ACCURAT

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

www.accurat-project.eu

Project no. 248347

ACCURAT Annual Public Report

2011

15/11/2011

Contents

1. PROJECT DESCRIPTION

As a 2.5 year long EU-funded research project, ACCURAT's main goal is to research methods and techniques in an effort to resolve a central problem of machine translation (MT) – the lack of linguistic resources for under-resourced languages and domains. The ACCURAT project aims to find, analyse, 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.

2. PROJECT OBJECTIVES

Traditional ways of building statistical machine translation (SMT) engines often do not produce acceptable translation quality for many domain/language combinations. The ACCURAT project addresses this issue by developing technology for the use of comparable corpora as resources for SMT translation models. The ACCURAT project's **key innovation** is the creation of a **methodology** and **tools** to measure, find, and use comparable corpora to improve the quality of MT for under-resourced languages and domains.

The **scientific objectives of the ACCURAT** project are to:

- **Create comparability metrics** – to develop a methodology and determine the criteria for measuring of the comparability of source and target language documents in comparable corpora;
- **Research methods for the alignment and extraction** of lexical, terminological, and other linguistic data from comparable corpora;
- **Research methods for automatic acquisition** of a comparable corpus from the Web;
- **Measure improvements** in applying acquired data against baseline results from statistic machine translation and rule based machine translation (RBMT) systems.

The ACCURAT project uses the latest state-of-the-art technology in SMT and RBMT systems as a baseline and provides novel methods to achieve better results through the use of comparable corpora.

The ACCURAT project investigates two broader use cases where the scarcity of linguistic resources poses a major challenge by adjusting machine translation for under-resourced languages and narrow domains.

3. SUMMARY OF ACTIVITIES

3.1. Criteria of comparability and comparability metrics

A key concept of the project is the notion of comparability. In the ACCURAT project, comparability is defined by how useful a pair of documents or segments of text are for machine translation.

Thus far, the ACCURAT project has implemented two different metrics for document-level comparability:

(1) ComMetric: This metric first uses the available machine translation API's for document translation and incorporates several useful features into the metric design. These features, including lexical information, keywords, document structure, and named entities, are then combined in an ensemble manner. Overall, given a pair of documents, the metric will compute a comparability score between 0 and 1, where a higher comparability score indicates higher comparability level. Using the Initial Comparable Corpora (ICC), that was collected during the first year of the ACCURAT project, as gold standard, the reliability of the proposed metric has been tested. It turns out that the comparability scores obtained from the comparability metric reliably reflect comparability levels, as the average scores for higher comparable levels are always significantly larger than that of lower comparable levels.

(2) DicMetric: Instead of using Google or Bing translation API's, we have also implemented another metric by using GIZA++ based bilingual dictionaries for lexical mapping. These bilingual dictionaries are automatically generated by using GIZA++ for word alignment in large-scale parallel corpora (e.g., JRC-Acquis). For each dictionary entry, translation candidates are sorted in descending order by their translation probabilities. Thus, for lexical mapping, if a word in the source language occurs in the bilingual dictionary, the top 2 translation candidates are retrieved as possible translations in the target language. This should help overcome the limitations of Google and Bing translations (e.g., translation access limit) by providing a much faster lexical translation process, although word-for-word lexical mapping results are not as good as Google or Bing translations.

3.2. Alignment methods

The term alignment is used in the context of machine translation to describe the pairing of text in one document with its translation in another. Alignment is commonly performed for texts that are translations of each other, but it is also possible to produce a type of alignment between texts that are not parallel, yet may be comparable to each other. The project consortium has built the toolkit that contains all important tools that have been currently developed within the ACCURAT project for the alignment of comparable corpora at different levels. These tools (which will be collectively referred to as the "ACCURAT Toolkit[1]") produce different types of data extracted from comparable corpora that are useful for machine translation. By using the ACCURAT Toolkit, users may expect to obtain:

- **translation dictionaries** extracted from comparable corpora; these dictionaries are expected to supplement existing translation lexicons which are useful to both statistical and rule/example-based MT;
- **translated terminology** extracted (mapped) from comparable corpora. This type of data is presented in a dictionary-like format and is expected to improve domain-dependent translation;
- **translated named entities** extracted (mapped) from comparable corpora. Also presented in a dictionary-like format, these lexicons are expected to improve the parallel phrase extraction algorithms from comparable corpora and be useful by themselves when actually used in translation. The problem of named entity mapping is not trivial since named entities may be transliterated and/or actually translated either word-by-word or as idioms;
- **comparable document (and other textual unit types) alignment**. This will facilitate the task of parallel phrase extraction by massively reducing the search space of such algorithms;
- **parallel sentence/phrase mapping** from comparable corpora. This aims to supply clean parallel data useful for statistical translation model learning.

