoloji sahələrdə istifadəyə təqdim etməkdir.

Söz bazasının yaradılması prosesi Azərbaycan dilinin milli korpusunun və konkordansların hazırlanması ilə sıx bağlıdır. Korpus dilçiliyi dili təbii mətnlər əsasında öyrənmək və təhlil etmək üçün mətnlərin elektron formada toplanmasını nəzərdə tutan müasir dilçilik istiqamətidir. Milli korpusda dilin müxtəlif janrları, üslubları, dialektləri və digər xüsusiyyətləri əhatə olunur. Bu tip korpuslar tədqiqatçılara operativ və etibarlı linqvistik məlumat təqdim edən mühüm alətlərdir.

Ənənəvi dilçilikdən fərqli olaraq, korpustəməlli yanaşma dili onun real istifadəsi şəraitində öyrənməyə imkan verir. Axtarış sistemləri və təhlil alətləri vasitəsilə tədqiqatçılar istənilən leksik və ya qrammatik vahidi tez və dəqiq müəyyən edə bilirlər. Korpusun səmərəliliyi onun əhatə etdiyi mətnlərin müxtəlifliyi və miqyasından asılıdır.

Beynəlxalq təcrübədə bu sahədə nüfuzlu nümunələr mövcuddur:

- Britaniya Milli Korpusu (British National Corpus – BNC) – həm yazılı, həm də şifahi mətnlərdən ibarət olub, təxminən 100 milyon söz-formanı əhatə edir;
- Amerika Milli Korpusu (American National Corpus – ANC) – 22 milyon söz-formalı korpusdur, müxtəlif növ mətni və nitq nümunəsini özündə birləşdir;
- Türkçe Ulusal Derlem (TUD) – Türkiyədə hazırlanmış, 50 milyon söz-formadan ibarət, 39 müxtəlif janrı (elmi məqalə, roman, məktub, bloq və s.) əhatə edən korpusdur. Layihə TUBİTAK tərəfindən dəstəklənmişdir.

Azərbaycan dilinin milli korpusu üzrə son illərdə mühüm işlər görülmüşdür. [2;5;8;9] Müasir ədəbi dil nümunələri, klassik ədəbiyyat və yazılı abidələr əsasında aparılan statistik-linqvistik tədqiqatlar bu sahədə əsaslı baza formalaşdırır. Tezlik və əks lüğətlərin hazırlanması da bu işin mühüm tərkib hissəsidir [1; 3;4; 10].

Söz bazası, korpus və konkordansların sintezi Azərbaycan dili üçün vahid linqvistik resursun formalaşmasını təmin edir. Konkordanslar konkret müəllif və ya mənbənin dil üslubunu, leksik ehtiyatını ortaya qoyduğu halda, söz bazası dilin funksional üslublar üzrə tezlik göstəricilərini və ümumi leksik mənzərəsini təqdim edir. Bu resurslar, həm nəzəri dilçilik, həm də tərcümə sistemləri, dil tədrisi və süni intellekt tətbiqləri üçün çoxşaxəli istifadə imkanları yaradır.

Konkordans layihələri

AMEA Nəsimi adına Dilçilik İnstitutunun Süni intellekt və kompüter dilçiliyi şöbəsində görkəmli Azərbaycan şairləri Hüseyn Cavid, Mikayıl Müşfiq və Sabir Rüstəmxanlının əsərlərinin, dilimizin möhtəşəm abidəsi "Kitabi-Dədə Qorqud"un mükəmməl konkordansları hazırlanmışdır [11]. Hüseyn Cavidin seçilmiş əsərlərinin beşcildliyində ümumi söz sayının 186864 (I cild – 30356, II cild - 44284, III cild – 42 558, IV cild – 19680, V cild – 49986), Mikayıl Müşfiqin əsərlərində 77106, "Kitabi-Dədə Qorqud"da 31902, Sabir Rüstəmxanlının 15 cildliyində 1777566 (I cild – 34583, II cild – 36647, III cild – 37136, IV cild – 61005, V cild – 137369, VI cild – 140240, VII cild – 161690, VIII cild – 160443, IX cild – 149281, X cild – 157171, XI cild – 174290, XII cild – 142615, XIII – 154080, XIV – 149840, XV – 71176) olduğu müəyyən edilib.

