Azərbaycan Milli Elmlər Akademiyası M. Füzuli adına Əlyazmalar İnstitutu, *ISSN 2224-9257*
Filologiya məsələləri № 8, 2025
Səh. 51-55

https://doi.org/10.62837/2025.9.51

GÜLNAR NOVRUZ QIZI SAYILOVA
AMEA Nəsimi adına Dilçilik İnstitutu
Eksperimental Fonetik Araşdırmalar Laboratoriyası
sailovagulnar1995@icloud.com

SÜNİ İNTELLEKT TEXNOLOGİYALARININ NİTQ MATERİALLARININ ARXİVLƏŞDİRİLMƏSİNDƏ TƏTBİQİ.

Xülasə

Bu məqalədə səsli və yazılı nitq materiallarının müxtəlif proqramlar vasitəsilə toplanması və çeşidlənməsi prosesinin süni intellekt vasitəsilə inteqrasiyasından bəhs olunur. Nitq materiallarının saxlanılması və arxivləşdirilməsi prosesində süni intellekt texnologiyalarından istifadə edilməsinin əsas məqsədi məlumatları sağlam şəkildə saxlamaq və dil irsini qorumaqdır. Arxivləşdirilmə süni intellekt texnologiyalarının inkişafı ilə də birbaşa əlaqəlidir. Çünki bu texnologiyalarından istifadə etməklə nitq materiallarını arxivləşdirmək daha əlverişli imkanlar qazanmış olur.

Məqalədə nitq prosesinin arxivləşdirilməsinin mərhələli şəkildə təsviri verilmişdir. Dil faktlarının arxivləşdirilməsi zamanı bir neçə proqramdan istifadə olunur ki, bunların içərisində ən geniş şəkildə istifadə olunanı ELAN proqramıdır. Bu proqram sayəsində istifadəçi materialı həm audio, həm video, həm də yazılı şəkildə saxlaya və nitq siqnalının akustik analizini apara bilər.

ELAN proqramında Azərbaycanın müxtəlif bölgələrində qeydə alınmış dilin fərqli dialekt və şivələrini əks etdirən nitq materialı, eləcə də ədəbi dil nümunələri tərəfimizdən bu proqrama daxil edilmişdir.

Açar sözlər: Dil korpusu, süni intellekt, ELAN, nitq materialının arxivləşdirilməsi, kompüter dilçiliyi.

Giriş.

Müasir dövrdə dünya dillərinin bir çoxu təhlükə altına düşdükcə, həmin dillərə aid materiallar gələcəkdə araşdırılacaq tədqiqat obyekti kimi alimlər və icmalar üçün əsas mənbəyə çevrilir. Sənədli materiallar insan dilinin potensialı haqqında biliklərimizi genişləndirmək üçün empirik əsas rolunu oynayır, müxtəlif mədəni və diskursiv ənənələrin reyestrini və ictimai irsin maddi qeydinin gələcək nəsillərə çıxışını təmin edir. Bununla belə, bu materiallar adətən qısa ömürlüdür. Audio və video lentlər qırılır, SD kartlar və sərt disklər yanğınlara, daşqınlara və dəyişən texnologiyaya qarşı həssasdır. Dilçilər və digər mütəxəssislər dil sənədlərinin nəticələrini qorumaq və gələcəkdə bu materialları əldə etmək üçün mühüm resurs kimi rəqəmsal arxivlərə daha çox müraciət edirlər.

Qeyd edək ki, verilənlərin arxivləşdirilməsi və sıxılması arasında fərq mövcuddur. Birincisi bir neçə faylı və hətta qovluqları bir faylda - arxivdə

birləşdirməyə, ikincisi isə artıqlığı aradan qaldırmaqla (itkisiz qablaşdırma, yəni orijinal faylları dəqiq bərpa etmək imkanı ilə) mənbə fayllarının ölçüsünü azaltmağa xidmət edir. Arxivləşdirmə böyük həcmli məlumatları bir anda saxlamaq və ya ötürmək üçün istifadə olunur.

