Linqvistik verilənlər bazalarının quruluş xüsusiyyətləri

Maya Heydərova
AMEA Nəsimi adına Dilçilik İnstitutu, Bakı, Azərbaycan
maya-heyderova@rambler.ru

Xülasə **— Dilçilikdəki tədqiqatların nəticələrinin cəmiyyətdə aktiv istifadəsi bu sahədə rasional şəkildə təşkil edilmiş informasiya resursları ilə birbaşa bağlıdır. Bu baxımdan da tədqiqatçıların lazımi məlumatlara girişini təmin etmək üçün elmdə və xüsusən dilçilikdə daha səmərəli informasiya sistemlərinin yaradılmasının zəruriliyi artır. Bu zaman informasiyanın daha effektiv saxlanması və emalı üçün verilənlər bazası (VB) texnologiyalarından istifadə etmək ən optimal variantdır. Verilənlər bazası kompüter texnologiyalarından istifadə edərək yaddaşda saxlanılan və idarə oluna bilən strukturlaşdırılmış verilənlər bazası yaratmağa imkan verir. Məqalədə fonetik, leksikoqrafik və qrammatik səviyyələrdə hazırlanmış linqvistik verilənlər bazaları və onların quruluş xüsusiyyətləri müzakirə olunur**.

Açar sözlər — kompüter dilçiliyi; verilənlər bazası; fonetika; leksika; qrammatika

I. GİRİŞ

Linqvistik tipologiyanın inkişafı dilçilik sahəsində linqvistik verilənlər bazasının yaradılmasına səbəb olmuşdur. Linqvistik tipologiya – aralarındakı genetik əlaqənin təbiətindən asılı olmayaraq dillərin struktur və funksional xüsusiyyətlərinin müqayisəli tədqiqidir [1]. Linqvistik verilənlər bazasında müxtəlif dillərin xüsusiyyətləri cəmlənir. Buna görə də müqayisəli və tiploloji dilçilikdə tədqiqatlar aparmaq üçün daha səmərəli vasitə olan linqvistik verilənlər bazalarının hazırlanmasına ehtiyac yaranmışdır.

Verilənlər bazası texnologiyası həm ənənəvi, həm də elektron lüğətlərin yaradılması prosesində istifadə olunur. Hal-hazırda tətbiqi dilçiliyin müxtəlif sahələrində informasiyanın kompüterdə verilənlər bazası şəklində təsvirindən istifadə olunur:

- Müqayisəli və tipoloji dilçilikdə, fonetik, leksik, qrammatik tədqiqatlarda;
- Tezaurusların tərtib edilməsi daxil olmaqla semantik tədqiqatlarda;
- Kompüter dilçiliyində və tətbiqi dilçilik məsələlərinin həllində, xüsusilə də:
  - Linqvodidaktika sahəsində tədqiqatlar;
  - Psixolinqvistika sahəsində tədqiqatlar;
  - Sosiolinqvistika sahəsində tədqiqatlar;
  - Kompüter vasitələrinin köməyi ilə dilin tədrisi;
  - Avtomatik tərcümə sistemlərinin yaradılması;
  - Avtomatik tanıma və nitqin sintezi;

– Dialektologiya sahəsində tədqiqatlar [2].

Hazırda dil verilənlərinin strukturlaşdırılmasına və sistemləşdirilməsinə imkan verən, onlara rahat əlçatanlığı təmin edən və axtarış imkanı verən verilənlər bazası idarəetmə sistemi (VBİS) geniş istifadə olunan vasitədir. Lakin bütün dilçilər VBİS-lə işləmə vərdişlərinə malik deyildir və nəticədə onlar heç də həmişə öz verilənlərini VBİS cədvəllərində saxlaya bilmir. Bunun üçün dilçi tədqiqatçıların istifadə edə biləcəyi interfeysdə verilənlər bazasını həm iyerarxik, həm də relyasiya modelinə uyğun hazırlamışlar.

Bu bazalar əsasən dilçiliyin üç səviyyəsini əhatə edir:

- Fonetik;
- Leksikoqrafik;
- Qrammatik.

