MARCELL

Multilingual Resources for CEF.AT in the legal domain

Multilingual Resources for CEF.AT in the Legal Domain

Tamás Váradi

Project Coordinator

Research Institute for Linguistics, Hungarian Academy of Sciences

varadi.tamas@nytud.mta.hu

Showcasing Language Technology
in H2020 and CEF Telecom projects

Brussels, 26/04/2019

Factsheet

- Objective:
  - to enhance the E-Translation system by preparing as training data the body of national legislation in the seven countries.
- Duration: 24 months (1.10.2018.–30.09.2020.)
- Budget: 1,883,958 EUR
- Consortium: Bulgaria, Croatia, Hungary, Poland, Romania, Slovakia and Slovenia.

The Challenges

- MT is data driven and data hungry technology
  - MARCELL languages are under-resourced
  - Parallel data are few and far between
- Demand for domain-specific data
  - Lexical items are inherently ambiguous out of context
  - Domains-specific data help improve precision
- Language models need to be trained on domain-specific data
- eTranslation requires new data not seen by MT@EC

The data

- National Legislative documents in seven languages
- Justification:
  - **Unseen by the EC**
    - After Legal harmonisation after accession, national legislation is not automatically sent to the Publication Office
  - **Valued by DGT Translation Services**
- multilingual but not parallel and still useful? Yes, because of
  - Back-translation
  - language modelling