Nitq materialının arxivləşdirilməsi süni intellekt dövründə mühüm əhəmiyyət kəsb edir. Onların istifadəsi avtomatik nitqin tanınması və sintezi texnologiyalarının inkişafı, dil modellərinin hazırlanması, linqvistik müxtəlifliyin qorunması kimi məsələlərin həllinə kömək edir (1, s.74).

Nitq materialının arxivləşdirilməsi prosesinin mərhələli təsvirini bu şəkildə vermək olar:

1. Məlumatların (səs materialının) toplanması.
2. Rəqəmsallaşdırma: Analoq medianın (kasetlər, çarxlar) rəqəmsal formata çevrilməsi.
3. Emal: Səs keyfiyyətinin təmizlənməsi və təkmilləşdirilməsi.
4. Transkripsiya: Danışıq dilinin yazılı formaya çevrilməsi.
5. Annotasiya: İşarələmə (fonetik, sintaktik və s.)
6. Saxlama: Arxivlərdə və verilənlər bazasında metadata ilə yerləşdirmə.

Dünyada nitq arxivlərinin müxtəlif nümunələri mövcuddur. Onlardan ELAR (Nəsli kəsilməkdə olan Dillər Arxivi) - Kiçik və nəsli kəsilməkdə olan dillərin sənədləşdirilməsi, TIMIT Fonetik şərhli ingilis nitqi, LibriSpeech Audiobook-a əsaslanan açıq ingilis dilli korpusu, Common Voice (Mozilla) Crowdsourced nitq korpusu, Lingua Libre (Wikimedia) - müxtəlif dillərdə sözlərin tələffüzlərinin qeydə alınması və arxivləşdirilməsi kimi layihələri qeyd etmək olar.

Nitqin arxivləşdirilməsi nitq məlumatlarının toplusunun yaradılması və strukturlaşdırılması prosesidir, o cümlədən: audio yazılar (canlı çıxışlar, müsahibələr, dialoqlar, hekayələr), nitq komponentləri olan videolar, transkripsiyalar (audio fayllara uyğun mətnlər), metadata (natiqlər haqqında məlumatlar, vaxt, yazının yeri, dil və s.), annotasiyalar (fonetik, morfoloji, sintaktik, praqmatik və s.) xüsusi qaydada toplanılır.

Arxivləşdirmədə əsas məqsəd dil irsinin qorunması, xüsusilə nəsli kəsilməkdə olan və azlıqların dilləri, dialektləri və etnoqrafik irslərinin qeydə alınmasıdır. Elmi tədqiqat baxımından bu tipli material dilçilik (fonetika, morfologiya, sosiolinqvistika), psixolinqvistika, etnoqrafiya, tarix, sosiologiya üçün maraqlıdır. Bu tip arxivlər canlı şifahi ənənələrin qeydə alınması baxımından xüsusi əhəmiyyət daşıyır.

Nitq materialının arxivləşdirilməsi süni intellekt texnologiyalarının inkişafı ilə əlaqədar daha da aktuallaşmışdır (2; 3). Nitqin tanınması, nitq sintezi, dialoq sistemləri, dil modellərinin yaradılmasında nitq materialı xüsusi mənbə kimi çıxış edir.

Nitq materialının arxivləşdirilməsi üçün bir sıra proqram təminatları işlənib hazırlanmışdır. Bunlardan ən çox istifadə olunan, qeyd etdiyimiz kimi, ELAN proqramıdır.

ELAN proqramı – Maks Plank Psixolinqvistika İnstitutu tərəfindən hazırlanmış audio/video verilənlər üçün linqvistik annotasiya vasitəsidir. O, audio və/və ya video yazıları qeyd etmək üçün nəzərdə tutulmuşdur - xüsusilə dilçilik, işarə dili tədqiqatları, jest təhlili, dil sənədləri və multimodal məlumatları əhatə edən digər sahələrdə faydalıdır.

