Report on the outcomes of a Short-Term Scientific Mission[1]

Action number: CA18209

Grantee name: Manjola Lumani Zacellari

Details of the STSM

Title: Building language corpora for Albanian as a low resource language for research and educational purposes

Start and end date: **02/10/2023 to 06/10/2023**

Description of the work carried out during the STSM

UD is being implemented in a large number of projects. For the Albanian language, so far there are two syntactically annotated data sets, which are not sufficiently of a good quality.

Together with the host researcher Dr. Besim Kabashi (University of Erlangen-Nuremberg, Germany) we decided to work on analysing the UD Albanian Treebank developed by Marsida Toska, under the supervision of Joachim Nivre at Uppsala University, Schweden, (https://universaldependencies.org/treebanks/sq_tsa/index.html) and to improve it as we came across of some flaws of the data set.

The existing UD Treebank for Standard Albanian (TSA) consists of 60 sentences corresponding to 922 tokens, which means that its size is extremely small, which makes difficult to train models based on it.

During my stay at University of Erlangen-Nuremberg we analyzed this existing treebank and decided – first of all – to refine it (with new features, i.e syntactic relations) and then to increase its size with more sentences.

This STSM in fact, where we analysed the existing UD Albanian treebank, has served me as a starting point for my ongoing collaborating with Dr. Kabashi to continue working for a new, more precise and larger enough, UD treebank for Albanian.

[1] This report is submitted by the grantee to the Action MC for approval and for claiming payment of the awarded grant. The Grant Awarding Coordinator coordinates the evaluation of this report on behalf of the Action MC and instructs the GH for payment of the Grant.

Funded by the European Union

We analysed the morphological and the syntactic relations. We concluded that there are several mistakes that should be edited, as for example in the sentence 33 of the treebank the verb "janë" considered in the treebank as an auxiliary verb, it is in fact a verb with full lexical meaning.

This undertaking is an ongoing process and requires a deep understanding of the Albanian language, its unique grammatical structures, and syntactic nuances. We meticulously analyzed the sentences, scrutinizing every word, phrase, and dependency relation to ensure that they align with the established linguistic rules and conventions.

Through a combination of manual verification and automated techniques, we sift through the treebank, identifying any inconsistencies, ambiguities, or inaccuracies. We aimed at methodically correct and enhance the annotations, resulting in a more robust and reliable resource for natural language processing tasks.

Beyond the mere correction of mistakes, this work also involved enriching the treebank with additional linguistic information. We went beyond the surface-level annotations and delved deeper into the underlying syntactic structures, capturing more intricate relations between words and uncovering hidden patterns that may have been overlooked initially.

The enhanced treebank not only empowers future natural language processing applications but also contributes to the broader field of linguistics, advancing our knowledge of Albanian syntax and paving the way for more comprehensive linguistic analyses.

Apart from that we also analysed another treebank of Albanian, the UD Gheg GPS Albanian treebank https://universaldependencies.org/treebanks/aln_gps/index.html (developed by Christian Ebert, Artan Islamaj, Adrian Kuqi, Barbara Sonnenhauser, Paul Widmer, Magdalena Plamada)

UD Gheg GPS contains 966 sentences from 64 recordings of Gheg speakers re-narrating the *Pear Stories* video. Due to the multilingual setting, the treebank contains many instances of code-switching (mostly Swiss-German). It also exhibits characteristics of (semi-)spontaneous speech, like disfluencies and corrections.

However, we started also working on this treebank as some morphological and syntactic features need to be verified and edited. Because this is a treebank consisting of spoken Gheg dialect we identified a lot of inconsistencies, mainly related to words that in fact don't represent the Gheg dialect! There are also not enough syntactic relations, or other features, which are crucial for the UD.

Description of the STSM main achievements and planned follow-up activities

Description and assessment of whether the STSM achieved its planned goals and expected outcomes, including specific contribution to Action objective and deliverables, or publications resulting from the STSM. Agreed plans for future follow-up collaborations shall also be described in this section.

As explained in the section above, because the UD treebank for Albanian has several inconsistencies, we worked on it and we are currently working to correct and perfect the treebank and plan to create a new one.

After that, I have more knowledge in the topic and I'm interested to continue working on that area. Dr. Kabashi and me are planning to work together an annotating such corpora, and of course consider to publish the results.

Therefore, during my stay at Erlangen University we have seen:

- how corpora are built
- Then we analysed the UD rules https://universaldependencies.org/introduction.html
- we analysed how the work with a small data set for Albanian in UD is performed https://universaldependencies.org/sq/index.html
- then we analysed e discussed selected sentences
- analysed a data set of Gheg UD treebank https://universaldependencies.org/aln/index.html
- on 4/10/2023 together with Dr. Besim Kabashi we have analysed and improved some of the mistakes of the treebank
- On Thursday and Friday, we worked on how to expand the treebank, by also completing the two treebanks with morphological and syntactic categories that were missing in both data sets.
- In the last two days, we have focused on finding new syntactic relations that were not represented in the current data sets.
- As a conclusion, we aim at creating a new data set (from scratch).

Literature:

https://universaldependencies.org/introduction.html

https://aclanthology.org/2020.udw-1.20/

Additional literature

https://aclanthology.org/people/b/besim-kabashi/

https://aclanthology.org/L16-1682/

https://aclanthology.org/L18-1412/

https://www.besim-kabashi.net/publications.html