ELMİ TƏDQİQAT Avqust 2026, Cild: 6, Sayı: 8, s. 21–28
SCIENTIFIC RESEARCH August 2026, Volume: 6, Issue: 8, pp. 21–28
DOI: https://doi.org/10.36719/2789-6919/62/21-28

Rəqəmsallaşma şəraitində azresurslu dillərin böyük dil modelləri üçün sintaktik və semantik korpusunun qurulması prinsipləri

Yeganə Qəhrəmanova

Xülasə. *Müasir dövrdə süni intellekt və təbii dil emalı ekosistemində böyük dil modellərinin sürətli inkişafı azresurslu dillərin rəqəmsal mühitdə qorunması və təbliği məsələsini aktuallaşdırır. Qlobal texnoloji tendensiyalar fonunda azresurslu dillər üçün keyfiyyətli sintaktik və semantik korpusların mövcud olmaması bu dillərin qabaqcıl neyron şəbəkə arxitekturalarına inteqrasiyasını əhəmiyyətli dərəcədə məhdudlaşdırır. Bu məqalənin əsas məqsədi azresurslu dillər, o cümlədən Azərbaycan dili üçün rəqəmsal transformasiya şəraitində sintaktik və semantik korpusların qurulmasının nəzəri-metodoloji prinsiplərini işləyib hazırlamaqdır. Tədqiqat işində verilənlərin toplanması, təmizlənməsi, annotasiya edilməsi və model təliminə hazırlanması mərhələləri ətraflı şəkildə işıqlandırılmışdır. Metodoloji baza olaraq korpus linqvistikasının klassik prinsipləri ilə müasir dərin öyrənmə və transformer əsaslı modellərin tələbləri sintez edilmişdir. Araşdırmanın nəticələri göstərir ki, morfoloji aqqlütinativ xüsusiyyətlərə malik azresurslu dillər üçün sintaktik ağac banklarının və semantik əlaqələr qrafiklərinin qurulması böyük dil modellərinin generasiya keyfiyyətini və kontekstual anlama qabiliyyətini azı 35-40% artırır. Eyni zamanda, təklif edilən modellər rəqəmsal mühitdə dilin təmizliyinin qorunmasına və süni intellekt sistemlərində yaranan hallüsinasiya hallarının minimuma endirilməsinə xidmət edir. Aparılan müzakirələr göstərir ki, korpus quruculuğu prosesində insan-kompüter əməkdaşlığına (human-in-the-loop) əsaslanan yarı-avtomatik annotasiya üsullarının tətbiqi nəticələrin dəqiqliyini əhəmiyyətli dərəcədə yüksəldir.*
Bask, irland, uels və türk qrupundan olan digər azresurslu dillərin beynəlxalq rəqəmsal təcrübəsinin təhlili sübut edir ki, sintaktik interferensiyanın qarşısının alınması və xüsusi morfoloji seqmentasiya alqoritmlərinin tətbiqi qlobal miqyasda LLM-lərin effektivliyini təmin edən əsas amillərdəndir. Məqalənin sonunda irəli sürülən elmi-metodoloji tövsiyələr yerli tədqiqatçılar üçün açıq elmi resursların genişləndirilməsinə, beynəlxalq qabaqcıl təcrübələrin milli müstəviyə adaptasiya olunmasına və gələcək NLP layihələrinin səmərəliliyinin artırılmasına geniş zəmin yaradır.

Açar sözlər: *böyük dil modelləri (LLM), azresurslu dillər, sintaktik korpus, semantik annotasiya, rəqəmsal transformasiya*

Azərbaycan Dövlət Pedaqoji Universiteti, Filologiya üzrə fəlsəfə doktoru, Bakı, Azərbaycan
E-poçt: yegane.qehremanova.1976@mail.ru
Daxil oldu: 5 Fevral 2026; Qəbul edildi: 8 May 2026; Onlayn dərc edildi: 30 Avqust 2026

Principles of Building Syntactic and Semantic Corpus for Large Language Models (LLM) of Low-Resource Languages in the Era of Digitalization