II. FONETİK VERİLƏNLƏR BAZASI

Fonetik "nitq korpusu" müəyyən bir dildə toplanmış nitq fraqmentləri toplusundan və onların əsasında yaradılmış proqram təminatı ilə təmin olunan verilənlər bazasıdır.

Fonetik verilənlər bazasının istifadəsi ilə həll olunan dilçilik məsələləri:

- Nitqin fonetik xüsusiyyətlərinin öyrənilməsi;
- Nitqin avtomatik tanınması və sintezi sistemlərinin işlənməsi;
- Səsinə görə şəxsiyyətin identifikasiyası sistemlərinin işlənməsi [3].

Fonetik verilənlər bazasına nümunə kimi aşağıdakıları göstərə bilərik:

TIMIT (Texas Instruments, Massachusetts Institute of Technology) akustik-fonetik korpusu geniş fonetik tədqiqatlar üçün, habelə ingilis dilinin Amerika versiyası çərçivəsində avtomatik fasiləsiz nitq tanıma sistemlərinin hazırlanması və sınaqdan keçirilməsi üçün nəzərdə tutulmuşdur. Korpusda səslərin qeyd olunmasında ABŞ-ın 8 regional dialekt zonasından 630 diktor iştirak edir [4].

RuSpeech (http://russpeech.spbu.ru/) müəyyən mətndə kəsilməz rus nitqinin fraqmentlərini, fonetik transkripsiyasını və informatorlar haqqında əlavə məlumatları özündə birləşdirən nitq verilənlər bazasıdır. Hazırda RuSpeech hər tələffüz edilən cümlənin fonetik işarələnməsi olan 50 mindən çox cümləni ehtiva edir. Korpusun hazırlanmasında hər biri orta hesabla 250 cümlə söyləyən 220 diktor dəvət edilmişdir [4].

III. LEKSİKOQRAFİK VERİLƏNLƏR BAZASI

Leksikoqrafik verilənlər bazasına müxtəlif tipli elektron lüğətlər daxildir.

Elektron lüğətlər - xüsusi kompüter formatında hazırlanmış və ya daha mürəkkəb kompüter proqramlarının (maşın tərcüməsi sistemləri) tərkib hissəsi kimi insanların istifadəsi üçün nəzərdə tutulmuş sözlərin toplusudur. Elektron lüğətlərin iki növü vardır: tətbiqi proqram şəklində olan offlayn lüğətlər və İnternetdə olan onlayn lüğətlər.

Tətbiqi proqram şəklində olan lüğətlərdən istifadə etmək istəyən istifadəçilər bu proqramları telefonuna, kompüterinə yükləyib İnternetə çıxışı

olmadan istifadə edir. Bu tip lüğətlər kommersiya məqsədilə geniş şəkildə istifadə olunur. Belə lüğətlərə misal olaraq, geniş istifadəçi auditoriyasına malik olan və özündə 150-dən çox lüğət bazasını birləşdirən "ABBYY Lingvo" lüğətini göstərə bilərik. Bu lüğət 1996-cı ildə istifadəyə verilmiş və qısa müddət ərzində istifadəçilər arasında geniş yayılmışdır. Son illər müxtəlif təşkilatlar tərəfindən hazırlanan ödənişli və ödənişsiz proqram versiyaları geniş vüsət almışdır. Azərbaycanda bu tipli proqramlara nümunə kimi, Poliqlot və Dilmanc avtomatik lüğətlərinin proqram versiyalarını göstərmək olar.

