1st UniDive Training Summer School 2024
The title: Kyrgyz Universal Dependency Treebank
Applicant's name and affiliation (including the country):
Bermet Chontaeva, Tübingen of University, Germany
A list of 3-4 key-words:
Kyrgyz, syntax, annotation, treebanks, Universal Dependencies, Turkic languages, Parallel treebank,...
Description of a resource related to the topics of the training school:
- Universal Dependency (UD) treebank
- low resource language

Explanation how the participation in the training school will be useful for the project:
Late nights poring over sentences and trying to decipher my GitHub comments have taught me a lot, but they've also made me realize something important: I need to up my tech game. As a General Linguistics master student and native Kyrgyz speaker, my goal is to create treebanks for Kyrgyz. But I've encountered an obstacle—I lack the necessary technical knowledge.
After analyzing around 100 sentences, it's clear to me that technical knowledge is one of the missing pieces of the puzzle. I'm progressing slowly, and what I really need is hands-on training, like what your summer school offers. From mastering GitHub to understanding validation processes, these workshops seem tailor-made for someone like me.
I'm not alone in this struggle. Working with a group of Kyrgyz linguists in Kyrgyzstan, we've all realized our collective lack of computer literacy. Bridging this gap isn't just about speeding up our projects; it's about understanding the modern world of linguistics (computer linguistics) better.
By attending the summer school, I'll not only be able to share from here, Europe what I learn with my team but also with students back in Kyrgyzstan. And it's not just about Kyrgyz—it's about helping researchers in all Turkic languages.
The opportunity to connect with like-minded individuals facing similar challenges but in different languages is incredibly exciting. I'm looking forward to exchanging experiences, making new acquaintances, and possibly even laying the groundwork for future collaborations. And beyond the professional benefits, I'm excited to finally put faces to the avatars and meet everyone in person at the summer school. Until then, take care!

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

- empty tokens.
- copula (simple copula sentences)
- copula as auxiliary in Kyrgyz: kir: ketken e-, ketse e-, ketet e-
- existentials: kir: jok/bar, az: yox/var, tr: yok/var
- obligue/object distinction
- compounds/compounds:lvs vs. object/oblique
- small words: kir: da, kerek, ele
- two-part words: alip bar- (kir/kaz), chygyp getti (kir/az/kaz)
- code-switching
- 'periphrastic' negative finite verb forms (kir/kaz: bargan joksun, tat: narganin yuq)
- MWE/segmentation
- semantic representation
- root in parataxis, compound sentences
- adpositions
- question particles: kir: 'бы' (translit: 'by'), tr: mi
- adpositions
- double causative
- passive/cusative

- pronominals
- zero Adj -> Noun derivation
- clausal argumets/ non-finite verbs
- -ki and double -ki : kir: üydögünü-kü - the one in the small house.
- inflection versus derivation

- language related structure can be not shown with current UD instruction.

- struggling to pass the validation, because of the directions of the errors. If we change it into validation. The language is not presented well.

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

The Kyrgyz-TueCL project has started during the summer semester of 2023 at the University of Tübingen under the guidance of Dr. Çağrı Çöltekin, as part of the Tools and Resources for Low-resourced Languages course. This initiative, spearheaded by me, a native Kyrgyz speaker, involves curating a corpus of approximately 120 sentences. These sentences, which include 20 Cairo sentences and around 100 suggestions from the UD Turkic Group, have manual annotation for part-of-speech (PoS) tags and dependency relations. Additionally, lemmas have been incorporated into the treebank. Originally written in Turkish with English translations, the sentences have been translated into Kyrgyz and Azerbaijani, forming a segment of the parallel Turkic Treebank.

The primary goal is to augment the forthcoming Kyrgyz treebank, scheduled for release on May 1st, 2024, by introducing new annotation layers and morphological annotations to enhance linguistic analysis. Furthermore, participation in the UniDive 2nd meeting in Istanbul has provided invaluable insights and connections, enabling the integration of this project into the broader Parallel Turkic treebank initiative. This collaboration has also catalyzed the creation of the second Kyrgyz language treebank.

Looking ahead, the project aims to expand its scope by venturing into the creation of a Kyrgyz-Russian Code-Switching treebank and developing parallel treebanks for other Turkic languages. Additional tasks on the agenda include augmenting the current version with more sentences, morphology, and addressing transliteration issues, given that Kyrgyz is currently written in Cyrillic script.

To achieve these objectives, the project will focus on tasks such as adding parallel sentences in other Turkic languages and establishing a new treebank dedicated to Kyrgyz grammar. Participation in the summer school promises to equip me with the necessary tools and insights to propel the project forward, enabling me to contribute meaningfully to the advancement of low-resourced language resources.