Türk dillərində milli korpuslar

Nərgiz Şakir qızı Ələkbərova
AMEA Nəsimi adına Dilçilik İnstitutunun
fəlsəfə doktoru proqramı üzrə doktorantı
E-mail: narguizli@gmail.com

Rəyçilər: filol.ü.e.d., prof. M.Ə. Mahmudov,
filol.ü.e.d., prof. K.A. Vəliyeva

Açar sözlər: korpus, altkorpus, maşın fondu, altfond, leksik vahid
Ключевые слова: корпус, подкорус, машинный фонд, подфонд, лексическая единица
Key words: corpus, subcorus, machine fund, subfund, lexical unit

1988-ci ildə Moskvada SSRİ çərçivəsində Sovet Türkoloqları Komitəsinin rəhbərliyi ilə təşkil olunmuş XIV plenum müasir türk dillərinin korpusunun yaradılması üçün həlledici rol oynamışdır. Sözügedən plenumda türk dillərinin maşın fondunun yaradılması ideyası irəli sürülmüş və bu istiqamətdə görüləcək tədbirlər planı müzakirə olunmuşdur. Qurultayda qərara alınmış planlardan biri türk dillərinə aid qrammatik, leksikoqrafik, üslubi-statistik məlumatların toplanması, müqayisəli-qarşılaşdırma metoduna əsasən sistemləşdirilməsi və türk dilləri haqqında istənilən məlumatı dolğun çatdıracaq qaydalar sisteminin işlənib hazırlanması olmuşdur. Türk dillərinin maşın fondunun təsis olunması üçün elmi tədqiqatlar mərkəzi Qazaxıstanın SSR Elmlər Akademiyasının Dilçilik institutu seçildi və ilkin olaraq tədqiqat obyekti kimi birhecalı sözlərin struktur-fonetik müxtəlifliyi ön plana çəkilmişdir. Burada əsas məqsəd türk dillərində olan sözlərin qrammatik formalarının yaranma proseslərini izləmək idi. Bu planın sonrakı mərhələsində həm ümumtürk dil sistemini, eyni zamanda hər bir konkret türk dilini modelləşdirməyə icazə verən, bir çox funksiyaları özündə ehtiva edən Türk Dillərinin Maşın Fondunun (TDMF) yaradılması nəzərədə tutulurdu.

Ayrı-ayrı türk dilləri üçün maşın fondunun yaradılması ideyasının irəli sürülməsi və inkişafında R.Q.Piotrovski, A.M.Şerbak, V.Q.Quzevin xidmətləri böyükdür (1, s. 38).

Türk Dillərinin Maşın Fondunun (TDMF) yaradılması qarşısında duran ilkin olaraq başlıca məsələ ulu türk dilinin rekonstruksiyası olmuşdur. Bu mürəkkəb prosesin araşdırılması aşağıdakı məlumatların toplanması zəruri idi:

- türk dillərinin birhecalı söz köklərinin müxtəlif növləri barədə struktur-fonetik informasiya;
- morfem siyahıları;
- sintaktik əlaqələri əks etdirən sxemlər;
- affikslərin qrammatik tezaurusu;
- konkret dillərin fonetik, qrammatik quruluşu barədə analitik göstəricilərin toplusu (1, s. 37).

Belə ki, birinci mərhələdə türk dillərinin linqvistik bankının yaradılması qarşıya qoyulan başlıca məqsədlərdən biri idi və bunun üçün ümumtürk və dillərarası fonomorfoleksik-semantik təhlil problemlərinin öyrənilməsi işin əsas həlledici mərhələsi idi.

İkinci mərhələdə isə bir çox funksiyaları özündə birləşdirən TDMF-in yaradılması nəzərdə tutulurdu. Sözügedən fond istifadəçilər üçün həm ümumtürk, həm də hər bir konkret dil sis-

temini modelləşdirərək yararlı bir sistem şəklinə salınmalı idi. TDMF-in əsas bloklarını leksik, qrammatik, fonoloji və morfonoloji bloklar təşkil edirdi.

Türk dilinin milli korpusu

Türkiyə türkcəsi 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. 2018-ci ildən oflayn oara fəaliyyət göstərir (2).

Bu korpus örnək olaraq dünyanın ən qabaqcıl korpusu sayılan Britaniya Milli Korpusunu (BMK) əsas götürmüşdür. Ümumiyyətlə, korpusun hazırlanmasında bir çox qabaqcıl elmi mərkəzlərdə hazırlanmış, mükəmməl korpusların qurulma təcrübəsindən yararlanılmışdır.