Konkordansların tərtibi sahəsində Azərbaycan dilçiliyində toplanmış təcrübəyə əsaslanaraq cari və gələcək səyləri yazılı tarixi abidələrin dilini, eləcə də klassik və müasir Azərbaycan müəlliflərinin ədəbi əsərlərinin dilini əhatə edən yeni konkordansların hazırlanması ilə bu tədqiqat sahəsinin inkişafına yönəltmək faydalı olardı. Növbəti illərdə Azərbaycan ədəbiyyatının görkəmli nümayəndələrinin (Seyid Əzim Şirvani, Qasım bəy Zakir, Mirzə Ələkbər Sabir, Mir Cəlal, Rəsul Rza və b.) seçilmiş əsərləri əsasında konkordansların tərtib olunması nəzərdə tutulur. Bu

sahədə tezliklə göstərilməli mühüm təşəbbüslərdən biri Mahmud Kaşqarinin "Divanü Lüğət-it-Türk" kimi məşhur ümumtürk yazılı abidəsinin (ərəb dilindən) Azərbaycan dilinə (professor Nadir Məmmədli tərəfindən tapılmış, redaktə edilmiş, şərhlər yazılmış və nəfis şəkildə nəşr edilmiş) tərcüməsi üçün konkordansın yaradılması ola bilər. Belə bir möhtəşəm abidənin orijinaldan Azərbaycan dilinə tərcüməsi üçün konkordansın hazırlanaraq rəqəmsal mühitə daxil edilməsi ümumtürk leksikoqrafiyası və tarixi dilçilik sahəsində gələcək tədqiqatlara yeni perspektivlər vəd edir.

Son illərdə süni intellekt, mətnin avtomatik emalı, səsləndirilməsi sistemləri, inteqrasiya olunmuş elektron lüğətlər korpusu kimi texnologiyalarla yanaşı, linqvistik analizatorların hazırlanması sahəsində də mühüm addımlar atılmışdır. Bu istiqamətdə tədqiqatların genişləndirilməsi xüsusi əhəmiyyət daşıyır.

Azərbaycan dilinin milli korpusunun yaradılması üçün ilkin elmi zəmin formalaşdığından, artıq bu işlərə daha geniş miqyasda başlamaq mümkündür. Bunun üçün isə genişhəcmli və müxtəlif sahələri əhatə edən mətn materialına ehtiyac vardır. Korpusun etibarlılığı onun həcmi və balanslı strukturu ilə sıx bağlıdır.

Azərbaycan dilinin milli korpusunun struktur modeli

Azərbaycan dilinin milli korpusunun yaradılması istiqamətində mövcud zəmin nəzərə alınaraq, daha geniş miqyaslı, sistemli və balanslı mətn bazasının formalaşdırılması zəruri hesab olunur. Bu məqsədlə korpusun əsas strukturu [6, 60-63; 7, 112-118] aşağıdakı istiqamətlər üzrə təklif edilir:

1. Əsas korpus

a) *Bədii mətnlər:* nəzm və nəsr nümunələri, dramaturgiya, folklor mətnləri, klassik və müasir ədəbiyyat;

b) *Publisistik mətnlər:* qəzet və jurnal materialları, bloq yazıları, internet məqalələri;

c) *Rəsmi və işgüzar üslub mətnləri:* dövlət sənədləri, çıxışlar, sərəncamlar, qanunvericilik aktları;

d) *Elmi və texniki mətnlər:* humanitar, ictimai, texniki və təbiət elmlərinə dair yazılar;

e) *Dini və fəlsəfi mətnlər:* müxtəlif dinlərə və ideoloji cərəyanlara aid yazılı mətnlər;

f) *Dialektoloji mətnlər:* dialekt və şivə xüsusiyyətlərini əks etdirən yazılı və şifahi nümunələr.

2. Elektron lüğətlər altkorpusu

Azərbaycan dilində mövcud olan müxtəlif tipli lüğətlərin elektron versiyaları bu altkorpusa daxil edilir: sinonimlər, antonimlər, omonimlər, frazeoloji, terminoloji, dialektoloji, ikidilli və çoxdilli, ensiklopedik, izahlı, orfoepik və statistik lüğətlər, tezlik və əks əlifba, orfoqrafiya lüğətləri.

3. Şifahi mətnlər altkorpusu

a) Məişət və gündəlik danışıq nümunələri
b) Rəsmi çıxışlar və müsahibələr
c) Dialektlərlə bağlı audio-yazılı materiallar

4. Paralel korpuslar

a) İkidilli: Azərbaycan-ingilis, Azərbaycan-rus, Azərbaycan-alman və s.
b) Çoxdilli: eyni mətnlərin bir neçə dildə tərcümələri
c) Müqayisəli təhlil üçün müxtəlif dillərdə tərcümə variantları

5. Tədris altkorpusu

a) Orta və ali məktəblər üçün dərs vəsaitləri
b) Tədris materialları, metodik vəsaitlər və imtahan testləri

6. Konkordanslar

a) Azərbaycan ədiblərinin və klassiklərin əsərləri əsasında hazırlanmış konkordanslar
b) Azərbaycan dilinin yazılı abidələrinin dili əsasında hazırlanmış konkordanslar