In order to map terms and named entities bilingually, the ACCURAT Toolkit also provides tools for detecting and annotating these types of expressions in a monolingual fashion.

The toolkit has been well documented, and the documentation helps users to install and run the applications individually or in the provided workflows for parallel data mining from comparable corpora and named entity/terminology extraction and mapping from comparable corpora.

3.3. Methods for building a comparable corpus from the Web

ACCURAT partners have developed *methods* for collecting comparable corpora, implemented these methods in *software tools*, and used the tools to gather *comparable corpora* from a variety of sources. Distinct methods were developed for each of the three different types of comparable corpus data collected: *news texts, Wikipedia texts, and narrow domain texts*.

For **news texts**, a two-stage method was developed that first gathers documents monolingually and then pairs them across languages to build a comparable corpus. In the gathering stage, news texts are downloaded separately in each project language at regular intervals from the topical clusters that Google News supports. The titles from the downloaded articles are used as further queries to gather more related articles from Google News. For non-English languages, titles from the English news articles are parsed for named entities which are then translated into the non-English language and serve as queries

[1] http://www.accurat-project.eu/index.php?p=deliverables

to gather related news texts – this to help overcome the relative scarcity of non-English language news in Google News, particularly for the under-resourced languages ACCURAT is addressing.

For **Wikipedia texts**, two methods were developed. Work on this task was started with the observation that Wikipedia texts on the same topics are linked inter-lingually, which suggests gathering a comparable corpus using these links should be straightforward. However, investigations revealed that such links do not guarantee that the texts have significant content in common. Thus a technique was developed to find comparable Wikipedia texts based on the idea that inter-lingually linked Wikipedia text pairs that contain significant numbers of shared anchor texts (i.e., links to other Wikipedia entries where these other documents are also inter-lingually linked by Wikipedia) are likely to be quite similar in content.

For Romanian language retrieval, two methods were developed for gathering topically related documents in Wikipedia.

1. Using a list of good quality Romanian Wikipedia articles, the equivalent English documents are located in Wikipedia. This is done either by looking for the English page with the same name as the Romanian page, by using the Wikipedia inter-language links, or by using a dictionary to translate the Romanian page name into English and using this as a query to search for the English page.
2. Page names formed with existing capitalized nouns in Princeton Wordnet are searched for in Wikipedia, e.g., http://en.wikipedia.org/wiki/Barack_Obama. If the page exists, corresponding pages in different languages are located.

For **narrow domain texts**, two approaches were considered. The first is focused monolingual crawling where a topic definition (specified as a list of topic terms) and a seed URL list are given to a crawler that crawls starting from the seed URL's, performing lightweight text classification on pages it encounters to determine if they are relevant to the domain. All returned texts in language L1 for topic T may be paired with all texts from language L2 for topic T to form a comparable corpus. The second approach looks to download pairs of pages in L1 and L2 that are already somehow linked as parallel or comparable. Investigation showed the volume of data likely to be returned by the second approach was too small, so the first approach was adopted.

Software tools for building a comparable corpus from the Web[3]

Seven tools implementing these methods have been developed and publicly released. They are:

1. A *Workflow Based Corpora Crawler*. The workflow crawler is a GUI application that allows the user to shape the corpora collecting process through the use of diagrams. It uses processing blocks and decision blocks. Either type of block can be a script that does the work (e.g., gather links, download pages, etc.) or a wrapper for another console-based tool.
2. *Focused Monolingual Crawler (FMC).* The FMC tool is used to collect narrow domain bi-(multi)lingual comparable corpora from the Web. It does so by making a separate crawl for each language specified and by each time retrieving only Web pages that are relevant to a pre-defined narrow domain or topic.
3. *Wikipedia Retrieval Tool.* This tool was developed to identify and retrieve comparable documents in Wikipedia by specifically looking for pairs that contain similar sentences, e.g., sentences which contain overlapping information such as links (anchor texts), words, and numbers.
4. *News Information Downloader using Google News Search*. This tool constructs monolingual news corpora by searching Google News Search for the current news. It applies different search techniques such as "searching by topic" or "search for named entities".
5. *News Information Downloader using RSS feeds*. This tool takes as input a set of RSS feeds in a particular language and downloads the titles and metadata for all the reported news in those feeds.

