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 2010

15/11/2010

Contents

1. PROJECT DESCRIPTION

ACCURAT is a 2.5 year long EU-funded research project that aims to research methods and techniques to overcome one of the central problems of machine translation (MT) – the lack of linguistic resources for under-resourced languages and domains. The main goals are 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.

2. PROJECT OBJECTIVES

Traditional ways of building SMT engines that produce acceptable translation quality are often not possible for many domain / language combinations. The ACCURAT project is addressing this issue by developing the technology for using comparable corpora as resources for SMT translation models. The **key innovation** of the ACCURAT project will be the creation of a **methodology** and **tools** to measure, find, and to 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 the methodology and determine criteria to measure 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** from applying acquired data against baseline results from statistic machine translation (SMT) and rule based machine translation (RBMT) systems.

The project will use the latest state-of-the-art in SMT and rule-based MT systems as a baseline and will provide novel methods to achieve much better results by extending these systems through the use of comparable corpora.

The ACCURAT project will investigate two broader use cases where the scarcity of linguistic resources poses a major challenge – 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 can be defined by how useful a pair of documents or segments of text are for machine translation. Therefore initially four levels of comparability were introduced:

- **Parallel corpora** – collections of traditional parallel texts that are either true and accurate translations of each other, or approximate translations with minor variations, which can be aligned on the sentence level.
- **Strongly comparable corpora** – collections of closely related texts reporting the same event or describing the same subject, which typically can be aligned on text level.
- **Weakly comparable corpora** – collections of texts in the same subject domain and genre, but describing different events. These corpora typically cannot be aligned on the text level, but still can contain collections of translation equivalents.
- **Non-comparable:** pairs of tests drawn at random from a pair of very large collections of texts (e.g. the web) in the two languages.

During the first months of the project we identified an initial set of criteria of comparability that can guide our procedure to construct comparable corpora. We primarily focused on comparability on higher levels (corpus and document comparability), with the task for selecting comparable corpora, texts and paragraphs for further alignment and use within MT. Features that can be used to identify the comparability level of a pair of documents are summarized in Table 1. These features are divided into

two categories: language dependent features which need some translation methods or other linguistic knowledge, and language independent features which do not.

Table 1. Features of comparability

Language-dependent Features	Language-independent Features
LM divergence (words, phrases, N-grams)	String match overlap or letter N-gram overlap (without translation)
Cosine similarities (words, phrases, N-grams)	Out-link overlap
Named entities LM divergence / cosine similarities	Number of links to each other
Named entity tags LM divergence / cosine similarities	Genre overlap (binary)
String match overlap or letter N-gram overlap	Domain overlap (binary)
Parts of speech (including POS N-grams, frequent discontinuous POS signatures)	Date proximity
	Document length difference / ratio
	URL character overlap
	URL slash overlap

All language dependent features identified in Table 1 appear to be good indicators of the comparability level of a document pair. In particular an identifiable gap in the values between parallel, strongly comparable, weakly comparable and non-comparable document pairs is present for most of the features. Only, the cosine similarity among term frequencies seems not to be able to separate strongly from weakly comparable.

Language independent features were useful for identifying parallel documents, but do not identify any of the other degrees of comparability well, displaying little difference in the values of these features between comparable and non-comparable documents. One exception is the Image Link Overlap, which appears to identify both strongly comparable and parallel documents.

The features are used for developing and implementing comparability metric, i.e., the software package, which allows to compute multidimensional features and feature combinations on the level of individual texts and the whole corpora that show their level of comparability. The comparability metrics processes corpora and texts in a modular way with the aim to extract and compare features in a unified standard format.

Identification of these successful features indicated the need to develop a more general framework for systematic testing and discovery of useful feature combinations across different dimensions in documents and in corpora. The framework, which we implemented as open software architecture, generalises the way how features are annotated and extracted from corpora and texts, and allows us to test novel feature combinations of comparability which work best for specific translation directions.

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 but may be comparable to each other.

