Milli Dil Korpusunda Lüğətlər Blokunun Proqram Təminatının İşlənilməsi Məsələləri

Rəna Məmmədova[1], Roza Şahverdiyeva[2], Leyla Əkbərova[3]
[1]AMEA Nəsimi adına Dilçilik İnstitutu, Bakı, Azərbaycan
[2,3]AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
[1]*rena.memmedova.1991@inbox.ru*, [2]*shahverdiyevar@gmail.com*, [3]*nihatrl@mail.ru*

***Xülasə*— Məqalədə Azətbaycan milli dil korpusunun lüğətlər blokunun hazırlanmasının aktuallığı göstərilmişdir. Milli dil korpusunun məzmunu açıqlanmışdır. Müasir kompüter dilçiliyi sistemləri təhlil olunmuşdur. Kompüter lüğətçiliyinin və onun proqram təminatının işlənilməsi məsələləri araşdırılmışdır.**

Açar sözlər— milli dil korpusu, lüğətlər bloku, kompüter lüğətiçiliyi, maşın tərcüməsi, proqram təminatı sistemi

I. GİRİŞ

Yaşadığımız əsr informasiya və yüksək texnologiyalar əsridir. İndi dünyada bu və ya digər hadisəni öyrənmək, problemlə tanış olmaq üçün daha çox elektron kitabxanalara, elmi tədqiqat əsərlərinin toplandığı elmi bazalara müraciət edilir. Məhz bu baxımdan milli dil korpuslarının yaradılması olduqca aktual məsələdir. İlk dəfə olaraq ölkəmizdə Azərbaycan dilinin milli korpusunun yaradılması, onun lüğətlər blokunun optimal strukturunun və müvafiq proqram təminatının işlənib hazırlanması nəzərdə tutulur. Tədqiqatın nəticələri Azərbaycan dili, eləcə də digər türk dilləri üçün milli korpusların yaradılması ilə bağlı elmi-praktik və texniki sistemlərdə istifadə oluna bilər. Bu istiqamətdə milli dil korpusunun daxilində lüğətlər bloku, onun optimal strukturu və müasir proqram təminatının işlənilməsi mühüm məsələdir.

II. AZƏRBAYCAN MİLLİ DİL KORPUSUNUN MAHİYYƏTİ VƏ MƏZMUNU

Milli dil korpusu dedikdə, hər hansı konkret bir dildə mətnlərin elektron formada toplanmasına istiqamətlənmiş məlumat axtarışı sistemləri başa düşülür. Konkret dilin milli korpusunda tarixi dövrün müəyyən mərhələlərində dil tam şəkildə bütün üslubları, janrları, ədəbi dili, dialektləri ilə təmsil olunur [1]. Lakin, korpusda mətnlər kor-koranə deyil, müəyyən qayda və nizamla yerləşdirilir. Elektron formatda olan və dili tam təmsil edən çap məhsulları elə yerləşdirilir ki, istifadəçi lazım gəldikdə ondan müəyyən informasiya əldə edə bilsin. Yəni, milli dil korpusları elektron kitabxanalardan və elmi bazalardan fərqli xüsusiyyətlərə malikdir. Milli korpus elektron kitabxanalar kimi "maraqlı", daha çox "oxunan" əsərlərin toplusu deyil. Milli korpus dilin tam və bütün üslubları ilə təmsil olunmasına istiqamətlənmişdir. Əlbəttə, sonrakı mərhələlərdə dilin daha geniş və bütünlüklə təmsil olunması, bütün folklor nümunələrinin, klassik ədəbiyyatın, dialekt materiallarının elektron variantlarının yaradılması da nəzərdə tutula bilər. Bütün variantlarda milli dil korpusu elə qurulmalıdır ki, istifadəçi ondan müəyyən lazımi informasiyanı əldə etmək imkanına malik olsun.

Konkret dillər üçün milli korpusun yaradılması əvvəllər də dünya miqyasında aktual məsələ kimi araşdırılırdı və həmin vaxtlarda maşın fondu adlandırılırdı. Elə indi də hər iki termin, yəni maşın fondu və milli korpus terminləri paralel işlənməkdədir.

Bəzi tədqiqatçı alimlərin işlərində bu məsələ ilə bağlı maraqlı ideyalar irəli sürülür. Onlar göstərirlər ki, informasiya texnologiyalarına dair terminologiyanın bugünkü durumundan çıxış etdikdə, "milli korpus" termini əvəzinə "kompüter fondu" işlənməsini daha məqbul hesab etmək olar. Bu göstərilən terminin hələ də tam sabitləşməməsindən xəbər verir [6].

