Orfoqrafiya lüğətinin internet resurslarına inteqrasiyası və mətnlərin avtomatik düzəlişi məsələləri

Raqif İsmayılov[1], Kəmalə Qurbanova[2]
[1,2,] İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
[1]*raqif.ismayilov@mail.ru,* [2]*kemalewamil@gmail.com*

Xülasə — Müasir zamanda informasiya kommunikasiya texnologiyalarının dinamik inkişaf etdiyi hazırkı qloballaşma dövründə milli dil korpusunun onlayn mühitə tətbiqinə tələbat yaranmışdır. Milli dil korpusu zəngin məlumat bazasına və geniş axtarış imkanları olan proqram təminatlarına malik olmalıdır. İşdə elektron linqvistik korpusların yaranma tarixinə nəzər salınmış və dil korpusunun məlumat axtarış sistemi olduğu qeyd olunmuşdur. Azərbaycan milli dil korpusunun yaradılmasının və bu platformanın inkişafının təmin olunmasının zəruriliyi əsaslandırılmışdır.

Açar sözlər — milli dil korpusu; orfoqrafiya lüğəti; kompüter dilçiliyi; mətn korpusu; linqvistik korpus

I. GİRİŞ

İstənilən dilin orfoqrafiya lüğəti ədəbi dildə işlənən sözlərin düzgün yazılış formalarını özündə ehtiva edən mötəbər bir mənbədir. Zaman keçdikcə yeni sözlər yaranır ki, bu təbii proses də dilin orfoqrafiya lüğətini zənginləşdirir. Eyni zamanda hər millətin orfoqrafiya lüğəti onun dil vahidinin etimologiyasını da əks etdirir.

Yaşadığımız qloballaşma dövründə orfoqrafiya lüğətinin internet resurslara inteqrasiyası hər millətin milli dil korpusunun yaranmasına zərurət yaratmışdır. Dil korpusu şifahi nitq və ya yazılı mətnlər şəklində müəyyən bir dilin təhlili üçün linqvistik məlumatların məcmusudur [1].

Azərbaycan dilinin lüğət tərkibində də ciddi dəyişikliklər özünü göstərmiş və ədəbi dilimiz zamanın tələbi ilə bir sıra yeni anlayışları ifadə edən söz və ifadələrlə xeyli zənginləşmişdir. Milli dilimizin orfoqrafiya lüğətinin internet resurslarına inteqrasiyası zamanın tələbidir.

II.AZƏRBAYCAN ORFOQRAFİYA LÜĞƏTİNİN İNKİŞAF MƏRHƏLƏLƏRİ

Azərbaycan dilinin ilk orfoqrafiya lüğəti 1929-cu ildə nəşr olunmuşdur. Linqvistika üzrə görkəmli alim Vəli Xuluflunun latın qrafikalı əlifba əsasında tərtib etdiyi bu orfoqrafiya lüğətindəki sözlərin sayı cəmi 11 min olmuşdur. Orfoqrafiya lüğətinin sonrakı təkmilləşdirilmiş nəşrləri 1940 (20 min söz), 1960 (40 min söz), 1975-ci ildə (58 min söz) işıq üzü görmüşdür. 2013-cü ildə nəşr olunmuş sonuncu orfoqrafiya lüğəti isə həcminə görə daha böyük olmuşdur. Lüğətdəki sözlərin sayı 104 mini keçmişdir [2].

Azərbaycan dilinin orfoqrafiya lüğəti 2021-ci ildə Azərbaycan Milli Elmlər Akademiyasının həqiqi üzvü, filologiya elmləri doktoru, professor Möhsün Nağısoylunun rəhbərliyi ilə yenidən işlənmiş 7-ci nəşr Azərbaycan Respublikası Nazirlər Kabinetinin 16 aprel 2019-cu il tarixli qərarı ilə təsdiqlənmiş və 3 noyabr 2020-ci il tarixli 438 nömrəli qərarı ilə müəyyən dəyişikliklər edilmiş "Azərbaycan dilinin orfoqrafiya normaları" əsasında hazırlanmış ilk nəşrdir. Lüğətə 90 minə yaxın söz daxil edilmişdir [3] .