We have studied and evaluated existing alignment strategies designed for parallel corpora, comparable corpora, and non-comparable corpora. We focused particularly on the appropriateness of these techniques to corpora of different levels of comparability and established guidelines for their applicability and the resources required by the techniques. A case study was performed making use of four different alignment methods (Giza++, Moses, cognate based alignment, co-occurrence based alignment) and applying them to corpora of different levels of comparability. We tested the accuracy of these methods for alignment of words or phrases by comparing the alignments produced to human word alignments. We showed that the most widely used existing alignment methods (Giza++ and Moses) are not well suited for use directly on strongly or weakly comparable texts, but for parallel corpora it is possible to 85% of the correct alignments using this method (Figure 1).

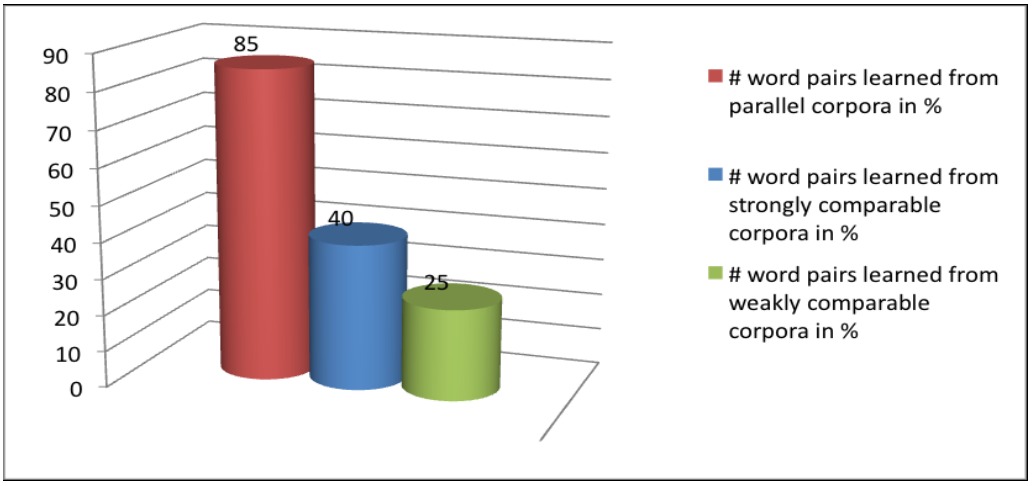

Figure 1. Results for alignment of all word tokens in test data of different levels of comparability using Giza++

Additionally, we showed that for weakly comparable corpora it is possible to correctly identify only around 35% of the alignments in text using word co-occurrence information. The results indicate that there is much room for improvement on alignment accuracy of strongly and weakly comparable texts and increasing this accuracy will major point of focus for this project.

3.3. Methods for building a comparable corpus from the Web

At first the initial comparable corpora (ICC) have been collected for ACCURAT 9 language pairs: Estonian-English, Latvian-English, Lithuanian-English, Greek-English, Romanian-Greek, Croatian-English, Romanian-English, Romanian-German and Slovenian-English, following common domain/genre distribution principles (Table 2).

Table 2. Domain and genre distribution of initial comparable corpora

Domain	Genre	Percent
International news	Newswires	20%
Sports	Newswires	10%
Admin	Legal	10%
Travel	Advice	10%
Software	Wikipedia	15%
Software	User manuals	15%
Medicine	For doctors	10%
Medicine	For patients	10%

Every language corpus in ICC consists of approximately one million words (Table 3 indicates number of words for under-resourced language), there are some small deviations in word count for lesser used language pair, i.e. Romanian-Greek. Parallel and strongly comparable documents are aligned at the document level.