Konkordanslar leksik vahidlərin həm semantik, həm də üslubi təhlili üçün əhəmiyyətli mənbədir

7. Linqvistik analizatorlar və proqram təminatı

a) Morfoloji, sintaktik, semantik və leksik səviyyələr üzrə təhlil modulları
b) Mətnin səsləndirilməsi və nitqdən-mətnə çevrilmə texnologiyaları
c) Axtarış sistemləri və istifadəçi interfeysi ilə təmin olunmuş platforma

Belə bir struktur Azərbaycan dilinin milli korpusunun təkcə dilçilik tədqiqatları üçün deyil, həm də süni intellekt, təbii dilin emalı, tədris texnologiyaları və avtomatik tərcümə sistemləri üçün də universal tətbiq imkanlarını təmin edir.

Azərbaycan dilinin söz bazasının yaradılması

Azərbaycan dilinin milli korpusunun əsas komponentlərindən biri olaraq leksik bazanın yaradılması statistik leksikoqrafiya, korpus dilçiliyi və təbii dilin emalı sahələrinin sintezində həyata keçirilir. Bu təşəbbüsün məqsədi müxtəlif funksional üslub və janrlarda istifadə olunan mətnlərə əsaslanaraq, tezlik və semantik təhlilə söykənən, statistik cəhətdən reprezentativ leksik inventar formalaşdırmaqdır. Leksik bazanın müasir dilin real istifadəsini əks etdirməsi üçün mətnlərin sistemli şəkildə toplanması və emalına xüsusi yanaşma tətbiq olunur. Bu məqsədlə uyğun proqram təminatı və texnoloji vasitələrdən istifadə edilir.

Mətn toplanması və ilkin emal. Leksik bazanın formalaşdırılmasında Azərbaycan dilinin leksik və üslubi zənginliyini əks etdirən genişspektrli resurslardan istifadə edilmişdir. Seçilmiş mənbələr rəsmi, hüquqi-inzibati, kütləvi, publisistik, bədii, elmi mətn və yazıları əhatə etməklə, korpusun tematik və funksional baxımdan balanslaşdırılmasına xidmət edir. Toplanmış mətnlər ilkin mərhələdə texnoloji vasitələrlə strukturlaşdırılır, təmizlənir və təhlilə uyğun formata salınır. Bu mərhələ

leksik və statistik göstəricilərin etibarlılığını təmin edir və əsas baza rolunu oynayır.

İşin ilk mərhələsində müxtəlif funksional üslubları əhatə edən yazılı mətnlər aşağıdakı mənbələrdən toplanmışdır:

1. Prezident Administrasiyası və rəsmi nəşrlər:

https://president.az (Azərbaycan Respublikası Prezidentinin rəsmi saytı);

https://xalqqazeti.az/az (“Xalq” qəzeti – rəsmi hökumət nəşri);

https://www.azerbaijan-news.az/az (“Azərbaycan” qəzeti – parlamentin rəsmi orqanı);

https://e-qanun.az (Ədliyyə Nazirliyinin Qanunvericilik Bazası)

2. Rəsmi dövlət orqanlarının saytları (nazirliklər və komitələr):

https://arx.com.az (Dövlət Şəhərsalma və Arxitektura Komitəsi);

https://culture.gov.az (Mədəniyyət Nazirliyi);

https://eco.gov.az (Ekologiya və Təbii Sərvətlər Nazirliyi);

https://justice.gov.az/ (Ədliyyə Nazirliyi);

https://fhn.gov.az (Fövqəladə Hallar Nazirliyi);

https://minenergy.gov.az (Energetika Nazirliyi);

https://mfa.gov.az (Xarici İşlər Nazirliyi);

https://mod.gov.az (Müdafiə Nazirliyi);

https://njustice.gov.az (Naxçıvan MR Ədliyyə Nazirliyi);

https://scara.gov.az (Dini Qurumlarla İş üzrə Dövlət Komitəsi);

https://science.gov.az (Azərbaycan Milli Elmlər Akademiyası);

https://sosial.gov.az (Əmək və Əhalinin Sosial Müdafiəsi Nazirliyi);

https://stat.gov.az (Dövlət Statistika Komitəsi).

3. Mədəniyyət, diaspor və ictimai təşkilatlar:

https://www.millikitabxana.az/ (Azərbaycan Milli Kitabxanası);

https://medeniyyet.az (Mədəniyyət xəbərləri);

https://diaspor.az (Xaricdəki azərbaycanlılar və diaspor fəaliyyəti);

http://azyb.az (Azərbaycan Yazıçılar Birliyi).

