Expanding the Macedonian Language Tree Bank

Applicant Introduction

This project is submitted by Emilija Mitovska from The Republic of North Macedonia.

A 22-year-old student at the Faculty of Electrical Engineering and Information Technologies, in the last year of studies for a major in the area of Computer Systems Engineering, Automation and Robotics.

Languages and linguistics have been a strong passion of mine throughout the years and natural language processing is the area which perfectly combines my love for engineering and technologies and language.

Key Words

Macedonian language, modernization, expansion, UD Tree Bank

Resource – expansion of the Macedonian Language Tree Bank

The Macedonian language is a relatively young language, codified less than a 100 years ago, with a complicated history and scarce historical resources. It is considered a relatively small language which numbers around 2 million speakers[1], most of which live in the motherland.

The digitalization of said language is pacing slowly due to the lack of effort from its speakers and institutional action which leads us to the problem at hand: the Macedonian Language Tree Bank[2]. The tree bank has been, so far, built using the Macedonian Language Digital Resources – MLDR, which consists of less than a 100 Macedonian novels and poetry books, dating from the 1950s with the newest one being written in 2002[3].

These books are some of the fundamental literary pieces which form the pillars of the language itself and the MLDR base is the only one of its kind in the country and the contribution of the authors and the academics from the Macedonian Academy of Sciences and Arts, who are responsible for creating that digital resource, indisputable. However, when speaking about

[1]

[2]

[3]

creating a Tree Bank of dependencies which reflect the entire nature of a spoken language in all its forms and uses, this corpus is far from sufficient.

I strongly believe that this Tree Bank should be expanded with entries which more closely relay the spirit of the spoken, not just literary language, specifically the language spoken in this day and age, which highly differs from the one depicted in the corpus. The Macedonian Language Tree Bank lacks modern and colloquial entries, and at the same time contains a lot of archaic entries which are not as used today with the frequency presented in the corpus.

As the language lacks and official digital database, I propose the use of available resources such as news media, elaborate social media posts from chosen influential users which abide by a certain language standard, private collections of organizations which publish short stories and poetry, children's media and other resources of similar form, written and consumed by a younger demographic (0 - 64). By scraping these available forms of the written language, the Tree Bank will not only be expanded in terms of number of entries but will also be more up to date with the language forms that characterize contemporary Macedonian.

Participation in the training school

Personally, I intend to pursue this project and set the grounds for further work in this area, as I am passionate about the subject and it is a necessary project for further personal work, and the work of many colleagues and associates. Participating in the training school will be a great, and likely only opportunity, for me to learn from professionals in the area and go forward with this project. I will gain the necessary skills to expand the database on my own, but also to teach and engage other enthusiasts and professionals to continue the work and create a community which will hopefully sustain the Tree Bank in the future.

Questions

I need help creating a list of potential data sources needed for expanding the Tree Bank, so I would love to get the help of the participants and maybe hear about their personal experiences and advice, the course of action for using these resources, getting permission, how to implement them in the project, and hopefully many more related (and unrelated) questions that will come out of the discussion.

Project State

The project is in the planning phase, with lots of space to be improved and upgraded. The key to moving forward with it is gathering knowledge, skill, and manpower.

References

1 - Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2024. Ethnologue: Languages of the World. Twenty-seventh edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com

2 – Universal Dependencies Macedonian Tree Bank. Contributor: Vladimir Cvetkoski

3 - Macedonian Language Digital Resources - MLDR, a.k.a 135 Volumes of Macedonian Literature, published by the Macedonian Academy of Sciences and Arts