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

TÜRK DİLİNİN MİLLİ KORPUSUNDA LEKSİKOQRAFİK BÖLÜM

Açar sözlər: milli korpus, türk dili, lüğətlərin tərtibi, elektron lüğət, optimal struktur, korpusların xüsusiyyətləri

Key words: national corpus, turkish language, compilation of dictionaries, electron dictionary, optimal structure, features of corpses

Ключевые слова: национальный корпус, турецкий язык, 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 türk dillərinin 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 [9,47]. 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 [2, 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 nümunələrin seçilib elektron mühitə yerləşdirilərək işarələnməsi yolu ilə yaranan iki milyon sözdən ibarət oflayn bir korpusdur [4].

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 [5].

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 [6].

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. Marcel Erdalın başçılığı ilə həyata keçirilən bir korpusdur. Korpusun veb səhifəsi Alman dilində verilmişdir. Korpusun içərisində Uyğur türkcəsi dövrünə aid mətnlər yer almışdır [7].

2008-2010-cu illər arasında TÜBİTAK tərəfindən dəstəklənən "Sözlü Türkçe Derlemi (STD)" - şifahi türk dilinin korpusu üz-üzə və ya müxtəlif kommunikasiya vasitələri ilə həyata keçirilən türkcə danışıqlardan yaranan 1 milyon sözdən ibarət məlumat bazasını linqvistik metodlarla təhlil edərək müasir türk dilinin kompüter mühitində izlənilə bilinməsini hədəfləyən bir onlayn korpusdur. 2010-cu ildə korpusun sınaq versiyası istifadəçilərə axtarış üçün təqdim edilib, 2013-cü ilin sonuna doğru isə danışıqların yazıya çevrilmiş halı ilə 400000 sözdən ibarət versiyası istifadəyə verilmişdir.

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 [8].

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 [3].

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 etiketlənməsi ilə istifadəçiyə böyük köməklik göstərən ümumi məqsədli balanssız bir korpusdur. Korpusa qeydiyyat menyusunda istifadəçi adı və şifrə yaradıb daxil olmaq mümkündür [6].

İngilis dilində hazırlanmış olan korpusda ana səhifənin sol tərəfində "standart axtarış (standard query), məhdud sorğu (restricted query), söz axtarışı (word look up), tezlik siyahıları (frequencylists) və açar sözlər" bölmələrindən ibarət olan "Korpus sorğuları" (Corpus queries)" vardır.

Sorğu interfeysində "yüz (üz)" sözünə girdikdə böyük yaxud kiçik hərf fərqi olmadan sadə axtarış (Simple query (ignore case)) ilə tam olaraq "yüz" sözünün işləndiyi 46.497 nəticə əldə ediləcəkdir. Bu axtarışı böyük yaxud kiçik hərf fərqilə (simple query case-sensitive) axtarış apardığımızda isə 41.656 kiçik, 4.413 böyük hərflə başlayanların nəticələri əldə ediləcəkdir.

Söz növü istiqamətində etiketlənmiş olan korpusda sorğu interfeysində "etiket" kodlarına girərək sözün növünü axtarış etmək mümkündür. Sorğu interfeysinə _Verb yazdığımızda korpusda feil olaraq işarələnən bütün

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 düşmüş forması 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 –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.

Qarşımıza çıxan ekranda sorğu interfeysinin sağ tərəfində (*) ilə başlayan durğu işarəsi və ḍ, ḫ, ė, Ŋ, ġ, ķ simvolları vardır. Söz axtarışlarımızı "Qaraxanlı türkcəsi, Uyğur türkcəsi və Orxon türkcəsi" dövrlərinin hamısını, ya da bu dövrlərdən birini əhatə edəcək formada və "əsərin adı" (mətn), "əsr və mətn növü" kontekstində məhdudlaşdıraraq edə bilərik.

Nəticə səhifəsində axtardığımız söz, sintaksis kontekstində mətn növü, əsri, yerləşdiyi əsərin adı və dövrü haqqında məlumatların da verildiyi bir strukturda təqdim olunur. Sözün işləndiyi cümlənin aid olduğu əsərdəki misra və ya sətrin nömrəsi cümlənin sol tərəfində verilir.

Durğu işarəsi ilə hər hansı bir söz və bu sözün şəkilçili növlərini də tapa bilərik. Məsələn, sorğu interfeysinə "ığac*" şəklində girdikdə "ığaçka" sözü də əldə ediləcəkdir.

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 [10].

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.

Dünyanın aparıcı ölkələrində və dillərində aparılmış analoji tədqiqatlar türk dillərində 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 funksiyası konkret dil haqqında müəyyən informasiyanın istifadəçiyə təqdim olunmasıdır. Sadəcə elə üsul, metod, qayda işlənib hazırlanmalıdır ki, onun köməyi ilə istifadəçi hər hansı dilə xas özəl cəhətləri əldə etmək imkanına malik olsun. Yəni korpusda dili əks etdirən cəhətlər xüsusi olaraq işlənmir. Dilə xas linqvistik informasiyanın istifadəçiyə təqdim olunma formaları işlənir.

Ədəbiyyat siyahısı

1. M.Mahmudov. Kompüter dilçiliyi. Bakı, "Elm və təhsil", 2013, 356 s.
2. Жубанов А.К. Казахское языкознание: прикладная лингвистика. Алматы, «КИЕ», 2012, 696 с.
3. http://dergi.kmu.edu.tr/userfiles/file/Mayis20142/30m.pdf
4. http://lib.metu.edu.tr/tr/odtu-tez-koleksiyonu-sorgulama-sayfasi
5. https://www.tubitak.gov.tr/
6. https://tscorpus.com/

7. http://derlem.cu.edu.tr/
8. http://www.dam.org.tr/index.php/tr/derlemler/66-soezlue-tuerkce-derlemi
9. http://azerbaycandili.az/Home/Index
10. http://vatec2.fkidg1.uni-frankfurt.de/

Mammadova Rana

Lexicographical section in Turkish national corpus

Summary

The information about problems in the field of the creation of the national corpus of the Turkish language 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 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.

Similar studies conducted in the languages of the leading countries of the world can not be applied to neither Turkish languages, nor Kazakh language. But the scientific formulation and the idea of the problem can be applied to certain languages of the world with certain modifications. Early studies in various areas of the language can also be used in the corpuses of the national language, and this is even necessary.

As you know, the most important function of the language corpus is to provide the user with certain information about a particular language. Such methods, means, rules should be developed that would allow the user to gain access to any language features. That is, in the case, the features reflecting the language are not processed separately.

Here is a form for submitting specific linguistic information to the user.

Мамедова Рена

Лексикографический раздел в турецком национальном корпусе

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.

Подобные исследования, проводимые в языках ведущих стран мира, не могут быть применены как на тюркские языки, так и на казахский язык. Но научная постановка и идея проблемы могут быть применены к определенным языкам мира с определенными модификациями. Ранние исследования по различным областям языка также могут быть использованы в корпусах национального языка, и это даже необходимо.

Как известно, наиболее важной функцией языкового корпуса является предоставление пользователю определенную информацию о конкретном языке. Должны быть разработаны такие методы, средства, правила, которые позволили бы пользователю получить доступ к любым языковым особенностям. То есть в корпусе особенности, отражающие язык, отдельно не обрабатываются.

Здесь предоставляется формы подачи конкретной лингвистической информации для пользователя.

Rəyçi: Nadir Məmmədli
Filologiya elmləri doktoru, professor