4. ***Xəbər (informasiya) agentlikləri:***

https://apa.az (APA – Azərbaycan Press Agentliyi);

https://azertag.az (AZƏRTAC – Dövlət İnformasiya Agentliyi);

https://report.az (Report İnformasiya Agentliyi;

5. Xəbər portalları:

https://aznews.az (AzNews xəbər portalı);

https://bakupost.az (BakuPost xəbər portalı);

https://lent.az (Xəbər və informasiya portalı);

https://telejurnal.az (Media və jurnalistika xəbərləri portalı).

6. ***Qəzetlər və jurnallar*** (***çap və ya onlayn media orqanları***):

https://525.az (“525-ci qəzet”);

https://baki-xeber.com (Bakı-Xəbər qəzeti);

https://science.gov.az/az/forms/archive/3873 (“Elm” qəzeti);

https://edebiyyatqazeti.az (“Ədəbiyyat” qəzeti);
https://adalet.az/az (“Ədalət” qəzeti);
https://hurriyyet.az/az (“Hürriyyət” qəzeti);
https://ikisahil.az (“İki Sahil” qəzeti);
https://musavat.com (Onlayn ictimai-siyasi qəzet);
https://respublika-news.az (“Respublika” qəzeti);
https://sesqazeti.az (“Səs” qəzeti);
https://www.yeniazerbaycan.com (“Yeni Azərbaycan” qəzeti);
“Azərbaycan” jurnalı;
“Ulduz” jurnalı;
“Qobustan” jurnalı.

7. Bədii ədəbiyyat və elmi irs: Azərbaycan yazıçı, şair və alimlərinin əsərlərinin toplanması istiqamətində işlər aparılmışdır:

HTML, PDF, DOC formatında əsərləri toplanan *yazıçı və şairlər:* Anar, Afaq Məsud, Aslan Guliyev, Bəxtiyar Vahabzadə, Cəfər Cabbarlı, Cəlil Məmmədquluzadə, Çingiz Abdullayev, Elçin, Əkrəm Əylisli, Əlağa Kürçaylı, Əli Əmirli, Əli Kərim, Əli Vəliyev, Əlibala Hacızadə, Ənvər Məmmədxanlı, Əzizə Cəfərzadə, Fərman Kərimzadə, Fikrət Qoca, Firuz Mustafa, Flora Xəlilzadə, Xanımana Əlibəyli, Hüseyn Cavid, Hüseyn Kürdoğlu, İlyas Əfəndiyev, İsa İsmayılzadə, İsa Hüseynov, İsmayıl Şıxlı, Qılman İlkin, Kamal Abdulla, Kərim Dünyamalı, Mahirə Nağıqızı, Maqsud İbrahimbəyov, Mehdi Hüseyn, Mədinə Gülgün, Məmməd Aslan, Məmməd İsmayıl, Mikayıl Müşfiq, Mir Cəlal, Mirzə İbrahimov, Mirzə Fətəli Axundzadə, Mövlud Süleymanlı, Musa Yaqub, Nəbi Xəzri, Nəriman Həsənzadə, Nigar Rəfibəyli, Nüsrət Kəsəmənli, Nurəddin Ədiloğlu, Ramiz Rövşən, Rəsul Rza, Rüstəm Behrudi, Rüstəm İbrahimbəyov, Sabir Ahmadli, Sabir Rüstəmxanlı, Sabit Rəhman, Salam Qədirzadə, Seyran Səxavət, Səməd Vurğun, Söhrab Tahir, Süleyman Rəhimov, Süleyman Rüstəm, Tofiq Bayram, Vaqif Səmədoğlu, Yusif Səmədoğlu, Zəlimxan Yaqub.

Alimlər: Ağamusa Axundov, Bəkir Nəbiyev, Əziz Mirəhmədov, İsa Həbibbəyli, Qəzənfər Kazımov, Qəzənfər Paşayev, Qulu Xəlilov, Mir Cəlal, Nadir Məmmədli, Nəsiman Yaqublu, Nizami Cəfərov, Şirindil Alışanov, Tehran Əlişanoğlu, Yaqub Mahmudov, Yaşar Qarayev, Yusif Seyidov, Ziya Bünyadov.

Ümummilli Lider Heydər Əliyevin “Müstəqilliyimiz əbədidir” (46 cild) və Prezident İlham Əliyevin “İnkişaf məqsədimizdir” (125 cild) çoxcildlik topluları müasir Azərbaycan dilində siyasi-ideoloji diskursu əks etdirən əsas mənbələr kimi mətn korpusuna daxil edilmişdir.

Müxtəlif üslubları ehtiva edən qəzet, sayt, onlayn media materialı son *10 illik dövrü* əhatə etmişdir.

