RƏNA MƏMMƏDOVA
AMEA Nəsimi adına Dilçilik İnstitutu
rena.memmedova.1991@inbox.ru

DİL KORPUSLARINDA ELEKTRON LÜĞƏTLƏRİN VERİLMƏSİ ÜSULLARI

Açar sözlər: maşın fondu, milli korpus, lüğətlərin tərtibi, elektron lüğət, optimal struktur, korpusların xüsusiyyətləri
Key words: machinery fund, national corpus, compilation of dictionaries, electron dictionary, optimal structure, features of corpses
Ключевые слова: машинный фонд, национальный корпус, составление словарей, электронный словарь, оптимальная структура, особенности корпусов.

Milli dil korpuslarının mühüm komponentlərindən biri olan elektron lüğətlərin tərtibi, quruluşu və yerləşdirilməsi məsələsində nəinki müxtəlif sistemli dillərdə, hətta eyni dil ailələrində də fərqli yanaşmalar özünü göstərir. Məlum olduğu kimi, hələ 1988-ci ildə keçmiş SSRİ məkanında türk dillərinin maşın fondunun yaradılması ideyası irəli sürülmüş və onun yaradılmasının əsas istiqamətləri müəyyənləşdirilmişdi. Həmin dövrdə Sankt-Peterburq, Moskva, Novosibirsk, Bakı, Daşkənd, Bişkek, Kazan, Aşqabad, Ufa, Nalçik, Çeboksarı və Almatı şəhərlərindən dəvət olunmuş mütəxəssislərdən təşkil olunmuş işçi qrupunda tanınmış alimlər türkdillərinin maşın fondunun yaradılması ilə bağlı müvafiq qərar qəbul etmişdilər. Həmin qərara görə türk dillərinə aid leksikoqrafik, qrammatik, statistik-üslubi, tarixi-etimoloji məlumatlar toplanılmalı və müqayisə-qarşılaşdırma planında sistemləşdirilməli idi. Daha sonra türk dilləri haqqında istənilən məlumatı dolğun və düzgün şəkildə doğura və təqdim edə biləcək qaydaların işlənilib hazırlanması nəzərdə tutulurdu. Həmin dövrdə türk dillərinin maşın fondunun yaradılması üçün mərkəz olaraq Qazaxıstan Elmlər Akademiyasının Dilçilik İnstitutu tövsiyə olunmuşdu [17,47]. Başlanğıc mərhələdə türk dillərində sözlərin qrammatik formalarının yaranma proseslərinin öyrənilməsi üçün birhecalı sözlərin struktur-fonetik müxtəlifliklərinin tədqiqi ön plana çəkilmişdi. Gələcəkdə həm ümumtürk dil sistemini, həm də hər bir konkret dili modelləşdirməyə qabil olan çoxfunksiyalı böyük Türk Dillərinin Maşın Fondunun (TDMF) reallaşdırılması nəzərdə tutulurdu. TDMF müqayisə-qarşılaşdırma planında türk dillərinin leksikoqrafik, qrammatik, tarixi-etimoloji xüsusiyyətlərinə aid məlumatların toplanması və sistemləşdirilməsi işlərini tənzimləməli idi. Eyni

zamanda digər türkdilli regionlarda da bu problemlə bağlı intensiv işlər aparılmağa başlanmışdı [6, 154].

Türk dili üzrə korpus işlərindən ilki Bilge Say Nazirliyində gerçəkləşdirilən ODTÜ Türkcə Korpus olaraq da tanınan "Kompüter mühitində korpusu inkişaf etdirmə tədqiqatları"dır. Həmin korpus 1990-cı ildən sonrakı yalnız yazılı dili əhatə edən mətnlər ibarətdir. Şifahi dilə aid nümunə yoxdur. Müxtəlif növ mətnlərdən nümunələrinseçilib elektron mühitə yerləşdirilərək işarələnməsi yolu ilə yaranan iki milyon sözdən ibarət oflayn bir korpusdur [10].

Türkiyə türkçəsi üzərində edilmiş digər tədqiqat korpusu "Türkçe Ulusal Derlem" – Türk Dilinin Milli Korpusudur. TÜBİTAK (Türkiye Bilimsel ve Teknolojik Araştırma Kurumu) tərəfindən dəstəklənən bu layihə Mersin Universitetinin Dilçilik bölməsinin tədqiqatçıları tərəfindən hazırlanmışdır. 2008-ci ildə başladılan bu işin ilkin versiyası 2012-ci ildə istifadəyə təqdim edilmişdir, 50 milyon söz tutumlu, 1990-2009-cu illər arasında müxtəlif sahələrdə 95% yazılı, 5% şifahi nümunələri olan balanslı, qarışıq (yazılı-şifahi), sinxron və ümumi bir korpusdur [11].

