Informatics

AZƏRBAYCAN DİLİ ÜÇÜN LİNQVİSTİK KORPUSLARIN FORMALAŞDIRILMASI PROBLEMLƏRİ

Yadigar Nəsib oğlu İmamverdiyev[1], Adil Elçin oğlu Əliyev[2]
[1]Azərbaycan Texniki Universiteti, Bakı, Azərbaycan
[2]ÖzünÖyrən MMC, Bakı, Azərbaycan

PROBLEMS OF FORMATION OF LINGUISTIC CORPORA FOR AZERBAIJANI LANGUAGE

Yadigar Nasib Imamverdiyev[1], Adil Elchin Aliyev[2]
[1]Azerbaijan Technical University, Baku, Azerbaijan: yadigar.imamverdiyev@aztu.edu.az
https://orcid.org/0000-0002-3710-1046
[2]OzunOyren LLC, Baku, Azerbaijan: adil@ozunoyren.com
https://orcid.org/0009-0008-5312-0336

Abstract. Currently, computational linguistics and natural language processing technologies are based on the extensive use of data – the corresponding linguistic corpora. Therefore, corpus linguistics, which studies the methods of creating and using such corpuses, has become one of the leading directions in modern linguistics in a relatively short time. Ensuring the widespread use of the Azerbaijani language in the context of globalization is a state policy, and its implementation requires the development and application of appropriate computational linguistics technologies. One of the important prerequisites for the creation of such technologies is the availability of suitable linguistic corpora that meet modern requirements. For this reason, this article analyzes the existing linguistic corpora for the Azerbaijani language, identifies the scientific and practical problems of the formation of these corpora, and gives recommendations for solving them.

Keywords: *corpus linguistics, linguistic corpus, national corpus, parallel corpus, linguistic annotation.*

Giriş

Azərbaycan dilinin qloballaşma şəraitində geniş istifadəsinin təmin edilməsi dövlət siyasətidir [1] və onun həyata keçirilməsi müvafiq kompüter linqvistikası texnologiyalarının işlənməsini və praktiki tətbiqini tələb edir. Belə texnologiyaların işlənməsi zamanı süni intellekt sahəsində əldə edilmiş elmi nailiyyətlər geniş istifadə edilir.

Son zamanlar dərin öyrənmə (ing. Deep Learning) üsulları geniş vüsət almış və süni intellekt sahəsinin inkişafında böyük bir sıçrayışa səbəb olmuşdur. Dərin öyrənmə bir neçə neyron şəbəkə layından ibarət olub, ənənəvi maşın təlimindən fərqli olaraq əlamətləri özü öyrənir. Burada hər bir növbəti lay daha abstrakt əlamətləri öyrənir. Neyron şəbəkə laylarını bu cür birləşdirməklə kifayət qədər mürəkkəb funksiyaları öyrənmək mümkün olur [2].

Dərin öyrənmə üsullarının ən uğurlu tətbiqlərindən biri də kompüter linqvistikası (ing. computational linguistics) sahəsindədir. Belə ki, dərin öyrənmə üsulları təbii dilin emalına (Natural Language Processing, NLP) aid olan, habelə sentiment analiz, dilin modelləşdirilməsi, nitq hissələrinin təyin edilməsi (ing. part of speech, POS tagging), xüsusi isimlərin tanınması (ing. named entity recognition) və s. kimi bir çox problemlərin həllində qənaətbəxş nəticələr göstərmişdir [3].

Bu məqalənin müəllifləri dərin öyrənmə üsullarını azərbaycandilli mətnlərin emalı sahəsində kiçik bir məsələyə tətbiq etməyə cəhd edərkən müvafiq dil korpusunun əlyetər olmaması problemi ilə qarşılaşmışlar. Azərbaycan dili sahəsində kompüter linqvistikasına aid işlərin analizi də bizi dilimiz üçün korpusların işlənməsi vəziyyətini analiz etməyə sövq etmişdir.