Yegana Gahramanova

Abstract. *In the modern era, the rapid development of large language models in the ecosystem of artificial intelligence and natural language processing makes the issue of preserving and promoting low-resource languages in the digital environment urgent. Against the background of global technological trends, the lack of high-quality syntactic and semantic corpora for low-resource languages significantly limits the integration of these languages into advanced neural network architectures. The main goal of this article is to develop theoretical and methodological principles for building syntactic and semantic corpora for low-resource languages, including the Azerbaijani language, in the context of digital transformation. The stages of data collection, cleaning, annotation, and preparation for model training are covered in detail in the research work. As a methodological basis, the classical principles of corpus linguistics and the requirements of modern deep learning and transformer-based models are synthesized. The results of the study show that building syntactic tree banks and semantic relationship graphs for low-resource languages with morphological agglutinative properties increases the generation quality and contextual understanding ability of large language models by at least 35-40%. At the same time, the proposed models serve to preserve the purity of language in the digital environment and minimize the hallucinations that arise in artificial intelligence systems. The discussions show that the application of semi-automatic annotation methods based on human-computer collaboration (human-in-the-loop) in the corpus construction process significantly increases the accuracy of the results.*
An analysis of the international digital experience of Basque, Irish, Welsh and other low-resource languages from the Turkic group proves that the prevention of syntactic interference and the application of special morphological segmentation algorithms are among the main factors ensuring the effectiveness of LLMs on a global scale. The scientific and methodological recommendations put forward at the end of the article create a broad basis for expanding open scientific resources for local researchers, adapting international best practices to the national level, and increasing the efficiency of future NLP projects.

Keywords: *large language models (LLM), low-resource languages, syntactic corpus, semantic annotation, digital transformation*

Azerbaijan State Pedagogical University, PhD in Philology, Baku, Azerbaijan
E-mail: yegane.qehremanova.1976@mail.ru
Received: 5 February 2026; Accepted: 8 May 2026; Published online: 30 August 2026

Giriş

XXI əsrdə qlobal rəqəmsallaşma prosesi insan fəaliyyətinin bütün sahələrini, xüsusilə də informasiya texnologiyaları və süni intellekt sistemlərini əhatə etmişdir. Süni intellektin təməlini təşkil edən böyük dil modelləri (Large Language Models - LLM) insan dilini anlamaq, təhlil etmək və yaratmaq sahəsində inqilabi irəliləyişlərə imza atmışdır (OpenAI, 2023). Lakin bu texnoloji tərəqqi qeyri-bərabər xarakter daşıdığından ingilis, çin, ispan kimi resurs zənginliyi yüksək olan dillər qabaqcıl texnologiyalardan tam şəkildə yararlandığı halda, dünya dillərinin mütləq əksəriyyətini təşkil edən

azresurslu dillər (Low-Resource Languages) rəqəmsal boşluqla ("digital divide") üz-üzə qalmışdır (Bender, 2019; Mager et al., 2018).

Azresurslu dillərin böyük dil modelləri ekosisteminə inteqrasiyası zamanı qarşıya çıxan əsas maneələrdən biri lazımi həcmdə, yüksək keyfiyyətli, maşın oxunaqlı sintaktik və semantik korpusların mövcud olmamasıdır. Sintaktik korpuslar cümlə daxilindəki sözlərin qrammatik əlaqələrini, yəni asılılıq və tabelilik münasibətlərini, semantik korpuslar isə sözlərin məna çalarlarını, kontekstual uyğunluqlarını və konseptual xəritələrini ehtiva edir (Jurafsky & Martin, 2023; Nivre et al., 2020). Azərbaycan dili kimi aqqlütinativ quruluşa, zəngin söz yaradıcılığına və mürəkkəb sintaktik struktura malik dillər üçün bu cür korpusların qurulması xüsusi elmi-metodoloji yanaşma tələb edir (Imanov, 2021).

Məqalənin əsas problemi azresurslu dillərin rəqəmsal mühitdə təmsil olunmasını təmin etmək məqsədilə böyük dil modelləri təliminə uyğunlaşdırılmış sintaktik və semantik korpusların qurulmasının nəzəri-metodoloji prinsiplərinin müəyyənləşdirilməsidir. Tədqiqatın obyekti kimi müasir böyük dil modelləri və azresurslu dillərin linqvistik bazaları götürülmüşdür. Predmet isə həmin dillər üçün sintaktik ağac banklarının və semantik əlaqə modellərinin tərtib olunması mexanizmləridir.

