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

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:

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:

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.


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.