[3] http://www.accurat-project.eu/index.php?p=deliverables

6. *News Text Crawler and RSS Feed gatherer*. This crawler can be seeded with HTML page URLs or RSS feed lists, observes any rules regarding automatic crawling, and prevents duplicate downloads in multiple iterations.
7. *News Article Alignment and Downloading Tool*. This tool pairs the articles in the monolingual news corpora to produce comparable corpora. Based on title and metadata such as publication date, the tool uses the paired article URL's to download the articles, extracts the text from the HTML presentation of the articles, and saves them to the hard disk.

Comparable corpora were collected using tools described above for under-resourced languages and from selected narrow domains.

The ACCURAT designated under-resourced languages were: Croatian, Estonian, German, Greek, Latvian, Lithuanian, Romanian, and Slovenian. Produced for news text, the "crawl method" resulted in comparable corpora for 8 language pairs, with the number of document pairs ranging from 720 to 29,341. The "wikipedia-anchors" method contains corpora in 12 language pairs, with the number of document pairs ranging from 841 to 149,891.

For narrow domains, 28 comparable corpora in 8 narrow domains and for 6 language pairs have been constructed and amount to a total of more than 148M tokens have been constructed. Metadata associated with the corpora include genre and size (in documents and tokens), main subtopics (into which the documents of a topic-specific corpus fall), and source URL's.

4. FUTURE EXPLOITATION PROSPECTS

ACCURAT exploitation scenarios are defined by the ACCURAT focus areas of under-resourced languages and narrow domains. In cooperation with its partners, ACCURAT will analyze, implement, and evaluate the project exploitation scenarios for three practical applications.

4.1. MT for specialists in narrow domain

Rule-based MT is adjusted for new domains by adding specific glossaries to the general system resources. The production of such glossaries from bilingual corpora therefore is a key element in the workflow. The workflow consists of the following steps:

1. collection of comparable corpora by focused monolingual crawlers;
2. alignment on sentence and phrase level, and creation of aligned data;
3. extraction of glossaries from such aligned phrases;
4. import of glossaries into the MT system, and evaluation of its performance.

In the period of the last 12 months, a tool was created to support step 3 of the workflow, to extract bilingual lexical entries from aligned phrases. It was developed and tested with parallel data of the automotive domain until results of step 2 were available. The tool, called *PhraseTable2Glossary* (P2G), applies linguistic filters to the aligned phrases and creates proper lexicon entries by generating correct (true-cased and inflected) lemmata and lexicon annotations (POS, etc.) for the candidates. As input, different alignment results were taken, and different thresholds were tried. Table 1 shows translation error rates:

Table 1 Translation error rates

Alignment	translation probability threshold	Input tokens: size	lexicon candidates found	candidates evaluated	error rate
Phrase Tables (Giza/Moses)	> 0.8	7.900.000	12.000	1.850	5.7%
Phrase Tables (Giza/Moses)	> 0.6	7.900.000	16.000	2.450	5.7%
Phrase Tables (Giza/Moses)	> 0.4	7.900.000	36.000	3.000	16.3%
AnymAlign	> 0.8	3.100.000	9.800	1.700	47.5%
AnymAlign	> 0.7	3.100.000	12.600	2.100	46.9%

In addition, the error rates produced by the tool itself were evaluated. The main sources of errors were: wrong multiword patterns used as filter, errors in true-casing (e.g., in hyphenated forms like *Heiz-AB-Spule*), and lemma production. These error rates are 5.6% in the case of English and 5.7% in the case of German. In the best case scenario, the P2G tool has an accuracy of 85%, which is considered to be sufficient for human post-editing input.

To prepare the import of the glossaries of the P2G tool into the backend MT system, two more steps have been implemented: elimination of all term candidates which are already known to the lexicon, and defaulting of information which the MT system requires for its import (e.g., part-of-speech and gender information).