Ümumiyyətlə, kompüter linqvistikası ilə məşğul olarkən ən vacib elementlərdən biri də geniş mətn korpusunun olmasıdır. Tədqiq olunan dilə aid mətn korpuslarının olması, nəzəri və empirik tədqiqatlar üçün geniş imkanlar açır [4,5]. Bir çox nəzəri və tətbiqi məsələlərdə istifadə olunan korpusun keyfiyyəti də böyük önəm daşıyır. Keyfiyyəti aşağı olan korpuslarda tipoqrafik, orfoqrafik, sintaktik və bir çox digər xətalara tez-tez rast gəlinir [6,7]. Qənaətbəxş keyfiyyətə malik olmayan və yoxlanılmamış korpuslar tətbiqi məsələlərdə diskriminasiyaya, irqçiliyə, cinsi ayrı-seçkiliyə, siyasi yanlışlıqlara və digər xətalara yol açа bilir [8-11].

Lakin korpusun yaradılması məsələsi heç də asan başa gəlmir. Belə ki, korpus yaradılmasının özünəməxsus prosedurları və onlarla bağlı problemləri vardır. Bəzi dillər üzrə linqvistik problemlər tədqiqatçılar tərəfindən geniş şəkildə araşdırılmış və onların müxtəlif korpusları hazırlanmışdır. Hazırda bu dillər üzrə tədqiqat aparmaq üçün kifayət qədər çoxsaylı linqvistik resurslar mövcuddur. Elə dillər də var ki, bu dillər üzrə linqvistik resursların qıtlığı və yetərli ilkin tədqiqatların aparılmaması (və ya dünya tədqiqatçılarına əlyetər olmaması) həm nəzəri, həm də praktik baxımdan çətinliklər yaradır. Tədqiqatçılar belə dilləri "az-resurslu dillər" (ing. low-resource languages) adlandırırlar. Təəssüf ki, bütün türk dilləri kimi Azərbaycan dili də bu kateqoriyaya aid edilir və bir çox digər dillər kimi onun üçün də linqvistik korpusların işlənməsinə böyük ehtiyac vardır. Bu məqalədə korpusların yaradılmasının texnoloji prosesləri analiz edilir və Azərbaycan dili üçün linqvistik korpusların formalaşdırılması zamanı meydana çıxan problemlər müəyyən edilir.

Linqvistik mətn korpusu nədir?

"Linqvistik mətn korpusu" anlayışının vahid tərifi yoxdur. Ümumiyyətlə, mətn korpusu dedikdə böyük həcmdə autentik mətn nümunələrinin toplusu başa düşülür [12] və eyni zamanda aid olduğu təbii dili ifadə edir [13]. Təbii dilin emalı və müxtəlif linqvistik araşdırmalarda, eləcə də tətbiqi sistemlərin işlənməsində və test edilməsində korpuslar mühüm rol oynayırlar və əsas alət kimi çıxış edirlər.

Korpusu mətnlərin sadə kolleksiyasından fərqləndirən əsas xarakteristikası korpusda ona daxil olan mətnlərin xassələri haqqında əlavə informasiyanın (markerlərin, nişanların, annotasiyanın) olmasıdır. Korpusda hər bir mətnin linqvistik və ekstra-linqvistik (metalinqvistik) nişanı olur. Ekstra-linqvistik nişanda mətnin formatlanması xüsusiyyətləri (başlıq, abzas, sətirbaşı boşluq və s.) haqqında, müəllif haqqında və mətn haqqında məlumat olur (müəllif, ad, nəşr yeri və ili, janr, tematika və s.). Korpusun markerlənməsi ona istifadə sadəliyi və multi-funksionallıq kimi üstünlüklər qazandırır.

Yuxarıda deyilənləri və ədəbiyyat mənbələrini nəzərə alaraq, linqvistik korpuslara belə tərif vermək olar [14]: Linqvistik mətn korpusu – dilçiliyin konkret məsələlərinin həlli üçün nəzərdə tutulmuş, böyük həcmdə, strukturlaşdırılmış, nişanlanmış, reprezentativ dil verilənlərinin elektron formada massividir.

Korpusların növləri

Korpusları müxtəlif əlamətlər (kriteriyalar) ilə təsnifatlaşdırmaq olar. Aydındır ki, klassifikasiya əlaməti mətnlərin dili (ingilis, alman, azərbaycan), istifadə (müraciət) forması (açıq, qapalı, kommersiya), ilkin materialın janrı (bədii, publisistik, sənədli, elmi) ola bilər.

