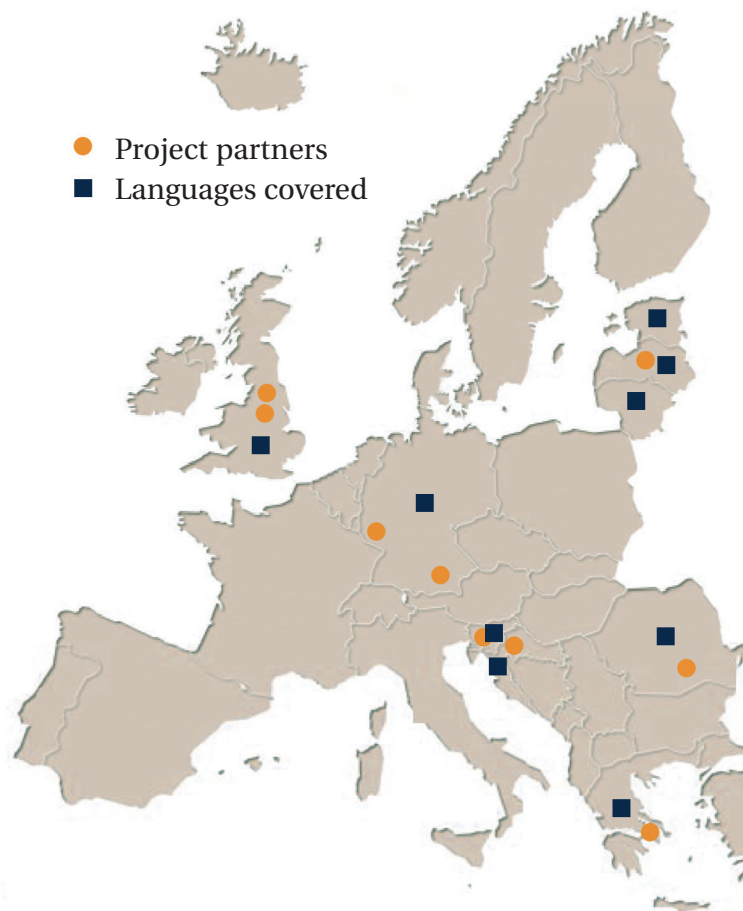




A true cross-European project connecting under-resourced languages, that are either EU official languages or are in the way to play that role soon, with well-resourced English and German.

Languages covered by the project: Latvian, Lithuanian, Estonian, Romanian, Greek, Croatian, Slovenian, English and German

www accurat-project.eu



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: Linguattec GmbH, Munich, Germany

ZEMANTA: Zemanta d.o.o., Ljubljana, Slovenia

Contact

Tilde, SIA
75a Vienibas Gatve
LV-1004 Riga
LATVIA

Contact person

Aivars Bērziņš
Project manager
P: + 371 67605001
F: +371 67605750
M: +371 29206344
E: aivars.berzins@Tilde.lv
W: <http://www.tilde.lv>



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

Do you have
ACCURAT
translation?

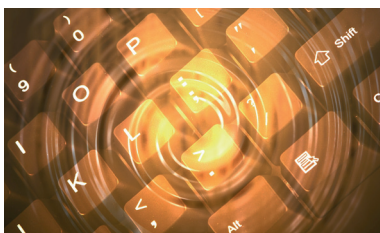
www accurat-project.eu



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

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 on order to compensate for the shortage of linguistic resources, and ultimately to significantly improve MT quality for under-resourced languages and narrow domains.



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 domains (e.g. Latvian and Croatian).

The goal of ACCURAT is to achieve increase in translation quality for under-resourced languages and narrow domains:

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



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.



Scientific objectives

- **Create comparability metrics** – to develop the methodology and determine criteria to measure the comparability of source and target language documents in comparable corpora;
- **Establish research methods for alignment and extraction** of lexical, terminological and other linguistic data from comparable corpora;
- **Develop research methods for automatic acquisition** of a comparable corpus from the Web;
- **Measure improvements** from applying acquired data against baseline results from SMT and RBMT systems.



Use cases

- Adjusting MT to narrow domain (automotive engineering, assistive technology, data processing domains, etc.)
- Develop application for Web authoring (blog and social networking)
- Use SMT in software localization (integration with CAT tools)

The increase of MT quality will be demonstrated on major and expected translation directions like English-Lithuanian, English-Croatian, German-Romanian, as well as minor and rare translation directions like Lithuanian-Romanian, Romanian-Greek and Latvian-Lithuanian.