Table 3. Size and comparability level of ICC

	Parallel		**Strongly comparable**		**Weakly comparable**		**Total**
	Words	**%**	**Words**	**%**	**Words**	**%**	
ET-EN	101 884	9,48	548 764	51,06	424 022	39,46	**1 074 670**
LV-EN	122 581	11,82	389 127	37,51	525 681	50,67	**1 037 389**
LT-EN	553 747	46,17	261 841	21,83	383 819	32	**1 199 407**
EL-EN	191 843	13,33	294 554	20,47	952 534	66,2	**1 438 931**
RO-EL	282 213	32,62	267 897	30,96	315 108	36,42	**865 218**
HR-EN	418 752	39,51	100 000	9,44	541 085	51,05	**1 059 837**
RO-EN	186 682	6,94	459 458	17,07	2 045 631	76	**2 691 771**

RO-DE	117 281	8,52	449 942	32,67	809 929	58,81	**1 377 152**
SL-EN	462 514	40,17	322 243	27,98	366 759	31,85	**1 151 516**
All language pairs	**2 018 745**	**20,49**	**2 993 826**	**26,01**	**5 823 483**	**53,5**	**11 895 891**

The ICC is intended as initial development data for the creation of comparability metrics and for the evaluation of existing alignment methods and will be expanded with content from the Web during the project lifetime.

For the German-English language pair ICC have been collected for automotive, medicine and software domains. The proportions of collected parallel, strongly comparable and weakly comparable corpora for the German-English language pair are provided in the table below (Nr. of words in German).

Table 4. Size and proportion of the German-English narrow domain corpus

Domain	Genre	Comparability			Total
		Parallel	Strong	Weak	
Automotive	Transmission	21 721	291522	288 044	601 287
Medicine	Pharmaceuticals (Emea)	11 066 466			11 066 466
Software	Wikipedia		723734		723 734
Software	User manuals	258 238	360062	12 983	631 283
Total		11 346 425	1 375 318	301 027	13 022 770

Since one of the central tasks of the ACCURAT project is to research methods for automatic acquisition of a comparable corpus from the Web, several novel approaches how to build a comparable corpus from the Web that are applicable to under-resourced languages have been researched and proposed. The proposed methods are focused on retrieving specific types of text (e.g. news, Wikipedia, narrow-topics, Twitter). The gathered documents can then be automatically processed further to choose the pairs most likely to be comparable.