Korpusların başqa təsnifatı da mövcuddur. Bütövlükdə, bütün məlum korpusların dörd variantı reallaşdırır:

- milli korpus – müxtəlif kommunikasiya sahələrində mətnləri əhatə edir (monolinqvistik korpus);
- müqayisəli (kontrastiv) korpus – analoji təşkil edilmiş (reprezentativliyi eyni və korpus meneceri ortaq olan) bir neçə milli korpus;
- paralel korpus – bir dildəki mətnləri və onların başqa dilə (dillərə) tərcüməsi olan korpus;
- danışıq nitqi korpusları – təkcə milli korpusun altkorpusu kimi deyil, ayrıca da mövcud ola bilən korpuslar.

Korpuslar həmçinin tematik və texnoloji baxımdan da təsnif etmək olar. Bu baxımdan tədqiqatlar zamanı heç də hazır korpuslar hər bir məsələ üçün işə yaramır. Müxtəlif hallarda müvafiq tədqiqatlar üçün yeni korpusların yaradılmasına ehtiyac duyulur.

Nümunə üçün maşın tərcüməsi üçün paralel korpuslardan istifadə olunur. Paralel korpusların yaradılması olduqca çətin işdir. Ona görə ki, açıq mənbələrdən paralel korpuslar kifayət qədər əlçatan deyil. Bəzi tədqiqatçılar müxtəlif dillərdə rəsmi tədbir protokolları [15], dini kitablar [16], elmi məqalə abstraktları [17] və bu kimi materiallardan istifadə etməklə paralel korpuslar yaratmışlar. Paralel korpuslar digər korpuslardan onunla fərqlənir ki, burada eyni mətnlərin hər iki dildə olan versiyaları olmalıdır. Bundan başqa, texnologiyadan asılı olaraq hədəf (tərcümədən sonra yekun) dilin ümumi korpusu, bəzən isə sözlərin lüğətləri və müəyyən qaydalar da bu korpusun hissəsi ola bilər.

Linqvistik korpusların yaradılması texnologiyası

Korpusun yaradılmasının texnoloji prosesini ümumi olaraq planlama və icra addımlarına bölmək olar. Planlama zamanı, bir qayda olaraq, korpusa qoyulan kriteriyalar və mənbələrin siyahısı müəyyənləşdirilir. İcra prosesi icazələrin əldə edilməsi (intellektual mülkiyyət baxımından), mətnlərin toplanması və rəqəmsallaşdırılması, mətnlərin ilkin emalı və kodlaşdırılması, mətnlərin linqvistik markerlənməsi (adətən, avtomatik yerinə yetirilir), avtomatik markerləmə nəticələrinin korreksiyası, mətnlərin saxlanılması və sənədləşdirilməsi, istifadəçilərin korpusa girişinin təmin edilməsi, korpusun müntəzəm təkmilləşdirilməsi mərhələlərinə (addımlarına) bölünür [18,19]. Əlbəttə, hər bir konkret halda mərhələlərin tərkibi və sayı yuxarıdakından fərqlənə və real texnologiya daha mürəkkəb ola bilər.

Korpusun yaradılması zamanı qarşıya qoyulan ümumi kriteriyaları aşağıdakı kimi təyin etmək olar:

- Korpus böyük həcmdə mətndən ibarət olmalıdır. Burada "böyük" sözü nisbidir və korpusun həcmi onun istifadə olunacağı məqsədə bağlıdır. Bəzi korpuslar 30~40 min sözdən ibarətdir, bəziləri isə milyardlarla sözdən təşkil edilib.
- Korpus dili və ya mənsub olduğu janrı ifadə etməlidir. Başqa sözlə, reprezentativ olmalıdır. Buraya yazı tərzi, orada istifadə olunan lüğət tərkibi, dəsti-xətt və s. aiddir.
- Kompüterin oxuya biləcəyi formatda olmalıdır. Təbii ki, əgər söbhət mətnin kompüterdə emal edilməsindən gedirsə, uyğun formatda olmalı və vahid kodlaşdırmaya malik olmalıdır.
- Adətən, əlavə linqvistik məlumatlarla annotasiya olunur. Burada da məqsəddən asılı olaraq müxtəlif annotasiyalardan söhbət gedə bilər.

Korpusların markerlənməsi

Korpusu sadə mətnlər korpusundan fərqləndirən əsas xarakteristika korpusda ona daxil olan mətnlərin xassələri haqqında əlavə informasiyanın olmasıdır. Bu informasiya korpusa markerləmə (işarələmə) mərhələsində daxil edilir [14].