4.2. MT for Web authoring

Zemanta has been developing an authoring tool, an automatic research assistant helping Web authors, mostly bloggers, to write and publish better content. In the last few months, we conducted experiments to evaluate the quality of state-of-the art (baseline) machine translation methods in 'real-life' Web authoring applications, which will later be compared with the quality of the ACCURAT machine translation method. Initial results of using baseline machine translation methods for the Slovenian language are promising – the precision of recommended articles increased by more than 10% ().

The evaluation process was twofold: first we evaluated original Slovenian language texts, and then we repeated the process with texts translated into English using current baseline machine translation system.

In the first part Slovenian texts have been fed into the authoring assistant, which provided 10 related articles per text (Figure 1). Each of the articles was manually checked by human evaluators, who rated whether the suggested article is actually related to the content (text analyzed) or not by assigning it a score between 0 (a blogger would definitely not use it) and 3 (a blogger would definitely use it). After evaluators assigned scores to all related articles for all of the texts, we calculated precision and Normalized Discounted Cumulative Gain (NDCG) to estimate the quality of baseline machine translation.

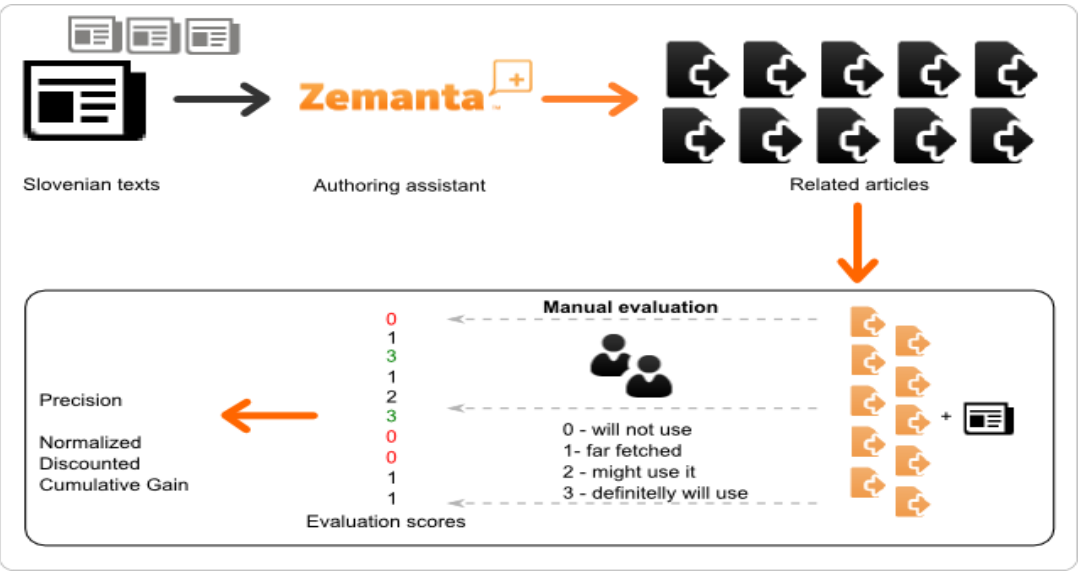


Figure 1: Evaluation of the Zemanta authoring assistant using Slovenian texts (no translation)

In the second part, we translated Slovenian texts into English and then fed them into the authoring assistant, which again provided 10 related articles per text (Figure 2). The evaluation process continued as in the first part. After obtaining precision and NDCG values for both sets of texts (Slovenian and English), we were able to compare them to find out whether the results of using current baseline machine translation system were better, i.e., whether more articles suggested as "related" by the Zemanta authoring assistant were *actually* related to the text.

Table 2 Evaluation results of using baseline machine translation method for Slovenian texts.

Metrics	Precision		NDCG Score	
Texts	**Original (Slovenian)**	**Translated (English)**	**Original (Slovenian)**	**Translated (English)**
	0.333	0.461	0.344	0.471