Qeyd edək ki, Britaniya milli korpusu (BNC) ilk dəfə 1980-ci ilin sonu 1990-ci ilin əvvəllərində Oksford Universiteti tərəfindən yaradılmışdır və müxtəlif janrlarda(məsələn, bədii ədəbiyyat, jurnal, qəzet və s.) mətnlərdən 100 milyona yaxın söz toplanmışdır [7].

Qeyd etmək lazımdır ki, milli dil korpusuna daxil edilən mətnlər Azərbaycan dilinin bütün üslublarını eyni səviyyədə və həcmdə ehtiva etməlidir. Korpusun həcmi nə qədər böyük olursa, korpus bir o qədər etibarlı və əhatəli hesab olunur. Müasir informasiya texnologiyaları milli dil korpuslarının həcmini istənilən qədər artırmağa imkan verir. Burada optimal yerləşdirmə üsullarından istifadə etmək məqsədəuyğundur. Korpusun strukturunu planlaşdırmaqdan çox şey asılıdır. Ümumiyyətlə, Azərbaycan dilində yazılmış və bu dilə aid istənilən mətn dili təmsil edirsə, korpusa əlavə oluna bilər.

III. TÜRK DİLLƏRİNİN MİLLİ KORPUSU

1988-ci ildə Moskvada Sovet Türkoloqları Komitəsinin XIV plenumunda keçmiş SSRİ ərazisində yaşayan türk xalqlarının dillərini (Azərbaycan, özbək, qazax, qırğız, türkmən, başqırd, tatar, qaqauz, yakut, çuvak, xakas, altay, qumıq, noqay, karaim və s.) əhatə edən maşın fondunun yaradılması sahəsində işlərə başlanılması qərara alındı [8]. Burada türk dilləri haqqında istənilən məlumatlar toplanıb sistemləşdirilməli idi. Türk dillərinin maşın fondunun yaradılması üçün Qazaxıstan Elmlər Akademiyasının Dilçilik İnstitutunu mərkəz olaraq seçmək qərara alınmışdı.

Türk dillərinin milli korpusunun yaradılmasında ilk növbədə aşağıdakı məlumatların toplanması zəruri hesab olunurdu:

- Türk dillərinin birhecalı söz köklərinin müxtəlif növlərini əhatə edən struktur-fonetik məlumat;

DOI: 10.25045/NCSoftEng.2017.48

- Morfem siyahıları;
- Sintaktik əlaqələri əks etdirən sxemlər;
- Affikslərin qrammatik tezarusu;
- Konkret türk dillərinin fonetik, qrammatik quruluşu barədə analitik göstəricilər toplusu.

Yuxarıda sadalananlar nəzərə alınmaqla türk dillərinin linqvistik bankının yaradılması nəzərdə tutulurdu [1].

IV. MÜASİR NLP SİSTEMLƏRİNİN YARADILMASI

Türk dillərinin müasir NLP (*National Language Processing*) sistemlərinin yaradılması işləri son zamanlar daha da aktuallaşmışdır. Bunu internetdə yer alan məlumatlardan da görmək olar. Nitqin tanınması, mətnin səsləndirilməsi, maşın tərcüməsi və müstəqil axtarış sistemlərinin yaradılması ilə bağlı ayrı-ayrı türk dillərində orijinal NLP sistemləri yaradılmışdır. Bu cür müasir NLP sistemlərinə aşağıdakıları misal göstərmək olar:

- *Türk dili üçün:*
  - SR (nitqin tanınması) [9];
  - MT (maşın tərcüməsi) [10];
- *Qazax dili üçün:*
  - SR (nitqin tanınması) [11];
  - MT (maşın tərcüməsi) [12].

Azərbaycanda NLP sistemlərinin yaradılması sahəsində tədqiqatlar "Dilmanc" layihəsi çərçivəsində aparılır [13]. Layihə əsasında bir çox işlər görülmüş və həmin işlərin davam etdirilməsi nəzərdə tutulmuşdur. Həmin layihədə dilin bütün üslublarını özündə əhatə edən ikidili korpusları və konkret dillər üçün birdilli korpusları xüsusi olaraq qeyd etmək lazımdır. Bu korpuslar və onların həcmi aşağıda göstərilmişdir:

- İngilis-Azərbaycan ikidilli korpusu – 2 milyon cümlə;
- Türk-Azərbaycan ikidilli korpusu – 277 min cümlə;
- Rus-Azərbaycan ikidilli korpusu – 4,5 milyon cümlə;
- Azərbaycan birdilli korpusu – 60 milyon cümlə;
- Türk birdilli korpusu – 322 milyon cümlə.