Son dövrlərdə informasiya kommunikasiya texnologiyaları (İKT) dinamik inkişaf edərək bütün sahələrə öz töhfəsini vermiş və həmin sahənin sürətlə inkişafına səbəb olmuşdur. Digər sahələr ilə yanaşı dil korpusunun onlayn mühitə inteqrasiyası istiqamətində aparılan işlər İKT-nin imkanları sayəsində yeni səviyyəyə yüksəlmişdir. Respublikamızda milli dil siyasətinin inkişafı Azərbaycan xalqının ümummilli lideri Heydər Əliyevin adı ilə bağlıdır. Azərbaycan Respublikası Prezidentinin imzaladığı proqramlar çərçivəsində milli dilimizin istifadəsinə və tədqiqinə dövlət qayğısı artırılmış və ölkədə dilçilik elmi əsaslı surətdə inkişaf etmişdir. Proqramda milli dilimizin öyrənilməsi və təbliği sahəsinə İKT-nin tətbiqi alimlərin qarşısına vəzifə kimi qoyulmuşdur. Ana dilimizin qlobal mühitə inteqrasiyası üçün avtomatik tərcümə sistemlərinin işlənməsi, linqvistik texnologiyaların yaradılması və inkişaf etdirilməsi Dövlət Proqramının icrasına dair tədbirlər planında əks olunmuşdur [4].

Milli dilimizdə olan məlumatların onlayn mühitdə axtarış imkanlarını genişləndirmək üçün dil korpusunun hazırlanması zəruri və aktual məsələdir. Dil korpusu zəngin məlumat bazasına və bu baza üzərindən geniş imkanlara malik proqram təminatlarına malik olmalıdır.

III.DÜNYADA YARADILAN ELEKTRON DİL KORPUSLARI

Dillərin öyrənilməsi və öyrədilməsi prosesində dil emalının yeni mütərəqqi tətbiqlərdən, nitqin tanınması, axtarış sistemləri və avtomatik tərcümə korpuslarından istifadə daha səmərəli nəticə verir.

Araşdırmalara görə dünyada elektron linqvistik korpusların yaranma tarixi 1960-cı illərə təsadüf edir. Birinci nəsil elektron linqvistik korpuslara aşağıdakıları misal göstərmək olar [8]:

- *The Brown Corpus.* Bu linqvistik korpus Amerika Braun (Brown) Universitetində 1961-ci ildən 1964-cü ilə qədər, 4 il müddətinə tərtib edilmişdir. Bir milyon ifadədən ibarət olan bu korpusun dili ingilis (Amerika) dilidir və yazılı mətnlərdən ibarətdir. Mətnlərin strukturu və kateqoriyaların seçimi çox yaxşı düşünülmüş olduğundan, korpus tədqiqatı üçün yararlıdır.

- *Lancaster-Oslo/Bergen (LOB) Corpus.* Layihə 1970-1978 illərdə Lankaster, Oslo universitetlərinin və Bergendəki elmi mərkəzin birgə fəaliyyəti nəticəsində ərsəyə gəlmişdir. Korpus ingilis (Britaniya) dilində bir milyon ifadədən ibarət idi. Strukturu *Brown Corpus*-a bənzəyir. Alimlərin

fikrincə bir milyon söz istifadəsi dilin aşağı tezlikli elementlərini təhlil etmək üçün kifayət etmirdi. Lakin buna baxmayaraq, o dövrdə yüzlərlə yüksək keyfiyyətli və maraqlı tədqiqatlar *Brown* və *LOB* korpusuna əsaslanmışdı.

- *London-Lund Corpus (LLC).* London Universiteti Kollecinin əməkdaşları tərəfindən 1975-ci ildə şifahi ingilis dilinə əsaslanan *LLC* korpusu yaradıldı. Korpus orfoqrafik transkripsiya, fonetik və prosodik (şifahi olmayan) işarələrdən ibarət idi və təxminən 500 min işarəni ehtiva edir. Əvvəlcə kağız üzərində hazırlanmış bu layihə İsveçin dilçi alimləri tərəfindən kompüter formasına çevrilmişdir.

- *The Cobuild Project / The Bank of English.* 1990-cı ildə Birminhem Universitetinin təşəbbüsü ilə ərsəyə gəlmişdir. Korpus daim yeniləndiyindən dəqiq ölçüyə malik deyil. 1997-ci ildə təxminən 300 milyon ifadə olan korpusda, 2005-ci ildə artıq 525 milyon ifadə var idi. Korpusun 25 faizi şifahi, 75 faizi yazılı ifadələrdir.

- *The International Corpus of English (ICE).* Bu korpus ingilis dilinin rəsmi dil olduğu ölkələrin (Avstraliya, Kanada, Yeni Zelandiya və s.) bir neçə universitetlərinin birgə layihəsi ilə 1996-cı ildə meydana gəlmişdir. Bu korpusun təhlili üçün xüsusi olaraq mürəkkəb proqram təminatı hazırlanmışdır.