Markerləmə (ing. tagging, annotation) – mətnlərə və onların komponentlərinə xüsusi nişanların əlavə edilməsidir. Markerləmənin müxtəlif növləri vardır, onları mətnlərin avtomatik analizinin inkişafı mərhələlərinə uyğun təqdim etmək daha anlaşıqlıdır:

Tokenizasiya (ing. *tokenization*): sonrakı analiz üçün mətndə minimal fraqmentlərin (tokenlərin) müəyyən edilməsi;

Lemmatizasiya (ing. *lemmatization*): morfoloji analiz metodudur, tokenlərin başlanğıc (lüğət) formalarına gətirilməsidir. Lemmatizasiya nəticəsində sözdən flektiv sonluqlar atılır və sözün əsas və ya lüğət forması alınır. İsimlər adlıq hala, fellər məsdər formasına (bəzən məsdər formasından məsdər şəkilçisi də atılır) və s. gətirilir. Qeyd etmək lazımdır ki, istənilən təbii dildə bəzi sözlərin lemmatizasiyası zamanı qeyri-müəyyənlik yarana bilər;

Nitq hissələrinin markerlənməsi (ing. *POS tagging*): sözün aid olduğu nitq hissəsinin müəyyən edilməsi. Hər bir söz müəyyən qrammatik əlamətləri olan nitq hissəsi kimi identifikasiya edilir;

Morfoloji markerləmə (ing. *morphological tagging*): söz-formasının morfoloji əlamətlərinin təyin edilməsi. Morfoloji analiz sonrakı analiz formaları – sintaktik və semantik analiz üçün əsas rolunu oynayır;

Sintaktik markerləmə (ing. *parsing*): sözlərə və sözlərin birləşmələrinə müəyyən sintaktik əlamətlər təyin edilməsi. Mətndə sözlərin qarşılıqlı əlaqələri – cümlələrdə mübtəda, xəbər, təyin, tamamlıq, müxtəlif nitq komponentləri müəyyən edilir;

Semantik markerləmə (ing. *semantic annotation*): sözün müəyyən leksik-semantik sinfə aid edilməsi. Semantik analiz zamanı sözə onun məna kateqoriyalarına və altkateqoriyalarına aidiyyətini əks etdirən teqlər verilir. Belə informasiya mətnlərin tonallığının analizi, avtomatik referatlaşdırma və digər məsələlər üçün çox qiymətlidir;

Bunlar markerləmə növlərinin tam siyahısı deyil. Anaforik, prosodik, struktur, diskurs və s. markerləmə növləri də vardır.

Linqvistik korpuslar sahəsində beynəlxalq standartlar

Korpusların təkrar istifadəsini, digər korpuslarla uyarlığını, hamılıqla qəbul edilmiş elmi nəzəriyyələrə və təsnifatlara uyğunluğunu, ümumi linqvistik prosessorlardan istifadəni, standart proqram vasitələrinin tətbiqini mümkün etmək üçün korpuslar yaradılarkən müəyyən standartlara əməl edilməlidir. Linqvistik korpuslar sahəsində standartlar əsas etibarı ilə TEI (Text Encoding Initiative) və ISLE (International Standards for Language Engineering) layihələri və EAGLES (Expert Advisory Group on Language Engineering Standards) tövsiyələri əsasında formalaşmışdır. Korpus standartları sırasından ilk növbədə CES (Corpus Encoding Standard), XCES (Corpus Encoding Standard for XML) və CDIF (Corpus Document Interchange Format) standartlarını göstərmək olar.

TEI layihəsinə 1988-ci ildə start verilmişdi, əsas məqsədi humanitar sahədə verilənlərin mübadiləsi üçün formatların işlənməsidir. TEI çərçivəsində mətnlərin elektron nəşri üçün bir sıra tövsiyələr işlənmişdir (identifikasiya, təsvir, analiz və interpretasiya, təsvir və kodlaşdırma üçün metadil). Markerləmə vasitələrinin işlənməsi üçün SGML (Standard Generalized Markup Language) və onun altçoxluğu XML (eXtensible Markup Language) istifadə edilir.