TUD türk dilini təmsil etmək gücünə malik, balanslaşdırılmış, türk yazılı və şifahi mətnlərini əhatə edən türk dilinin ilk məlumat-sorğu korpusu hesab oluna bilər. TUD layihələndirildiyi zaman korpus tərtibində təməl olaraq qəbul edilmiş prinsiplər əsas götürülmüşdür. Bu prinsiplər bunlardır: korpusun dili təmsil etmə gücü, mətnlərin balanslaşdırılması, mətnlərin seçildiyi qaynaq və örnəklər, dildə baş verən dəyişikliklərin nəzərə alınması, mətnlərin məlumatvericilik cəhətinə diqqət yetirilməsi. TUD 20 illik zaman kəsiyində dildə mövcud olan 50 milyon söz və ya söz-formanı təmsil edir. Burada şoxlu sayda müxtəlif sahələrə məxsus, fərqli janrlarda yazılmış şifahi və yazılı mətnlər əks olunmaqdadır. Bəhs etdiyimiz mətn nümunələri 4442 qaynaqdan götürülmüş, 9 fərqli mövzu sahəsini, 39 dil növünü (elmi əsərlər, məktublar, romanlar, bloqlar və s.) əhatə edir.

Qeyd etmək lazımdır ki, bəzən korpusların yaradılması tam sistemli və mərkəzləşdirilmiş şəkildə aparılmır. Komputer mühəndisləri və ya həvəskarlar internet resursları əsasında və ya online ensiklopediyalardan yararlanaraq dil korpusları yaratmağa cəhd göstərirlər. Hazırda internet məkanında yüzlərlə belə korpuslara rast gəlmək mümkündür.

Yuxarıda qeyd olunan korpuslardan biri də dilçilik və sosial elmlərlə bağlı elmi məqalələrin yayımlandığı Vikipediya saytından götürülmüş, 42 milyon sözün əsasında yaradılmış Turkish Wac korpus-sorğu sistemidir.

Bundan əlavə, TurCo adlı internet qaynaqlarına əsaslanan başqa bir korpus da vardır ki, on fərqli internet səhifəsindən yararlanır. Bu korpus 44 milyon söz və söz birləşmələri bazası əsasında yaradılmışdır.

Boun korpusu digər korpuslardan fərqli olaraq Türkiyə ərazisində ən çox oxucu kütləsinə sahib 3 fərqli qəzeti ehtiva edən 4 müxtəlif altkorpusdan və 423 milyon sözdən ibarət bir korpusdur (3).

Yuxarıda qeyd etdiyimiz kimi Türkiyə dilçi alimləri ən gözəl örnək olaraq korpus yaradıcılığında Britaniya Milli korpusundan (British National Corpus) yararlansalar da Azərbaycan və bir sıra Keçmi SSR-i tərkibində olan dövlətlər Rus Dilinin Mili korpusunu örnək kimi götürürlər. Bu həm də o səbəbdən qaynaqlanır ki, vaxtilə SSR-i tərkibinə daxil olan türk xalqlarının maşın fondu ilə bağlı bir neçə layihələr həyata keçirilmiş və Rusiya tərkibində yaşayan xalqlar üçün indi də bu sahədə işlər Rusiya Elmlər Akademiyası tərəfindən görülməkdədir.

Qazax dilinin milli korpusu

SSRİ məkanında türk dillərinin maşın fordunun yaradılması ilə bağlı bir neçə ciddi müzakirələrin getdiyi bir dövrdə artıq Qazax dilinin milli korpusunun formalaşması ilə bağlı bir neçə əsaslı addımlar atılmışdır. Bundan əlavə, türk dillərinin maşın fondunun yaradılması ilə

bağlı keçirilmiş II ümumittifaq konfransda qazax dilçiləri formal təhlil sitemləri üzrə apardıqları nəzəri və təcrübi araşdırmalar əsas götürülərək bu işlər üzrə təşkilatçılıq Qazaxıstan Elmlər Akademiyasının Dilçilik İnstitutuna həvalə olunmuşdur. O vaxtdan qazax dilçiləri korpus dilçiliyi sahəsində bir çox səmərəli işlər görmüş və bir çox uğurlara imza atmışlar.