- *Gigaword corpora.* 2014-cü ildə yaranan bu nəhəng korpus Avropa İttifaqı tərəfindən maliyyələşdirilib və ingilis, ərəb, çin və digər dillərin linqvistik monitorinqini apara bilir. Korpus 1 milyard söz bazasından ibarətdir.

Texnologiyanın dinamik inkişaf etdiyi son zamanlarda mətnlərin emalı, axtarışı və saxlanması istiqamətində əldə olunmuş yeniliklər yüz milyon və ya daha çox sözdən ibarət korpus yaratmağa imkan verir. Buna misal olaraq 2019-cu ildə görkəmli alim Mark Davies tərəfindən yaradılmış English-Corpora.org platformasını misal göstərmək olar. Bu layihə tərkibində ən çox istifadə olunan onlayn korpusları birləşdirir və tədqiqatçılar tərəfindən müxtəlif məqsədlər üçün uğurla istifadə olunur. English-Corpora.org platforması müxtəlif sahələrdə, xüsusən də texnologiya və dil öyrənmə sahəsində böyük müəssisələr tərəfindən istifadə edilmişdir (Şək. 1) [9].

Şəkil 1. English-Corpora.org platforması

IV.ORFOQRAFİYA LÜĞƏTİNİN İNTERNET RESURSLARINA İNTEQRASİYASI

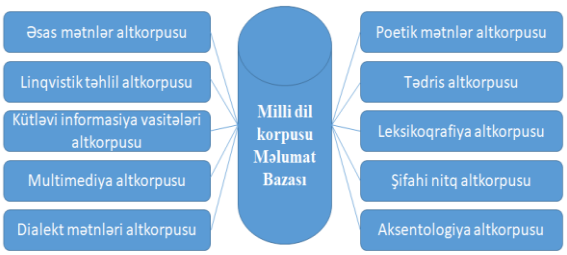
Orfoqrafiya lüğətinin internet resurslarına inteqrasiyası üçün Azərbaycan milli dil korpusunun yaradılması geniş imkanlar yarada bilər. Milli dil korpusu hər hansı bir konkret dildə mətnlərin elektron

formada toplanmasına istiqamətlənmiş məlumat axtarış sistemidir [5].

Dil korpusunda təmsil olunan dilin qrammatikası orfoqrafiyası və lüğəti ehtiva olunur. Eyni zamanda, dil korpusunda müxtəlif istiqamətli mətnlər (bədii, texniki və s.) də əks olunur. İstənilən dil korpusunda həmin dilə müraciət üsulları müəyyən qaydalar üzrə nizamlanır. İstifadəçi korpusdan maraq dairəsinə uyğun olan məlumatı asanlıqla əldə edə bilir [6, 7].

Azərbaycan milli dil korpusunun yaradılması üçün zəngin məlumat bazasının və bu baza üzərindən geniş imkanlara malik proqram təminatlarının hazırlanması zərurəti yaranır.

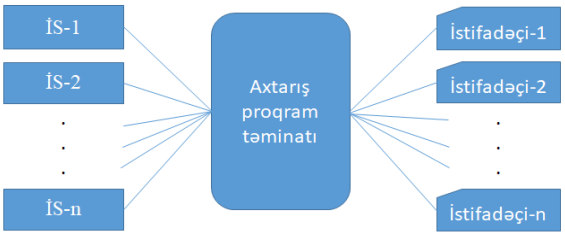
Milli dil korpusu ümumiləşmiş məlumat bazasından ibarət olmalı və bu baza alt korpuslardan təşkil olunmalıdır (Şək. 2.).



Şəkil 2. Azərbaycan milli dil korpusunun baza strukturu

Milli korpus ümumi və alt sistemləri əhatə etməklə yanaşı mövcüd bazalar üzərindən analitik təhlillərin aparılması üçün geniş imkanlara malik olmalıdır, digər mühitlərdə və sistemlərdə istifadə olunacaq proqram servislərini özündə saxlamalıdır.

Azərbaycan milli dil korpusunun yaradılması üçün ikinci əsas məsələ olaraq veb-proqram təminatı hazırlanmalı, məlumat bazasının zənginləşdirilməsi məqsədi ilə veb-servis proqram təminatlarının tərtib edilməsi və istifadəçilər üçün rahat mühit təqdim edilməlidir (Şək. 3.).