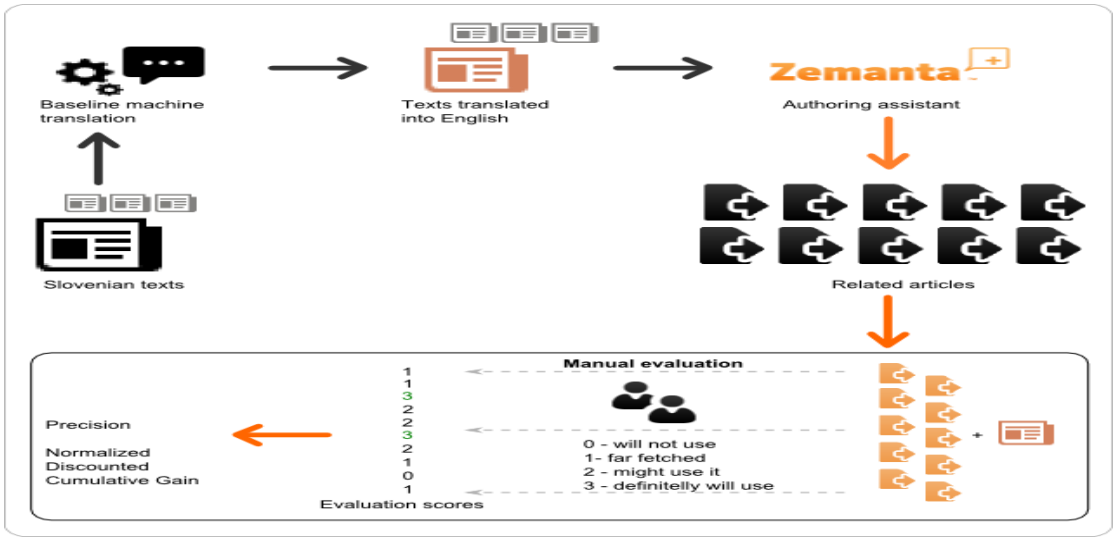


Figure 2: Evaluation of baseline machine translation results (Slovenian → English) using Zemanta authoring assistant

4.3. MT for localization services

One of the ACCURAT target use-cases is application in localization services for under-resourced languages. For this ACCURAT English-Latvian baseline SMT system is integrated into SDL Trados CAT-tool ().

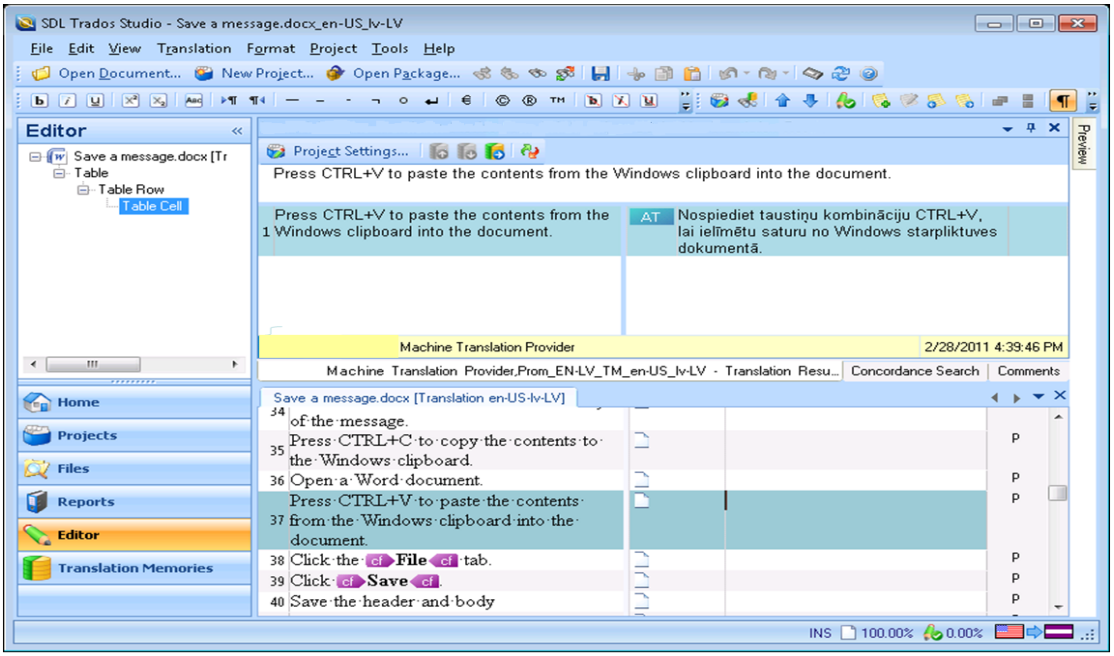


Figure 3 SMT system integrated into SDL Trados

The system is used to provide translation recommendations for those translation segments that do not have exact match or close match in the translation memory. Localization specialists are able to choose these translations and post-edit them for a professional result. Suggestions coming from the MT are clearly marked to allow translators to pay more attention to these suggestions.