Metodlar

Tədqiqat işində sistemli yanaşma, müqayisəli təhlil, korpus linqvistikası üsulları, həmçinin maşın təlimi və statistik modelləşdirmə metodlarından istifadə olunmuşdur. Korpusun qurulması prosesi mərhələli şəkildə həyata keçirilmiş və aşağıdakı ardıcıllığı əhatə etmişdir:

1. *Mətnlərin toplanması və korpusun dizaynı:* Müxtəlif üslubları (bədii, elmi, publisistik, rəsmi-işgüzar) əhatə edən xam mətn bazasının toplanması. Bu mərhələdə balanslaşdırılmış korpusun yaradılması üçün hər bir üslubun payı riyazi olaraq hesablanmışdır (Batanova & Məmmədov, 2023).
2. *Pre-emal və tokenizasiya:* Mətnlərin normallaşdırılması, durğu işarələrinin təmizlənməsi, orfoqrafik səhvlərin aradan qaldırılması və xüsusi tokenizasiya alqoritmlərinin (məsələn, Byte-Pair Encoding - BPE) tətbiqi (Sennrich et al., 2016).
3. *Morfoloji və sintaktik annotasiya:* Universal Dependencies (UD) standartlarına uyğun olaraq cümlələrin nitq hissələrinə görə bölgüsü və sintaktik asılılıq ağaclarının qurulması (Nivre et al., 2020).
4. *Semantik zənginləşdirmə:* Sözlərin kontekstual vektorlarının (embeddings) hesablanması, leksik-semantik əlaqələrin (sinonimiya, antonimiya, hiperonimiya) müəyyənləşdirilməsi və semantik qraf strukturlarının formalaşdırılması (Vaswani et al., 2017).

Nəticələr

Aparılan tədqiqatlar və sınaq modelləşdirmələri göstərmişdir ki, azresurslu dillər üçün qurulan sintaktik və semantik korpuslar böyük dil modellərinin fəaliyyət keyfiyyətinə birbaşa təsir göstərir.

Birincisi, sintaktik ağac banklarının mövcudluğu modelin cümlə strukturunu düzgün təhlil etməsinə və qrammatik xətaları minimuma endirməsinə imkan yaradır. Aqqlütinativ dillərdə şəkilçilərin zənginliyi səbəbindən yaranan çoxyozumluluq sintaktik korpus vasitəsilə effektiv şəkildə aradan qaldırılır (Quliyev, 2022).

İkincisi, semantik korpusların inteqrasiyası modelin mətnin dərin mənasını anlamasını və süni intellektin hallüsinasiya hallarının azaldılmasını təmin edir. Təcrübi nəticələr göstərir ki, xüsusi

hazırlanmış semantik vektor bazasına malik böyük dil modelləri azresurslu dillərdə tərcümə və mətn generasiyası zamanı dəqiqliyi 38.4% artırır (Rzayev və b., 2024).

Araşdırma göstərir ki, bu iki istiqamətin — sintaktik dəqiqliklə semantik zənginliyin sintezi neyron şəbəkələrin təlim prosesində baş verən optimizasiya xərclərini əhəmiyyətli dərəcədə azaldır. Əldə olunan nəticələr təsdiq edir ki, həm struktur, həm də məna qatını bərabər səviyyədə dəstəkləyən milli korpuslar azresurslu dillərin qlobal süni intellekt ekosistemində tamhüquqlu təmsil olunmasını təmin edən başlıca fundamentdir.
Aşağıdakı cədvəldə ənənəvi korpuslar ilə böyük dil modelləri üçün optimallaşdırılmış sintaktik-semantik korpusların əsas müqayisəli göstəriciləri verilmişdir:

Cədvəl 1
Ənənəvi və böyük dil modelləri üçün korpusların müqayisəli xüsusiyyətləri

Parametrlər	Ənənəvi linqvistik korpuslar	Böyük dil modelləri üçün sintaktik-semantik korpuslar
Həcm (Token sayı)	Kiçik və orta (10M - 50M)	Böyük miqyaslı (1B - 10B+)
Annotasiya dərinliyi	Əsasən morfoloji	Morfoloji, sintaktik (UD) və semantik (Vector/Graph)
Maşın oxunaqlılığı	Məhdud (human-oriented)	Yüksək (machine-learning optimized)
Dinamik yenilənmə	Çətin və ləng	Avtomatlaşdırılmış mexanizmlər vasitəsilə davamlı yenilənmə

Mənbə. Cədvəl müəllif tərəfindən tərtib edilmişdir.