Məlumdur ki, hər bir dil xüsusi və fərdi quruluş və struktura malikdir. Ona görə də, hər hansı dünya dilləri üzərində keçirilən korpus yaradıcılığı ilə bağlı təcrübələri hər dilə tətbiq etmək mümkünsüz görünür. Lakin məsələnin elmi qoyuluşunu və ideyasını bütün dünya dillərinə modifikasiyalarla tətbiq etmək olar. Onu da qeyd etmək lazımdır ki, korpus istifadəçiyə elə təqdim edilməlidir ki, istifadəçi ondan istifadə etməklə həmin dilə xas özəl cəhətləri əldə etmək imkanına malik olsun və korpus həmin linqvistik informasiyanın istifadəçiyə təqdim olunma formasıdır. Qazax dilinin milli korpusu zəngin mənbələr əsasında formalaşmışdır.

Qazax dilçiliyində qazax dili mətnlərinin statistik işlənməsi ilə bağlı A.K.Jubanovun tədqiqatları elmi ictimaiyyət tərəfindən maraqla qarşılanmışdır. Tədqiqatçı görkəmli qazax yazıçısı M.Ayezovun "Abay jolı" romanını linqvostatistik baxımdan tədqiq etmişdir və belə qənaətə gəlmişdir ki, "Abay jolı" romanında təxminən 466 sözişlətmə (söz-forma) işlənmişdir. 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 (1, s. 55).

Bununla yanaşı, Qazax dilçilərinin məhz statistik lüğətçilik sahəsində göstərdikləri əzmkarlıq məxsusi olaraq qeyd olunmalıdır. Statistik lüğətlərdə toplanan maraqlı məlumatlar, sözsüz ki, milli korpusda elə şəkildə yerləşdirilmişdir ki, istifadəçilər onları əldə etməkdə heç bir maneə ilə qarşılaşmasınlar.

Qazaxıstan milli korpusunun formalaşmasında görkəmli qazax yazıçısı Auezovun 20 cildlik kitabının müstasna rolu olmuşdur. Belə ki, həmin kitabın hər cildi üçün əlifba tezlik lüğətləri hazırlanmış və elektron formata salınmışdır. Tədqiqatın da fərqli cəhəti də ondan ibarətdir ki, istənilən dilçi istənilən söz-formanı kontekstdə, elektron və ya kağız formatda əldə edə bilər. Bu cür konkordanslar istifadəçilərdə həmin yazıçının dili haqqında daha geniş təsəvvürlər formalaşdırır. Onu da qeyd etmək lazımdır ki, sözügedən 20 cildlik kitabda 3 milyon söz-forma toplanmışdır.

Bundan əlavə, qazax dilinin milli korpusunda qeyd edilməli əsas həlledici məqamlardan bir də 10 cildlik qazax dilinin izahlı lüğətidir. Lüğətə daxil edilmiş hər bir lingvistik vahid leksik mənanı, morfoloji strukturu, sintaktik funksiyanı linvistik işarələnmə vasitəsi ilə istifadəçiyə çatdırır.

Yaxın gələcəkdə qazax dilinin milli korpusunu ehtiva edəcək xüsusi lüğət blokunun yaradılması nəzərdə tutulur. Buraya M. Auezerin 20 cildlik kitabı, 10 cildlik izahlı lüğət, bundan əlavə qarammatik lüğət, fərqli janrlar üzrə tezlik lüğətlərinin daxil edilməsi nəzərdə tutulur.

Qırğız və qazax mətnlərinin statistik metodlarla tədqiqi sahəsində T.Sadıkov və B.Şarşembayevin tədqiqatları da böyük maraq doğurur. Bu iki tədqiqatçının çoxillik zəhmətinin nəticəsi olan və 2011-ci ildə Ankarada Türk Dil Kurumu Yayınları tərəfindən çap olunmuş 1647 səhifədən ibarət "Manas destanı. Kırgızca-türkçe büyük dizin" sözlüyünü xüsusi qeyd etmək olar. Bu araşdırma türk dillərinin müştərək dil korpusunda yerləşdiriləcək material kimi önəmlidir. Bir milyondan çox misradan ibarət dastanın sözlüyünün tərtibi yeni texnoloji vasitələrin köməyi olmadan mümkün olmazdı. Dastanda işlənmiş sözlərin qrammatik formadan lüğət vahidi formasına gətirilməsi işləri də yerinə yetirilmişdir (4, s. 18).

İstənilən bir dilin milli korpusunda olduğu kimi, qazax dilinin də milli korpusu yaradılar-

kən zəngin proqram təminatına əsaslanması nəzərdə tutulur. Həmin proqram təminatı əsaslı linqvistik təhlil aparmağa imkan verməlidir. Bəhs etdiyimiz həmin linqvistik təhlilliərə avtomatik morfoloji, sintaktik, semantik təhlillər daxildir. Tam təhlilin həyata keçməsini təmin edən sistem linqvistik prosessor funksiyasını həyata keçirir.