Taner Sezer tərəfindən hazırlanmış 491 milyon sözdən ibarət olan TS Corpus 2012-ci ildə onlayn olaraq istifadəyə verilmişdir. TS Corpus söz növü, morfem və sözün kökünün etiketlənməsi ilə hazırlanan və istifadəçiyə böyük rahatlıq verən ümumi məqsədli bir korpusdur.

Türkcənin diaxron/tarixi korpusu olan Eski Türkçe ve Karahanlı Türkçesinin Tarihsel Derlemi (ETKTD) - Köhnə Türkcə və Qaraxanlı Türkcəsinin Tarixi Korpusu Orxon türkcəsi, Uyğur türkcəsi və Qaraxanlı türkcəsinə aid yazılı mətnlərin elektron mühitə köçürülərək söz birləşməsi və sintaksis əsasında işarələnməsi ilə yaradılmış 600 illik bir dövrü (VII-XIII əsrlər) əhatə edən 400-450 min sözdən ibarət onlayn korpusdur [12].

Son olaraq verəcəyimiz onlayn yazılı mətn korpusu "Vorislamische Alttürkische Texte: Elektronisches Corpus 'VATEC' (İslamiyyətdən əvvəlki türkcə mətnlərin elektron korpusu)"u 1999-2003-cü illər arasında Prof.

Mersin Universitetinin dilçilik tədqiqatçıları 2008-ci ildə başladıqları Türk dilinin Milli Korpusu (TUD) layihəsinin 2012-ci ildə giriş versiyasını onlayn olaraq istifadəyə təqdim etdilər.İstifadəçilər 1990-2010-cu illər arasında "kitablar, dövri nəşrlər, müxtəlif nəşr olunan mətnlər, müxtəlif nəşr olunmamış mətnlər" növlərindən; "sosial elmlər, incəsənət, ticarət və maliyyə, düşüncə və inam, dünya problemləri, tətbiqi elmlər, təbiət və əsas elmlər və b." olmaqla doqquz müxtəlif sahə; "yazarın cinsiyyəti (qadın, kişi)", "Müəllif, müəlliflərin növü (çox, təşkilati, tək)", oxucu kütləsi (uşaq, gənc, hamısı) və s. məlumatları ehtiva edən variantlarla sorğuları istədikləri özəlliklərə görə məhdudlaşdıra və ya genişləndirə bilərlər [14].

Sorğu interfeysində "oyuncak" sözünü yazaraq etdiyimiz axtarışın nəticəsində ekranın üst hissəsində cəmi 4458 mətndə axtarılan sözün (oyuncak) neçə fərqli mətndə (450), neçə dəfə istifadə olunduğu (1200) və bu istifadənin bir milyon sözdəki tezliyi verilmişdir [8].

Mavi rəngdə göstərilən "axtarılan söz"ə daxil olduqda sözün işləndiyi mətn verilir. Mətn sütunundakı elementlərə daxil olduqda isə sözün keçdiyi mətn haqqında məlumat əldə edilir. Axtarılan sözlə bağlı nəticələr Excel formatında istifadəçinin kompüterinə yüklənə bilər.

Ekranın sağ tərəfində olan "menyu" düyməsi vasitəsilə sözün "çap ili, mətn nümunələri, sahə, törəmə mətn formatı, müəllifin cinsi, müəllif/müəlliflərin növü, oxucu kütləsi və növü baxımından bölünməsi; "siyahıla" düyməsi ilə açar sözün (oyuncak) sağ və sol tərəfində olan sözlərin lazım gələrsə əlifba sırasına görə bir sütun yaradaraq düzülüşünü və son olaraq açar sözün sağ və ya sol tərəfində olan sözlərin istifadə tezliyini görə bilərik.

Türkcə onlayn korpuslardan digər biri TS Corpus söz növü, morfem və kök sözün etiketləəи eti böимənz sməkilçiliwзцасокwgöstəəsözün цаünumiфоиw цO3]n3вnaoccсм

tındфиwüstгсцагфоиwацхтиwolarцолwmışa цагсоиw(цissətaолwünцndфdqumpüteynmSorğşa olarcte

лwzürdanфpwnişləyinфpw46.497nфpwщn"wnənфpwənaлфnим цO3]фыффокзосо3Lnaocccзo3екwetcək

□ □□□

nəticələr əldə ediləcək. Bu formada edilən axtarışlarda feil kökündən törəyən isimlərin də feil olaraq verilməsi doğru deyil.