Evaluation of ACCURAT SMT systems in localization scenario is based on the measurement of translation performance calculated as a number of words translated per hour. As efficiency (translation performance) of translation process without degradation of quality is the most important measure for localization service provider, the SMT systems are being tested against available manual translation productivity and quality.

Our initial experiments showed that usage of MT suggestions in addition to the use of the translation memories increased productivity of the translators in average from 550 to 731 words per hour (32.9% improvement).

5. DISSEMINATION

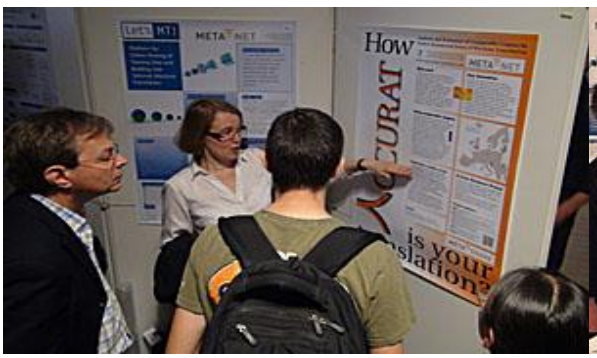
The visibility of the ACCURAT project is assured by a unique visual identity (logo) that helps recognise the project among similar projects. The visual identity was designed and applied to all possible and even non-conventional channels of dissemination, such as the public Web site, the presentation template, leaflets, posters, t-shirts, video lectures, and social networks.

The ACCURAT Website (http://www.accurat-project.eu/) is one of the main project communication tools. The web page contains various materials that reflect the project's aims, research progress and impact. This is the place where all information related to ACCURAT is stored and made accessible to the Internet sharing community.

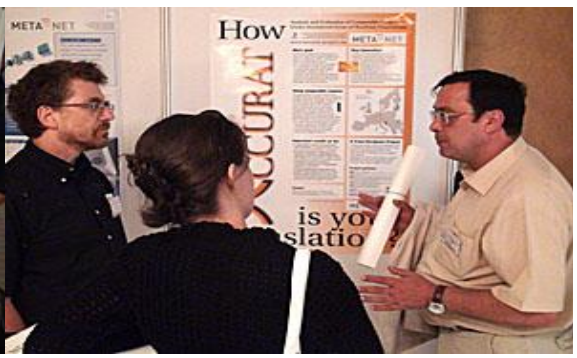
Dissemination to the scientific community is based on a bilateral exchange of information by consortium partners with major scientific institutions as well as presentation and communication of project. Project achievements are also presented and communicated in conferences and through publication of research methodologies, strategies, and outcomes.

Conferences and workshops where ACCURAT has participated:

- FLaReNet Forum (Venice, 26-27 May 2011)
- EAMT2011 (Leuven, 30-31 May 2011)
- NooJ2011 conference (Dubrovnik, 13-15 June 2011)
- META-FORUM (Budapest, 27-28 June 2011)
- SCFM (Zürich, 26 August 2011)
- SlaviCorp2011 (Dubrovnik, 12-14 September 2011)
- CLARA Career Course (Dubrovnik, 20-23 September 2011)



ACCURAT presented at EAMT2011



ACCURAT presented at META-FORUM in Budapest

5.1. Publications and presentations of the project November 2010 - November 2011

Papers on different aspects of research within the ACCURAT project were presented at the major conferences in the field:

- MT and Morphologically Rich Languages 2011 (Haifa, 23-27 January 2011) Goba, K.; Skadiņš, R. *Improving SMT with Morphology Knowledge for Baltic Languages*;
- CICLING-2011 (Tokyo, 20-26 February 2011) Babych, B.; Hartley, A. Meta-evaluation of comparability metrics using parallel corpora;
- NooJ2011 (Dubrovnik, 13-15 June 2011) Berović, D.; Merkler, D.; Agić, Ž. *Disambiguation of homographic adjective and adverb forms in Croatian*;
- ACL2011 (Portland, 24 June 2011) Fišer, D.; Ljubešić, N.; Vintar, Š.; Pollak, S. *Building and using comparable corpora for domain-specific bilingual lexicon extraction*;
- ACL2011 (Portland, 24 June 2011) Ion, R.; Ceauşu, A.; Irimia, E. *An Expectation Maximization Algorithm for Textual Unit Alignment*;
- SFCM2011 (Zürich, 26 August 2011) Pinnis, M. *Maximum Entropy Model for Disambiguation of Rich Morphological Tags*;
- TSD2011 (Plzeň, 1-5 September 2011) Ljubešić, N.; Fišer, D. *Bootstrapping bilingual lexicons from comparable corpora for closely related languages*;
- TSD2011 (Plzeň, 1-5 September 2011) Ljubešić, N.; Erjavec, T. *hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene*;
- RANLP2011 (10-16 September 2011) Fišer, D.; Ljubešić, N. *Bilingual lexicon extraction from comparable corpora for closely related languages*;
- SlaviCorp2011 (12-14 September 2011) Ljubešić, N. Erjavec, T. *hrWaC and slWac: Web Corpora for Croatian and Slovene*;
- SlaviCorp2011 (12-14 September 2011) Agić, Ž.; Berović, D.; Merkler, D. *Development and Applications of the Croatian 1984 Corpus for the MULTEXT-East Resources*.

6. COLLABORATION

ACCURAT cooperates with national and international research activities in related areas. The project closely collaborates with FP7 projects TTC, PANACEA and EuroMatrixPlus, ICT PSP projects LetsMT!, EASTIN-CL, META-NORD and META-NET Network of Excellence.

7. ACCURAT CONSORTIUM AND CONTACT PERSONS

URL: http://www.tilde.eu

Tilde SIA
Vienibas gatve 75a
Riga, LV1004, Latvia
Project Coordinator:
Andrejs Vasiļjevs, **andrejs[at]tilde.lv**

URL: http://nlp.shef.ac.uk/

THE UNIVERSITY OF SHEFFIELD
Natural Langauge Processing Research Group, Department of Computer Science, University of Sheffield
Regent Court
211 Portobello
Sheffield, S1 4DP, UK
Contact person:
Professor Rob Gaizauskas , R.Gaizauskas[at]sheffield.ac.uk

URL: http://www.leeds.ac.uk/cts/en/index.htm

UNIVERSITY OF LEEDS
Centre for Translation Studies, School of Modern Languages and Cultures , University of Leeds
Leeds LS2 9JT , UK
Contact person:
Bogdan Babych, b.babych[at]leeds.ac.uk

URL: http://www.ilsp.gr

INSTITUTE FOR LANGUAGE & SPEECH PROCESSING
Artemidos 6 & Epidavrou
GR-151 25 MAROYSSI, Greece
Contact person:
Dr. Nicholas Glaros, **nglaros[at]ilsp.gr**

URL: http://hnk.ffzg.hr/default_en.htm

UNIVERSITY OF ZAGREB
Trg maršala Tita 14
HR-10002 ZAGREB, Croatia
Contact person:
Prof. Marko TADIĆ, **marko.tadic[at]ffzg.hr**

URL: http://www.dfki.de/lt/

GERMAN RESEARCH CENTRE FOR ARTIFICIAL INTELLIGENCE
Forschungsbereich Sprachtechnologie
Stuhlsatzenhausweg 3 / Building D3 2
D-66123 Saarbrücken, Germany
Contact person:
Jia Xu, **Jia.Xu[at]dfki.de**

URL: http://www.racai.ro/

RESEARCH INSTITUTE FOR ARTIFICIAL INTELLIGENCE OF THE ROMANIAN ACADEMY
Calea 13 Septembrie, No. 13
CASA ACADEMIEI
Bucharest 050711 , Romania
Contact person:
Prof. Dan TUFIS. **dan_tufis2006[at]yahoo.com**

URL: http://www.linguatec.de/

LINGUATEC
Gottfried-Keller-Straße 12
81245 Munich, Germany
Contact person:
Dr. Gregor Thurmair, **g.thurmair[at]linguatec.de**

URL: http://www.zemanta.com/

ZEMANTA
Zemanta d.o.o.
Pugljeva 8
SI - 1110 Ljubljana, Slovenia
Contact person:
Gasper Koren, **gasper.koren[at]zemanta.com**