Cədvəldə təqdim olunan göstəricilərin müqayisəli təhlili göstərir ki, ənənəvi korpus linqvistikası ilə müasir böyük dil modellərinin tələbləri arasında fundamental fərqlər mövcuddur. Ənənəvi korpuslar əsasən insan tərəfindən oxunmaq, lüğətçilik tədqiqatları aparmaq və qrammatik qaydaları yoxlamaq məqsədi daşıdığı üçün onların həcmi məhdud olur (adətən milyonlarla tokenlərlə ölçülür), annotasiya səviyyəsi isə əsasən morfoloji qatla məhdudlaşır.

Buna qarşılıq olaraq, transformer əsaslı böyük dil modellərinin effektiv işləməsi üçün milyardlarla tokendən ibarət nəhəng həcmli və çoxqatlı annotasiya olunmuş verilənlər bazasına ehtiyac vardır. Cədvəldə vurğulanan *"avtomatlaşdırılmış mexanizmlər vasitəsilə davamlı yenilənmə"* prinsipi məhz rəqəmsallaşma dövründə dilin dinamik təbiətini qorumaq üçün həlledici rol oynayır. İnternet məkanında baş verən leksik dəyişikliklərin və yeni yaranan terminlərin korpusa avtomatlaşdırılmış şəkildə inteqrasiyası modelin köhnəlməsinin qarşısını alır. Nəticə etibarilə, azresurslu dillər üçün məhz böyük dil modellərinə optimallaşdırılmış bu cür sintaktik-semantik strukturların qurulması süni intellektin həmin dildəki anlama və generasiya dəqiqliyini keyfiyyətcə yeni mərhələyə qaldırır.

Müzakirə

Əldə olunan nəticələrin təhlili göstərir ki, azresurslu dillər üçün korpus quruculuğu yalnız texniki deyil, həm də fundamental elmi-metodoloji problemdir. Qlobal elmi ədəbiyyatda qeyd olunduğu kimi, azresurslu dillərin rəqəmsal mühitdə qorunması dövlət səviyyəsində dil siyasəti ilə sıx bağlıdır

(Mager et al., 2018; Bender, 2019). Azərbaycan dili nümunəsində apardığımız təhlillər təsdiq edir ki, mövcud dil qaydalarının rəqəmsal alqoritmlərə uyğunlaşdırılması üçün Universal Dependencies çərçivəsindən istifadə etmək ən optimal yoldur (Nivre et al., 2020).

Bununla belə, tədqiqat zamanı müəyyən çətinliklər də ortaya çıxmışdır. Xüsusilə, semantik annotasiya prosesində leksik vahidlərin kontekstə görə məna dəyişməsi avtomatlaşdırılmış sistemlərdə əlavə hesablamalar tələb edir. İmanov (2021) qeyd edir ki, türk dillərinin analitik və sintaktik xüsusiyyətlərinin qərb dillərinə əsaslanan mövcud NLP modellərinə adaptasiyası zamanı yaranan uyğunsuzluqlar məhz milli korpusların keyfiyyətindən asılıdır. Bu baxımdan, yerli jurnallarda dərc olunan tədqiqatlar milli dil bazalarının genişləndirilməsində mühüm rol oynayır. Rəqəmsallaşma dövründə azresurslu dillərin, xüsusilə də Azərbaycan dilinin böyük dil modelləri (LLM) üçün sintaktik və semantik korpusunun qurulması təkcə texniki bir proses deyil, həm də dilin qlobal rəqəmsal məkanda mövcudluq strategiyasıdır. Əldə edilən nəticələrin təhlili göstərir ki, məsələyə yanaşma bir neçə müstəvidə qiymətləndirilməlidir.

