UniDive Training School Application

UNIVERSAL DEPENDENCIES FOR UZBEK

Arofat Akhundjanova
arak00001@stud-uni.saarland.de
Department of Language Science and Technology
Saarland University, Germany

Key words: Uzbek, annotation, tagset, UD treebank, parsing

1. Description of a resource related to the topics of the training school

Morphological and syntactic tagsets, along with tagged corpora, are essential resources in language processing. Among the many languages with a growing treebank inventory, Uzbek has been one of the less fortunate languages, lacking annotated corpora. My graduate thesis project seeks to fill this gap by initiating the Universal Dependencies (UD) treebank for Uzbek. This endeavor is crucial for advancing computational linguistics in Uzbek, a low-resource language with limited NLP research and resources.

Creating a standard annotation framework serves both theoretical and practical purposes. Theoretically, it facilitates comparative linguistic studies and crosslinguistic typology. Practically, it enables the development of fully automatic taggers and multilingual NLP systems.

In addition to introducing the treebank, my project will explore linguistic constructions in Uzbek that challenge the UD framework, providing language-specific documentation to explain main linguistic features of this language.

2. Explanation of how the participation in the training school will be useful for the project

Participating in the training school will greatly benefit my project in several ways:

1. Enhanced Understanding of Annotation Principles: The training will provide a valuable opportunity to learn more about annotation principles shared by Turkic languages. I aim to gain insights into annotating complex linguistic structures such as compound words, verbal phrases, and other ambiguities specific to Uzbek. This knowledge will be essential in ensuring the accuracy and consistency of the UD treebank I am developing.

2. Language-Specific Guideline Development: I'll gain skills necessary to tailor annotation guidelines for Uzbek, pivotal for effective treebank annotation and documentation.

3. Familiarization with Automatic Annotation Tools: Exploring various automatic annotation tools and format validators will streamline the annotation process and ensure adherence to the UD standards.

4. Networking and Collaboration Opportunities: The school offers a platform for networking with experts and peers engaged in treebank development and facilitates knowledge exchange, and potential collaborations, enriching the project's scope and impact.

3. Open questions related to the project which could be addressed during the brainstorming hackathon

1. How can we leverage transfer learning techniques to enhance the accuracy and efficiency of annotation for low-resource languages like Uzbek?
2. What strategies can be employed to optimize the integration of automatic annotation tools with manual annotation efforts, balancing accuracy and efficiency?
3. How can we develop robust evaluation metrics and benchmarks to assess the performance and effectiveness of the UD treebank for Uzbek?

4. Short statement of the project phase (planning, started, in the process of creation)

The project is in the process of creation.

Accomplished steps:

1. Data selection and preprocessing: 500 sentences – 5747 tokens
2. Development of Uzbek morphological tagset
3. Semi-automatic morphological annotation with UzMorphAnalayzer tool
4. Development of Uzbek dependency relations tagset

Next steps:

1. Automatic sytactic parsing with Stanza tool
2. Manual correction and disambiguation
3. Development of language-specific guidelines and documentation
4. Pass content and format validation
5. Release of the treebank