Linqvistik markerləri SGML/XML-ə əsaslanan mövcud korpuslar ən müxtəlif kodlaşdırma sistemlərini istifadə edirlər. Məsələn, BNC korpusu CDIF; American National Corpus, Croatian National Corpus və s. XCES; ICE (International Corpus of English), Czech National Corpus və Hungarian National Corpus isə geniş yayılmış TEI standartını istifadə edirlər.

Linqvistik korpuslar üzrə standartlar ISO/TC 37 komitəsinin rəhbərliyi altında işlənib hazırlanır. Ümumi "Language resource management" adına malik olan belə standartlardan aşağıdakıları göstərmək olar:

- ISO 24612:2012 – Linguistic Annotation Framework (LAF);
- ISO 24611:2012 – Morpho-syntactic annotation framework (MAF);
- ISO 24613:2008 – Lexical markup framework (LMF);
- ISO 24615-1:2014 – Syntactic annotation framework (SynAF);
- ISO 24617-1:2012 – Semantic annotation framework (SemAF) – Hissə 1 (SemAF-Time, ISO-TimeML);
- ISO 24617-4:2014 – Semantic annotation framework (SemAF) – Hissə 4: (SemAF-SR).

Azərbaycan dili üçün mövcud korpuslar

Kompüter linqvistikası sahəsində Azərbaycan dili ilə əlaqədar müxtəlif tədqiqatlar aparılmışdır [20-22]. Eyni zamanda kompüter linqvistikası sahəsində lazım olan texnologiyaların yaradılması barədə müvafiq normativ akt da təsdiq edilmişdir [1]. Azərbaycan dili üçün yaradılmış və istifadə edilən bəzi korpuslar barədə aşağıda məlumat verilir.

Dilmanc. Dilmanc Maşın Tərcümə sistemi Azərbaycan dilindən ingilis və əksinə tərcümə etmək imkanı verən proqram təminatıdır [23]. Dilmanc komandası tərcümə sistemi üçün Azərbaycan dili korpusu üzərində çalışmış və xüsusi korpus hazırlamışdır.

Bu korpus azərbaycandilli bir çox veb-saytlardakı mətnlərdən istifadə etməklə hazırlanmış və təqribən 300 milyon tokendən ibarətdir. Eyni zamanda Azərbaycan-ingilis paralel korpusunun hazırlanması üzərində də çalışmışlar.

AzBookCorpus. AzBookCorpus korpusu səkkiz klassik Azərbaycan ədəbiyyatı mənbələrindən istifadə edilməklə yaradılmış və 723055 tokendən ibarətdir.

AzWebCorpus. AzWebCorpus isə 500 Azərbaycan veb-saytından mətnlərin əldə edilməsi hesabına yaradılmış, 492842 tokendən ibarətdir [20].

AzWebCorpus veb-sayt mətnlərindən istifadə etdiyi üçün tərkibində digər dillərdə olan mətnlər də mövcuddur. Müəllifin məqalədə verdiyi cədvəllərə əsasən demək olar ki, korpusdakı tokenlər sözlərin başlanğıc forması kimi deyil, elə cümlədə olduğu kimi istifadə edilir [20]. Ona görə də 723055 token dedikdə, bu qədər söz deyil, sözlərin şəkilçilərlə (qrammatik şəkilçilər daxil olmaqla) birlikdə olan versiyaları başa düşülməlidir.

azWaC (Azerbaijani corpus from the web). Azərbaycan dilli müxtəlif internet resursların mətnlərindən istifadə edilməklə xarici müəlliflər tərəfindən yaradılmış azWaC korpusu 94000 tokendən ibarətdir [24]. Təəssüf ki, bu korpusa daxil olan mətnlərin xüsusiyyətləri barədə ətraflı məlumat verilmir.

Vikipediya. Vikipediya açıq internet ensiklopediyası müxtəlif dillərdə olduğu, eyni zamanda orada olan mətnlərin istifadəyə açıq olduğu üçün bir çox tədqiqatçı kompüter linqvistikası məsələlərində vikipediya mətnlərindən geniş istifadə edir. Buna görə də vikipediya mətnlərindən də xüsusi halda korpus kimi istifadə edirlər. Əsasən söz vektorlarının öyrənilməsində vikipediyadan istifadə geniş yayılmışdır [25,26].

Tanzil. OPUS layihəsi açıq mənbəli bir çox paralel korpuslar hazırlamışdır [16]. Bu layihədə Azərbaycan dili üçün də maraqlı bir korpus vardır.

