How ACCURATE is your translation?

Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation

The project has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement n° 248347

www.accurat-project.eu

Main goal

To develop methods and techniques to overcome one of the central problems of Machine Translation (MT) – the lack of linguistic resources for under resourced areas of machine translation. The main goal is to find, analyze and evaluate novel methods that exploit comparable corpora in order to compensate for the shortage of linguistic resources, and ultimately to significantly improve MT quality for under-resourced languages and narrow domains.

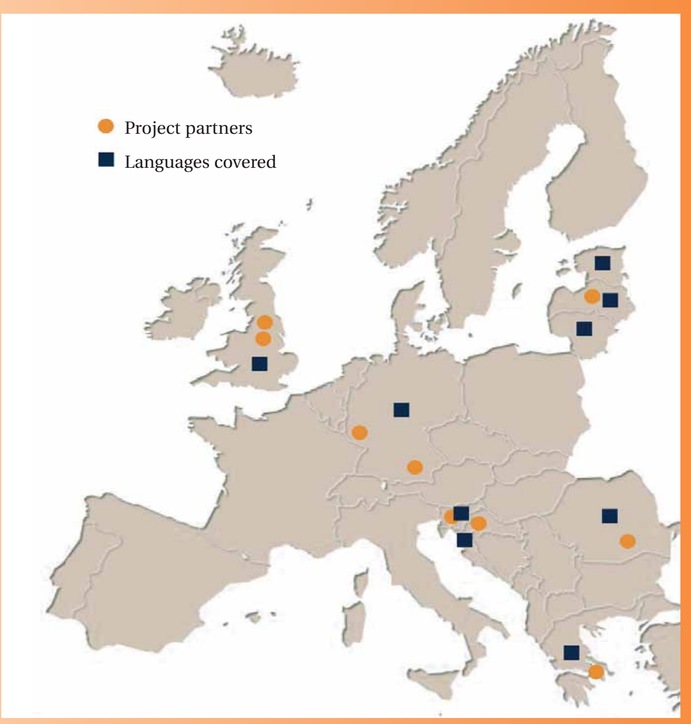
Key innovation

Creation of methodology and tools to measure, to find and to use comparable corpora to improve the quality of MT for under-resourced languages and domains. Thus the ACCURAT project will significantly contribute not only to the theory of MT, but also to corpus linguistics, information extraction and natural language processing in general and will strongly advance theoretical foundations and methodology for research in corpus linguistics.

Using comparable corpora

The applicability of current data-driven methods directly depends on the availability of large quantities of parallel corpus data. For this reason the translation quality of current data-driven MT systems varies dramatically from quite good for language pairs with large corpora available (e.g. English and French) to almost unusable for under-resourced languages and narrow domains where little data is available (e.g. Latvian and Croatian).

- for under-resourced languages covering Latvian, Lithuanian, Estonian, Greek, Croatian, Romanian and Slovenian;
- narrow domains (data processing, automotive engineering etc.).



Important results so far

- **Comparability metrics and associated tools developed** – identified features that can measure comparability of corpora;
- **Research methods for alignment and extraction of lexical, terminological and other data** from comparable corpora based on available techniques for parallel corpora;
- **Developed research methods for automatic acquisition of a comparable corpora** from the Web;
- **Several multilingual comparable and parallel corpora gathered** from the Web;
- **Measure improvements** from applying acquired data against developed baseline results from SMT and RBMT systems.

A Cross-European Project

Connecting under-resourced languages, that are either EU official languages or are in the process of becoming one soon, with well-resourced languages such as English and German.
Languages covered by the project: Latvian, Lithuanian, Estonian, Romanian, Greek, Croatian, Slovenian, English, German.

Project partners

TILDE: Tilde SIA, Riga, Latvia
USFD: University of Sheffield, Computer Science Department, NLP Group, Sheffield, UK
CTS: University of Leeds, Centre for Translation Studies, Leeds, UK
ILSP: Athena Research and Innovation Center in Information Communication & Knowledge Technologies, Institute for Language and Speech Processing, Athens, Greece
FFZG: University of Zagreb, Faculty of Humanities and Social Sciences, Department of Linguistics, Zagreb, Croatia
DFKI: Deutsches Forschungszentrum für künstliche Intelligenz, LT Lab, Saarbrücken, Germany
RACAI: Romanian Academy, Research Institute for Artificial Intelligence, Bucarest, Romania
LT: Linguatec GmbH, Munich, Germany
ZEMANTA: Zemanta d.o.o., Ljubljana, Slovenia

Contact

Tilde, SIA | 75a Vienibas Gatve | LV-1004 Riga | LATVIA
Contact person | Andrejs Vasiljevs | p: + 371 67605001 |
f: +371 67605750 | e: andrejs@tilde.lv | w: http://www.tilde.lv

ACCURAT results will also be available through META-SHARE