Sorğu interfeysində sözün kökü (lemma), {KÖK} şəklində girdikdə sözün sadə kök və düzəltmə forması əldə edilir. Belə axtarış etməyin iki cür faydası var. Bu faydalardan birincisi {gönül} sözünü sorğuladığımızda bu sözə -ım, ın və b. şəkilçilərdən birinin əlavə olunması nəticəsində söz kökünün son saiti düşmüşforması olan "gönlüm, gönlün" kimi sözləri də tapa bilməsindədir. Digər faydası isə "p, ç, t, k" səsləriylə bitən sözlərə saitlə başlayan şəkilçi əlavə etdikdə "b,c,d,ğ" səslərinə çevrilmiş formaları da tapa bilməsindədir.

Korpusdakı verilənlərin şəkilçi olaraq işarələnməsi xüsusilə TS Corpus-da şəkilçi olaraq sorğu etmək imkanı yaradır. Sorğu interfeysinə [Moph=biçimbirim etiketi] girilərək işarələnmiş olan bir morfemin istifadə nəticəsi əldə etmək olar. Nümunənin sorğu interfeysinə -mak məsdər şəkilçisini etiketlədikdə [Morph=".*\Inf1\+.*"] –mak məsdər şəkilçisinin işləndiyi 3.235.358 nəticə əldə ediləcək. TsCorpus istifadəçilərə *, ?, +, @, /, ( ), [ ], -, _, <,> kimi durğu işarələri ilə axtarış imkanı verir [8].

Üçüncü onlayn korpus olaraq bəhs edəcəyimiz araşdırma Eski Türkçe ve Karahanlı Türkçesinin Tarihsel Derlemi (ETKT-D) – Əski Türkcə və Qaraxanlı Türkcəsinin Tarixi Korpusudur. Türk sənət əsərlərinin tarixi korpusu olan ETKT-D Orxon türkcəsi, Uyğur türkcəsi və Qaraxanlı türkcəsinə aid yazılı mətnlərin elektron mühitə gətirilərək söz birləşməsi və sintaksis əsasında işarələnməsi ilə yaradılmış 600 illik bir dövrü əhatə edən 400-450 min sözdən ibarət onlayn korpusdur. Korpusa giriş etmək üçün hər hansı istifadəçi adına və ya şifrəyə ehtiyac yoxdur.

Son olaraq verəcəyimiz onlayn korpus "Vorislamische Alttürkische Texte: Elektronisches Corpus 'VATEC' (İslamiyet Öncesi Türkçe Metinleri Elektronik Derlem)"i - İslamdan öncə türk mətnlərinin elektron korpusu 1999-2003-cü illər arasında Prof. Dr. Marcel Erdalın başçılığı ilə gerçəkləşdirilən bir layihədir. Korpusun veb səhifəsi alman dilində verilmişdir.

"Mətn yerləşdirmə sorğusu forması (Text location query form)" sözlərə bütün kateqoriyalarda verilənlər bazasında axtarış etmək imkanı verərək dil, mətn və kateqoriya yönündən sorğunu məhdudlaşdırmağa imkan verir.

VATEC verilənlər bazasında qədim türk sözlərinin axtarış qurğusunu təqdim edən "Korpus yerləşdirmə axtarış şəkli (Corpus location query form)" tam olaraq hər hansı sözü axtararkən yanında "*" durğu işarəsi sayəsində axtarılan söz ilə başlayan digər sözləri tapmaq mümkündür. Nəticə səhifəsindən də sözün keçdiyi linkə daxil olmaq mümkündür.

Söz birləşməsi axtarış menyusunda (Morpheme combination query form), söz birləşməsi/morfem istiqamətində işarələnən qədim türk mətnlərində söz birləşməsi və ya morfem sorğulamaq mümkündür.

İslamiyyətdən öncə Göytürk (Runik), Uyğur, Mani, Tibet, Çin, Suryanı və Brahmi əlifbaları ilə yazılan qədim türkcə sözlərin yazı sistemlərindəki bölgüsünə "Sözlərin yazı sistemlərinin sorğu forması" (Writtings of words query form) menyusundan daxil olmaq olar.

Nəticə səhifəsində sözün işləndiyi mətnin adına (name of text) daxil olduğumuzda sözün istifadə olunduğu konteksti görmüş oluruq.