Proqram təminatı ilə emal. Mətn bazasının toplanmasından sonrakı mərhələdə aşağıdakı əməliyyatlar proqramlaşdırılmış şəkildə həyata keçirilmişdir:

1. *Söz-formaların ayrılması və tezlik sıralaması:* 2.918.910 söz-forma əsasında tezlik siyahısı hazırlanmışdır.

2. *Səhv sözlərin və texniki qalıqların təmizlənməsi:* təxminən 126.000 səhv söz-forma ayırd edilib silinmişdir.

3. *Əlavələr bloku üzrə onomastik vahidlərin bazadan silinməsi:*

- Əlavə 1 – Azərbaycan coğrafi adları: 46.346 söz-forma
- Əlavə 2 – Dünya coğrafi adları: 13.878 söz-forma
- Əlavə 3 – Azərbaycan şəxs adları: 24.009 söz-forma
- Əlavə 4 – Beynəlxalq şəxs adları: 5.392 söz-forma

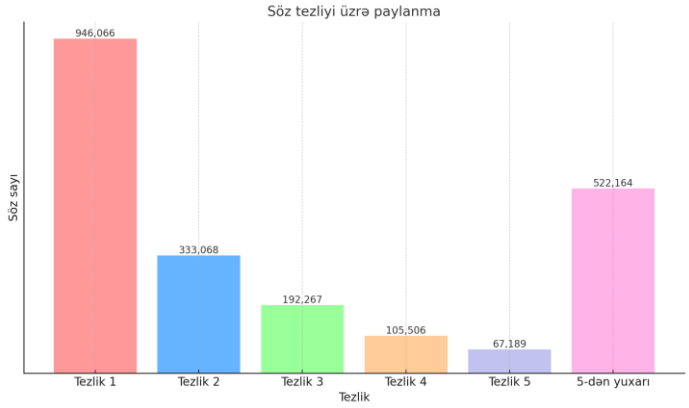
4. *Rəqəm və işarələrin avtomatik silinməsi:* rəqəmlər (22.299.902), hərfli kodlar (2.479.123)

5. *Defis ilə yazılmış sözlərin düzəldilməsi və əlavə edilməsi:*

- Əvvəlində defis olan sözlərdən: + 5.122 söz-forma
- Sonunda defis olanlardan: + 14.014 söz-forma

6. Sözügedən silinmə və əlavəolunma əməliyyatlarından sonra söz bazasında 2.915.817 söz-forma qalmışdır.

Kök və əsaslara ayrılma. Proqram təminatının imkan verdiyi qədər söz kökləri və əsaslar avtomatik ayrılmış, mümkün olmayanlar isə əl üsulu ilə təhlil edilmişdir. Bu mərhələdə söz-formaların leksemlərə çevrilməsi prosesi davam etdirilmiş və dəqiqləşdirilmiş tezlik göstəriciləri yenidən siyahıya salınmışdır. Üzərində əl ilə təmizlənmə və düzəlişlər aparılandan sonra Söz bazasının statistik göstəriciləri aşağıdakı qrafikdə (Bax: Şəkil 1) göstərildiyi kimi olmuşdur. Bazadakı söz-formaların ümumi sayı 2.171.260 təşkil edir. Buraya həm leksik, həm də qeyri-leksik vahidlər (xarici sözlər, səhvlər, abreviaturlar və s.) daxil ola bilər.



Şəkil 1

Tezliyi 1 olan sözlər ümumi söz bazasının (2.171.260) təxminən 43.6%-ni təşkil edir. Bu, korpuslarda tez-tez rast gəlinən haldır və çox vaxt ya nadir sözləri, ya da səhvləri (yazı səhvləri, transliterasiya fərqləri və s.) əhatə edir.

Tezliyi 5-dən yuxarı olan sözlər ümumi söz bazasının dörddə birindən çoxunu (24.03 %) təşkil edir. Bu kateqoriya daha çox funksional və ümumi leksik vahidləri əhatə edir.

Söz bazasından çıxarılan sözlər və formalaşdırılmış sözlüklər. Söz bazasının emalının son mərhələsində texniki və linqvistik təmizləmə əməliyyatları daha da dərinləşdirilmiş və nəticədə baza strukturlaşdırılmış formaya salınmışdır. Bu mərhələdə əsas məqsəd qeyri-standart, texniki mənşəli, çox aşağı tezlikli sözlərin bazadan çıxarılması olmuşdur.

Aşağıdakı tezlik göstəricilərinə malik olan sözlər təhlil edilərək bazadan silinmişdir (Bax: Şəkil 2):

Şəkil 2

Bu təmizləmə nəticəsində ümumilikdə 1.036.247 söz-forma bazadan silinmişdir.