Başqırd dilinin milli korpusu

Başqırd dilinin maşın fondunun formalaşması 2003-cü ilə təsadüf edir.

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 (5, s. 87).

Başqırd dilinin korpusu 20 milyon sözü ehtiva edib "online" olaraq https://bashcorpus.ru/ saytında yerləşdirilmişdir. Sözügedən saytda korpusa belə tərif verilir: "Korpus, elektron şəkildə axtarış imkanı olan mətnlər toplusudur." Burada sözlər qrammatik xüsusiyyətləri və rus dilindəki tərcümələrinə görə axtarış edilə bilər. Bundan əlavə saytda Başqırd dilinin şifahi korpusu, Başqırd poetik korpusu, Başqırd dilinin dil ehtiyatlarının kataloquna link vasitəsilə keçid üçün də imkanlar yaradılmışdır (6).

Başqırd dilinin maşın fondu 7 məlumat bazasını özündə birləşdirən altfondlardan ibarətdir:

- əsas kartotekalar altfondu;
- leksikoqrafik altfond;
- qrammatik altfond;
- əlyazmaların kataloqu altfondu;
- qədim çap kitablarının kataloqu altfondu;
- təcrübi-fonetik altfond;
- dialektoloji altfond (7, s. 57).

Əsas kartotekaların altfondu 100000 söz və düzəltmələrdən ibarətdir ki, bu kartoteka dilin leksik bazası haqqında ən zərurui informasiyanı verir. İddia edilir ki, həmin katoteka dilin bütün qatlarını ehtiva edir. Bu kartotekada sözün aid olduğu nitq hissəsi, etimologiyası, tarixiliyi və ya arxaikliyi, və ya neologizm olub-olmaması, dialekt və ya ədəbi dilə məxsusluğu, şəxsiz və ya şəxsli, ümumi və ya xüsusui olması haqqında ətrafı informasiya yerləşdirilmişdir.

Leksikoqrafiya altfondu başqırd dilinin müasir lüğət fondu əsasında yaradılmışdır. Bu altfondda 500000 lüğət vahidi ilə bağlı çap olunmuş məqalələr öz əksini tapmışdır. Bunlara tezlik lüğətləri, ikidilli, terminoloji, sinonim, omonim, frzeoloji, onomastik və s. lüğətlər aiddir.

Təcrübi-fonetik altfond başqırd dilinin sait və samitlərinin artikulyasiya səciyyələrini əks etdirir. Burada 8000 fonetik vahidin əks olunduğu fonetik lüğət yer almışdır.

Qədim çap kitablarının kataloqu altfondu 200 dən çox vahidi əks etdirən iki kataloqla təmsil olunmuşdur. Sözügedən kataloqlar əlyazmaların başlığını (rus dilinə tərcümə ilə), başlığın transliterasiyasını, əsərin müəllifini, müəllifin adını (transliterasiya olunumuş şəkildə), həmin əlyazmanın üzünü köçürən şəxs barədə məlumatı, əsərin köçürüldüyü ili, həcmini, formatını, səciyyəsini, annotasiyasını, əlyazmanın kim tərəfindən tapıldığını və təhvil verildiyini, dilini (ərəb, qədim türk, osmanlı və s.), paleoqrafiyasını, harada saxlanıldığını, şifrəsini və s. kimi məlumatları özündə əks etdirir.

Dialektoloji altfond 3 müstəqil bazanı özündə ehtiva edir. Bunlara leksik, kartoqrafik, tekstoloji bazalar aiddir.

Başqırd dilinin maşın fondunun yaradılması ilə bağlı qazanılmış təcrübələr və toplanmış dəyərli materiallar korpusun hazırlanmasında çox böyük əsaslı dəstəyi təmin etməklə yanaşı, həm də, bununla bağlı aparılan işləri daha da sürətləndirdi.

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. Hazırda başqırd dilinin milli korpusunun periodika (qəzet, jurnal), folklor mətnləri, rəsmi-işgüzar və elmi mətnləri əhatə edən altkorpuslarının yaradılması işlərinə başlanılmışdır (7, s. 50).

Dilçilərin bir dili hərtərəfli öyrənmələri üçün ilk növbədə korpus lazımdır. Ancaq korpusun faydası təkcə bundan ibarət deyil. Birincisi, korpus dili qorumağın bir yoludur. Sözlüklər və qrammatika kitabları sözlərin işlənmə yerini və stilistikanın bütün nüanslarını nəzərə ala və təsvir edə bilmirlər. Ancaq korpus belə nüansları əks etdirir.

