# İnformasiya Texnologiyalarının Korpus Linqvistikasında Tətbiqi

Səyyar Abdullayev[1], Südabə Abasova[2], Xanım Kərimova[3]
[1,2,3] AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
[1]*depart5@iit.science.az,* [2]*s.abasova@iit.sciense.az,* [3]*xanim_b@mail.ru*

***Xülasə*— İşdə informasiya texnologiyalarının aparat, sistemli və tətbiqi proqram təminatının, linqvistik verilənlərin avtomatik emalı və linqvistin avtomatlaşdırılmış iş yerlərində tətbiqinin zəruriliyi göstərilmişdir. Korpus linqvistikasının müxtəlif aspektləri və onların qaydaları haqqında məlumat verilmış, milli korpusların yaradılmasının üstünlükləri önə çəkilmişdir.**

***Açar sözlər*— *linqvistika, linqvistik korpus, korpus dilçiliyi, milli korpus***

## I. GİRİŞ

Linqvistik verilənlər üzərində, həmçinin linqvistik modelləşdirmə üçün böyük hesablamaların yerinə yetirilməsi məqsədilə hesablama maşınlarından (və ya kompüterlərdən) istifadə etmək daha münasibdir. Kompüterin hər hansı bir işi yerinə yetirməsi üçün aparat təminatı (hardware) ilə yanaşı, ona göstərişlər toplusu, yəni proqramlar lazımdır. Kompüterin proqram təminatı (software) kompüter sisteminin ayrılmaz bir hissəsi olub, kompüterin texniki təminatının məntiqi davamını təşkil edir. Kompüterin konkret tətbiq sahəsi onun proqram təminatı ilə müəyyən olunur. Kompüterlərin istifadəsində informasiya texnologiyalarının aparat (hardware) və proqram təminatı (software) bir-biri ilə sıx əlaqəlidir [1] .

Proqram təminatı (PT) – müvafiq sənədləşdirmə əsasında və informasiya üzərində müxtəlif əməliyyatların yerinə yetirilməsi və ya aparat vasitələrinin idarə olunması üçün maşın dili komandaları əsasında ardıcıl yazılan kompüter proqramlarıdır. Sistem və tətbiqi proqram təminatı proqram vasitələrinin təyinatından asılı olaraq bir-birindən fərqlənir. Sistem proqramları apparat vasitələrinin işinin idarə olunmasına və əməliyyat sistemlərinin, utilitlərin, drayverlərin və bir sıra digər proqramların yüklənməsinə xidmət göstərir. Tətbiqi proqram son istifadəçi üçün nəzərdə tutulur və ona informasiya üzərində müxtəlif əməliyyatların yerinə yetirməsində: mətni (mətn redaktorları), qrafik təsvirləri (qrafik redaktorlar) yaratmaq və emal etməyə, səsli və video informasiyalar uzərində işləməyə (multimedia proqramları), statistik məlumatların emalı üçün elektron cədvəlləri (elektron cədvəllər) və s. yaratmağa imkan yaradır.

Elektron tərcümə və lüğətlər, həmçinin multimedia tədris proqramları kimi tətbiqi proqramlar da linqvistika üçün xüsusilə yararlıdır. Bəzi tədqiqatçılar informasiya texnologiyalarının aparat və proqram təminatı ilə bərabər bütün linqvistik resursları (qrammatik məlumatlar, lüğətlər, ensiklopediyalar, linqvistik verilənlər bazası və s.) ifadə edən linqvare (və ya linquvare) anlayışından istifadə edirlər. Linqvistik verilənlərin avtomatik emalı üçün zəruri olan aparat, proqram və linqvistik vasitələr birlikdə linqvistin Avtomatlaşdirilmiş İş Yerləri (AİY) kimi adlandırılır. Ana dili və tədqiq olunan xarici dil, həmçinin müxtəlif linqvistik kompüter resursları, əməliyyat və baza tətbiqi Proqram Təminatları (PT) və kompüter linqvistin AİY-nin tərkib hissəsidir. Xüsusiyyətindən asılı olaraq linqvistin AİY-ni xarici dilin öyrənilməsi və tərcüməsi ilə əlaqəli olan tətbiqi proqramlar və linqvistik rasurslar ilə təkmilləşdirmək olar. Öz AİY-nı daimi aktuallaşdırmaqla bərabər, aparat və proqram təminatının müasir vəziyyətinin saxlanması, xüsusi linqvistik resurslar bazasının mütəmadi olaraq təkmilləşdirilməsi, yəni axtarışı, saxlanması, linqvistik məlumatların, lüğətlərin və verilənlər bazasının əldə edilməsi və yaradılması bu sahədə təhsil alanların əsas vəzifəsidir [2, 3].