V. KOMPÜTER LÜĞƏTÇİLİYİ VƏ ONUN PROQRAM TƏMİNATININ İŞLƏNİLMƏSİ

Kompüter dilçiliyi korpus dilçiliyinin obyekt və predmetini ehtiva edir. O.S.Rublyova rus dilinin milli korpusu sahəsində aparılan tədqiqatları kompüter dilçiliyinin predmetinə aid edir. Kompüterlərdə lüğət materialları maşın kartotekalarında saxlanılır və istifadə olunur. Kompüter texnologiyası inkişaf edib təkmilləşdikcə daha mükəmməl leksikoqrafik məlumatlar bazası yaradılmağa başladı. O.S.Rublyovanın fikrincə, "Rus dilinin milli korpusu"nu belə leksikoqrafik bazalardan biri hesab etmək olar [14]. Həmin vaxtdan etibarən lüğətlərin elektronlaşdırılması geniş vüsət aldı.

Dünya dilçiliyində, o cümlədən, rus dilçiliyində korpus elektron daşıyıcıda saxlanan, müxtəlif dil hadisələri və aspektləri ilə bağlı linqvistik tədqiqatlara material verən, müəyyən nizamla düzülmüş təbii dil mətnləri çoxluğudur. Korpusu təşkil edən mətnlər və onlara müraciət müəyyən qaydalar əsasında aparılır. İlk çoxfunksiyalı məlumat bazaları və lüğətlər 1956-cı ildə ABŞ-da yaradılmışdır. Elektron məlumatlar bazası əsasında bir çox lüğətlər hazırlanmışdır. Buna Webster's English dictionary nümunə göstərilə bilər [15]. İlk korpuslar da bu dövrdə (The Brown Corpus, Lancaster-Oslo/Bergen Corpus, London-Lund Corpus) yaranmağa başladı. Lakin bu korpuslarda sözlərin sayı məhdud idi.

Kompüter lüğətçiliyi kompüter dilçiliyinin bir sahəsi kimi həmin kontekstdə öyrənilməlidir. Bu termin vaxtılə dəbdə olan və çox işlənən statistik leksikoqrafiya terminini də ehtiva edir.

Kompüter lüğətçiliyi adi lüğətlərdən aşağıdakı səciyyələrinə görə fərqlənir:

- lüğətə müraciət və ondan müəyyən sorğular əsasında məlumatların alınması proseduru daha sadə və sürətlidir;
- eyni zamanda bir neçə lüğətə müraiət edib, sözün mənasının qısa zamanda dəqiqləşdirilməsi və ümumi nəticənin çıxarılması mümkündür;
- bu lüğətlərin hər birində hər hansı bir dili ön plana çəkmək və başqa lüğətlərlə müqayisə aparmaq imkanı var;
- adi lüğətlər çap olunduqdan sonra təkrar nəşrə qədər dəyişməz qalır, sözlərin sayı və izahı dəyişdirilə bilməz. Kompüter lüğətləri açıq, dinamik sistemlərdir. Belə lüğətlərə yeni sözləri əlavə etmək və köhnəlmiş sözləri çıxarmaq mümkündür.

Azərbaycanda kompüter lüğətçiliyinin tarixi ötən əsrin 70-ci illərindən başayır. Həmin illərdə ilk dəfə Azərbaycanda qəzet dilinin tezlik lüğəti o vaxtkı elektron hesablama maşınlarının köməyi ilə tərtib olunmuşdur. Sonrakı illərdə həm kompüter dilçiliyi, həm də kompüter lüğətçiliyi sahəsinin proqaram təminatının hazırlanması üzrə bir çox işlər görülmüşdür. Hazırkı vaxtda da həmin işlər müasir proqram mühəndisliyi mexanizmlərindən və texnologiyalarından istifadə etməklə inkişaf etdirilir [16-18].

VI. AZƏRBAYCAN DİLLİ LÜĞƏTLƏR VƏ MAŞIN TƏRCÜMƏ SİSTEMLƏRİ

Bu ikidilli və birdilli korpuslar "Dilmanc" layihəsində fəaliyyət göstərən mətnin formal linqvistik təhlili və maşın tərcüməsi sistemlərində uğurla istifadə olunmaqdadır. İkidilli paralel mətn korpusları Azərbaycan, türk, rus və ingilis dilləri arasında avtomatik tərcümə vasitəsi və lüğət kimi istifadə olunur. Birdilli korpuslar isə tərcümə olunmuş mətnlərin düzgünlüyünü yoxlamaq baxımından əhəmiyyətilidir.