Belə ki, OPUS layihəsi Tanzil adlı başqa bir layihənin hazırladığı kontentdən istifadə etmişdir ki, bu da Quranın üzərində qurulub. Quran, İncil, Tövrat və digər dini kitablarda müəyyən bir struktur vardır və həmin strukturlar tərcümə olunduğu istənilən dillərdə bir-biri ilə uyğun gəlir. Tanzil layihəsi Quranın bir çox dillərdə tərcümələrini toplamış, OPUS isə o tərcümələrdən istifadə edərək bir çox dillər üçün paralel korpuslar yaratmışdır (http://opus.nlpl.eu/Tanzil.php).

Apertium layihəsi. Apertium layihəsi açıq-kodlu maşın tərcümə sistemidir [27], bir çox dilləri dəstəkləyir. Bu layihənin eksperimental qollarından biri də Apertium-aze alt-layihəsidir ki, bunun üçün də korpus yaratmağa cəhd etmişlər. Hazırlanmış korpus 2012-ci ilə qədər dərc edilmiş bəzi qəzet materialları və Quran əsasında yaradılmışdır.

AZ-SRDat (Azerbaijani language Speaker Recognition DATa). Korpusa ofis şəraitində 86 şəxsdən (21 kişi və 65 qadın) toplanmış azərbaycandilli nitq nümunələri daxildir və səsə görə şəxsin tanınması sahəsində eksperimentlərin aparılması üçün yaradılmışdır [28]. Ondan nitqin tanınması, dilin və aksentin identifikasiyası üçün də istifadə etmək olar.

Azcorpus. Kitablar, jurnallar, mətbuatda dərc olunmuş məqalələr və vikipediya əsasında hazırlanmış korpusdur. Korpus 1.9 milyon sənəd və ümumilikdə 18 milyon cümlədən ibarətdir [29].

Digər korpuslar. Bir sıra tətbiqi məsələlərin həlli və tədqiqatı üçün də kiçik korpuslar yaradılmışdır [30,31]. Təbii ki, bizə məlum olmayan digər korpuslar da vardır. Şübhəsiz ki, Google Translate, Yandex Translate kimi maşın tərcümə sistemləri də Azərbaycan dilində tərcümə etdiyi üçün müəyyən korpuslara malikdir. Lakin bu barədə açıq mənbələrdən məlumat əldə edə bilməmişik.

Azərbaycan dili üçün korpusların formalaşdırılması problemləri

Mövcud linqvistik korpusların analizi nəticəsində Azərbaycan dilində linqvistik korpusların formalaşdırılmasında meydana çıxan aşağıdakı problemləri diqqətə çatdırmaq mümkündür:

- Azərbaycan dili üçün yaradılmış korpusların əksəriyyəti birinci tərtib – "çiy" korpuslardır, onların əksəriyyətində standartlara uyğun keyfiyyətli linqvistik nişanlanma aparılmayıb.
- Azərbaycan dili ilə bağlı NLP məsələləri üzrə tədqiqatlar yetəri qədər geniş deyil. Eyni zamanda, mövcud NLP alətlərinin azərbaycan dilinin xüsusiyyətlərinə uyğunlaşdırılması və ya yenilərinin yaradılması məsələsi də aktual olaraq qalır.
- Linqvistik tədqiqatlarda korpuslardan demək olar ki, istifadə edilmir.
- Linqvistik korpusların yaradılması təşəbbüsləri ilə çox zaman informasiya texnologiyaları üzrə mütəxəssislər çıxış edirlər və bu təşəbbüslər davamlı olmur.
- Linqvistik korpusların yaradılması sahəsində fəaliyyət koordinasiya edilmir.
- Azərbaycan dilinin milli korpusunun yaradılması üzrə hər hansı miqyaslı layihə həyata keçirilməyib.

Nəticə

Böyük həcmli verilənlərlə işləmək dilçilik sahəsində tədqiqatların avtomatlaşdırılmasını tələb edir. Buna görə də hazırda kompüter linqvistikası ilə korpus linqvistikasının sıx qarşılıqlı əlaqəsi gerçəkləşir. Hazırda bir çox dil üçün həcminə və əhəmiyyətinə görə çox mühüm korpuslar yaradılmışdır.