## II. KORPUS LİNQVİSTİKASI

Linqvistikanın mühüm məsələlərindən biri də, linqvistik tədqiqatlar üçün materialların mənbəyinin toplanması və saxlanılmasıdır. Hal-hazırda bu məsələlərin həlli üçün elektron şəkildə saxlanılması daha münasib olan ən müxtəlif böyük həcmli mətnlərin toplusundan istifadə olunur. Kompüterlərin və xüsusi telekommunikasiya şəbəkələrinin istifadəsi yalnız böyük həcmli mətnlərin elektron şəkildə saxlanması deyil, həm də onların üzərində axtarışın həyata keçirilməsinə, onların emal edilməsinə və s. imkan yaradır. Mətnlərin və ya korpusların elektron şəkildə toplanması məsələsi müasir linqvistika üçün o qədər vacibdir ki, hətta bu elektron mətnlərin toplanması tətbiqi linqvistikanın xüsusi bölməsinin tədqiqat obyektinə - korpus linqvistikasına cevrilmiş olur. Korpus linqvistikası dilçiliyin bölməsi olub, linqvistik korpusların kompüterin vasitəsi ilə istifadəsi və ümumi prinsiplərinin işlənib hazırlanması ilə məşğul olur. Belə müəyyən etmək olar ki, korpus linqvistikası ozünü aşağıda göstərilən iki aspektdə büruzə verir [4, 5]:

- avtomatik alətlərdən istifadə etməklə mətn korpuslarının yaradılması;
- müxtəlif tipli korpuslar bazasında dilin müxtəlif səviyyələrinin araşdırılması üsullarının işlənib hazırlanması.

Müasir tədqiqatçı-dilçilər öz şəxsi yaratdıqları və digər tədqiqatçılar və onların kollektivlərinin yaratdıqları korpusların (hər kəsə əlçatar olan) əsasında lazım olan tədqiqatları apara bilərlər. Elmi tədqiqatlardan başqa, korpuslardan aşağıda göstərilən hallarda istifadə oluna bilər:

- linqvistik lüğətlərin yaradılması, çoxmənalı sözlərin müəyyən olunması və s. üçün;

DOI: 10.25045/NCSoftEng.2017.81

- qrammatikada morfemlərin tezliyini, söz birləşmələrinin tipini, cümlələri və s. müəyyən etmək üçün;
- linqvistikada mətnlərin növlərini bir-birindən fərqləndirmək üçün abzaslararası və abzasların daxilindəki əlaqəni təyin etmək və s.;
- bir sıra tərcümə ekvivalentinə malik olan sözün kontekstinin axtarışı üçün mətnlərin avtomatik tərcüməsində, paralel mətnlərdə ekvivalent tərcümələrin axtarışı və s.;
- tədris məqsədi ilə sitatların seçimi, əsərlərin fraqmentləri, tədris məşqələlərinin təşkili üçün misallar, tədris vəsaitlərinin yaradılması və s.;
- nitqin avtomatik təhlili və sintezi proqramlarının testləşdirilməsində və s. [6, 7, 8].

Korpus linqvistikasının əsas anlayışı - linqvistik korpus - axtarış sistemi ilə təmin olunmuş və müxtəlif linqvistik parametrlərlə işarələnmiş, xüsusi seçilmiş mətnlərin toplusu kimi təyin olunur. Beləliklə, korpusu "Korpus = mətnlər + onların işarələnməsi" kimi xarakterizə etmək olar. Daha geniş mənada desək, korpus mətnlərin istənilən toplusu deməkdir. Bu mənada izah olunmuş və ya izah olunmamış mətnlərin korpusları bir-birindən fərqlənir. Bu kimi izahı verilməmiş korpuslar kifayət qədər tədqiqat və tədris məqsədləri üçün nəzərdə tutulan, periodik nəşrlərin və ya xəbərlər silsiləsinin elektron arxiv versiyası, virtual kitabxanaya məxsus olan mətnlərin elektron toplusu kimi baxılır. Yalnız özündə axtarış alətləri saxlayan izahı verilməmiş mətnlər toplusunun istifadəsi informasiya həcmini artırır və həmin informasiya tədqiqatçı üçün münasib olmamaqla yanaşı, çətinlik yaradır. Korpus linqvistikasının bu mövzusu ilə əlaqədar olaraq izah olunmuş mətnlərin korpusu üstünlük təşkil edir. İlk mərhələdə korpusun yaradılması mətnlərin seçimi ilə başlanır. Bu halda funksional üslublu mətnlərin və konkret janrların hansı ildə nəşr olunması və hansı sayda korpusa daxil edilməsi haqqında düşünmək lazım gəlir. Korpusun yaradılmasında mətnlərin seçimi zamanı aşağıdakı tələblərə diqqət etmək vacıbdir.