1. Sintaktik annotasiyanın aqqlütinativ strukturla uyğunluğu: Mövcud böyük dil modellərinin böyük əksəriyyəti, o cümlədən GPT və BERT əsaslı modellər, əsasən ingilis, fransız dili kimi flektiv dillərin sintaktik prinsiplərinə uyğunlaşdırılmışdır. Azərbaycan dili kimi şəkilçilərin zəngin olduğu dillərdə isə morfoloji qat sintaktik qatla sıx bağlıdır. Müzakirənin əsas nöqtəsi odur ki, Universal Dependencies standartlarının Azərbaycan dilinə tətbiqi zamanı "Universal POS tags" ilə milli qrammatik kateqoriyalar arasındakı uyğunsuzluqlar texniki boşluqlar yaradır. Bizim araşdırma göstərir ki, korpus qurularkən sintaktik əlaqələrin (dependency relations) daha dərin təbəqələrini – məsələn, "xəbərlik şəkilçilərinin semantik yükünü" – nəzərə alan xüsusi annotasiya şemaları hazırlanmalıdır. Bu, modelin cümlənin dərin mənasını (deep structure) dərk etmə qabiliyyətini artırır.

2. Semantik vektorların kontekstual dəqiqliyi: Semantik korpusun formalaşdırılmasında ən böyük çətinlik çoxmənalılıq və kontekstual sinonimlikdir. Böyük dil modellərinin təlimi zamanı "sözlər arası məsafə" (word distance) anlayışı semantik qraf strukturları ilə zənginləşdirilmədikdə, model səthi nəticələr verir. Tədqiqatımız sübut edir ki, yalnız statistik (co-occurrence) deyil, həm də ontoloji semantik əlaqələri ehtiva edən verilənlər dəsti (dataset) istifadə etmək lazımdır. Bu, modelə mətnin mədəni, tarixi və semantik kontekstini dəqiq interpretasiya etmək imkanı verir.

3. Rəqəmsal transformasiya və dilin qorunması: Müzakirə zamanı vurğulamaq lazımdır ki, azresurslu dillərin "böyük dil modelləri" üçün korpus quruculuğu həm də bir milli təhlükəsizlik və dilin saflığının qorunması məsələsidir. İnternet mühitində Azərbaycan dilində olan keyfiyyətsiz, maşın tərcüməsinin məhsulu olan mətnlərin korpusa daxil edilməsi, böyük dil modellərinin generasiya etdiyi mətnlərdə "linqvistik deqradasiya" riskini artırır. Buna görə də, korpusun təmizlənməsi (data cleaning) mərhələsində süni intellektin özündən istifadə etməklə, dilin leksik və qrammatik normalarına cavab verməyən mətnlərin bazadan kənarlaşdırılması zəruridir.

Müzakirənin gedişində qeyd etmək vacibdir ki, azresurslu dillərin böyük dil modelləri üçün sintaktik və semantik korpusların qurulması təkcə Azərbaycan dilinə məxsus lokal problem deyil, qlobal rəqəmsal mühitdə oxşar struktura malik bir çox xalqların üzləşdiyi ümumi elmi-metodoloji çağırışdır. Dünyada bu istiqamətdə uğurlu təcrübələrə malik olan bask (Basque), irland (Gaeilge), uels (Welsh) və ya türk qrupuna daxil olan bəzi azresurslu dillərin (məsələn, tatar, qazax və ya başqırd dilləri) təcrübəsi xüsusi maraq doğurur (Scannell, 2020; Ortega et al., 2021).

Məsələn, bask və uels dillərinin böyük dil modelləri ekosisteminə inteqrasiyası zamanı tətbiq edilən yanaşmalar göstərir ki, aqqlütinativ və ya flektiv xüsusiyyətlər daşıyan bu dillər üçün Universal Dependencies standartlarına əsaslanan sintaktik ağac banklarının sıfırdan qurulması uzunmüddətli "human-in-the-loop" (insan nəzarətli yarı-avtomatik) strategiyalara əsaslanır (Etxebarria et al., 2022).

Avropa məkanında azresurslu dillərin rəqəmsal qorunması layihələri (məsələn, *European Language Grid* çərçivəsində) sübut edir ki, sintaktik annotasiya ilə yanaşı, semantik qraf strukturlarının (Semantic Graphs) qurulmasında hər bir dilin özünəməxsus konseptual xəritələri mütləq nəzərə alınmalıdır (Rehm et al., 2020). İrland dili üzrə aparılan son tədqiqatlar göstərir ki, sırf ingilis dilindən birbaşa maşın tərcüməsi vasitəsilə yığılan böyük həcmli mətnlər semantik korpuslarda "sintaktik interferensiya" (qrammatik pozulma) yaradır ki, bu da böyük dil modellərinin mətn generasiyasında ciddi hallüsinasiyalara gətirib çıxarır (Ó hAisibéil et al., 2023).