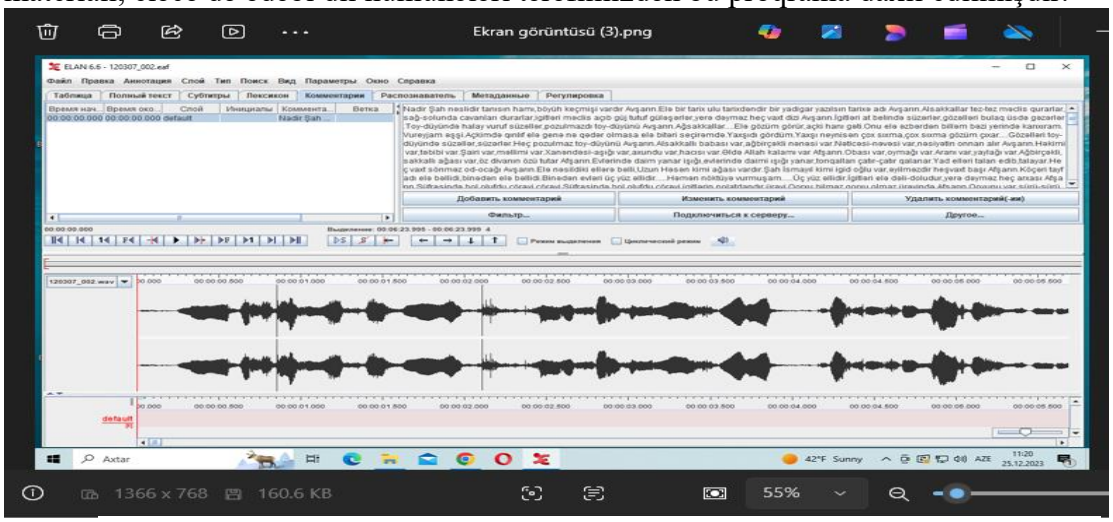
ELAN ilə istifadəçi audio və/və ya video yazılarına məhdudiyyətsiz sayda mətn annotasiyaları əlavə edə bilər. Annotasiya bir cümlə, söz və ya ifadə, şərh, tərcümə və ya mediada müşahidə olunan hər hansı bir xüsusiyyətin təsviri ola bilər. Annotasiyalar səviyyə adlanan bir neçə təbəqədə yaradıla bilər. Səviyyələr iyerarxik olaraq bir-birinə bağlı ola bilər. Annotasiya ya zamanla mediaya uyğunlaşdırıla bilər, ya da digər mövcud annotasiyalara istinad edə bilər. Annotasiyaların məzmunu Unicode mətnindən ibarətdir və annotasiya sənədləri XML formatında (EAF) saxlanılır.

ELAN-ın əsas xüsusiyyətləri bunlardır:

- annotasiyalara baxmaq üçün bir neçə yol təqdim edir, hər bir görünüş media qrafikinə qoşulur və onunla sinxronlaşdırılır;
- çoxsəviyyəli və təksəviyyəli iyerarxiyaların yaradılmasını dəstəkləyir;
- Nəzarət olunan lüğətləri dəstəkləyir;
- annotasiya sənədi ilə 4-ə qədər video faylı əlaqələndirməyə imkan verir;
- media dəstəyi;
- Windows Media Player, QuickTime və ya VLC kimi mövcud, doğma media çərçivələri üzərində qurulur;
- audio və video formatlarının dəstəklənməsi əməliyyat sistemindən asılıdır, adətən yüksək performanslı media oxunmasına nail olmaq olar.

Dil korpusunun yaradılmasında ELAN annotasiya proqramının tətbiqi dillər üçün standartlaşdırılmış transkripsiya sistemlərinin və yazılı formaların olmaması səbəbindən çətinliklər meydana çıxır (4; 5). Azərbaycan Respublikasında süni intellektin inkişafını sürətləndirmək, süni intellekt üzrə informasiya texnologiyalarının və onların idarə edilməsi mexanizmlərinin təkmilləşdirilməsini təmin etmək məqsədi ilə imzalanmış "Azərbaycan Respublikasının 2025–2028-ci illər üçün süni intellekt Strategiyası"nın təsdiq edilməsi haqqında Azərbaycan Respublikası Prezidentinin 19 mart 2025-ci il tarixli Sərəncamı bu sahənin inkişafına şərait yaratmağı nəzərdə tutur.