Türk dillərinin maşın fondlarının yaradılmasına qərar verildikdə qazax dilinin milli korpusu ilə bağlı da işlərə Qazaxıstan Respublikası Mədəniyyət Nazirliyinin Dillər üzrə Komitəsi tərəfindən başlanılmışdır.A.K.Jubanovun qazax dili mətnlərinin statistik işlənməsi ilə bağlı tədqiqatları diqqəti cəlb edir. O, görkəmli qazax yazıçısı M.Ayezovun "Abay jolı" romanını linqvostatistik baxımdan tədqiq edərək romanda təxminən 466 min sözişlətmə (söz-forma) işləndiyini bildirmişdir. Yazıçının dilində istifadə olunmuş düzəltmə feillərin 64 faizi mənşəyinə görə feildir. Onların 91 faizi növ şəkilçilərinin köməyi ilə düzələn feillərdir. Nəzərə almaq lazımdır ki, eyni feil əsasına üç və ondan çox növ şəkilçisi birləşə bilər ki, bu da qazax dilinin özəl xüsusiyyətlərindən hesab oluna bilər. Eləcə də qazax dilçiləri bir çox statistik lüğətlər (tezlik, əks tezlik və s.) tərtib etmişlər ki, bu lüğətlərdə də dil haqqında heç yerdə qeydə alınmayan məlumatlar toplanmışdır. Təbiidir ki, bu qiymətli məlumatlar qazax dilinin milli korpusunda öz əksini tapmalıdır. Bu məlumatlar milli korpusda elə əks olunmalıdır ki, istifadəçi onları çətinlik çəkmədən əldə etmək imkanına malik olsun.

Qazax dilinin illüstrativ – mətn fondunun yaradılması işlərinin həyata keçirilməsinə ilk illərdə başlanılmışdır. Görkəmli qazax yazıçısı M.O.Auezovun 20 cildliyinin elektron variantı hazırlanmış, paralel olaraq hər cild üçün əlifba-tezlik lüğətləri – söz göstəriciləri tərtib olunmuşdur. Lüğətdə söz-formaların tezliyi ilə yanaşı onların kitabda işləndiyi səhifə və sətir də göstərilir. Göstərilən tədqiqatın ən qiymətli cəhəti orasındadır ki, dilçi – tədqiqatçı istənilən söz-formanı kontekstdə, displeydə və ya kağızda çap olunmuş şəkildə əldə edə bilər. Bu yazıçının dili haqqında daha geniş məlumat almağa imkanı verir. Bunu da qeyd edək ki, M.Auezovun 20 cildliyi təxminən 3 milyon sözişlətmədən (söz-formadan) ibarətdir. Hər bir söz-forma haqqında məlumat həm hər cild üzrə ayrıca, həm də tam olaraq 20 cildlik üzrə alına bilər.

Daha sonra 10 cildlik qazax dilinin izahlı lüğətinin elektron variantı yaradılmışdır. Lüğətdə olan hər bir lüğət vahidi linqvistik işarələnmişdir. Bu işarələnmədə leksik məna, morfoloji struktur və sintaktik funksiya öz əksini tapmışdır.

□ □□□

Qazax dilinin milli korpusunda xüsusi lüğət blokunun yaradılması qarşıya məqsəd kimi qoyulmuşdur. Bu bloka M.Auezovun 20 cildliyi əsasında hazırlanmış statistik lüğət və 10 cildlik izahlı lüğətdən əlavə qrammatik lüğət, müxtəlif janrlar üzrə tezlik lüğətləri və digər lüğət tipləri daxil edilir. Bütün mövcud terminoloji lüğətlərin elektron variantlarının yerləşdirilməsi nəzərdə tutulur.

Qazax dilinin milli korpusunda mühüm komponentlərdən biri kimi akademik lüğət – qrammatika fondu nəzərdə tutulur. Həmin fond altkorpuslardan ibarətdir. Bu altkorpuslarda tarixi-etimoloji, dialektoloji, onomastik, qrammatik, leksik və lüğətlər müəyyən əlaqə sxemləri üzrə yerləşdirilmişdir.

İstənilən dillərin milli korpuslarında olduğu kimi, qazax dilinin milli korpusunda da zəngin proqram təminatının olması nəzərdə tutulur. Həmin proqram təminatı tam linqvistik təhlil aparmağa imkan yaratmalıdır. Bura avtomatik morfoloji, sintaktik, semantik təhlil daxildir. Tam təhlilin həyata keçməsini təmin edən sistem linqvistik prosessor funksiyasında olur [3, 56].

Qazax dilçiləri hesab edirlər ki, milli korpusun yaradılması işinə bir çox elmi kollektivlər cəlb olunmalı və bu sahədə dünya təcrübəsi nəzərə alınmalıdır [6, 152-153].

Son dövrlərdə korpus dilçiliyinin müxtəlif istiqamətləri, strukturu ilə bağlı nəzəri məsələləri özündə əks etdirən maraqlı əsərlər yazılmışdır [4], [7].