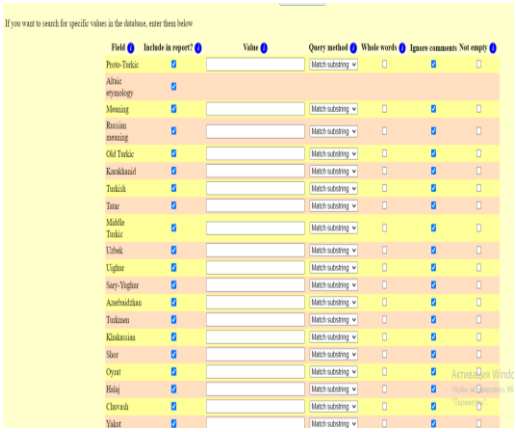
İnternet lüğətlərindən internetə birbaşa çıxışı olan hər bir istifadəçi yararlana bilər. İnternet lüğətləri proqram lüğətləri ilə müqayisədə daha çox istifadə olunur. Bir çox axtarış sistemlərində (portallarda) onlayn tərcümə lüğətləri yerləşdirilmişdir. Buna misal olaraq, google.com, yandex.ru, rambler.ru, mail.ru axtarış sistemlərini göstərmək olar. *Google* və *Yandex* kimi tərcümə lüğətləri 100-dən çox dildə, o cümlədən, Azərbaycan dilindən digər dillərə və yaxud da digər dillərdən Azərbaycan dilinə tərcümə həyata keçirən pulsuz onlayn lüğətdir. Avtomatik tərcümə lüğətlərinin müəyyən çatışmazlıqları da vardır. Məsələn, *Google* tərcümə sistemi istifadəçiyə xarici dildə olan mətnin ümumi məzmununu anlamağa kömək edir, lakin mətni dəqiq tərcümə edə bilmir və nəşr standartlarına uyğun məzmun təqdim etmir, məsələn, bir çox hallarda sözləri mətndən ayrı, kontekst nəzərə alınmadan tərcümə edir və qrammatik qaydalara əməl olunmur.

Azərbaycan dilinin ikidilli onlayn lüğətlərindən olan Dilmanc hazırda Azərbaycanda ən çox istifadə edilən tərcümə lüğətidir. Bundan başqa ikidilli və izahlı lüğətlərin elektron versiyasını özündə ehtiva edən Azleks, Azərbaycan dilinin orfoqrafiya lüğəti əsasında hazırlanan Azərdict, Azərbaycan dilinin orfoqrafiya və izahlı lüğətlərinin, şəxs adları və ixtisarlar lüğəti əhatə edən "Azərbaycan dilinin elektron lüğətlər korpusu" onlayn versiyada hazırlanmışdır (korpus.azerbaycandili.az) [2].

IV. QRAMMATİK VERİLƏNLƏR BAZASI

Etimologiya (sözün mənşəyini öyrənən dilçilik sahəsi) sahəsində tədqiqatlar aparmaq üçün hazırlanan "Babil qülləsi" linqvistik verilənlər bazasını nümunə kimi göstərə bilərik [5]. Bu verilənlər bazası müqayisəli-tarixi dilçiliyə həsr edilmişdir və internet üzərində onlayn olaraq fəaliyyət göstərir. Bazada Avrasiyanın bütün əsas dil ailələrinin materiallarını birləşdirən çoxlu etimologiyalar var. "Babil qülləsi"ndə bütün etimoloji bazalara giriş eyni qaydada qurulur, saytın açılan səhifəsində bazalar sadalanır, bazaların yaradıcıları və yaradılma vaxtı göstərilir (Şək.1). İstifadəçinin hər birinin bazaya (onun tərtibçiləri, lüğət mənbələri, bazalar arasındakı əlaqələr) və bazadakı materiallara ardıcıl baxmaq imkanı vardır. İstifadəçi müvafiq formanı doldurараq xüsusi bir sorğu yarada və bütün türk dilləri üçün məlumat ala bilər (Şək.2).

Şəkil 1. "Babil qülləsi" linqvistik verilənlər bazası



Şəkil 2. "Babil qülləsi" LVB-nin istifadəçi interfeysi

Tipoloji tədqiqatlar aparmaq üçün hazırlanan tipoloji verilənlər bazaları (TVB) müxtəlif təbii dillərin müəyyən olunmuş xarakteristikaları haqqında verilənləri strukturlaşdırılmış sistemli şəkildə özündə saxlayır. Tipoloji verilənlər bazalarında müxtəlif dillərin qrammatik xüsusiyyətləri və xarakteristikaları haqqında informasiyalar saxlanılır, məsələn, müxtəlif dillərdə müşahidə olunan qrammatik cinslərin sayı və s. Qrammatik cəhətdən cinslərə bölünməyən, iki cinsə (kişi, qadın) bölünən, üç cinsə (kişi, qadın, orta) və daha çox cinslərə bölünən dillər vardır. TVB-yə daxil olan verilənlər tədqiqatçılar tərəfindən müvafiq dillərin qrammatik xüsusiyyətlərinə uğunlaşdırılmışdır. TVB-də nəzərdən keçirilən dillərin hər biri bütövlükdə və ya ayrı-ayrı dil hadisələri kəmiyyətinə və tərkibinə, təbiətinə, onlara daxil olan əlamətlərin tərkibi və quruluşuna görə xarakterizə olunub, fərqlənə bilər.