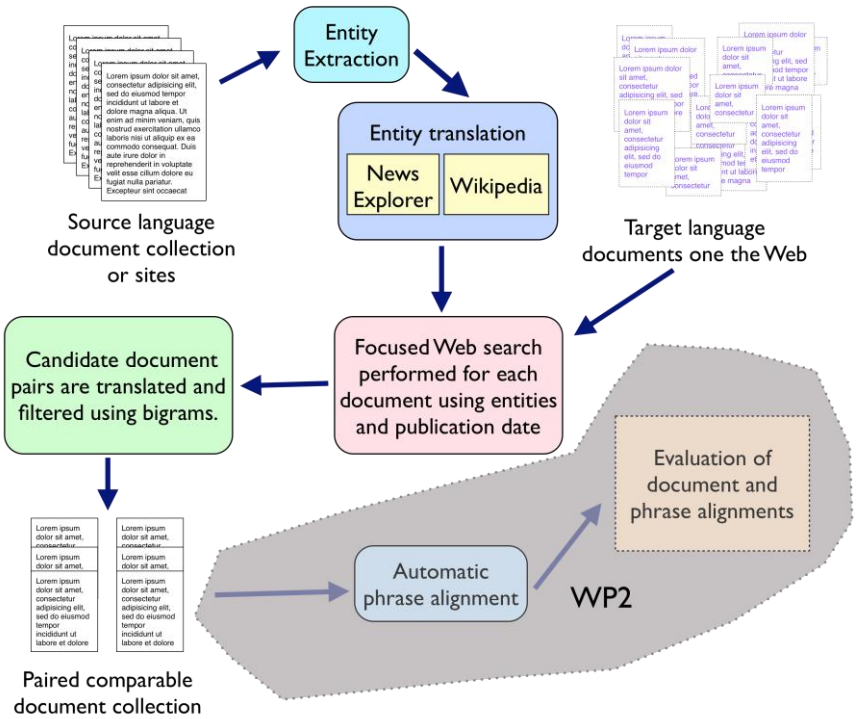

Figure 2. Retrieval of Comparable Documents from News Sites

3.4. Comparable corpora in MT systems

To evaluate efficiency and usability of above mentioned methods and techniques for under-resourced languages and narrow domains research results will be integrated into ACCUARAT baseline MT systems. Therefore the ACCURAT baseline SMT systems have been be set up for 17 translation routes: English⇨Latvian, English⇨Lithuanian, English⇨Estonian, English⇨Greek, English⇨Croatian, Croatian⇨English, English⇨Romanian, English⇨Slovenian, Slovenian⇨English, German⇨English, German⇨Romanian, Romanian⇨German, Lithuanian⇨Romanian, Romanian⇨Greek, Greek⇨Romanian, Romanian⇨English, Latvian⇨Lithuanian. The SMT systems are created using existing SMT techniques – Moses decoder, language models and translation models trained on available parallel corpora e.g. JRC-ACQUIS Multilingual Parallel corpus, SETimes corpus.
In order to facilitate the access to MT functionality both for research and applications, a software infrastructure that provides a client-server architecture in which MT engines for many language pairs, many users, and many domains can be easily accessed through a uniform framework is being developed.
This framework, which we call "MT Serverland" is available as an open-source software under a BSD-style license. Figure 3 shows the translation result for theEnglish-Latvian SMT system on Serverland.

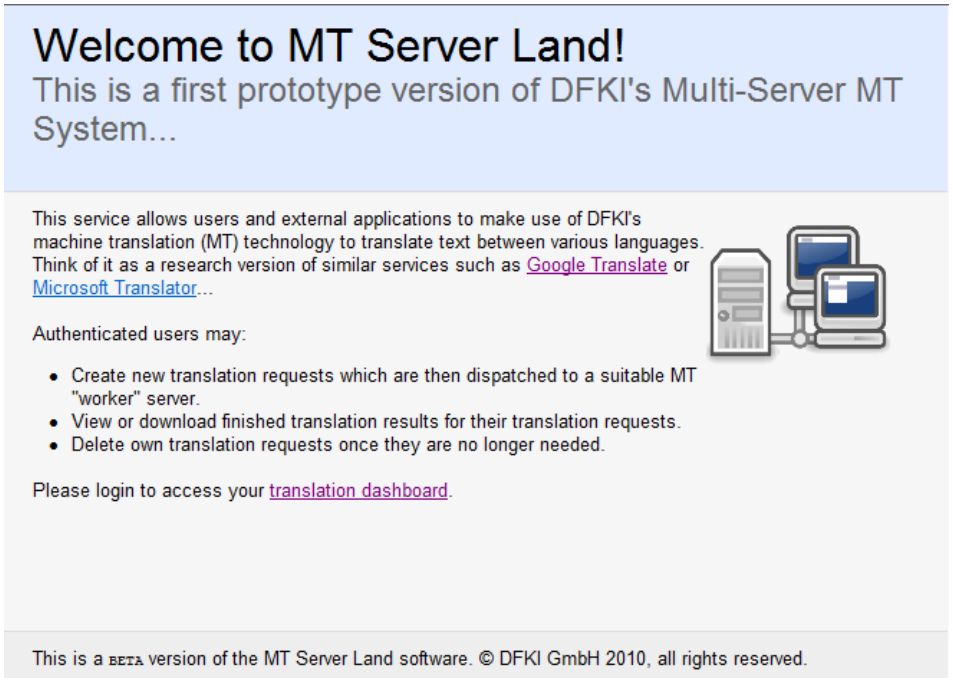

Figure 3 MT Serverland

4. FUTURE EXPLOITATION PROSPECTS

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

4.1. MT for specialists in narrow domain

In a market which is characterized by a large number of MT providers, and the possibility of getting free access to MT services provided by large multinationals like Google, one possible option is to take a step outside the general market, and focus on the high quality segment. This step requires tuning the MT engine towards particular domains and text areas, mainly by creating application-specific dictionary resources. Manual creation of dictionary resources has become less and less profitable, in particular in specialized domains where few expert translators / lexicographers are available. In addition, such dictionaries are already outdated when they reach the market. Corpus-based techniques allow for an up-