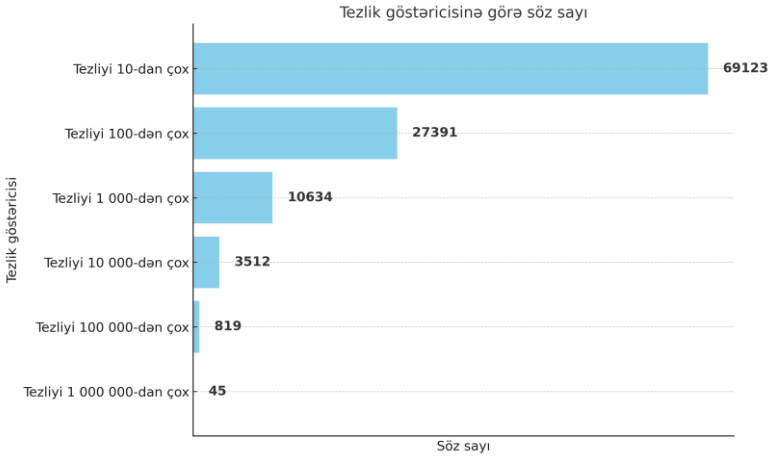
Təhlil və təmizləmə işlərinin yekun mərhələsində aşağıdakı dörd əsas sözlük formalaşdırılmışdır:

1. ***Tezlik siyahısı üzrə sözlük (Söz bazası – I)*:**
   - 175.521 sözdən ibarətdir;
   - Sözlər tezliyə görə azalan sıra ilə düzülmüşdür;
   - Bu sözlük yüksəktezlikli sözlərin seçilməsi və dil resurslarının hazırlanmasında baza rolunu oynayır.
2. ***Əlifba siyahısı üzrə sözlük (Söz bazası – II):***
   - Eyni sayda (175.521) söz ehtiva edir;
   - Sözlər əlifba sırasına görə düzülmüşdür və qarşısında tezlik göstəricisi qeyd olunmuşdur.
3. ***Orfoqrafiya lüğətində olmayan sözlərin tezlik siyahısı (Sözlük – III)*:**
   - Bu siyahıda 93.287 söz vardır;
   - Həmin sözlər Azərbaycan dilinin orfoqrafiya lüğətində (Azərbaycan dilinin orfoqrafiya lüğəti. Bakı: Elm, 2021) yer almayan, lakin söz bazasında rast gəlinən vahidlərdir.

4. ***Orfoqrafiya lüğətində olmayan sözlərin əlifba siyahısı (Sözlük – IV):***
   - Eyni sayda – 93.287 söz ehtiva edir;
   - Burada sözlər əlifba sırasına görə düzülmüş və qarşısında tezlik göstəriciləri verilmişdir.

Sözügedən III və IV sözlüklər orfoqrafiya və izahlı lüğətlərin gələcək nəşrləri üçün etibarlı baza rolunu oynaya bilər.

Azərbaycan dilinin söz bazasının yaradılması təkcə bir lüğətçilik təşəbbüsü deyil, həm də milli dilin rəqəmsal transformasiyası üçün fundamental baza yaradan strateji layihədir. Bu baza əsasında yaradılacaq korpuslar, lüğətlər, tədris vasitələri və dil texnologiyaları Azərbaycan dilinin həm elmi, həm də texnoloji müstəvidə inkişafına xidmət edəcəkdir. Məhz bu baza əsasında Azərbaycan dilində funksional sözlərin əhatə dairəsini müəyyənləşdirmək mümkündür. Bu baza bizə müxtəlif tezlikli sözlərin inventarını müəyyən etməyə imkan verir. Belə bir imkan aşağıda verilmiş qrafikdə aşkar nümayiş olunur (Bax: Şəkil 3):



Şəkil 3

Azərbaycan dilinin söz bazasının hazırlanması AMEA-nın prezidenti akademik İsa Həbibbəyli və AMEA Nəsimi adına Dilçilik İnstitutunun direktoru, filologiya elmləri doktoru, professor Nadir Məmmədlinin rəhbərliyi altında icra edilmişdir.

Söz bazasının materiallarının toplanması, emalı, təsnifi, mərhələlər üzrə əllə təmizlənməsi, proqram vasitəsilə təmizlənməsinə məsləhətlər verilməsi, söz-formaların cədvəllər üzrə və linqvistik baxımdan təhlili işlərini Nəsimi adına Dilçilik İnstitutunun Süni intellekt və kompüter dilçiliyi şöbəsinin müdiri, filologiya elmləri doktoru, professor Məsud Mahmudov və Hind-Avropa dilləri şöbəsinin müdiri, filologiya elmləri doktoru, professor İlham Tahirov yerinə yetirmişlər.

