Collecting and Processing Legal Corpora for Seven Under-resourced Languages

MARCELL
Multilingual Resources for CEF.AT in the Legal Domain

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

What?

MARCELL's purpose is to enable enhancement of the European Commission service in Automatic Translation (eTranslation) on the body of national legislation (laws, decrees, regulations, etc.) in seven countries and in seven EU official, yet under-resourced, languages with the following partners:

- **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

Why?

National legislation texts are not automatically available to eTranslation and existing MT systems could be improved if they had access to national legislative texts for translation and language model refinement. This will improve the quality of translation in the legal domain.

In addition to the expected overall improvement of the MT system in the seven languages concerned, the Action will have an impact both on the eJustice and the Online Dispute Resolution DSIs as the resources focus on national legislation, which is of direct relevance to both DSI's.

How?

Three language resources available in all seven languages will be used:

- corpora of national legislation in the respective languages
- multilingual ontology-based thesaurus EUROVOC
- IATE term collection

Results achieved

1. Seven large-scale monolingual corpora of national legislation documents:

- compiled and processed with linguistic processing chains (LPCs) including tokenization, PoS/MSD-tagging, NERC, dependency parsing (where available);
- classified with top-level EUROVOC descriptors;
- EUROVOC and IATE terms annotated in texts.

2. A comparable corpus of seven languages aligned at 21 top-level domains identified by EUROVOC descriptors thus representing valuable in-domain training data.

3. Seven Dockerised sustainable document processing pipelines that provide the periodic flow of new legal documents as they appear in seven languages to CEF.AT for further training.

4. Collected Croatian-English parallel corpus consisting of 1800 national legislative documents which are translated and aligned at the sentence level and manually checked. This result is much needed since Croatian is deficient in resources for training MT systems because it became the EU official language only in 2013.

Funding

EU Connecting Europe Facility (CEF) Telecommunications Programme

Total eligible costs: 1,883,714.67 €

Estimated CEF contribution: 1,412,786.00 €

Duration

2018-10-01 – 2021-03-31 (30 months)

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

www.marcel-project.eu