TVB-nin iki əsas növü vardır:

- xarakterik TVB – geniş spektrli dil hadisələrini təşkil edir;
- ixtisaslaşdırılmış TVB – ayrıca dil hadisələrinin öyrənilməsi üçün hazırlanmışdır [6].

Xarakterik tipoloji verilənlər bazaları sxeminin əsası nisbətən sadə "matris" strukturudur, verilmiş n sayda dillərin m sayda əlamətləri n*m kimi göstərilir. Başqa sözlə desək, n sayda dillər m qədər əlamətlə xarakterizə olunur. Sxemin qurulmasında oxşar prinsiplər olan TVB-ni şərti olaraq xarakterik adlandırmaq olar. TVB-nin əsas fərqləndirici xüsusiyyətlərindən biri müqayisə edilən dillərin siyahısı və onları təsvir etmək üçün əlamətlər dəstidir. Bu əlamətlər dəsti öyrənilən dil hadisəsinin miqyasına tədqiqatın məqsəd və vəzifəsinə uyğun olaraq müəyyən edir. Əgər dil hadisələrinin məlumatları qlobal miqyasda tədqiq olunursa, o zaman dillərdən nümunə götürülməsi adətən maksimum areal (dilin yayıldığı ərazi) və genetik təmsilçilik

nəzərə alınmaqla aparılır, yəni tədqiqatçılar TVB-yə mümkün qədər fərqli dil qruplarının və regionların nümayəndələrini daxil etməyə çalışırlar. TVB nisbətən kiçik dil qruplarını da əhatə edə bilir, məsələn roman dil qrupu (cins prinsipinə görə) və ya Afrikan dil qrupu (coğrafi prinsipə görə).

Uzun illər ərzində işlənən, bir çox dil qruplarının xüsusiyyətlər massivini dəstəkləyən iddialı iki əsas TVB layihəsi vardır:

- TVB olan Dil Strukturlarının Dünya Atlası (World Atlas of Language Structures – WALS) maksimum 192 əlamətdə 2679 dil haqqında informasiyaları saxlayır [7]. Bu 192 əlamət hər bir dil üçün deyil, məsələn rus dili üçün bu əlamətlərin sayı 156-dır [8]. Bu baza 2005-ci ildə kitab şəklində, 2008-ci ildən isə onlayn versiyada istifadəçilərə təqdim olunmuşdur. Burada dillərin qrammatik, sintaktik, həmçinin leksik və fonetik xüsusiyyətləri əks olunur.
- Dillərin xüsusiyyətlərini özündə saxlayan ikinci böyük TVB Dünya dilləri (Языки Мира – ЯМ) 315 dilin 3821 əlamətini özündə saxlayır [9].

Struktur parametrik modellərə əsaslanan çoxdilli Linqvistik verilənlər bazalarının (LVB) səmərəliliyini artıran digər funksiya coğrafi informasiya sistemləri ilə birləşməsidir. Belə sistemlərə nümunə kimi Picin və Kreol Dillərinin Strukturları Atlası (Atlas of Pidgin and Creole Language Structures – APiCS) çoxdilli linqvistik verilənlər bazasını göstərmək olar. ApiCS-də kreol və picin dilləri üçün xüsusi parametlər vardır. Məsələn: bir çox dillərdə işlənən "uşaq" və ya "balaca" sözü portuqal dilində "pequeno", yəni "kiçik" sözünə çevrilir [10]. Hal-hazırda bu LVB hər biri 130 -ə qədər fonetik, leksik və qrammatik parametrlər ilə xarakterizə olunan 76 dili təsvir edir və dil hadisələrinin coğrafi bölgüsünü nəzərdən keçirməyə imkan verən coğrafi informasiya sistemi ilə inteqrasiya olunmuşdur.