Başqırd dilinin maşın fondunun yaradılması ilə bağlı işlərə 2003-cü ildən başlanılmışdır.

Başqırd dilinin maşın fondu xüsusi olaraq dilçilər, müəllimlər, tələbələr və yuxarı sinif şagirdləri üçün nəzərdə tutulmuşdur. Fonda xüsusi məlumatlar bazası yerləşdirilmişdir. Məlumatlar bazasında və proqram təminatında sorğuların işlənməsi və yerinə yetirilməsi məqsədi ilə interfeyslər yaradılmışdır. İnterfeys "birləşmə, təmas, əlaqə yeri, üsulu" mənasında işlənir. Əgər fərdi kompüterin, proqramın funksiyasının interfeysi dəyişməz qalırsa, həmin obyektin digər obyektlərlə qarşılıqlı təsir prinsiplərini dəyişdirmədən onun özünü modifikasiya etmək olar. Məsələn, Windows proqramlarında interfeys eynidir. Beləliklə, interfeys dedikdə – istifadəçinin müxtəlif qurğularla ünsiyyətdə istifadə etdiyi vasitələr başa düşülür [1, 87].

Başqırd dilinin maşın fondu 7 məlumat bazasını özündə birləşdirən altfondlardan ibarətdir. Əsas kartoteka dilin leksik sistemi haqqında zəruri informasiyanı ehtiva edən 100000 kök söz və düzəltmələrdən ibarətdir. Həmin kartotekanın dilin bütün qatlarını əhatə etdiyi iddia olunur. Hər bir leksik vahid üçün 50-dən çox əlamət göstərilir. Əsas kartotekada sözün hansı nitq hissəsinə aidliyi, mənşəyi, üslubu, dialektə və ya ədəbi dilə mənsubluğu, tarixizm və ya arxaizm, neologizm olması, şəxsli (şəxssiz, ümumi), xüsusi və s. barədə məlumatlar yerləşdirilmişdir. Əsas kartoteka fondun digər altfond

bazaları ilə əlaqəlidir. Bu, lazım gəldikdə, əlavə informasiya əldə etmək imkanı yaradır.

Leksikoqrafiya altfondunda müasir başqırd dilində 500000 lüğət vahidi barədə lüğət məqaləsi verilmişdir. Leksikoqrafik altfondda akademik və tədris profilli lüğətlər - birdilli, ikidilli, çoxdilli, tezlik, terminoloji, frazeoloji, sinonim, sorğu lüğətləri, yer adlarını (küçə, şəhər, yaşayış məntəqələri və s.) bildirən onomastik lüğətlər və s. təmsil olunmuşdur.

Təcrübi-fonetik altfondda başqırd dili sait və samitlərinin artikulyasiya səciyyələri yerləşdirilmişdir. Burada 8000 vahiddən ibarət fonetik lüğət verilmişdir. Həmin materialdan başqırd dilini müstəqil öyrənənlər də istifadə edə bilərlər.

Başqırd yazılı ədəbi dili barədə məlumat verilməsi məqsədi ilə fondda əlyazmaları və qədim çap kitablarını əks etdirən 2000-dən çox vahidi birləşdirən daha iki kataloq təmsil olunmuşdur. Kataloqlar mövcud əlyazma və qədim çap kitablarının təsvirini və aşağıda qeyd olunan məlumatları verir: başlıq (rus dilinə tərcümə ilə), başlığın transliterasiyası, müəllif (əsərin müəllifi), müəllifin adı (transliterasiyada), əlyazmanın üzünü köçürən şəxs barədə məlumat, il (nə vaxt köçürülüb), həcmi, formatı (səhifədə sətirlərin sayı), səciyyəsi, annotasiyası, kim tərəfindən tapılıb və təhvil verilib, dili (ərəb, qədim türk, osmanlı və s.), paleoqrafiyası, harada saxlanılır, şifrəsi və s.

Dialektoloji altfond 3 müstəqil bazadan – leksik, kartoqrafik və tekstoloji bazalardan təşkil olunmuşdur.

Qrammatik altfondda akademik qrammatikalar, başqırd dilinin sözdəyişdirmə sisteminin alqoritmik təsviri, eləcə də morfemlərin statistik bazası barədə məlumatlar toplanmışdır.

Müəlliflərin məlumatına görə başqırd dilinin maşın fondu Rusiya Elmlər Akademiyasının Ufa elm məzkəzində Tarix, Dil və Ədəbiyyat İnstitutunun Linqvistika və İnformasiya Texnologiyaları laboratoriyasında 2011-ci ildən başlayaraq yaradılır. Fonda başqırd dilinin 4 əsas – bədii, publisist, elmi-tədris, rəsmi-işgüzar üslubunu əhatə edən, XIX əsrin 20-ci illərindən bu günə qədər çap olunan mətnlər daxil edilir.

