An accessible method for collecting corpora in under-resourced languages

MARCELL
Multilingual Resources for CEF.AT in the Legal Domain

Marko Tadić, University of Zagreb, Faculty of Humanities and Social Sciences, marko.tadic@ffzg.hr
Tamás Váradi, MTA Research Institute for Linguistics, varadi.tamas@nytud.hu

LT4ALL
UNESCO, Paris
2019-12-05

Context

Language technology crucially depends on large amounts of texts. Digitally published text is a natural source for fast production of the fundamental language resources – corpora. However, clean, openly and freely available texts are difficult to come by. Even national official languages suffer from scarcity of quality language data. We are presenting a project that can serve as a role model for the collection of large monolingual corpora for under-resourced languages.

The approach could be applicable to any linguistic community that publishes legislative texts in their own language in digital form, to quickly build very big corpora. Although the texts are all of the same (legal) domain, they could also be classified following a predefined scheme (we used top-level domains of EUROVOC thesaurus), resulting in a set of subcorpora representing various domains.

The tools used in this project all belong to the Basic Language Resources Kit, BLARK (Krauwer, 2003) and should be developed for each language community that tries to build language technologies for its language anyway.

What?

MARCELL's first purpose was 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:** MTA Research Institute for Linguistics
- **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.

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

Expected results

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

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

2. A comparable corpus of seven languages aligned at the top-level domains identified by EUROVOC descriptors.

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

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 – 2020-09-30 (24 months)

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

www.marcel-project.eu