to-date and cost-effective alternative, and therefore the semi-automatic creation of dictionary resources from parallel, or even comparable, corpora is a clear option. Linguatec will exploit the technology of multilingual term extraction, as developed in ACCURAT, to tackle various market segments, analyzing its current user distribution (e.g. in automotive, in IT, and others), to make efforts to improve MT quality by offering domain-specific language resources, starting with its best-selling language directions.

4.2. MT for web authoring

Zemanta has developed an authoring assistant, to help its users to publish enriched content. Zemanta's tool acts as an automatic research assistant that comes up with actionable content for the writer. Currently Zemanta's tools support only the English language. During project implementation for web authoring we will incorporate the ACCURAT translation solution into an "Authoring assistant" tool for German⇨English, Slovenian⇨English and Croatian⇨English language.

4.3. MT for localization services

In recent years the localization industry has started to use MT to reduce the human workload in the translation business. However, until now due to the constraints of under-resourced areas MT application has been limited to the larger languages and major domains only. Therefore one of the ACCURAT target use-cases is application in localization services for under-resourced languages. The goal is to integrate ACCURAT results within the localization process to increase the efficiency of the translation process, increasing the output of translators in comparison to the results available by manual translation. For this the following SMT systems will be integrated into SDL Trados CAT-tool: English⇨Latvian, English⇨Lithuanian and English ⇨Estonian.

5. DISSEMINATION

5.1. Dissemination Strategy

The visibility of the ACCURAT project is assured by a unique visual identity (logo) that helps in recognising the project among the similar projects. The visual identity was designed and applied to all possible channels of dissemination such as public web site, presentation template, leaflets, posters, but also t-shirts and more non-conventional channels such as e.g. video lectures, Wikipedia article or social networks.

5.1.1. Web site

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. Beside classic dissemination web-site content (project description, publications, deliverables etc.), we also provide some newer means of dissemination information about the project such as video recorded lectures.
Currently the Web site provides descriptions of project in all of the project languages including English, Latvian, Greek, Croatian, German, Romanian and Slovenian.

5.1.2. Dissemination to the scientific community and the industry

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 achievements in conferences and through publication of research methodologies, strategies and outcomes.
ACCURAT conference and workshops presence:

- LTDays at Luxembourg, 22-24 March 2010;
- LREC2010 Valletta, 17-23 May 2010/ EU Projects Village;

- LREC2010 3rd Workshop on Building and Using Comparable Corpora 22 May 2010 (vide lectures available on ACCURAT web page);
- LREC2010 Valletta, Workshop on Methods for the automatic acquisition of Language Resources and their evaluation methods 23 May 2010
- 14th EAMT2010 annual conference, Saint-Raphaël (France), 27-28 May 2010;
- FASSBL7 conference, Dubrovnik, 4-6 October 2010;
- Baltic HLT2010 conference, Riga, 7-8 October 2010.