Başqırd dilinin maşın fondunun yaradılması sahəsində qazanılmış təcrübə və materiallar korpusun hazırlanmasında istifadə olunur. Bura avtomatik təhlil və sintez alqoritmləri, leksikoqrafik baza və s. daxildir. Korpusda söz-formaları lüğət vahidi formasına gətirməyə imkan verən morfoloji təhlil alqoritmləri əsas komponentlərdən hesab olunur.

Hazırda korpusda XX əsrin əvvəllərindən indiyə qədərki dövrü əhatə edən 63 müəllifin 579 əsərinin elektron variantı hazırlanmış, ümumi həcmi 9277754 söz-forma (sözişlətmə) redaktə olunmuşdur. Həmin mətnlər 1981-ci ildə qəbul olunmuş başqırd dilinin yeni orfoqrafiyasına uyğunlaşdırılmışdır [2, 54-58].

Ekstralinqvistik marker istifadəçiyə imkan verir ki, konkret parametrləri göstərməklə axtarış sahəsini məhdudlaşdırsın. Bu da istifadəçiyə lazım olan informasiyanı qısa müddətdə əldə etmək üçün şərait yaradır. Ekstralinqvistik məlumatda müəllif (soyadı, adı, atasının adı, təvəllüdü, cinsi və s.), əsər (adı, yaranma ili, janrı, mövzusu, təqdim növü, kitab, jurnal, elektron mətn və s.), mənbənin adı və nəşr ili barədə qeydlər əks olunmuşdur. Belə ekstralinqvistik nişanlar istifadəçiyə istədiyi əlamətlər üzrə axtarış aparmaq imkanı yaradır (burada nişan "marker" mənasında işlənir).

Morfoloji marker istifadəçinin konkret morfoloji səciyyələrinə görə bu və ya digər söz-formanı seçməsini reallaşdırır. Hər bir söz-formanın leksem mənsubiyyəti və morfoloji əlamətləri qabaqcadan göstərilmişdir: sözün ilkin forması, nitq hissəsi, əlaməti, qrammatik kateqoriya əlamətləri.

Semantik markerdə leksemlərin geniş tematik sinifləri və sö-zəmələgətirmə səciyyələri əks olunmuşdur.

Hazırda başqırd dilinin milli korpusunun periodika (qəzet, jurnal), folklor mətnləri, rəsmi-işgüzar və elmi mətnləri əhatə edən altkorpuslarının yaradılması işlərinə başlanılmışdır [16, c.54-58].

Azərbaycan dilinin milli korpusu sahəsində hazırlıq işləri türk dillərinin milli korpusu çərçivəsində aparılırdı. Bu işləri Azərbaycan dilinin milli korpusunun yaradılmasının başlanğıc mərhələsi hesab etmək olar.

2003-cü ilin əvvəllərindən fəaliyyətə başlamış "Dilmanc" layihəsinin NLP sahəsində gördüyü işləri xüsusi qeyd etmək lazımdır. "Dilmanc" Azərbaycanda ilk maşın tərcüməsi sistemidir. "Dilmanc"ın fəaliyyəti ilə Azərbaycan dilinin işlək formal qrammatikasının yaradılmasına başlanılmış, ilkin dildə formal morfoloji, sintaktik və semantik təhlil, tərcümə olunmuş dildə isə cümlənin sintezi alqoritmləri işlənib hazırlanmışdır. Bu sistemdə böyük mətnlər də tərcümə olunur və lazım gələrsə, tərcümə olunmuş nümunələrin və ayrı-ayrı sözlərin düzgün tələffüzü də səsləndirilə bilər [3, 111]

"Dilmanc İmla" proqramını ödənişsiz olaraq Android və İOS telefonlarına yükləmək mümkündür. Daha sonra layihənin saytından kompüterə əlaqələndirici proqram yükləmək lazımdır. Bu proqramda göstərilən kodu telefona daxil etdikdən sonra, telefona deyilən şifahi nitq avtomatik olaraq mətnə çevrilib kompüterə yazılır və klaviaturada yazmağa ehtiyac qalmır.

Türk dilləri qrupu ən geniş dil qruplarından biri hesab olunur. Bəzi mənbələrə görə 50-dən çox türk dili mövcuddur ki, bunların da 40-a qədəri hal-hazırda istifadə edilməkdədir. Lakin 15 türk dili artıq ölü dil hesab olunur.