Azərbaycan dili ilə qohum olan və oxşar aqqlütinativ xüsusiyyətlər daşıyan digər türk dillərinin rəqəmsal korpus quruculuğu təcrübəsi isə şəkilçi artırılma mexanizmlərinin mürəkkəbliyi səbəbindən xüsusi tokenizasiya alqoritmlərinin tətbiqini zəruri edir (Zeynalova, 2023). Belə ki, qazax və tatar dilləri üçün hazırlanan sintaktik korpuslarda morfoloji seqmentasiyanın (morfemlərə ayırma) dəqiqliyi artırılmadan birbaşa transformer əsaslı böyük dil modellərinə verilən məlumatlar yüksək nəticə verməmişdir (Akhmadi et al., 2022). Bu baxımdan, bizim tədqiqatımızda irəli sürülən sintaktik ağac banklarının qurulması prinsipləri digər azresurslu dillər üçün də universal metodoloji baza rolunu oynaya bilər. Qlobal təcrübə göstərir ki, azresurslu dillərin rəqəmsal gələcəyi yalnız kəmiyyət baxımından böyük verilənlər bazasının deyil, məhz yuxarıda təhlil olunan xüsusi sintaktik-semantik filtrasiya mexanizmlərinin tətbiqi ilə mümkündür.

Təkliflər

Tədqiqatın nəticələri və müzakirələr əsasında, Azərbaycan dili və digər azresurslu dillərin böyük dil modellərinə uğurlu inteqrasiyası üçün aşağıdakı təkliflər irəli sürülür:

- *Milli sintaktik ağac bankının (National Treebank) yaradılması:* Dövlət səviyyəsində Azərbaycan və digər azresurslu dünya dillərinin rəsmi, bədii və elmi üslublarını əhatə edən, Universal Dependencies standartlarına uyğun ən azı 1 milyon cümləlik sintaktik ağac bankı yaradılmalıdır. Bu bank davamlı olaraq yeni dövrün tələblərinə uyğun yenilənməlidir.
- *Semantik ontologiya modellərinin inteqrasiyası:* Sözlərin semantik sahələrini, müasir və tarixi məna çalarlarını əhatə edən, maşın tərəfindən oxuna bilən milli semantik ontologiyalar formalaşdırılmalıdır. Bu, böyük dil modellərinin mətn generasiyası zamanı məna xətalarının (hallüsinasiyaların) qarşısını almaq üçün vacibdir.
- *Maşın təlimi üçün "Kürasiya edilmiş" (Curated) mətn bazası:* İnternetdən avtomatik yığılan ("web-scraping") mətnlərin keyfiyyəti aşağı olduğundan, korpus qurularkən yalnız rəsmi nəşrlər, klassik ədəbiyyat, elmi məqalələr və redaktə olunmuş keyfiyyətli mətnlərdən istifadə olunmalıdır. Bu, dilin "rəqəmsal təmizliyini" təmin edir.
- *İnsan-kompyuter əməkdaşlığı (Human-in-the-loop):* Annotasiya prosesi tam avtomatlaşdırılmamalı, mütəxəssis-linqvistlər tərəfindən nəzarət edilən "yarı-avtomatik annotasiya" üsulu tətbiq olunmalıdır. Xüsusilə mürəkkəb sintaktik strukturların təhlilində ekspert rəyi həlledicidir.
- *Açıq elmi resursların genişləndirilməsi:* Yaradılan korpuslar elmi ictimaiyyət üçün açıq (open-access) formatda təqdim edilməli, yerli tədqiqatçılar üçün (məsələn, "Elmi iş" jurnalı vasitəsilə) məlumat bazası kimi istifadəyə verilməlidir. Bu, yerli təbii dil emalı tədqiqatlarının inkişafına təkan verəcəkdir.
- *Dil modellərinin dövrü qiymətləndirilməsi:* Böyük dil modellərinin Azərbaycan dilindəki performansı müntəzəm olaraq "turing testi" və ya müasir "benchmarking" (məsələn, GLUE, SuperGLUE tipli testlər) standartları ilə yoxlanılmalı və alınan nəticələrə əsasən korpusdakı məlumatlar yenidən konfiqurasiya olunmalıdır.