İnternetdə Azərbaycan dili lüğətlərinin biblioqrafiyasında aşağıdakı elektron lüğətlərə də rast gəlini:

1. AzerDict – Azərbaycanın ən böyük pulsuz onlayn azərbaycanca-ingiliscə, ingiliscə-azərbaycanca lüğəti;

2. İntelsoft – rus dilindən Azərbaycan dilinə tərümə sistemi;

DOI: 10.25045/NCSoftEng.2017.48

3. Google Translate – ingiliscə-azərbaycanca, azərbaycanca-ingiliscə maşın tərcüməsi;

4. Azərbaycanca-türkcə sözlük və b.

Bütün bu göstərilən elektron lüğətlər və maşın tərcüməsi sistemləri gələcəkdə yaradılacaq Azərbaycan dilinin milli korpusunun komponentləri hesab oluna bilər.

NƏTİCƏ

Aparılan araşdırmalar və təhlillər göstərir ki, müasir Azərbaycan milli dil korpusunun yaradılması aktual bir məsələdir. Eyni zamanda milli dil korpusu daxilində lüğətlər blokunun işlənilməsi məsələsi kompüter lüğətçiliyi istiqamətinin tərkib hissəsi kimi mühüm əhəmiyyətə malikdir. Ona görə də lüğətlər blokunun optimal strukturunun formalaşmasında və onun proqram təminatının işlənilməsində müasir İKT-nin imkanlarından, o cümlədən proqram mühəndisliyi elementlərindən, texnologiya və mexanizmlərindən effektiv istifadə edilməlidir.

ƏDƏBİYYAT

[1] M.Mahmudov, Kompüter dilçiliyi. Bakı, “Elm və təhsil”, 2013, 356 s.

[2] M.Mahmudov, R. Fətullayev və b., Azərbaycan dili üçün NLP sistemləri və milli korpusun yaradılmasının nəzəri və tətbiqi məsələləri. Türkoloiya, N4, Bakı, 2016, s. 15-28.

[3] M.Mahmudov, R.Məmmədova, Azərbaycan dilinin riyazi-statistik metodlar və yeni texnoloji vasitələrlə öyrənilməsi məsələləri: problemlər, perspektivlər. Tədqiqlər, N1. AMEA Nəsimi adına Dilçilik İnstitutu, Bakı, 2016, s.18-29.

[4] Л.А. Бускунбаева, З.А. Сиразетдинов. О проблемах национального корпуса башкирского языка. Материалы «Современное казахское языкознание: актуальные вопросы прикладной лингвистикн». Алматы, 2012, с. 54-55.

[5] http://corpus.byu.edu/bnc/

[6] У.Ю.Шарапова, Г.Н.Бабшанова, Компьютерная лексикография как одно из направлений современной прикладной лингвистики В сборнике: Актуальные проблемы лингвистики - 2013. материалы Всероссийской научно-практической конференции студентов, аспирантов и молодых учёных, 2013, с. 176-179

[7] http://www.natcorp.ox.ac.uk/

[8] http://www.turcologica.org/rossijskij-komitet-turkologov

[9] http://www.sestek.com/tr/konusma-tanima

[10] http://cevirsozluk.com/

[11] http://uniline.kz/wordpress/?p=665

[12] https://sozdik.kz/ru/dictionary/translate və s.

[13] http://dilmanc.az/

[14] http://www.ruscorpora.ru/

[15] http://www.webster-dictionary.org/

[16] Vəliyeva K.A. Kompüter dilçiliyinin müasir istiqamətləri. İnformasiya cəmiyyəti problemləri, 2016, №2, s.98–107

[17] R.M.Əliquliyev, Ə.M.Qurbanova, Terminoloji informatika: formalaşma mərhələləri və inkişaf istiqamətləri. Ekspress-informasiya. İnformasiya cəmiyyəti seriyası. Bakı: “İnformasiya Texnologiyaları” nəşriyyatı, 2014, 71 səh.

[18] R.M.Əliquliyev, Ə.M.Qurbanova, Azərbaycanda terminoloji informasiya sisteminin yaradılmasının konseptual əsasları. İnformasiya cəmiyyəti problemləri, Bakı, 2011, №1, səh. 3-8.

DOI: 10.25045/NCSoftEng.2017.48