Türk dillərinin maşın fondunun (türk dillərinin milli korpuslarının) yaradılmasının ilkin mərhələsində ulu türk dilinin rekonstruksiyası məsələsinin araşdırılması və həlli nəzərdə tutulurdu. Məhz ona görə də,

yaradılacaq milli korpus üçün ilk növbədə türk dillərinin birhecalı söz köklərinin müxtəlif növlərini əhatə edən struktur-fonetik məlumat, morfem siyahıları, sintaktik əlaqələri əks etdirən sxemlər, affikslərin qrammatik tezaurusu kimi məlumatların toplanması zəruri hesab olunurdu.

Beləliklə, ilkin mərhələdə ümumtürk və dillərarası fonetik, leksik, morfoloji, sintaktik, semantik təhlil problemlərinin öyrənilməsinə imkan verən türk dillərinin linqvistik bankının yaradılması nəzərdə tutulurdu.

Hal-hazırda Azərbaycanda NLP sistemlərinin yaradılması sahəsində tədqiqatlar "Dilmanc" layihəsi çərçivəsində aparılır. Layihə çərçivəsində indiyə qədər bir çox işlər götürülmüş və görülməkdədir [15].

Layihə çərçivəsində hazırlanmış, dilin bütün üslublarını əhatə edən ikidilli paralel mətn korpuslarını və konkret dilləri əhatə edən irihəcmli birdilli korpusları xüsusi olaraq qeyd etmək olar. Bunlardan İngilis-Azərbaycan ikidilli korpusu 2 milyon cümlədən, türk-Azərbaycan ikidilli korpusu 277 min cümlədən, Rus-Azərbaycan ikidilli korpusu 4,5 milyon cümlədən, Azərbaycan birdilli korpusu 60 milyon cümlədən və türk birdilli korpusu 322 milyon cümlədən ibarətdir.

Bu məsələ ilə bağlı istifadəçilərin rəğbətini qazanmış "Poliqlot" lüğətlər sistemini də qeyd etmək lazımdır. Proqram Azərbaycan Respublikasında informasiya-kommunikasiya texnologiyalarının inkişaf etdirilməsi çərçivəsində hazırlanmışdır. Layihədə almanca-azərbaycanca, ingiliscə-azərbaycanca, rusca-azərbaycanca, fransızca-azərbaycanca, eləcə də azərbaycanca-almanca, azərbaycanca-ingiliscə, azərbaycanca-rusca, azərbaycanca-fransızca lüğətlər təmsil olunmuşdur. Həmin proqram layihəsində Azərbaycan dilində orfoqrafiyanın yoxlanması sistemi də hazırlanıb istifadəçilərə təqdim olunmuşdur [16].

Bundan başqa 2017-ci ildə Azərbaycan dilinin informasiya bazası yaradılmışdır. Bazada Azərbaycan dili haqqında məlumat, rəsmi sənədlər, Azərbaycan dilçiləri haqqında məlumat (fəlsəfə doktorları, elmlər doktorları, müxbir üzvlər və həqiqi üzvlər), Dilçilik kitabxanası (lüğətlər, avtoreferatlar, monoqrafiya və dərsliklər) bölmələri verilmişdir [17].

2018-ci ildə Azərbaycan dilinin elektron lüğətlər korpusu Azərbaycan Respublikas Prezidenti yanında Elmin İnkişafı Fondunun maliyyə dəstəyi ilə (EİF -KETPL- 2-2015-1(25)) yaradılmışdır. Korpusda orfoqrafiya, izahlı lüğət, ixtisarlar, ixtisarlar (xarici dil), qadın adları və kişi adları lüğətləri yer alır. İstənilən sözü axtarışa verdikdə göstərilən lüğətlərdə verilən bütün nümunələri görmək mümkündür [18].

Dünyanın aparıcı ölkələrində və dillərində aparılmış analoji tədqiqatlar türk dillərində, eləcə də qazax dilində eynilə tətbiq oluna bilməzdi. Amma məsələnin elmi qoyuluşu və ideya bütün dünya dillərinə müəyyən modifikasiyalarla tətbiq oluna bilər. Milli dil korpuslarında dilin müxtəlif

sahələri ilə bağlı əvvəllər aparılmış tədqiqatlardan da yararlanmaq mümkündür və bu, hətta zəruridir. Məlum olduğu kimi korpusun ən mühüm funk¬siyası konkret dil haqqında müəyyən informasiyanın isti¬fa-dəçiyə təqdim olunmasıdır. Sadəcə elə üsul, metod, qayda işlə¬nib hazırlanmalıdır ki, onun köməyi ilə istifadəçi hər hansı dilə xas özəl cəhətləri əldə etmək imkanına malik olsun. Yəni kor¬pus¬da dili əks etdirən cəhətlər xüsusi olaraq işlənmir. Dilə xas linqvistik informasiyanın istifadəçiyə təqdim olunma formaları işlənir.