Azərbaycan dili materialları əsasında ELAN proqramında Azərbaycanın müxtəlif bölgələrində qeydə alınmış, dilin fərqli dialekt və şivələrini əks etdirən nitq materialı, eləcə də ədəbi dil nümunələri tərəfimizdən bu proqrama daxil edilmişdir.



Toplanılan materiallar gələcəkdə Azərbaycan dilinin milli korpusuna əlavə edilə və nitq texnologiyaları ilə bağlı müxtəlif layihələrdə istifadə oluna bilər. Dilçi və informasiya texnologiyaları sahəsində çalışan alimlər arasında davamlı fənlərarası əməkdaşlıq bu istiqamətin inkişafına töhfəsini verə bilər.

ƏDƏBİYYAT

1. Zhang et al., "Speech Data Collection and Annotation for AI", 2022.
2. AUSTIN, Peter K. Language Documentation & Legacy Text Materials. Asian and African.
3. HIMMELMANN, Nikolaus P. Linguistic Data Types and the Interface between Language. Documentation and Description. Language Documentation & Conservation. vol. 6. p. 187-207. 2012.
4. Languages and Linguistics. n. 11. p. 23-44. 2017
5. Moore et al., "The Role of Speech Corpora in AI", 2021.
6. Common Voice Project by Mozill

Gulnar Sayilova

APPLICATION OF ARTIFICIAL INTELLIGENCE TECHNOLOGIES IN ARCHIVING SPEECH MATERIALS.

Summary

This article discusses the integration of the process of collecting and sorting audio and written speech materials through various programs into artificial intelligence. The main goal of using artificial intelligence technologies in the process of storing and archiving speech materials is to preserve information in a healthy way

and preserve the linguistic heritage. Archiving is also directly related to the development of artificial intelligence technologies. Because it is considered more convenient to archive speech materials using these technologies. The article provides a step-by-step description of the archiving of speech processes. Several programs are used in the archiving of linguistic facts, the most widely used of which is the ELAN program. Thanks to this program, the user can save the material in both audio and written form and use it conveniently. The ELAN program includes speech material recorded in various regions of Azerbaijan, reflecting different dialects and accents of the language, as well as examples of literary language.

Key words: Language corpus, artificial intelligence, ELAN, speech archiving, computational linguistics

Гульнар Сайилова

ПРИМЕНЕНИЕ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ПРИ АРХИВИРОВАНИИ РЕЧЕВЫХ МАТЕРИАЛОВ

Резюме

В данной статье рассматривается интеграция процесса сбора и сортировки аудио- и письменных речевых материалов с помощью различных программ в искусственный интеллект. Основной целью использования технологий искусственного интеллекта в процессе хранения и архивирования речевых материалов является сохранение информации в здоровом виде и сохранение языкового наследия. Архивирование также напрямую связано с развитием технологий искусственного интеллекта. Поскольку считается, что архивировать речевые материалы с помощью этих технологий удобнее. В статье дается пошаговое описание архивирования речевых процессов. При архивировании языковых фактов используется несколько программ, наиболее широко используемой из которых является программа ELAN. Благодаря этой программе пользователь может сохранять материал как в аудио-, так и в письменной форме и удобно им пользоваться. Программа ELAN включает в себя речевой материал, записанный в различных регионах Азербайджана, отражающий разные диалекты и акценты языка, а также примеры литературного языка.

Ключевые слова: Корпус языка, искусственный интеллект, ELAN, архивирование речи, компьютерная лингвистика.

Rəyçi: f.ü.f.d., dosent Nəzakət Qazıyeva