Bunlardan başqa, daha bir neçə linqvistik verilənlər bazasını misal göstərmək olar: Tipoloji Verilənlər Bazası Sistemi (Typological Database System – TDS), Fonoloji Seqment İnventar Bazası (Phonological Segment Inventory Database), Dünya Alınma Söz Bazası (The World Loanword Database – WOLD) və s. [13]

İxtisaslaşdırılmış tipoloji verilənlər bazası geniş profilli xarakterik TVB-lərdən fərqli olaraq, çox geniş və ya məhdud şəkildə dil materiallarını, yəni tədqiqatçıların müvafiq dilləri xarakterizə edib, təsnif etdikləri təhlilə əsasən bəzi dizaynları və ya mətn parçalarını əhatə edir. Konkret layihələr üzərində işləyən tədqiqatçıların məqsədlərinə və nəzəri əsaslara uyğun olaraq hazırlanmış bu TVB-lər daha çətin struktura malik olur. Məsələn, Şəxs və işarə əvəzlikləri TVB-si ümumi dil məlumatları ilə yanaşı, özündə tipoloji fərqli 109 dilin dəstini saxlayır. Eyni zamanda bu əvəzliklərin paradiqmalarını, həmçinin hər bir dil üçün əvəzliklər sisteminin xülasəsini özündə ehtiva edir [11]. Bununla yanaşı hər bir əvəzlik verilən dəstin morfoloji və sintaktik xüsusiyyətləri ilə təsnif olunur.

Bu cür TVB-lərin məqsədi, bir qayda olaraq, tipoloji tendensiyaları və tipoloji

hadisələri aşkara çıxarmaqdır. Məsələn, Şəxs və işarə əvəzlikləri TVB-nin təhlili nəticəsində müəyyən edilmişdir ki, onda olan məlumatlar morfoloji və sintaktik atributların universalilərinə uyğun gəlir. Bununla yanaşı göstərilir ki, məsələn, tay dilinin əvəzliyi digər dillərin əvəzliklərindən fərqlənir. Bu fərqlilik o qədər ciddidir ki, bu əvəzliklərin şəxs əvəzliklərinə və yaxud isimlərə aid olması sual yaradır. Beləliklə, Şəxs və işarə əvəzlikləri TVB-si ilə iş tədqiqatçılara belə bir fundamental sual yaradır, həqiqətən də şəxs əvəzlikləri mövcuddurmu və bütün dillərdə bu əvəzliklər varmı?

Dil hadisələrini və obyektlərini təsvir edən daha bir TVB Surrey Universitetinin (Böyük Britaniya) morfoloji qrupu tərəfindən hazırlanmışdır. Bu TVB müxtəlif dillərdə sinkretizm haqqındadır. Sinkretizm, adətən eyni formaya malik olan, morfoloji və sintaktik xüsusiyyətləri fərqli olan sözlər arasındakı əlaqəyə deyilir [12].

Bu TVB-də sinkretizmi təsvir edən on mümkün xarakteristika dəsti diqqət çəkir: cəm, hal, cins, müəyyənlik, şəxs, zaman, feil şəkilləri və növləri, tərz, inkarın olub, olmaması. Bu xarakteristikaların mənaları ayrıca cədvəldə cəm şəklində qruplaşdırılır, sinkretizm özü isə bu birləşmələrin cütü şəklində təsvir olunur. TVB müxtəlif dillərdə sinkretizm haqqında yalnız binar (ikili) sinkretizmlərin qeyd edilməsini təmin edir: əgər ikidən çox birləşmənin ümumi forması varsa, onda belə sinkretizm bir neçə ikiliyə bölünür.