Şəkil 3. İnternet mühitində mövcud olan zəngin informasiya sistemlərindən istifadə

V.MƏTNLƏRİN AVTOMATİK DÜZƏLİŞİ MƏSƏLƏLƏRİ

Azərbaycan milli dil korpusunun yaradılması mətnlərin yazılışında qrammatik və orfoqrafik səhvlərin avtomatik düzəlişi, durğu işarələrinin mətndəki yeri və söz seçimi məsələsini asanlaşdıra bilər.

Microsoft Word proqramında mətnlərdə qrammatik və orfoqrafik səhvləri göstərən *When correcting spelling and grammar in Word* funksiyası mövcuddur. Bu tip problemləri həll etməyə yardımçı olan digər proqram təminatlarını misal göstərək:

- *Key Switcher* – müxtəlif qrammatik və orfoqrafik səhvləri aşkar edən və düzgün variantı göstərən proqram vasitəsi;
- *Punto Switcher* – bir yazı sistemindəki hərflərin başqa yazı sisteminin hərfləri ilə

verilməsi əməliyyatını icra edən proqram vasitəsidir;

- *G*rammarly – mətndə durğu işarələrinin mövqeyi və söz seçimi ilə bağlı yaranmış səhvləri aşkar etmək funksiyasına malikdir. Veb saytlarda müştəri ilə əlaqəni təkmilləşdirmək məqsədi ilə istifadə olunur. Süni intellekt texnologiyası istifadəçilərinə yazı üslubunu təkmilləşdirmək imkanı yaradır.

NƏTİCƏ

İnternetin yaratdığı qloballaşma dövründə informasiya texnologiyalarının və proqram təminatının dəstəyi ilə dilin qorunması və inkişafı üsulları ilə bağlı problemlər aktual məsələlərdən birinə çevrilmişdir. Müasir mühitdə linqvistik texnologiyaların davamlı şəkildə təkmilləşdirilməsi hesabına milli dilimizdə gedən prosesləri intellektual analiz etmək mümkündür.

Azərbaycan milli dil korpusu yaradılmalı və bu platformanın inkişafı təmin olunmalıdır. Belə bir korpusun yaradılması və veb-servis proqram təminatlarının tərtib edilməsi milli dilin təkmilləşdirilməsinə, məlumatların toplanmasına, analitik təhlilinə, qiymətləndirilməsinə və onlardan məqsədyönlü şəkildə istifadəyə şərait yaradacaqdır.

İSTİNADLAR

[1] К.Н. "Волченкова Параллельный корпус как справочная база данных в работе переводчика", Проблемы и перспективы развития образования в России, №. 33, с. 32-35, 2015.

[2] M.Nağısoylu, "Azərbaycan dilinin orfoqrafiya lüğəti", Bakı, "Şərq-Qərb" nəşriyyat evi, s. 840, 2013.

[3] M.Nağısoylu, "Azərbaycan dilinin orfoqrafiya lüğəti", Bakı, "Elm" nəşriyyat evi, s. 756, 2021.

[4] İ.Əliyev, "Azərbaycan dilinin qloballaşma şəraitində zamanın tələblərinə uyğun istifadəsinə və ölkədə dilçiliyin inkişafına dair Dövlət Proqramı"nın təsdiq edilməsi haqqında Azərbaycan Respublikası Prezidentinin Sərəncamı, Bakı, 9 aprel 2013-cü il, № 2837, https://e-qanun.az/framework/25537.

[5] M.Mahmudov, "Kompüter dilçiliyi", Bakı, "Elm və təhsil", 356 s., 2013.

[6] R. Məmmədova, R.Şahverdiyeva, L.Əkbərova, "Milli Dil Korpusunda Lüğətlər Blokunun Proqram Təminatının İlənilməsi Məsələləri", Proqram mühəndisliyinin aktual elmi-praktiki problemləri‖ I respublika konfransı, Bakı, 181-183, 17 may 2017.

[7] R.M.Əliquliyev, R.M Alıquliyev, Y.M.Cəfərov, F.F. Yusifov, Ə.M. Qurbanova, "Kosmopolit e-dövlətdən milli e-dövlətə doğru: virtual mühitdə azərbaycan dili xidmətlərinin yaradılması perspektivləri" İnformasiya cəmiyyəti problemləri, №2, s. 3-25, 2021.

[8] М. И. Солнышкина, Г. М. Гатиятуллина, "История развития корпусной лингвистики (на примере англоязычных корпусов)", Вестник Томского государственного университета. Филология, №. 63, с. 132-160, 2020.

[9] English-Corpora.org, ://www.english-corpora.org/byu-corpora.asp