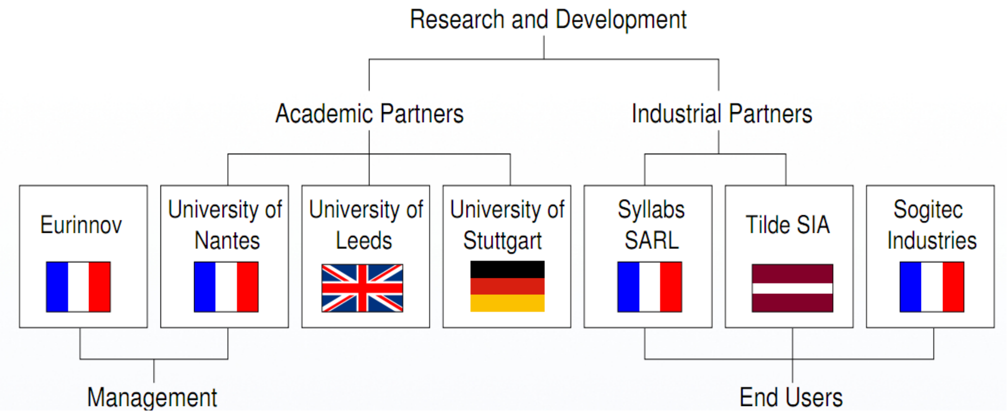
TTC: Terminology Extraction, Translation Tools and Comparable Corpora

Béatrice Daille beatrice.daille@univ-nantes.fr
Université de Nantes – LINA
www.ttc-project.eu

Description

Consortium



The TTC project leveraged machine translation (MT) systems, computer-assisted translation (CAT) tools and multilingual content (corpora and terminology) management tools by developing methods and tools that allow users to generate bilingual terminologies automatically from comparable (non-parallel) corpora in seven languages: five European languages (English, French, German, Spanish, Latvian) as well as Chinese and Russian, and twelve translation directions. The TTC project has developed generic methods and tools for the automatic extraction and alignment of terminologies, in order to break the lexical acquisition bottleneck in both statistical and rule-based MT. It has also developed and adapted tools for gathering and managing comparable corpora, collected from the web, and managing terminologies. In particular, a topical web crawler and the MyEuroTermBank open terminology platform have been developed.

The key output of the project is the **TTC web platform**. It allows to create thematic corpora given some clues (such as terms or documents on a specific domain), to expand a given corpus, to create a comparable corpora from seeds in two languages, to choose the tools to apply for terminology extraction, to extract monolingual terminology from such corpora, to translate bilingual terminologies, and to export monolingual or bilingual terminologies in order to use them easily in automatic and semi-automatic translation tools.

For generating bilingual terminologies automatically from comparable corpora innovative approaches have been researched, implemented and evaluated that constituted the specificities of the TTC approaches: (1) t**opical web crawling** which will gather comparable corpora from domain-specific Web portals or using query-based crawling technologies with several types of conditional analysis; (2) for **monolingual term extraction**, different techniques, a knowledge-rich and a knowledge-poor approaches were followed; a massive use of morphological knowledge to handle morphologically complex lexical items; (3) for **bilingual term extraction**, an unified treatment for single word term and multi-word term was designed as well as an hybrid method that used both the internal structure and the context information of the term.

Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.)
Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 449.