Easily Accessible Language Technologies for Slovene, Croatian and Serbian

Nikola Ljubešić,[1,2] Tomaž Erjavec,[1] Darja Fišer,[1,3] Tanja Samardžić[4], Maja Miličević[5], Filip Klubička[2], Filip Petkovski[6]

[1] Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana
[2] Faculty of Humanities and Social Sciences, University of Zagreb
[3] Faculty of Arts, University of Ljubljana
[4] CorpusLab, University of Zürich
[5] Faculty of Philology, University of Belgrade
[6] Freelance developer, Macedonia

Overview

Projects

- ReLDI (Regional Linguistic Data Initiative) – SNSF
- JANES (Jezikovna Analiza NEstandardne Slovenščine) – ARRS

Overview

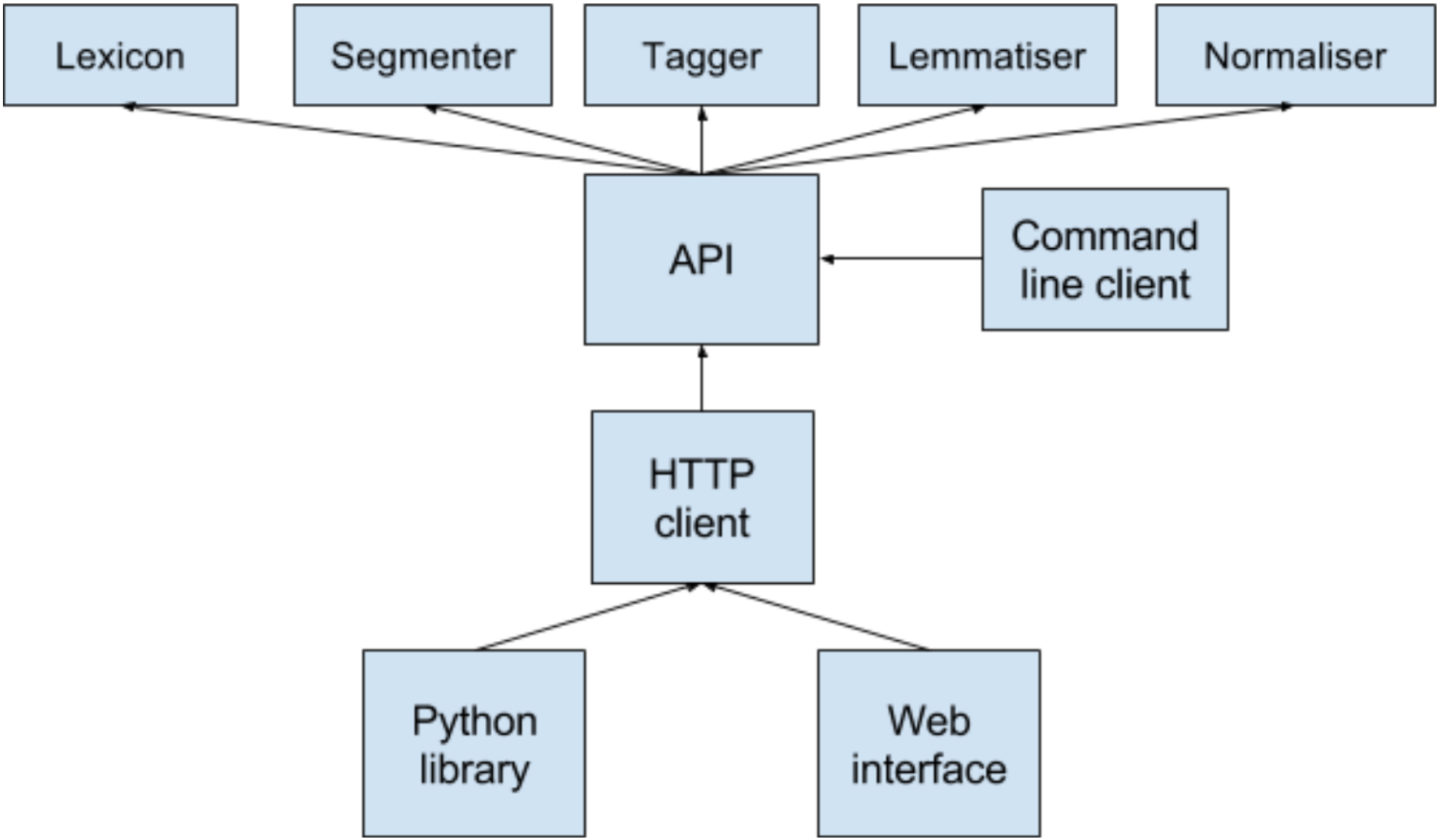
Projects

- ReLDI (Regional Linguistic Data Initiative) – SNSF
- JANES (Jezikovna Analiza NEstandardne Slovenščine) – ARRS

Included technologies

- Morphological lexicons (Sloleks 101k lexemes, hrLex 187k lexemes, srLex 193k lexemes)
- Segmentation
- Diacritic restoration (99% standard, 98% non-standard language)
- Morphosyntactic tagging (92-93% accuracy)
- Lemmatisation (98% accuracy)

Architecture



The near future

Technologies to follow

- Dependency parsing (`http://universaldependencies.org`)
- Named entity recognition (PERS, LOC, ORG, MISC)
- Text normalisation via character-level SMT – non-standard and historical language (error reduction of 30-80%, downstream processing error reduction of (up to) 50%)
- Semantic Role Labeling (Slovene-Croatian bilateral project)

The near future

Technologies to follow

- Dependency parsing (`http://universaldependencies.org`)
- Named entity recognition (PERS, LOC, ORG, MISC)
- Text normalisation via character-level SMT – non-standard and historical language (error reduction of 30-80%, downstream processing error reduction of (up to) 50%)
- Semantic Role Labeling (Slovene-Croatian bilateral project)

Migration to CLARIN

- All web services are developed to be easily included in WebLicht
- Received funding from CLARIN to add existing and emerging technologies, workshop in Ljubljana in November

Easily Accessible Language Technologies for Slovene, Croatian and Serbian

Nikola Ljubešić,[1,2] Tomaž Erjavec,[1] Darja Fišer,[1,3] Tanja Samardžić[4], Maja Miličević[5], Filip Klubička[2], Filip Petkovski[6]

[1] Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana
[2] Faculty of Humanities and Social Sciences, University of Zagreb
[3] Faculty of Arts, University of Ljubljana
[4] CorpusLab, University of Zürich
[5] Faculty of Philology, University of Belgrade
[6] Freelance developer, Macedonia