Innovative aspects: topic classification

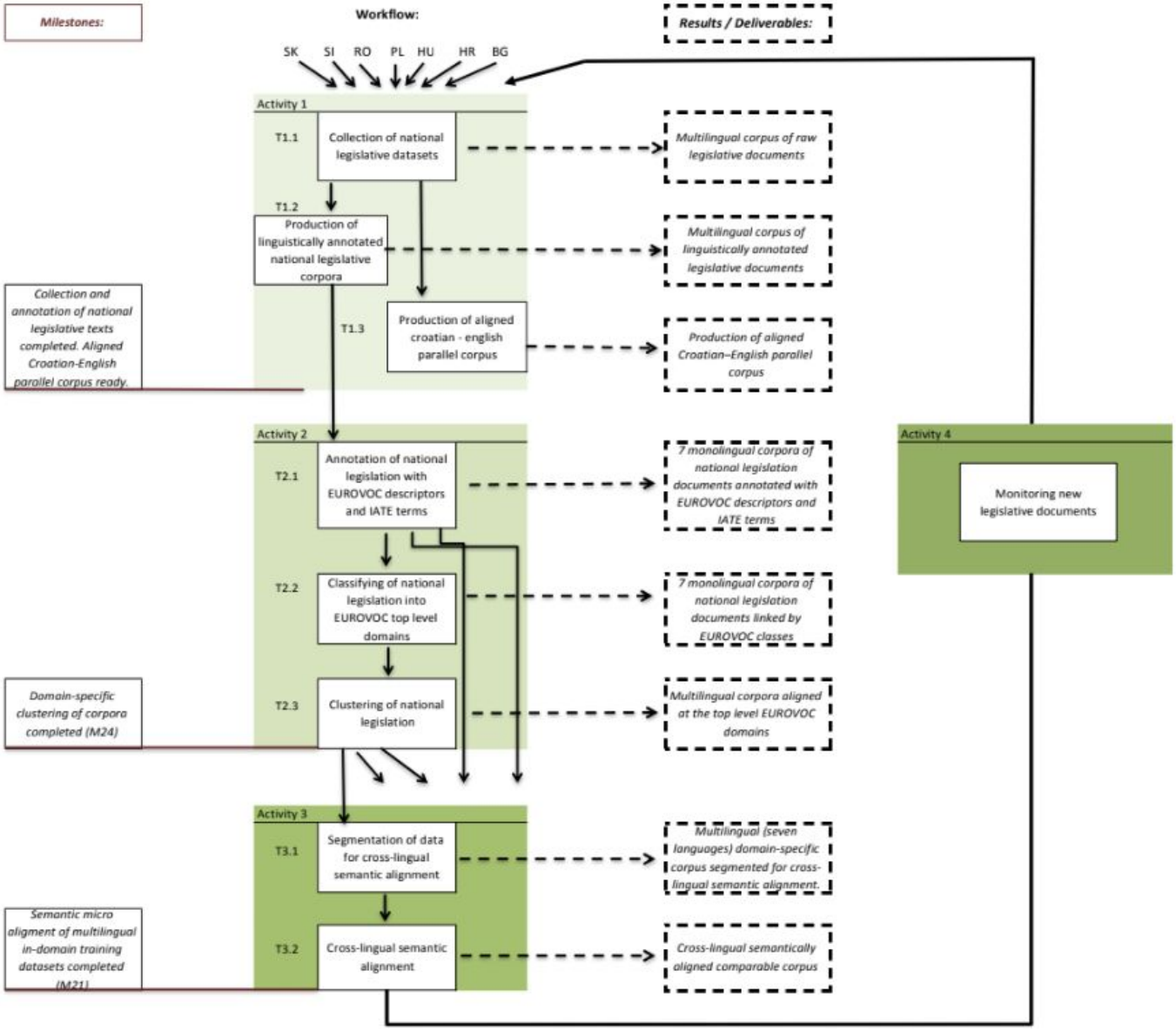
Topic classification

- law as a domain?
  - sure, but at the same time, far too diverse
    (laws on labour relations, family protection etc.)
- define **subdomains** matching EUROVOC top level categories

Terminology enrichment

- focus on domain specific terminology
- terms from IATE terminology database identified

Workflow


Milestones:
Workflow:
Results / Deliverables:
SK SI RO PL HU HR BG
Activity 1
T1.1
Collection of national legislative datasets
Multilingual corpus of raw legislative documents
T1.2
Production of linguistically annotated national legislative corpora
Multilingual corpus of linguistically annotated legislative documents
T1.3
Production of aligned croatian - english parallel corpus
Production of aligned Croatian–English parallel corpus
Collection and annotation of national legislative texts completed. Aligned Croatian-English parallel corpus ready.
Activity 2
T2.1
Annotation of national legislation with EUROVOC descriptors and IATE terms
7 monolingual corpora of national legislation documents annotated with EUROVOC descriptors and IATE terms
T2.2
Classifying of national legislation into EUROVOC top level domains
7 monolingual corpora of national legislation documents linked by EUROVOC classes
T2.3
Clustering of national legislation
Multilingual corpora aligned at the top level EUROVOC domains
Domain-specific clustering of corpora completed (M24)
Activity 3
T3.1
Segmentation of data for cross-lingual semantic alignment
Multilingual (seven languages) domain-specific corpus segmented for cross-lingual semantic alignment.
T3.2
Cross-lingual semantic alignment
Cross-lingual semantically aligned comparable corpus
Semantic micro aligment of multilingual in-domain training datasets completed (M21)
Activity 4
Monitoring new legislative documents

Activities

1. Collection of raw national legislative texts & preprocessing
2. Domain-specific classification of national legislative corpora
3. Semantic alignment of the multilingual corpora
4. Sustainability
5. Dissemination
6. Management

In progress so far

- 1. Collection of raw national legislative texts and preprocessing
  - Multilingual corpus of raw legislative documents

Partner	Size	Format	Scope
ILS	22 328	XML	22328 National, Official Section 45 M tokens
IBL	101 496	XML	50655 Unofficial Section State agencies and Municipalities 15745 Unofficial Section Calls and Messages 12963 Unofficial Section Courts 196 Official Section Constitutional Court 6594 Official Section Ministerial Council 7949 Official Section Ministries and Other Departments 4488 Official Section National Assembly 2906 Official Section President of the Republic
JSI	21 557	JSON/HTML	National legislation

