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

Action number: CA18209

Grantee name: Enriketa Sogutlu

Details of the STSM

Title: Building language corpora for educational and research purposes

Start and end date: 02/10/2023 - 06/10/2023

Description of the work carried out during the STSM

On the first day of the visit, the researcher was introduced to the building of linguistic corpora for research purposes and to the steps involved in creating language corpora in general. Universal Dependencies (https://universaldependencies.org/) tree bank was explored in terms of component parts: tokenization and word segmentation, POs features, morphology, and syntax. As syntax constituted the focus of the visit, the researcher became familiar with its description in the Albanian tree bank (https://universaldependencies.org/sq/index.html) and analysed the six components included in it. Then the researcher was introduced to annotation of linguistic corpora focusing on syntactic annotation. After reading the total of 60 sentences in the Albanian tree bank (https://github.com/UniversalDependencies/UD_Albanian-TSA/blob/master/sq_tsa-ud-test.conllu), 10 sentences were randomly selected. The researcher identified and selected the sentences containing wrong annotations, which were then double checked in cooperation with the host researcher. Initially the type of mistake performed in the existing annotation was identified then the improved version of the annotation was considered. During evaluation of annotations, the researcher noted down all types of wrong annotations as well as challenges and issues encountered during evaluation of the selected sentences.

After provision of the improved version of all the selected sentences containing mistakes, double and peer checking was performed and the results were discussed among the researchers. In addition, other identified issues and challenges were discussed along with how they were addressed, or how they could be addressed in future work. A discussion and comparison of results demonstrated that the most common issues were: insufficient syntax content and categories and explanation in terms of sentence and clause types; wrong annotation of the verb to be (its classification as auxiliary even when used as a main verb). Other issues constituted the need for a detailed classification of different types of determiners, in particular particles preceding adjectives.

Part of the evaluation included discussion of utilization of the results for educational and research purposes. The guest researcher will integrate examples from the results (including challenges and how to address them) in the instruction of syntactic dependencies to English language majors in her institution, mostly focusing on comparative

[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.

analysis of English and Albanian thus making use of the Albanian tree bank and other languages included in the UD,

Finally, the researchers discussed the possibility of continuing analysis and evaluation of the other sentences in the tree bank, and most importantly the need to update the syntax content with all the types of missing information. It was discussed to consider further work with the remaining datasets/sentences in the UD Albanian treebank in order to enhance it and furthermore to consider publication of results. The possibility for the creation of a new dataset from scratch was also discussed.

List of links to the UD and resources/references

1. Universal dependencies (https://universaldependencies.org/

2. UD for Albanian (https://universaldependencies.org/sq/index.html

3. UD Albanian TSA (https://github.com/UniversalDependencies/UD_Albanian-TSA/blob/master/sq_tsa-ud-test.conllu

Toska, M., Nivre, J., & Zeman, D. (2020, December). Universal dependencies for Albanian. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)* (pp. 178-188). https://aclanthology.org/2020.udw-1.20.pdf

https://www.besim-kabashi.net/publications.html https://aclanthology.org/people/b/besim-kabashi/

Kabashi, B., & Proisl, T. (2016, May). A proposal for a part-of-speech tagset for the Albanian language. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 4305-4310). https://aclanthology.org/L16-1682.pdf

Kabashi, B., Herbst, T., & Götz-Votteler, K. (2007). Pronominal clitics and valency in Albanian: A computational linguistics perspective and modelling within the LAG-framework. *Valency. Theoretical, Descriptive and Cognitive Issues*, 339-352.

Kabashi, B. (2016). Building an Albanian text corpus for linguistic research. *Kumtesë në konferencën "Corpus-Based Approaches to the Balkan Languages and Dialects*, 5-7.

Reppen, R. (2010). Building a corpus. *The Routledge handbook of corpus linguistics*, 31-37. https://www.torosceviri.info/wp-content/uploads/2022/03/ch3.pdf

Misini, A., Canhasi, E., & Krrabaj, S. (2020, September). Albanian syntactic parsing. https://repository.ukim.mk/bitstream/20.500.12188/9490/1/albanian-syntactic-parsing.pdf

Ramadani, N. (2020). CONTRASTIVE ANALYSIS OF RELATIVE CLAUSES IN ALBANIAN AND ENGLISH–STRUCTURE AND USAGE. *ANGLISTICUM. Journal of the Association-Institute for English Language and American Studies*, *9*(11), 31-35. https://www.anglisticum.org.mk/index.php/IJLLIS/article/view/2138/2559

Dhima, T. (2005). *Gjuha shqipe: sintaksa: tekst për studentët e Fakultateve të Gjuhëve të Huaja dhe të Shkencave të Edukimit*. Shblu. Tirane.

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

The STSM achieved its planned goals and expected outcomes by making a specific contribution to the Action's objectives and deliverables in the framework of both research coordination and capacity building. The researcher became familiar with corpora building and syntactic annotation, as well as analysis and evaluation of existing syntactic annotations. The Albanian tree bank in Universal Dependencies was used as the main focus of the visit.

The main achievements of the mission can be summed up as:

Fostering of knowledge and experience exchange and interdisciplinary network

Exploration of and familiarity with corpus building and syntactic annotations.

Familiarity with linguistic aspects and processes involved in corpus building and syntactic annotation of the Albanian tree bank in Universal Dependencies

Analysis and evaluation of some syntactic annotations in the Albanian tree bank

Discussion of possible future cooperation between the host and guest researchers in further analysis and improvement of the Albanian tree bank in the UD.

As follow-up activities:

Use of the analysed sentences from Albanian tree bank and their comparison to other languages (English mostly) in the UD in the instruction of syntax by the guest researcher.

Further analysis and improvement of the Albanian tree bank in the UD corpus.