Azerbaijani-TueCL: Universal Dependency Treebank for Azerbaijani Language

Soudabeh Eslami

University of Tübingen

soudabeh.eslami@student.uni-tuebingen.de

Introduction

Azerbaijani language, also known as Azeri belongs to the Turkic branch of the Altaic language family and is classified under the Oghuz sub-group. Serving as the official language of the Republic of Azerbaijan, Azerbaijani dialects are also spoken in parts of Iran, Georgia, Russia, Turkey, and Iraq. The language exhibits notable dialectal variations, primarily between North Azerbaijani, spoken in the Republic of Azerbaijan, and South Azerbaijani, prevalent in Iran. Azerbaijani is characterized as an agglutinative language, with affixes attached as suffixes to the stem, and dialects differ in pronunciation, vocabulary, and grammar.

Given Azerbaijani's classification as a low-resource language concerning treebanks, the project aims to address this gap by creating a Universal Dependency treebank. Initially, the project began with annotating approximately 100 Azerbaijani sentences, including 20 from the Cairo corpus and additional sentences recommended by the UD Turkic Group, as a foundational step towards achieving this goal.

Motivation

As a graduate student in Computational Linguistics, I am enthusiastic about enhancing my skills in various aspects of linguistics. Azerbaijani-TueCL, the first Azerbaijani treebank, currently passed validation and will be realized on May 15th, 2024. I am looking forward to this training as it will help me improve the initial version of the treebank. I will learn how to add more annotation layers and morphological annotations, which will further enhance the linguistic analysis. Additionally, I am excited to increase my knowledge about Universal Dependencies (UD) and become familiar with Surface-Syntactic Universal Dependencies (SUD). I am also excited to gain experience with tools such as Grew-Match and different data formats like ConNLL-U and CUPT. This will aid me in developing my technical skills and enhancing my understanding of the subject.

Attending the training school presents an excellent opportunity for me to connect with experts and new members of the UD community. Our past networking efforts, during the UniDive 2nd meeting in Istanbul, have been successful in kickstarting the creation of the Azerbaijani Treebank and laying out a plan for the future as a part of the Parallel Turkic Treebank project. This upcoming training will build on that experience, helping me to further refine the project goals and strategies. Additionally, the advanced skills gained from the training will better equip me for the next steps in this project, enabling me to address complex research challenges more effectively. Ultimately, the knowledge and expertise gained from the training school will directly contribute to the success of our project, allowing me to share my knowledge with others and mentor new members, thus ensuring the project's sustainability and growth.

In summary, taking part in the training school enhances our ongoing project and sets me up for future success in linguistic annotation and analysis.

Challenges

While annotating Azerbaijani sentences, we faced several challenges that made the annotation process difficult. Our latest study proposed strategies for annotating pronominalized locatives in Turkic Universal Dependency Treebanks. However, there are still many issues that need to be resolved, such as how to effectively deal with empty token instances, like omitted copulas. We also need to determine the best strategies for identifying and annotating these instances in the treebank. Additionally, we need to find effective methods for annotating and analyzing zero derivation from adjectives to nouns. We also need to determine the specific linguistic features or patterns that characterize existential sentences in Azerbaijani. Furthermore, we need to find methods to accurately segment multi-word expressions within sentences, among many other challenges.

Project Description

The project was initiated during the summer semester of 2023 at the University of Tübingen under the supervision of Dr. Çağrı Çöltekin. In this phase, we manually annotated approximately 110 sentences with part-of-speech (PoS) tags and dependency relations. This includes 20 sentences from Cairo and about 90 sentences recommended by the UD Turkic Group. The lemmas have also been included in the treebank. Azerbaijani-TueCL currently passed validation and will be realized on May 15th, 2024.

Looking ahead, the project envisions several future tasks. These include the creation of an Azerbaijani-Persian Code-Switching treebank, expansion of the parallel treebank for Turkic languages, and augmentation of the current version with more sentences. Notably, Azerbaijani is written in three distinct alphabets: the Persian alphabet in the South, and the Cyrillic and Latin alphabets in the North. Currently, the treebank comprises sentences in the Latin script, with plans to incorporate more sentences written in the Arabic-based Persian alphabet. Potential working tasks for the project include the addition of sentences with Cyrillic script to diversify the linguistic corpus and establishing a new treebank for North Azerbaijani. These tasks will contribute to the enrichment and diversity of the treebank, thereby enhancing its utility and applicability for linguistic research and analysis.

References

Furkan Akkurt, Bermet Chontaeva, Çağrı Çöltekin, Mehmet Oguz Derin, Gulnura Dzhumalieva, Soudabeh Eslami, Tunga G¨ung¨or1, Sardana Ivanova, Murat Jumashev, Aida Kasieva, Aslı Kuzgun, B¨u¸sra Mar¸san, Balkız Ozt¨urk, Chihiro Taguchi, Susan ¨Uskudarlı, Jonathan Washington and Olcay Taner Yıldız. *Unifying the Annotations in Turkic Universal Dependencies Treebanks*, 2nd General Meeting of UniDive (COST Action CA21167: Universality, diversity and idiosyncrasy in language technology), University of Naples L'Orientale, Italy, 7-9 February 2024.

Jonathan Washington, Çağrı Çöltekin, Furkan Akkurt, Bermet Chontaeva, Soudabeh Eslami, Gulnura Jumalieva, Aida Kasieva, Aslı Kuzgun, Büşra Marşan and Chihiro Taguchi. *Strategies for the Annotation of Pronominalised Locatives in Turkic Universal Dependency Treebanks,* Joint International Conference on Computational Linguistics, Language Resources, and Evaluation (LREC-COLING 2024), Turin, Italy, 20-25 May 2024.

Lee, Sooman Noah. (1996). *A grammar of Iranian Azerbaijani*. University of Sussex, UK.