Onların bəzilərində milyardlarla söz vardır. Təəssüf ki, ayrı-ayrı tədqiqatçıların və tədqiqatçı qruplarının cəhdlərinə baxmayaraq, azərbaycan dili üçün reprezentativ korpuslar hələlik mövcud deyil. Bunun bir çox səbəbləri vardır. Mətnlərin korpusa daxil edilməsi və emalı xüsusi səylər və bir sıra mürəkkəb məsələlərin, o cümlədən texniki və təşkilati məsələlərin həllini tələb edir. Azərbaycan dilinin müxtəlif funksional janrları əhatə edən 100 milyonlarla sözdən ibarət milli korpusunun yaradılması bu korpus üçün mətnlər, texnologiyalar, təşkilati və maliyyə resursları təqdim edən şəxslərin və təşkilatların yalnız əlaqələndirilmiş əməkdaşlığı şəraitində mümkündür.

ƏDƏBİYYAT

1. Azərbaycan dilinin qloballaşma şəraitində zamanın tələblərinə uyğun istifadəsinə və ölkədə dilçiliyin inkişafına dair Dövlət Proqramı. Azərbaycan Respublikası Prezidentinin 2013-cü il 9 aprel tarixli Sərəncamı, [Onlayn]. Available: https://president.az/articles/7744.
2. LeCun Y., Bengio Y. and Hinton G. Deep learning. Nature, vol. 521, 2015, no. 7553, pp. 436-444.
3. Xiao R.Z. Well-known and influential corpora. In Corpus Linguistics: An International Handbook. Handbooks of Linguistics and Communication Science, Berlin, Mouton de Gruyter, 2008, pp. 383-457.
4. Walker T., Weinreich U., Labov W., Herzog M.Y. Theory-driven and corpus-driven computational linguistics, and the use o corpora. In Corpus Linguistics: An International Handbook. Handbooks of Linguistics and Communication Science, Berlin, Mouton de Gruyter, 2008, pp. 68-96.
5. Wilks Y. Corpus linguistics and computational linguistics. International Journal of Corpus Linguistics, 2010, vol. 15, no. 3, pp. 408-4011.
6. Grouin C. Certification and Cleaning up of a Text Corpus: Towards an Evaluation of the "Grammatical" Quality of a Corpus. LREC, 2008.
7. Sekiguchi Y., Yamamoto K. Improving quality of the web corpus. Proceedings of The First International Joint Conference on Natural Language Processing (IJCNLP-04), 2004, pp. 201-206.
8. Gumusel E., Malic V.Q., Donaldson D.R., Ashley K., Liu X. An Annotation Schema for the Detection of Social Bias in Legal Text Corpora. Information for a Better World: Shaping the Global Future, 2022, pp. 185-194.
9. Binns R. Fairness in machine learning: Lessons from political philosophy. Conference on fairness, accountability and transparency, 2018, pp. 149-159.
10. Bolukbasi T., Chang K.-W., Zou J.Y., Saligrama V., Kalai A.T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems, 2016.
11. Nissim M., R. van Noord, R. van der Goot. Fair Is Better than Sensational: Man Is to Doctor as Woman Is to Doctor. Computational Linguistics, vol. 46, 2020, no. 2, pp. 487-497.
12. Manning C.D. Computational Linguistics and Deep Learning. Computational Linguistics, 2015, vol. 41, no. 4, pp. 701-707.
13. Bowker L., Pearson J. Working with Specialized Language A Practical Guide to Using Corpora. London, Routledge, 2002, p. 9.
14. McEnery T., Hardie A. Corpus Linguistics: Methods, Theory and Practice. Yearbook of Corpus Linguistics and Pragmatics, 2013, vol. 1, p. 275–277.
15. Koehn P. Europarl: A Parallel Corpus for Statistical Machine Translation. Proceedings of Machine Translation Summit X: Papers, 2005, pp. 79-86.
16. Tiedemann J. Parallel Data, Tools and Interfaces in OPUS. Proceedings of the Eighth International Conference on Language Resources and Evaluation, 2012, pp. 2214-2218.
17. Dogru G., Martín-Mor A., Aguilar-Amat A. Parallel Corpora Preparation for Machine Translation of Low-Resource Languages: Turkish to English Cardiology Corpora. In Proceedings of the LREC 2018 Workshop 'MultilingualBIO: Multilingual Biomedical Text Processing, Japan, 2018.
18. Wynne M. In Developing Linguistic Corpora: A Guide to Good Practice, Oxbow Books, 2005, pp. 8-12.
19. Burnard L. Reference Guide for the British National Corpus (XML Edition). Chapter 1.3, 2007.
20. Adamov A. Text analysis case study: Determining word frequency based on Azerbaijan top 500 websites. Proc. of the 9th International Conference on Application of Information and Communication Technologies (AICT), 2015, pp. 76-79.
21. Mammadova S., Azimova G., Fatullayev A. Text corpora and its role in development of the linguistic technologies for the Azerbaijani language. Proc. of the 3rd International Conference Problems of Cybernetics and Informatics, 2010, pp. 67-70.
22. Mahmudov M.Ə. Mətnin formal təhlili sistemi. Bakı: Elm, 2002.
23. Fatullayev R., Abbasov A., Fatullayev A. "Dilmanc" is the 1st MT system for Azerbaijani. Proceedings of Swedish Language Technology Conference, 2008, pp. 63-64.
24. Baisa V., Suchomel V. Large corpora for Turkic languages and unsupervised morphological analysis. Proceedings of the Eighth International Conference on Language Resources and Evaluation, 2012, pp. 28-32.