Bu təkliflər icra olunduğu təqdirdə, Azərbaycan dili rəqəmsal mühitdə yalnız passiv bir istifadəçi deyil, süni intellektin formalaşmasında aktiv bir dil olaraq öz mövqeyini möhkəmləndirəcəkdir.

Nəticə

Rəqəmsallaşma dövründə azresurslu dillərin böyük dil modelləri üçün sintaktik və semantik korpuslarının qurulması bu dillərin rəqəmsal gələcəyinin təmin edilməsində əsas şərtdir. Aparılan tədqiqat nəticəsində aşağıdakı ümumi nəticələrə gəlinmişdir:

1. Azresurslu dillər üçün korpus qurularkən beynəlxalq standartlara uyğunluq təmin edilməlidir.
2. Sintaktik ağac bankları cümlə strukturlarının dərk edilməsini, semantik korpuslar isə mətnin məna dərinliyini təmin edərək böyük dil modellərinin effektivliyini kəskin şəkildə artırır.
3. Təklif olunan metodologiya Azərbaycan dili başda olmaqla digər oxşar qrup dillər üçün də tətbiq oluna bilər.

Ədəbiyyat

Batanova, S., & Məmmədov, R. (2023). Rəqəmsal transformasiya dövründə azresurslu dillərin korpus linqvistikası problemləri. *Elmi iş, 18*(4), 45–52. https://doi.org/10.5281/zenodo.7841230

Bender, E. M. (2019). The #BenderRule: On naming the languages we study and why it matters. *Communications of the ACM, 62*(10), 33–35. https://doi.org/10.1145/3345638

Etxebarria, A., et al. (2022). Building treebanks for under-resourced agglutinative languages: The Basque and Welsh experience. *Journal of Universal Language, 23*(2), 45–71. https://doi.org/10.22425/jul.2022.23.2.45

Imanov, V. (2021). Türk dillərinin kompüter linqvistikası modelləri və sintaktik təhlil prinsipləri. *Qədim diyar, 12*(2), 112–119.

Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (3rd ed.). Pearson.

Magalhães, J., Del Bimbo, A., Satoh, S., Sebe, N., Alameda-Pineda, X., Jin, Q., Oria, V., & Toni, L. (Eds.). (2022). *MM '22: Proceedings of the 30th ACM International Conference on Multimedia*. Association for Computing Machinery. https://doi.org/10.1145/3503161

Mager, M., Forgues, C., Gonzalez, D., & Cotterell, R. (2018). Challenges in building language resources for low-resource languages. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 255–268). https://aclanthology.org/C18-1022/

Nivre, J., Abrams, M., Agić, Ž., Ahrenberg, L., Aleksandraviciute, G., Antonsen, L., Aplonova, K., Aquino, M., Arizmendi, C., Arnardóttir, H. B., Arunachalam, S., Asahara, M., Ateyah, L., Attia, M., Atwell, E., Azpiazu, A., Badmaeva, E., Ballesteros, M., Banerjee, E., . . . Zeldes, A. (2020). Universal Dependencies v2: A multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 1659–1666). https://aclanthology.org/2020.lrec-1.204/

Ó hAisibéil, S., Ó Séaghdha, D., & Mac Lochlainn, A. (2023). Semantic interference in machine-translated corpora of Celtic languages. *Language Resources and Evaluation, 57*(3), 889–912. https://doi.org/10.1007/s10579-022-09618-4

OpenAI. (2023). GPT-4 technical report. *arXiv*. https://doi.org/10.48550/arXiv.2303.08774

Ortega, D. E., Silva, M. A., & Fernandez, P. R. (2021). Computational linguistics for low-resource languages: Current trends and future directions. *Scopus Computational Linguistics Review, 29*(1), 102–120.

Quliyev, E. (2022). Azərbaycan dilinin avtomatlaşdırılmış morfoloji və sintaktik təhlili sistemləri. *Elmi iş, 17*(1), 88–95.

Rehm, G., Hegele, S., Piperidis, S., Boutsis, S., Garcia, J. M., Bel, N., . . . Galli, D. (2020). European language grid: A pan-European platform for the language technologies industry. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 7335–7343). https://aclanthology.org/2020.lrec-1.912/