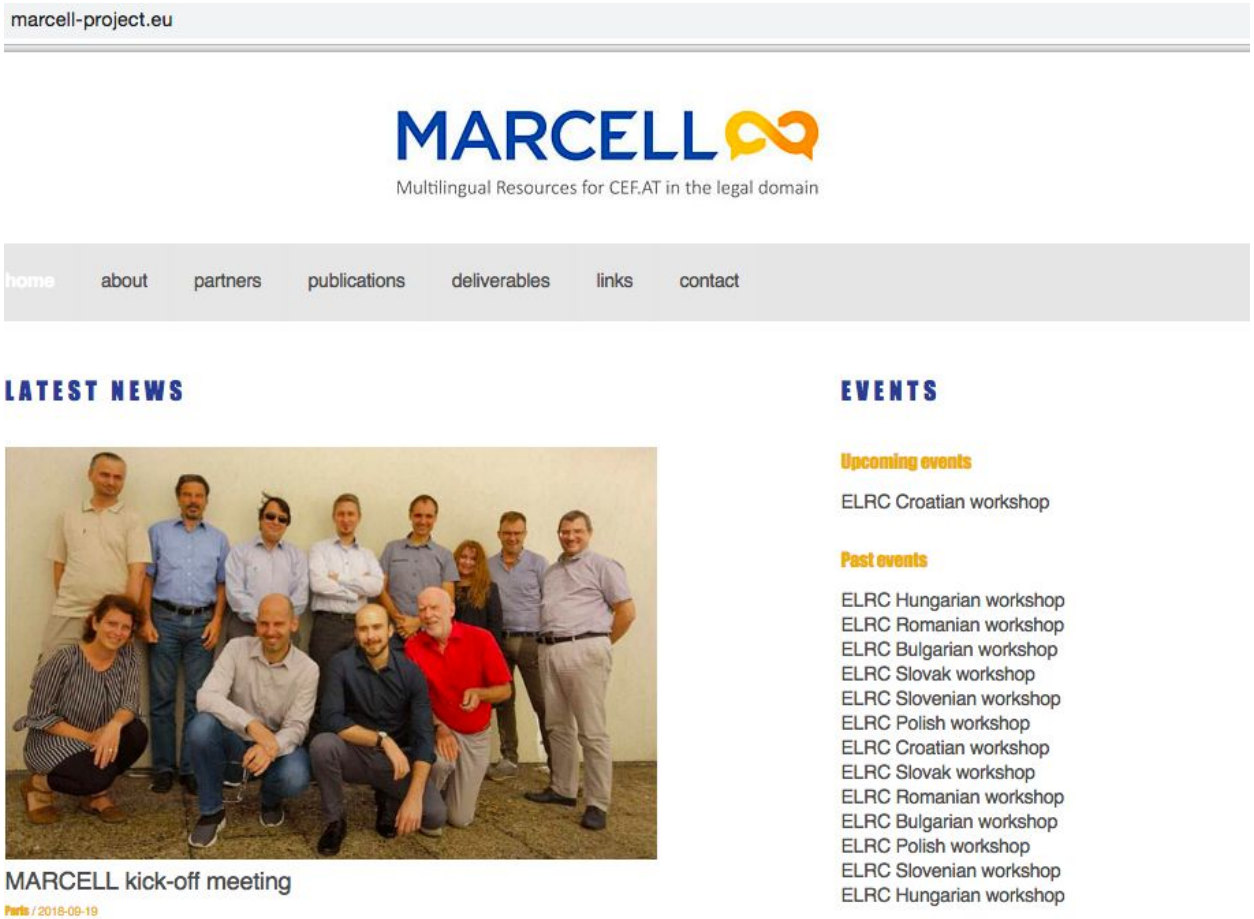
Partner	Size	Format	Scope
IPIPAN	23119	source format: PDF; extracted text stored in the database	23119 National
RILMTA	21 598	Source format: HTML	National Law 5204 Local (Budapest) 3851 Local (Debrecen) 1857 Local (Budapest VII.) 1706 Local (Szolnok) 1261 Local (Miskolc) 1235 Local (Budapest XVI) 1214 Local (Budapest XII) 759 Local (Sopron) 721 Local (Budapest IX) 719 Local (Göd) 468 Local (Budapest X) 392 Local (Nagykovácsi) 388 Local (Budapest IV) 381 Local (Paks) 359 Local (Budapest XV) 335 Local (Szombathely) 279 Local (Héviz) 268 Local (Mosonmagyaróvár) 145 Local (Salgotarján) 57
RACAI	148 289	txt	148289 National

In progress so far

- 4. Dissemination
  - Dissemination plan
    - Logos
    - PPT template
    - Web site
    - Plan of T-shirts
    - List of events and scientific journals
    - Social networks: Twitter account

In progress so far

- 6. Management
  - Project website
    - http://marcell-project.eu/



Next steps (6 months)

- Multilingual corpus of linguistically annotated legislative documents.
- Seven large-scale monolingual corpora of national legislation documents annotated with EUROVOC descriptors and IATE terms.
- Sustainability plan

MARCELL

Thank you
for your attention!

varadi.tamas@nytud.mta.hu

This action received funding from the CEF Telecommunications Programme.