İkincisi, korpus bir istinad mənbəyidir.

Üçüncüsü, korpus bir dil öyrətmək üçün istifadə edilə bilər. Nümunə olaraq, yuxarıda adı çəkilən saytda (https://bashcorpus.ru/) rus dili milli korpusunun təhsildə istifadəsi ilə bağlı yaradılan saytına keçid üçün link təqdim olunur .

Dördüncüsü, korpus dil ilə müasir intellektual qurğu sistemlərinin qarşılıqlı əlaqəsini təmin edir: orfoqrafiya yoxlama sistemləri, faktların avtomatik çıxarılması sistemləri və s.

Ümümiyyətlə, korpus filoloji baxımdan istənilən milli dillə bağlı ən etibarlı məlumat əldə etməyə imkan verən ən dolğun mənbə sayılır. Korpusun inkşafı müasir kompüter texnologiyalarının inkşafı ilə bağlı paralel aparılmalı və korpus dilçiliyi ilə bağlı görülmüş işlər hər bir dövlətin milli dilinə olan ehtiramını əks etdirdiyindən hər bir dövlət bu işdə daha səriştəli olmalı və daha çox maliyyə vəsaiti ayırmalıdır.

Məqalənin aktuallığı. Korpus linqvistikasının gələcək inkişafı onun bütün kompüter dilçiliyi məsələlərini özündə birləşdirə biləcəyini istisna etmir. Müasir dövrdə türk dillərini ehtiva edən hər bir milli korpusun yaradılması ən ümdə məsələlərdəndir.

Məqalənin elmi yeniliyi. Başqırd, Qax və Türk dillərinin korpuslarının formalaşmasının başlıca prinsipləri və inkşaf xüsusuiyyətləri araşdırılır. Bundan əlavə bu sahənin inkşafı ilə bağlı müəyyən tövsiyələr verilir.

Məqalənin praktik əhəmiyyəti və tətbiqi. Tədqiqatdan əldə edilən nəticələr türk dillərinin korpusları ilə bağlı araşdırmlarda tətbiq oluna bilər. Bundan əlavə, məqalədən bu dillərinin tədrisi zamanı da doktorant, dissertatnt və magistrantlar yararlana bilərlər.

Ədəbiyyat

1. Mahmudov M. Türk dillərinin milli korpusları. Bakı, Elm və təhsil, 2018, 392 s.
2. https://www.tubitak.gov.tr/
3. http://www.boun.edu.tr/
4. Mahmudov M., Fətullayev Ə., Abbasov S., Fətullayev R., Abdullayev N. Azərbaycan dili üçün NLP sistemləri və milli korpusun yaradılmasının nəzəri və tətbiqi məsələləri. Türkologiya, 2016, №4, s.15-28.
5. Əliquliyev R., Şükürlü S., Kazımova S. Elmi fəaliyyətdə istifadə olunan əsas terminlər. Bakı, İnformasiya Texnologiyaları, 2009, 201 s.

6. https://bashcorpus.ru/
7. M.Mahmudov. Kompüter dilçiliyi, 2013, 356 s.

Н.Ш. Алекперова

Национальные корпусы на турецких языках

Резюме

Корпусная лингвистика - одна из самых актуальных тем в лингвистике в последние годы в связи с развитием компьютерных технологий. В статье освещается работа ряда тюркских языков (турецкий, казахский, башкирский) в области языка. Рассмотрена работа, проделанная в связи с созданием корпуса на этих языках, и история шагов, предпринятых в этом направлении. Кроме того, обсуждается важность корпусного творчества в настоящем и будущем. В статье также объясняется основа, на которой строится каждый корпус, и конкретные характеристики этих оснований. В то же время национальный корпус каждого языка нуждается в дальнейшем улучшении, и теперь повседневное развитие компьютерных технологий предоставляет лингвистам широкие возможности для серьезных шагов в этом направлении.

N.Sh. Alakbarova

National corpora in Turkish languages

Summary

Corpus linguistics is one of the most relevant topics in linguistics in recent years due to the development of computer technology. The article highlights the number work done by Turkic languages (Turkish, Kazakh, Bashkir) in the field of language corpora. It looks at the history of the work done to establish corps in these languages and the steps taken in this direction. In addition, the importance of corpus creativity in the present and future prospects are discussed. In addition, the article explains the base on which each corpus is established and the specific characteristics of these bases. At the same time, the national corpus of each language needs to be further improved, and now the day-to-day development of computer technology provides ample opportunities for linguists to take serious steps in this regard.

Redaksiyaya daxil olub: 23.08.2021