Curated Multilingual Language Resources for CEF.AT

CURLICAT
Curated Multilingual Language Resources for CEF.AT

META-FORUM 2020
virtual
2020-12-01/03

What?

The overall objective of the Curated Multilingual Language Resources for CEF AT Action (CURLICAT) is to compile curated datasets in seven languages targeted by the consortium (**Bulgarian, Croatian, Hungarian, Polish, Romanian, Slovak and Slovenian**) in domains of relevance to the European Digital Service Infrastructures (DSIs) with a view to enhance the Automated Translation.

The prime source of data are the national corpora of the above-mentioned languages. The data will cover domains relevant for some of the CEF DSIs, such as eHealth, Europeana and eGovernment in general.

How?

The Action will deliver at least 14 Million sentences (estimated to contain at least 140 Million words) from domains including culture, education, health and science. Moreover, the Action will address the gap in machine translation technology, which crucially depends on the provision of domain specific quality language resources for the under-resourced languages.

Why?

By delivering seven large size monolingual datasets, which themselves will facilitate the improvement of the CEF Automated Translation core service platform, the Action will enable international users to access information about the relevant EU Member States, including information about local companies and investment opportunities. Thus, the Action will also support the economic growth in Europe, by supporting the CEF AT core service platform for exchanging information in multiple languages.

Planned activities

1. Aggregation and data preparation: Collecting the relevant parts from the national monolingual corpora for the seven languages covered by the consortium. The final corpus will contain at least 20 million words per language.

2. Additional collection and IPR clearance: clarification of the legal status of the data and if necessary obtaining the Intellectual property rights (IPR) to distribute the texts in the original unedited format, when possible. Also the identification of the unbalanced domain distribution across the targeted languages and collection of additional text data.

3. Anonymisation: removing or anonymizing personal and sensitive data from the language resources collected in previous activities.

4. Terminology enrichment: annotation of the documents in the seven monolingual corpora with IATE terms and the recognition of words and multiword expressions which fulfill the criteria for domain-specific terms.

5. Metadata harmonisation: homogenizing all the individual metadata schemes used by the seven large-scale monolingual corpora by finding a set of common attributes describing the relevant text properties e.g. text style and text domain. Based on this, specific translation models will be trained and the domains will be adapted for specific translation tasks.

Project consortium

- **Bulgaria:** Bulgarian Academy of Sciences, Institute for Bulgarian language Lyubomir Andreychev
- **Croatia:** University of Zagreb, Faculty of Humanities and Social Sciences
- **Hungary:** Hungarian Research Institute for Linguistics (coordinator)
- **Poland:** Polish Academy of Sciences, Institute of Computer Science
- **Romania:** Romanian Academy, Institute for Research in Artificial Intelligence
- **Slovakia:** Slovak Academy of Sciences, Linguistic Institute Ľudovit Štúr
- **Slovenia:** Jozef Stefan Institute

Funding

EU Connecting Europe Facility (CEF)
Telecommunications Programme

Total eligible costs: 959,999.72 €

Estimated CEF contribution: 719,999.79 €

Duration

2020-06-01 – 2022-05-31 (24 months)

This action received funding from the Connecting Europe Facility Telecommunications Programme.

www.curlicat-project.eu