TNA ACTIVITY REPORT

MACEDONIAN POETRY CORPUS

Author: Nikolche Mickoski

Current position: Research Associate

Affiliation: Macedonian Academy of Sciences and Arts

Host institution: LINHD-UNED

Mentor(s): Salvador Ros

Period of stay: May 10th – June 16th, 2022

Project description

Miladinov Brothers were Macedonian folklorists and representatives of Macedonian revival of the 19th century. The Collection of Miladinov Brothers was the first published collection of Macedonian songs. The Collection was published on June 24, 1961 in Zagreb, Croatia. It contains over 600 folk songs, traditions, games, riddles, proverbs and names. The Collection of Miladinov Brothers is regarded as the first big collection of Macedonian folk poetry. Songs of the Collection are available in new (as well as old) Cyrillic orthography on Wikisource (https://mk.wikisource.org/wiki/Зборник_на_Миладиновци) while the original book is available at the Macedonian National Library website (http://www.dlib.mk/bitstream/handle/68275/257/82213130.pdf).

The aim of the project was to create the first Macedonian poetry corpus from the folk songs collected by Miladinov brothers in the Collection and to try to annotate the corpus with Part-Of-Speech tags by using the available resources for Macedonian. In addition, as an added value of the project, the songs of the corpus should be parsed with the LINHD tool Averell and added to the POSTDATA (Poetry Standardization and Linked Open Data) catalogue.

Description of the corpus

Macedonian Poetry Corpus is available at the following link. It is consisted of 678 folk songs with the total of 144502 words. For this project, the songs from the Collection of Miladinov Brothers were prepared with unified TEI headers containing the following data:

CLS INFRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004984

<fileDesc> – description of the file
<title> – title of the song

<sourceDesc> – description of the source
<bibl> – bibliographical information:
<author> – collectors of the song, Miladinov brothers (Macedonian: Браќа Миладиновци)
<title> – title of the collection, Collection of Miladinov brothers (Macedonian: Зборник на Миладиновци)
<pubPlace> – place of publishing of the Collection
<date> – date of publishing of the Collection

<notesStmt> – additional information about the song
<note n="subcollection"> – title of the sub-collection to which the song belongs

<settingDesc> – the place of origin of the song

The TEI header for the first song of the Collection called *Јован Попов и самовила* looks like this:

```
<teiHeader>
  <fileDesc>
   <titleStmt>
    <title>Јован Попов и самовила</title>
   </titleStmt>
   <sourceDesc>
    <bibl>
     <author>Браќа Миладиновци</author>
     <title>Зборник на Миладиновци</title>
     <pubPlace>Загреб</pubPlace>
     <date>1961</date>
     <num>1</num>
    </bibl>
   </sourceDesc>
  </fileDesc>
  <notesStmt>
   <note n="subcollection">Самовилски</note>
  </notesStmt>
  <settingDesc>
   <setting />
  </settingDesc>
  <name>Струга</name>
 </teiHeader>
```

Annotation, tagging, and platform

Part-of-Speech tagging for Macedonian in general is quite a challenge because there is no official PoS tagger. Macedonian is not available in UDPipe. The only dataset available for Macedonian is the spaCy model. The attempt to annotate the corpus with the available tagset was unsuccessful because the tagset was not suitable to the corpus. The tagset was trained with newspaper articles, and the words used in the songs are rather archaic and not present in the tagset. So, apart from the conjunctions, some adverbs and punctuations, the PoS information for other words were missing. In addition, syntax information was irrelevant because the syntax of the folk songs is quite different than the regular newspaper syntax.

For example, the only correct PoS tags for the first line (Кинисал ми Јо'ан Попов,) of the first song (Јован Попов и самовила) are the tags for “ми” and the comma as punctuation mark and the PoS tagging looks like this:

```
<l id="s-0" >
 <tok id="w-0" form="Кинисал" lemma="Кинисал" upos="PROPN" feats="" reg="кинисал" ner="B" deprel="ROOT" head="Кинисал" />
 <tok id="w-1" form="ми" lemma="ми" upos="PRON" feats="" reg="ми" ner="O" deprel="dep" head="Кинисал" />
 <tok id="w-2" form="Јо'ан" lemma="јо'а" upos="ADJ" feats="" reg="јо'ан" ner="B" deprel="dep" head="Кинисал" />
 <tok id="w-3" form="Попов" lemma="Попов" upos="ADJ" feats="" reg="попов" ner="I" deprel="punct" head="Кинисал" />
 <tok id="w-4" form="," lemma="," upos="PUNCT" feats="" reg="," ner="O" deprel="punct" head="Кинисал" />
```

That is the reason why we opted not to include PoS information in the corpus at this point and the annotation and the tagging of the corpus will be done in the future. Regularization was not performed because some of the words are archaic and the corpus and the songs will lose authenticity.

Since Macedonian poetry corpus will be live corpus that will be additionally annotated and tagged, TEITOK was selected as a platform for corpus management, since it offers improvement and update of the corpus. As and added value of this project, the TEITOK platform interface was localized into Macedonian.

Research visit

The project was implemented during the research visit to LINHD (International research center in Digital Humanities) of the UNED and was mentored by prof. Salvador Ros Muñoz. LINHD team provided all the necessary help and support for the implementation of the project. During the research visit, the team was visited by Maarten Janssen, the creator of the TEITOK platform, who helped with the implementation of the corpus on the platform. In addition, I participated in the inaugural CLS INFRA Training School at Charles University, Prague and got familiar with some corpus management tools and NLP tools.

Considerations over future work

Lack of official PoS tagger proved to be a stumbling block for preparing an annotated poetry corpus. However, TEITOK as a platform provides the necessary infrastructure for adding the PoS tags in the future. Currently, the corpus is hosted at the LINDAT, an infrastructure project, but it will be transferred to the server of Macedonian Academy of Sciences and Arts.
Finally, the songs of the corpus will be parsed with the LINHD tool Averell and added to the POSTDATA (Poetry Standardization and Linked Open Data) catalogue.