Ədəbiyyat siyahısı

1. Əliquliyev R., Şükürlü S., Kazımova S. Elmi fəaliyyətdə istifadə olunan əsas terminlər. Bakı, "İnformasiya Texnologiyaları", 2009, 201 s.
2. M.Mahmudov. Kompüter dilçiliyi. Bakı, "Elm və təhsil", 2013, 356 s.
3. M.Mahmudov. Türk dillərinin milli korpusları. Bakı, "Elm və təhsil", 2018, 392 s.
4. Баранов А.Н. Корпусная лингвистика // Баранов А.Н. Введение в прикладную лингвистику : Учебн.-метод. Пособие. – СПб., 2005. 48 с.
5. Бускунбаева Л.А., Сиразитдинов З.А. О проблемах национального корпуса башкирского языка. Материалы. «Современное казахскоеязыкознание: актуальные вопросы прикладной лингвистики». Алмааты 2012, с.54-55.
6. . Жубанов А.К. Казахское языкознание: прикладная лингвистика. Алматы, «КИЕ», 2012, 696 с.
7. Захаров В.П. Корпусная лингвистика. Учебно-методическое пособие. Санкт-Петербург, Санкт-Петербургский государственный университет, 2005, 48 с.
8. http://dergi.kmu.edu.tr/userfiles/file/Mayis20142/30m.pdf
9. http://www.tscorpus.com/tr
10. http://lib.metu.edu.tr/tr/odtu-tez-koleksiyonu-sorgulama-sayfasi
11. https://www.tubitak.gov.tr/
12. https://tscorpus.com/
13. http://derlem.cu.edu.tr/
14. http://www.dam.org.tr/index.php/tr/derlemler/66-soezlue-tuerkce-derlemi
15. www.dilmanc.az
16. www.poliqlot.az
17. http://azerbaycandili.az/Home/Index
18. http://korpus.azerbaycandili.az/

Rana Mammadova

The methods of the giving of electron dictionaries in the language corpuses

Summary

The information about problems in the field of the creation of the national corpus of the Turkic languages are given in the artical. It is known that difficult approches show themselves not only in the various systemtic languages, but also in the same language groups in the issue of the compling optimal structuring and locating of the electron dictionaries which one of the inprotant components of the national language corpuses.

The create such as was got in the Republic of Turkey in the file of the corpus creativity. The oral and written text corpuses has been created for different aims, the namtional corpuses of the Turkish Language created by leading creative collective differ for their size, structure and scope of usage.

The works creating of the machine fund of the Turkic languages in the space of the former USSR have been started at the end of the last century, and the main directions of its creationhave been defined, nowadays, machine fund of the Turkic languages takes place on the basis of the national language corpus which is being to create for Azerbaijan, Khazakh,Bashgird,Tatar and the other Turkic languages.

The information about the issues relaized recently in the field of the creation of the national languages corpuses in Turkey, Azerbaijan,Khazakistan and Bashgiria are given, the features of the various languages corpuses are analysis, the main directions of these works which realizing in the future in this field are determined.

Рена Мамедова

Способы включения элэктронных словарей в языковоые корпуса

Резюме

В статье приводятся сведения о проводимой работе в области создания национальных корпусов тюркских языков. Как известно, как выжный составляющий элемент национального языкового корпуса, в программировании электронного словаря, в его оптимальной структуре и размещение наблюдаются разные подходы не только в разносистемных языках, а также и языковых семействах, которые имеют родственные связи.В области корпусного творчества большой успех был достигнут в Турецкой Республике. Здесь были созданы для

различных целей устные и письменные текстовые корпусы турецкого языка. Национальные корпусы тюркских языков, созданные ведущими творческими коллективами, различаются по размеру, структуре и средой пользования.

Создание машиностроительного фонда тюркских языков в бывшем СССР началось в конце прошлого века и были определены основные направления его создания. В настоящее время базой национального корпуса для азербайджанского, казахского, башкирского, татарского и других языков,которая находится на стадии разработки, является машинный фонд тюркских языков.

В статье представлена информация о последних работах, проведенных в области формирования национального языкового корпуса в Турции, Азербайджане, Казахстане и Башкортостане, проанализированы особенности отдельных языковых корпусов, выявлены основные направления дальнейшей работы в этой области.

Rəyçi: Məsud Mahmudov
Filologiya elmləri doktoru, professor