www.accurat-project.eu
accurat-project@tilde.com

ACCURAT Toolkit
for Multi-Level Alignment and Information Extraction from Comparable Corpora

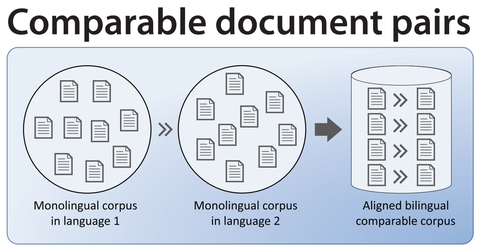
Mārcis Pinnis[1], Radu Ion[2], Dan Ştefănescu[2], Fangzhong Su[3], Inguna Skadiņa[1], Andrejs Vasiļjevs[1], Bogdan Babych[3]
[1]Tilde, Riga, Latvia • [2]Research Institute for Artificial Intelligence, Romanian Academy • [3]Centre for Translation Studies, University of Leeds

The ACCURAT Toolkit for Multi-Level Alignment and Information Extraction from Comparable Corpora is a collection of tools for analysing comparable corpora and extracting parallel data to improve the quality of statistical and rule based MT systems.

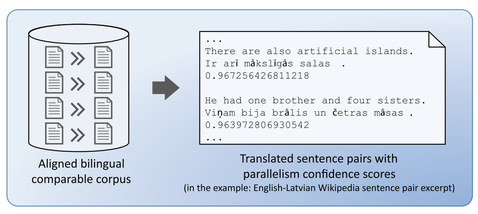
PDMWF – parallel data mining workflow

The workflow aligns comparable corpora in document level and extracts parallel sentences. Output of the workflow:

Comparable document pairs



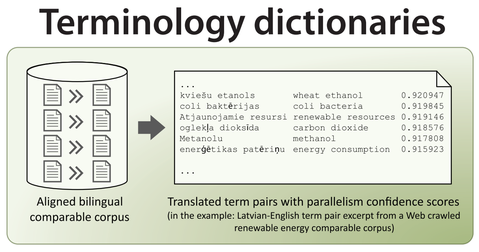
Parallel sentences



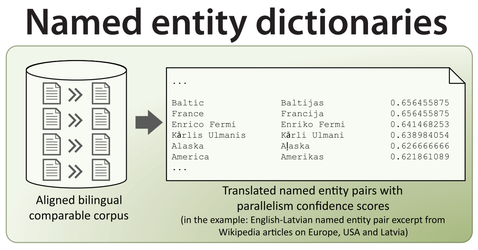
NERTEWF – named entity and term extraction and mapping workflow

The workflow monolingually tags terms and named entities in aligned comparable corpora and extracts translated named entity and term pairs. Output of the workflow:

Terminology dictionaries



Named entity dictionaries



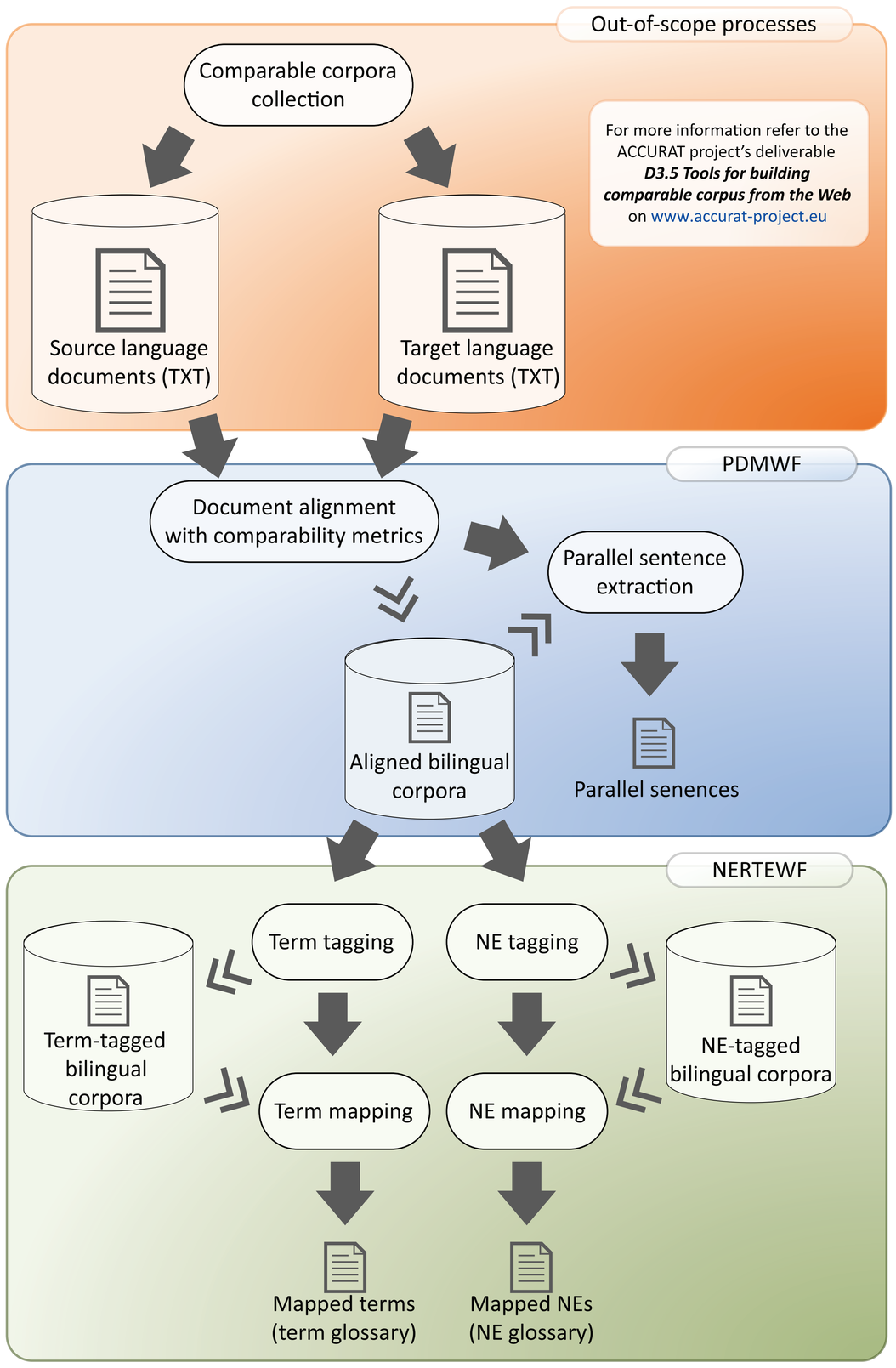
Language coverage of the tools

The tools for comparability estimation, parallel sentence extraction, and term and named entity mapping **are language independent**, however, may require language specific resources. Such resources are: probabilistic dictionaries, stopwords, word endings, and phrase markers.

Out-of-the-box support in *PDMW* is provided for: *English-Croatian, English-Estonian, English-Greek, English-Latvian, English-Lithuanian, English-Romanian, English-Slovenian*, and *Latvian-Lithuanian*.

Out-of-the-box support in *NERTEWF* is provided for: *English-Latvian, English-Lithuanian, English-Romanian*, and *Latvian-Lithuanian*.

Conceptual design of the ACCURAT Toolkit



Download the ACCURAT Toolkit

The ACCURAT Toolkit is **open source and freely available** for download after completing a registration form. The toolkit is released under the Apache 2.0 licence. www.accurat-project.eu

The research within the project ACCURAT leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement no 248347.