25. Bojanowski P., Grave E., Joulin A., Mikolov T. Enriching word vectors with subword information. Transactions of the Association of Computational Linguistics, 2017, vol. 5, no. 1, pp. 135-146.
26. Pennington J., Socher R., Manning C.D. GloVe: Global vectors for word representation. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2014, pp. 1532-1543.
27. Forcada M.L., Tyers F.M., Ramírez-Sánchez G. The Apertium machine translation platform: five years on /. Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation, 2009, pp. 3-10.
28. Имамвердиев Я.Н., Сухостат Л.В. AZ-SRDAT – речевая база данных для азербайджанского языка. İnformasiya texnologiyaları problemləri, 2013, vol. 1, no. 7, pp. 67-73.
29. Kishiyev H., Isbarov J., Suleymanli K., Heydarli K., Eminova L., Zeynalov N. Azcorpus - The largest open-source NLP corpus for Azerbaijani (1.9M documents, ~ 18M sentences). 2023. [Online]. Available: https://huggingface.co/ datasets/ azcorpus/azcorpus_v0.
30. Bannayeva A., Aslanov M. Development of the N-gram Model for Azerbaijani Language. 2020 IEEE 14th International Conference on Application of Information and Communication Technologies (AICT), 2020, pp. 1-5.
31. Huseynov K., Suleymanov U., Rustamov S., Huseynov J. Training and Evaluation of Word Embedding Models for Azerbaijani Language. Digital Interaction and Machine Intelligence, 2021, pp. 37-48.

AZƏRBAYCAN DİLİ ÜÇÜN LİNQVİSTİK KORPUSLARIN FORMALAŞDIRILMASI PROBLEMLƏRİ

Y.N.İmamverdiyev, A.E.Əliyev

Xülasə. Hazırda kompüter linqvistikası və təbii dilin emalı texnologiyaları verilənlərdən – müvafiq linqvistik korpuslardan geniş şəkildə istifadə edilməsinə əsaslanırlar. Bu səbəbdən belə korpusların yaradılması və istifadəsi metodlarını öyrənən korpus linqvistikası qısa müddətdə müasir dilçiliyin aparıcı istiqamətlərindən birinə çevrilmişdir. Azərbaycan dilinin qloballaşma şəraitində geniş istifadəsinin təmin edilməsi dövlət siyasətidir və onun həyata keçirilməsi müvafiq kompüter linqvistikası texnologiyalarının işlənməsini və tətbiqini tələb edir. Belə texnologiyaların yaradılmasının vacib ilkin şərtlərindən biri isə müasir tələblərə cavab verən müvafiq dil korpuslarının mövcud olmasıdır. Bu səbəbdən bu məqalədə Azərbaycan dili üçün mövcud linqvistik korpuslar analiz edilir, bu korpusların formalaşdırılmasının elmi-praktiki problemləri müəyyən edilir və onların həlli istiqamətində tövsiyələr verilir.

Açar sözlər: *korpus linqvistikası, linqvistik korpus, milli korpus, paralel korpus, linqvistik annotasiya*

Accepted: 13.11.2023