- təmsil olunmuş (korpusda sıxlığın yaranması onun təbii dilinin sıxlığına uyğun olmalıdır);
- dolğunluq (təmsil olunmuş ideyaya uyğun gəlməsə belə, informasiya korpusa daxil edilməlidir);
- kifayət qədər həcm (əgər ilkin korpusların həcmi milyon sözə çatıbsa, onda müasir korpusların həcmi milyon və milyardlarla hesablanır, məsələn, ingilis dilinin korpus həcmi Bank of English 2,5 milyard sözü keçir);
- səmərəlilik (problemli sahənin tədqiqi zamanı mətnlərin korpusu əvvəlcədən elə qurulmalıdır ki, tədqiqatçının vaxtına qənaət olunsun, həmçinin problemli sahə mətnlərinin altçoxluğu yanız ciddi deyil, həm də imkan daxilində daha "qənaətli" olmalıdır;
- materialın strukturlaşdırılması (korpusda ona adekvat olan saxlanma vahidləri qeyd olunmalıdır);
- kompüter dəstəyi (mətnlərin korpus dəstəyi verilənləri emal etmək üçün kompleks proqram olaraq, sözlərin kontekstinin təyini, statistik inventarlaşdırma, lüğətin avtomatik emalı və s.).

Korpusun yaradılması üçün ən vacib mərhələ onun işarələnməsidir. İşarələnmə (ingiliscə tağğinğ, annotation) mətnə və onun komponentlərinə xüsusi işarənin yazılmasıdır. Bu işarələrin vasitəsi ilə xarici (elektrolinqvistik) olaraq müəllif və mətn haqqında, daxili olaraq struktur və ya xüsusi linqvistika haqqında məlumat daxil edilir. Xarici işarələnmədə müəllif, mətnin adı, nəşrin çap olunduğu il və yeri, janrı, tematikası haqqında məlumat daxil edilir. Müəllif haqqında yalnız onun adı deyil, həm də onun yaşı, cinsi, tərcümeyi halı və bir çox başqa informasiyalar da daxil oluna bilər. İnformasiyanın belə kodlaşdırılması metaişarələnmə adlanır. Struktur işarələnmə hər bir vahidin (başlıq, abzas, cümlə, sözün forması) statusu haqqında informasiya daşıyır, xüsusi linqvistika isə leksik, qrammatik və mətnin digər elementlərinin xarakteristikasını əks etdirir. Linqvistik təsvirin səviyyəsinə uyğun olaraq morfoloji (danışığın bir hissəsi və morfoloji kateqoriyalar), sintaksis (sintaksis əlaqələrin təyini), semantik (kateqoriyalar, sözün mənasını xarakterizə edən), anoforik (peferent əlaqələrin xarakteristikası, məsələn, əvəzlik), prosodik (vurğunun xarakteristikası və intonasiyalar), diskursiv (pauzaların təyin olunması, təkrarı, şifahi danışığın düzəlişi) və bəzi işarələr bir-birindən fərqlənir [9, 10].

