Title:

Construction of a UD Treebank for Mukri Kurdish and Modification of Existing UDs for Sorani and Kurmanji Kurdish and Persian with Emphasis on Multi-Word Expressions

Applicant:

Hiwa Asadpour, Goethe University Frankfurt, Germany

Keywords:

Mukri Kurdish, Sorani Kurdish, Kurmanji Kurdish, Persian, Universal Dependencies, Treebank Annotation, Multi-Word Expressions

Description of the Resource:

The proposed project aims to address the dearth of linguistic resources for Mukri Kurdish, an Indo-European language variety spoken by a significant minority population. Mukri Kurdish is spoken in northwestern Iran and is in contact with Jewish and Christian Neo-Aramaic, Azeri Turkish, Armenian, Kurmanji Kurdish, all of which under superstratum of Persian as the official language of Iran. The language is spoken by 2 million people; however, it is under intensive influence from Persian and the neighboring languages. Mukri has some similarities with Kurmanji and Persian, however, phonologically, morphologically and syntactically there are very distinctive differences. For example, synthetic structures and a very complex cliticization features of this language variety make complex MWEs which cannot be made in Kurmanji Kurdish or Persian.

Despite its cultural and historical significance, Mukri Kurdish lacks a standardized Universal Dependencies (UD) Treebank, which is crucial for advancing research in computational linguistics, natural language processing (NLP), and machine learning. Through detailed annotation of syntactic and morphological structures, the UD Treebank will provide a comprehensive linguistic resource for Mukri Kurdish, facilitating various NLP tasks such as parsing, machine translation, and sentiment analysis. Thus it will be a great typological input to the UD by adding this language variety which is a low-resourced language but spoken by a large speech community.

In addition to constructing the UD Treebank for Mukri Kurdish, the project seeks to enhance existing UD Treebanks for Sorani Kurdish, Kurmanji Kurdish, and Persian by refining the annotation of Multi-Word Expressions (MWEs). Multi-Word Expressions, including compound verbs, idiomatic phrases, and collocations, pose significant challenges in natural language processing due to their complex syntactic and semantic properties. By detailed pilot annotating MWEs in the target languages, the project aims to improve the accuracy and coverage of existing UD Treebanks, thereby enhancing their utility for a wide range of linguistic and computational applications.

During discussions with Dan Zeman and other colleagues in online meetings and in the last general meeting in Naples, Italy, several challenges related to Multi-Word Expressions were highlighted, particularly concerning their classification within the UD framework and the development of annotation guidelines. For example, in Kurmanji Kurdish, verbs such as "dihatin xwendin" (to be learning) or "dibin dixwin" (to be seeing eating) present unique challenges in annotation, as they consist of multiple lexical elements with distinct syntactic and semantic roles. Similarly, in Sorani Kurdish, Kurmanji Kurdish and Persian, compound verbs and idiomatic expressions exhibit complex morphological and syntactic structures that require careful analysis and annotation. By addressing these challenges, the project aims to produce high-quality annotated data sets that reflect the rich linguistic diversity of the target languages and contribute to the advancement of research in computational linguistics and NLP in contribution to the instructors and in close working relation with other colleagues in the training school.

How Participation in the Training School Will Be Useful:

Participation in the UniDive training school will provide invaluable opportunities to collaborate with leading experts in UD Treebank annotation and Multi-Word Expression annotation. Engaging with instructors like Sylvain Kahane and Francis Tyers will offer insights into best practices and methodologies for constructing UD Treebanks, particularly for low-resource languages. This is a great opportunity for me because in the Naples, I was in a talk with Sylvain to start a project on annotation of Kurdish. Thus it is a unique opportunity to initiate the work in person while online and remote meetings are not easy and optimal in the first initiation stages. Additionally, interactions with Verginica Mititelu and Voula Giouli will facilitate the development of robust annotation strategies for capturing the intricacies of Multi-Word Expressions, ensuring the quality and consistency of the annotated data sets. I had close contact with Voula during our sub-tasks meetings and we discussed various points. Thus in an in person meeting, tailored for such training and task is great to improve and develop a Treebank for the sample languages in my project.

Furthermore, the brainstorming hackathon will provide a forum for addressing unresolved issues and exploring innovative solutions to challenges encountered during the annotation process. By leveraging the collective expertise and resources available through the training school, the project aims to produce comprehensive linguistic resources that benefit both academic research and practical NLP applications and it will be beneficial for the sub-tasks related to linguistic diversity and typology and low-resource languages with focusing on annotation of MWEs.

Open Questions for the Brainstorming Hackathon:

- How can we develop effective annotation guidelines for Multi-Word Expressions that accommodate the syntactic and semantic variations observed across different language varieties esp. for languages with lesser amount of resources or not a well-established guidelines in their annotation?
- What strategies can be employed to address the challenges of annotating Multi-Word Expressions in low-resource languages, where linguistic expertise and annotated data are scarce for example for Mukri Kurdish which is a language without any standard variation?
- How can we ensure the interoperability of the proposed UD Treebanks with existing linguistic resources and tools, such as parallel corpora and lexical databases from various genres and sources? Here I can also bring examples of personal field data which I turned it into a corpus.
- Are there existing annotation tools or frameworks that can streamline the annotation process for complex linguistic phenomena, such as compound verbs and idiomatic expressions, in the target languages?

Project Phase:

In the Process of Creation