Bu araşdırmada Azərbaycan dilinin söz bazasının hazırlanması ilə bağlı həyata keçirilən nəzəri və praktiki mərhələlər sistemli şəkildə təhlil edilmiş, dil

resurslarının formalaşdırılması üçün zəruri olan texnoloji və linqvistik yanaşmalar əsasında söz bazasının qurulması prinsipləri ərsəyə gətirilmişdir. Söz bazası: a) *Azərbaycan dilinin milli korpusunun əsas struktur komponentlərindən biri* kimi nəzərdən keçirilmişdir; b) leksik vahidlərin *mənbə əsasında avtomatik və yarı-avtomatik emalı* təcrübəsi əsasında yaradılmışdır; c) *fərqli funksional üslubları, çoxsaylı mənbələri və böyükhəcmli materialları əhatə etməsi* baxımından, mövcud korpustəməlli təşəbbüslərlə müqayisədə daha geniş miqyaslıdır. Layihə çərçivəsində 175 mindən artıq lüğət vahidindən ibarət tezlik və əlifba sözlükləri, Azərbaycan dilinin hazırkı orfoqrafiya lüğətindən kənarda qalan 93 mindən çox sözün siyahısı müəyyən edilmişdir. Araşdırma göstərmişdir ki, perspektiv planda hazırkı Söz bazasının milli lüğət strategiyası ilə inteqrasiyası təmin edilməli, gələcəkdə nəşr ediləcək orfoqrafik, izahlı, terminoloji, ikidilli və çoxdilli lüğətlər üçün istinad mənbəyi kimi istifadə edilməlidir. Digər tərəfdən, süni intellekt və təbii dilin emalı texnologiyalarının tətbiqi o mənada genişləndirilməlidir ki, bu baza Azərbaycan nitqinin tanınması, avtomatik tərcümə, çatbot sistemləri və dil modeli təlimləri üçün strateji əhəmiyyət kəsb edir. Hazırkı Söz bazasının tezlikəsaslı tədris resursları üçün istifadəyə verilməsi, lüğət-minimumların, danışıq və tədris materiallarının söz bazasının məlumatlarına əsaslanaraq hazırlanması prosesi də mühüm məsələlərdən sayıla bilər. Söz bazasının əlavə təmizlənmə, təkmilləşmə və morfoloji etiketləmə mərhələsinin həyata keçirilməsi də diqqətdən kənarda qalmamalıdır. Növbəti proseslərdə leksik vahidlərin fonetik, morfoloji, sintaktik xüsusiyyətlərinin sistemə daxil edilməsi faydalı olardı. Mühüm məsələlərdən biri Söz bazasının açıq elmi resurs kimi istifadəsi üçün veb platformanın hazırlanması, ictimai və elmi istifadə üçün axtarış sistemli veb-interfeys, sözün axtarış və statistika panelləri ilə təchiz olunmuş versiyasının ictimaiyyətə açıq edilməsi əhəmiyyətli ola bilər.

ƏDƏBİYYAT

1. “Kitabi-Dədə Qorqud” dilinin statistik təhlili. /tərtibçilər: K.A.Vəliyeva, M. Ə. Mahmudov, V. Y. Pines, C. Ə. Rəhmanov V. S. Sultanov/. – Bakı: Elm, – 1999, – 248 s.

2. Azərbaycan dili üçün NLP sistemləri və milli korpusun yaradılmasının nəzəri və tətbiqi məsələləri. /M. Mahmudov, R.Fətullayev, Ə Fətullayev, S.Abbasov, N. Abdullayev/ – Bakı: Türkologiya. – Bakı, – 2016, № 4, – s.15-28.

3. Azərbaycan dilinin əks əlifba lüğəti /tərtibçilər: M.Mahmudov, Ə. Fətullayev/. – Bakı: Nurlan, – 2004, – 524 s.

4. Azərbaycan dilinin tezlik lüğəti (söz kökləri). I cild. /tərtibçilər – M. Mahmudov, Ə. Fətullayev və b./, – Bakı: Elm, – 2010, – 464 s.

5. Mahmudov, M. Azərbaycan dilinin milli korpusunun yaradılmasının ilkin şərtləri və optimal strukturu. // Время собирать камни…, Məqalələr toplusu. – Bakı: Mütərcim, – 2017, – s. 155-172.

6. Mahmudov, M. Kompüter dilçiliyi. – Bakı: Elm və təhsil, 2013, 356 s.

7. Mahmudov, M. Türk dillərinin milli korpusları. – Bakı: Elm və təhsil, 2018, 392 s.

8. Mahmudov, M., Tahirov, İ., Ayda-zadə, K., Talıbov, S. İnteqrasiya olunmuş elektron lüğətlər sistemi Azərbaycan dilinin milli korpusunun yaradılmasında bir mərhələ kimi // – Bakı: Türkologiya, – 2019, №1, – s. 66-80.

