Review of the Belgrade European Language Grid Workshop

Bojana Bašaragin
bojana.basaragin@ivi.ac.rs
Research and Development Institute for Artificial Intelligence of Serbia
Belgrade, Serbia

PAPER SUBMITTED: 19 April 2022
PAPER ACCEPTED: 16 May 2022

March 11, 2022 was marked by the first in a series of dissemination events within the EU project European Language Grid (ELG)[1]. This was an online ELG Workshop organized by the Faculty of Philology in Belgrade and the Society for Language Resources and Technology JeRTeH. This event announced the recent and successful completion of the project in mid-2022.

The ELG project (2019-2022) was created with the idea of joining the academic community and industry in the field of language technologies and strengthening Europe's position in this regard compared to other continents. The result of the project is a unique ELG platform that collects resources and tools for processing European languages, aiming to make them more visible and accessible to academia, business partners, NGOs, and the public sector. One of the project goals is to help address the issue of digital language vulnerability by providing help to those European languages that are not sufficiently supported through language technologies. According to data from January 2022, the platform gathers over 10,000 services, tools, data sets, resources, and language models for 87 languages, and that number is continually increasing. Following the completion of the project, the plan is for the platform to continue its life, providing access to both non-commercial and commercial language technologies.

The workshop lasted three hours and was held online, via the Zoom platform, while the recording of the workshop can be found on the JeRTeH Society YouTube channel[2]. The official language of the workshop was English. After the welcoming speech of the organizer, prof. Cvetana Krstev, the project coordinator prof. Georg Rehm gave the introductory presentation. Prof. Rehm made a detailed review of the ELG project and its results, as well as perspectives for further growth and development of the platform. He

[1] European Language Grid project (ELG).

[2] JeRTeH YouTube channel.

also presented the problem of language inequality in terms of language technologies for most European languages, except Spanish, German, and French (Figure 1). He addressed the situation of the Serbian language, which, according to the European Language Equality (ELE) project report for Serbian[3], still belongs to the languages with weak language technology support. When asked what he sees as the next step in the development of resources and tools for the Serbian language, he mentioned the development of a comprehensive language model, which could then be applied to various tasks of language processing and generation.

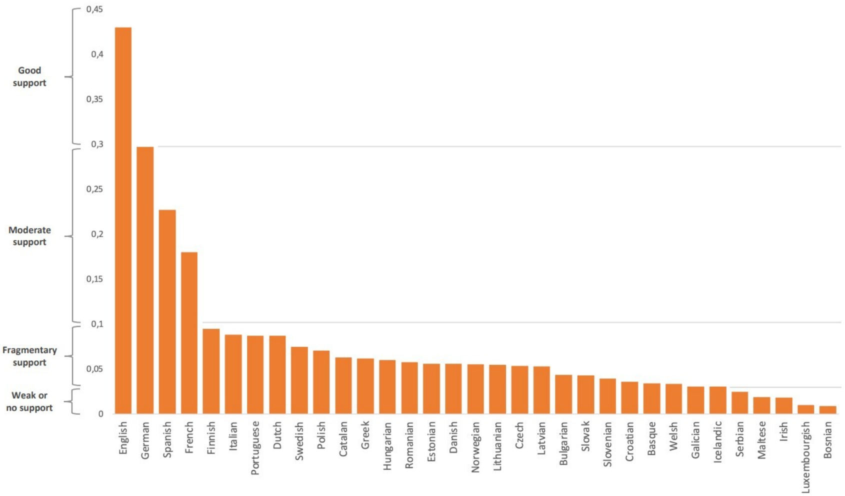

Figure 1. Language technology support for different European languages

The introductory part was followed by a practical and detailed demonstration of the platform use given by Penny Labropoulou, Senior Research Associate at the Institute for Language and Speech Processing of Greece. The presence of foreign lecturers allowed using the results of ELITR[4], a recently completed European project, for the needs of the workshop. ELITR participants developed a system for automatic translation and subtitling of

[3] ELE report for Serbian.

[4] European Live Translator project (ELITR).

meetings and conference presentations, as well as note-taking. Automatic captions of all workshop presentations and their translation into Serbian (or some other European) language could be followed in a separate window in real time.

After the introductory presentations came the examples from the domain of language technologies for the Serbian language. In the spirit of the project, the workshop was designed to bring together examples of successful applications of language technologies in the industry and the latest achievements in the scientific field. Thus, the first two presentations came from two large Serbian companies – the telecommunication service provider Yettel and the IT group ComTrade. Đorđe Hirš, Data Science Team Lead at Yettel Serbia, presented the method of unsupervised sentiment analysis that Yettel applies to the database of customer service voice complaints to improve their services. Dušan Jovičić, a programmer, presented the regional experience of the company ComTrade in developing conversational agents (chatbots).

Examples from the industry were followed by the presentation of resources and tools for the Serbian language developed in cooperation between the University of Belgrade (BU) and the JeRTeH Society. First, prof. Ranka Stanković spoke about the corpus of the modern Serbian language SrpKor, which at the end of 2021 got a new version – SrpKor2021. Compared to the previous version, this corpus is enriched with new texts from various domains, featuring ELTeC, a corpus of literary texts from the 19th and early 20th century, created as part of the COST action Distant Reading[5]. This corpus, composed of over 100 POS tagged literary works was the topic of the final workshop presentation. The ELTeC corpus was presented by prof. Cvetana Krstev and Milica Ikonić Nešić as a completely free resource that can be used for analyses and training models for the Serbian language. This corpus can currently be downloaded from the ELTeC GitHub. SrpKor2021, as well as the ELTeC corpus, can be searched through the NoSketch system, available to registered users through the JeRTeh website[6].

Biljana Rujević gave a practical demonstration of the *Leximirka* platform, developed for the unified supplementation, maintenance, and use of electronic dictionaries, corpora, and other lexical resources and tools created in cooperation with BU and JeRTech. This platform can also be accessed from the JeRTech website after registration. Mihailo Škorić presented BEaST, a newly developed POS tagger for the Serbian language that shows improved performance compared to the taggers it is based on: TreeTagger,

[5] Distant Reading for European Literary History COST action.

[6] JeRTeh website.

spaCy, and Stanza. This tagger, used to annotate the ELTeC text corpus, is available for download via GitHub. Branislava Šandrih Todorović spoke about the web portal *NER & Beyond*, developed within the COST action Distant Reading. The portal was designed to allow named entity recognition and tagging using various available tools but also mapping between the annotations produced by these tools. This platform is available through the JeRTeh website.

The ELG workshop was a real refreshment. On the one hand, it provided an opportunity to learn about the ELG project, as an important step towards unification and greater visibility of tools and resources for the Serbian language. On the other hand, the participants gained insight into some of the latest research results in the field of language technologies for the Serbian language and their applications. The dynamic structure within which the presentations of services, resources, and tools lasted 15 minutes each made it possible to get an overview of a large number of projects in a short time, with enough information about each of them. I look forward to the next opportunity to attend such a comprehensive and well-organized event, with a belief that every such effort contributes to improving the situation of language technologies for the Serbian language.