Verilən janrın (Hannover 2010) tədqiqi üzrə ölkələrarası tədqiqat layihəsi həyata keçirilərkən Tvitterin xəbərlər korpusunu işarələmək üçün əsasən növbəti növ işarələr istifadə olunub: (standart yazılış, yalnız sətri işarələrdən istifadə, yalnız baş hərflərdən istifadə, ikiqat qrafema, qrafemin düşməsi, əlavə qrafemanın yazılışı) və s. Korpusda mətnlərin toplanması xüsusiyyətlərindən, onların işarələnməsindən və digər faktorlardan asılı olaraq korpusların növləri fərqləndirilir. Müxtəlif milli dillər üçün yaradılmış korpusların daha vacib növü universal milli korpus hesab olunur. Universal milli korpusların yaradılması və genişləndirilməsi, korpus linqvistikasının ən vacib məsələlərindən biri hesab olunur. Universal milli korpus – dilin ən müxtəlif vəziyyətlərini tədqiq etmək üçün bütün dillərlə münasibətdə nüfuzlu olan, konkret təbii dil mətnlərinin toplusudur [11]. Britaniya milli korpusu (BMK) hamı tərəfindən qəbul olunan universal milli korpusdur [12]. Rus dili üçün belə nüfuzlu korpus Milli rus dili korpusu (MRDK) adlanır[13]. Slavyan dilləri korpusları arasında Praqada Karlova Universitetində yaradılmış Çex milli korpusu digərlərindən fərqlənir [14]. Milli korpuslar həmçinin çin, fin, alman və digər dillər üçün də mövcuddur. Ən ilkin məşhur korpuslardan biri 1960-cı ildə Braun Universitetində tez-tez işlədilən amerika variantlı ingilis dili lüğəti üçün yaradılmış Braun korpusu olmuşdur. Onun həçmi 1 mln. sözdən ibarət olmuşdur. Korpusun yaradıcıları (U.Frensis və Q.Kuçer) mətnlərin seçimində ciddi prosedur işləyib hazırlamışlar. Amerikalı müəlliflər tərəfindən yaradılmış korpusa 500 mətn fraqmentləri daxil olunmuş və 1961-ci ildə hər bir fraqment üzrə 2000 söz işlətməklə çap olunmuşdur. Mətnlər 15 ən çox yayılan informativ janrı və bədii nəsri təmsil edir [15]. Korpusda axtarış istifadəçinin sorğusu əsasında korpus meneceri adlanan xüsusi proqramlar ilə təmin olunur. Korpus meneceri statistik informasiyaların və nəticələrin istifadəçiyə münasib şəkildə verilməsini təmin edən, o cümlədən korpusda verilənlərin axtarışı üçün proqram vasitəsindən ibarət olan xüsusi axtarış sistemidir. Axtarışın nəticələri adətən qramem, ayrı-ayrı dil vahidlərinin xarakteristikasının tezliyi ilə, axtarış

DOI: 10.25045/NCSoftEng.2017.81

vahidi kimi təqdim olunan konkordans (razılaşdırılmış) şəkildə verilir (korpus menecerləri həm də konkordanslar adlanır). Beləliklə, korpus həm tətbiqi, həm də tədqiqat məqsədləri üçün geniş imkanlar yaradan, həcmi 100 mln. sözdən ibarət işarələnmiş mətnlərin toplusunu təmsil edir.

## NƏTİCƏ

İnformasiya texnologiyalarının linqvistikada tətbiqinin yuxarıda göstərilən vasitələrindən (şifahi nitqin analizi və sintezi, mətnlərin avtomatik daxil edilməsi, mətnin avtomatik emalı, mətn korpularından istifadə, dilin kompüter tədrisi və s.) əlavə olaraq, informatika və linqvistikanın: mətnlərdən biliyin əldə edilməsi, sənədlərin avtomatik indeksləşdirilməsi və hissələrə bölünməsi, linqvistikada hipermətn texnologiyaları və s. kimi bu və ya digər sahələr üzrə kəsişməsi mövcuddur.

## ƏDƏBİYYAT


[1] Аппаратное и программное обеспечение компьютера-единое целое, http://www.compgramotnost.ru/vvedenie/chto-takoe-computer

[2] А. Н. Степанов, Информатика: учеб. пособие. СПб.: Питер, 2006.

[3] Л.Ю. Щипицина, Информационные технологии в лингвистике : учеб. пособие, Л.Ю. Щипицина, М. : ФЛИНТА : Наука, 2013.

[4] Г.Г. Белоногов, Компьютерная лингвистика и перспективные информационные технологии. М.: Русский мир, 2004.

[5] А.В. Зубов, И.И. Зубова, Информационные технологии в лингвистике: учеб. пособие для студ. вузов. М.: Академия, 2004.

[6] Е.М. Чухарев, Компьютерные технологии в лингвистических исследованиях: указания по выполнению домашнего задания. Архангельск, 2009.

[7] Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие. М.: МИЭМ, 2011.

[8] Б.И. Большакова, Компьютерная лингвистика: методы, ресурсы, приложения, Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие. М.: МИЭМ, 2011.

[9] https://ru.wikipedia.org/wiki/Корпусная_лингвистика

[10] А.В. Всеволодова, «Компьютерная обработка лингвистических данных», учеб. пособие. 2-е изд., испр. М.: Флинта: Наука, 2007.

[11] Национальный_корпус:, https://ru.wikipedia.org/wiki

[12] Британский_национальный_корпус:, https://ru.wikipedia.org/wiki/

[13] Национальный корпус русского языка:, www.ruscorpora.ru

[14] Чешский_национальный_корпус:, https://ru.wikipedia.org/wiki/

[15] Браун корпус:, http://www.essex.ac.uk/linguistics/external/

[16] clmt/w3c/corpus_ling/content/corpora/list/private/brown/brown.html


DOI: 10.25045/NCSoftEng.2017.81