9. Mahmudov M. Süni intellektin linqvistik problemləri. – Bakı: Elm və təhsil, – 2024, – 376 s.

10. Məhəmməd Füzulinin nəzm əsərlərinin əlifba-tezlik sözlüyü. /tərtibçilər: K. A. Vəliyeva, M. Ə. Mahmudov, C. Ə. Rəhmanov, V. S. Sultanov/, – Bakı: Elm, 2004, –548 s.

11. https://korpus.azerbaycandili.az/concordance

NATIONAL CORPUS AND LEXICAL INFRASTRUCTURE: THE EXPERIENCE OF BUILDING THE AZERBAIJANI LANGUAGE LEXICAL DATABASE

ABSTRACT

The article presents a comprehensive scientific and technological analysis of the project, which aims to build the lexical database of the Azerbaijani language, developed within the framework of corpus linguistics and statistical lexicography. The lexical database: a) is considered as one of the main structural components of the Azerbaijani language national corpus; b) is created based on the experience of automatic and semi-automatic processing of lexical units based on the source; c) is more extensive than existing corpus-based initiatives in terms of covering different functional styles, multiple sources, and large volumes of materials. The authors believe that such a database, providing structured lexical inventories, is necessary for both academic research and technological applications. During the research, a corpus of 520 million word forms from various functional styles was collected, cleaned, and structured using specialized software. As a result of automatic processing, 2,918,910 word-forms have initially been identified; through several subsequent stages of technical and lexical filtering, the database was refined. Ultimately, 175,521 lexical units were compiled into structured word lists, sorted by frequency and in alphabetical order. Additionally, 93,287 words not found in the Azerbaijani language's orthographic dictionary were identified and listed separately. The article elaborates on the structure of the national corpus, its components (fiction, publisistic, scientific, official, spoken, and educational texts), concordances, and linguistic analyzers. According to the authors' viewpoint, the lexical database is a strategic resource for both theoretical research and practical applications such as natural language processing, automatic translation, text-to-speech systems, and artificial intelligence models. The study demonstrates significant scientific and practical outcomes in the development of digital lexical resources for the Azerbaijani language.

Keywords: *artificial intelligence, corpus linguistics, national corpus, lexical database, natural language processing, statistical lexicography, concordance*

НАЦИОНАЛЬНЫЙ КОРПУС И ЛЕКСИЧЕСКАЯ ИНФРАСТРУКТУРА: ОПЫТ СОЗДАНИЯ ЛЕКСИЧЕСКОЙ БАЗЫ АЗЕРБАЙДЖАНСКОГО ЯЗЫКА

РЕЗЮМЕ

В статье представлен комплексный научно-технический анализ проекта по созданию первой лексической базы данных азербайджанского языка, разработанной в рамках корпусной лингвистики и статистической лексикографии. Лексическая база данных: а) рассматривается как один из основных структурных компонентов национального корпуса азербайджанского языка; б) создана на основе опыта автоматической и полуавтоматической обработки лексических единиц на основе источника; в) является более обширной, чем существующие корпусные инициативы, с точки зрения охвата различных функциональных стилей, множественных источников и больших объемов материалов. Авторы считают, что такая база данных, предоставляющая структурированные лексические инвентари, необходима как для академических исследований, так и для технологических приложений. В ходе исследования был собран, очищен и структурирован с помощью специализированного программного обеспечения корпус из 520 миллионов словоформ различных функциональных стилей. В результате автоматической обработки было первоначально выявлено 2 918 910 словоформ; после нескольких последующих этапов технической и лексической фильтрации база данных была уточнена. В конечном итоге 175 521 лексическая единица была составлена в структурированные списки слов, отсортированные по частоте употребления и в алфавитном порядке. Кроме того, были выявлены и перечислены отдельно 93 287 слов, не встречающихся в орфографическом словаре азербайджанского языка. В статье подробно рассматривается структура национального корпуса, его компоненты (художественная литература, публицистика, научная литература, официальные, устные и учебные тексты), конкордансы и лингвистические анализаторы. По мнению авторов, лексическая база данных является стратегическим ресурсом как для теоретических исследований, так и для практических приложений, таких как обработка естественного языка, автоматический перевод, системы преобразования текста в речь и модели искусственного интеллекта. Исследование демонстрирует значительные научные и практические результаты в разработке цифровых лексических ресурсов азербайджанского языка.

Ключевые слова: *искусственный интеллект, корпусная лингвистика, национальный корпус, лексическая база, обработка естественного языка, статистическая лексикография, конкорданс*