Surrey qrupu müxtəlif dillər üçün supletivizm üzrə TVB hazırlayıb. Bu tip dillərdə bir leksik vahidin iki flektiv formaya malik olmasının aydın fonoloji əlaqəsi yoxdur. Supletivizm hadisəsi odur ki, söz kökünün dəyişməsi yolu ilə yeni sözlər əmələ gəlir. İltisaqi dil olan Azərbaycan dilində bu tip nümunələrə rast gəlinməsə də, flektiv dillərdə bu hadisəyə tez-tez rast gəlinir (məsələn, good – better – best, bad – worse – worst, человек – люди, итди – шел). Supletivizm nizamsız hadisə olduğundan ona nizamlı model qurmaq olmur. Müəlliflər bu TVB-nin təsviri üçün bir çox hallarda sinkretizmi təsvir edən TVB-nin sxemini təkrarlayan verilənlər sxemindən istifadə edir. Burada mənaları ayrıca cədvəldə cəm şəklində qruplaşdırılan dil xarakteristikalarını saxlayan on cədvəl vardır. Sonra daha yüksək səviyyəli cədvəldə leksik vahidlərin kökləri malik olduqları birləşmələrə bağlanır və daha sonra yuxarı səviyyəli bir cədvəldə xüsusi bir leksem supletivizimlə öz aralarında əlaqəli olan müxtəlif köklərə birləşdirilir. Beləliklə, TVB-də təsvir olunan hadisələr ağaca bənzər məntiqi struktura malik olur və verilənlər bazası relyasiya modelinin cədvəli şəklində təsvir olunur [12].

NƏTİCƏ

Nəticə olaraq qeyd etmək lazımdır ki, linqvistik verilənlər bazası ölkəmizin dilçilik tədqiqatlarında böyük əhəmiyyət kəsb etsə də, hələ də kifayət qədər öyrənilməyib. Bu səbəbdən də onun hərtərəfli öyrənilməsi və Azərbaycan dilinin bu tip dialektoloji və etimoloji bazalarının yaradılması dilçilik baxımından mühüm məsələlərdəndir.

İSTİNADLAR

[1] Р.Дж. Абдрахманова. Лингвистическая Типология: учебное пособие. Бишкек: Изд-во КРСУ, 2014. 108 с.

[2] M. Heydərova. Elektron məkanda dil materiallarının verilmə üsulları. Lənkəran Dövlət Universitetinin Elmi Xəbərləri. 2019 №1, s. 35-44

[3] M. Heydarova. Compiling of phonetic database. Path of Science: International Electronic Scientific Journal. Vol 7, № 4, 2021, s. 4001-4006

[4] Р. К. Потапова, В. В Потапов. Речевые базы данных как часть мультимодальных корпусов в интернете. Вестник МГЛУ. Гуманитарные науки, 2018. Вып. 6 (797). С. 99-116

[5] The Tower of Babel [Elektron resurs]. URL: http://starling.rinet.ru [istinad tarixi: 01.02.2023].

[6] Г.Кружков. Информационные ресурсы контрастивных лингвистических исследований: типологические базы данных, Системы и средства информ., 2015, том 25, выпуск 1, 198–212

[7] M.Haspelmath. The typological database of the World Atlas of Language Structures. The use of databases in cross-linguistic studies. Empirical approaches to language typology. Berlin -New York: Mouton De Gruyter, 2009. Vol. 41. P. 283-300.

[8] WALS: Language Russian. [Elektron resurs]. https://wals.info/languoid/lect/wals_code_rus [istinad tarixi: 01.02.2023].

[9] В.Д.Совольев Типологические базы данных: перспективы и использования. Вопросы языкознания, 2010. № 1. с. 94-110.

[10] Language Structures Online [Elektron resurs]. URL: http://apics- online.info/ [İstinad tarixi: 01.02.2023].

[11] H. Bliss, E. Ritter. 2009. A typological database of personal and demonstrative pronouns. The use of databases in cross-linguistic studies. Empirical approaches to language typology. Berlin -New York: Mouton De Gruyter. 41:77-116.

[12] P. H. Matthews The Concise Oxford Dictionary of Linguistics. Oxford: Oxford University Press, 1997. 432 p.

[13] A. Dimitriadis, M. Windhouwer, A. Saulwick, R. Goedemans and T. Bíró 2009. The Typological Database System. The use of databases in cross-linguistic studies. Empirical approaches to language typology. Berlin -New York: Mouton De Gruyter. 41:155-207.