Figure 4. ACCURAT at LREC 2010

Figure 5. ACCURAT at FASSBL7 conference

5.2. Publications and presentations of the project team in the period January to November 2010

5.2.1. Papers

- Eisele A., Xu J. Improving Machine Translation Performance Using Comparable Corpora // Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, European Language Resources Association (ELRA), La Valletta, Malta, pp 35-41, May 2010.
- Ion, R., Tufiş, D., Boroş, T., Ceauşu, A., Ştefănescu, D. On-Line Compilation of Comparable Corpora and their Evaluation, // Proceedings of the 7th International Conference Formal Approaches to South Slavic and Balkan Languages (FASSBL7), Croatian Language Technologies Society – Faculty of Humanities and Social Sciences, University of Zagreb, Dubrovnik, Croatia, pp 29-34, October 2010.
- Skadiņa I., Vasiļjevs A., Skadiņš R., Gaizauskas R., Tufis D, Gornostay T. Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation // Proceedings of the 3rd Workshop on Building and Using Comparable Corpora. European Language Resources Association (ELRA), La Valletta, Malta, pp 6-14, May 2010.
- Skadiņa, I., Aker, A., Giouli, V., Tufis, D., Gaizauskas, R., Mieriņa M., Mastropavlos, N. A Collection of Comparable Corpora for Under-resourced Languages. // Proceedings of the Fourth International Conference Baltic HLT 2010, IOS Press, Frontiers in Artificial Intelligence and Applications, Vol. 219, Riga, Latvia, pp 161-168, October 2010.
- Šojat, K., Agić, Ž., Tadić, M. Verb Valency Frame Extraction Using Morphological And Syntactic Features Of Croatian // Proceedings of the 7th International Conference Formal Approaches to South Slavic and Balkan Languages (FASSBL7), Croatian Language Technologies Society – Faculty of Humanities and Social Sciences, University of Zagreb, Dubrovnik, Croatia, pp 119-126, October 2010.
- Vučković, K., Agić, Ž., Tadić, M. Sentence Classification and Clause Detection for Croatian // Proceedings of the 7th International Conference Formal Approaches to South Slavic and Balkan Languages (FASSBL7), Croatian Language Technologies Society – Faculty of Humanities and Social Sciences, University of Zagreb, Dubrovnik, Croatia, pp 131-138, October 2010.

5.2.2. Invited Talks

- Vasiljevs, A., ACCURAT – Analysis and evaluation of comparable corpora for under resourced areas of machine translation, Language Technology Days, Luxembourg, March 22-23, 2010.
- Eisele, A., From corpora to resources and tools – towards a proper treatment of Eastern European languages, The Fourth International Conference Human Language Technologies — the Baltic Perspective, Riga, Latvia, October 7–8, 2010.

5.2.3. Presentations at conferences

- Eisele A., Xu J. Improving Machine Translation Performance Using Comparable Corpora // *The 3rd Workshop on Building and Using Comparable Corpora, LREC2010, La Valletta, Malta, 22 May 2010.*
- Eisele, A. ACCURAT poster, 14th Annual Conference of the European Association for Machine Translation, Saint-Raphaël, France, 27-28 May 2010.
- Ion, R., Tufiş, D., Boroş, T., Ceauşu, A., Ştefănescu, D. On-Line Compilation of Comparable Corpora and their Evaluation, // *The 7th International Conference Formal Approaches to South Slavic and Balkan Languages (FASSBL7), Dubrovnik, Croatia, 4-6 October 2010.*
- Skadiņa I., Vasiļjevs A., Skadiņš R., Gaizauskas R., Tufis D, Gornostay T. Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation // *The 3rd Workshop on Building and Using Comparable Corpora, LREC2010, La Valletta, Malta, 22 May* 2010.
- Šojat, K., Agić, Ž., Tadić, M. Verb Valency Frame Extraction Using Morphological And Syntactic Features Of Croatian // *The 7th International Conference Formal Approaches to South Slavic and Balkan Languages (FASSBL7), Dubrovnik, Croatia, 4-6 October 2010.*
- Štefanec, V., Vučković, K., Dovedan, Z. Towards Parsing Croatian Complex Sentences: Dependent Noun Clauses // *NooJ2010 Conference, Komotini, Greece, 27-29 May 2010.*
- Vasiļjevs, A. ACCURAT: Metrics for the evaluation of comparability of multilingual corpora, The Workshop on Methods for the Automatic Acquisition of Language Resources and their Evaluation Methods, LREC2010, *La Valletta, Malta, 23 May* 2010.
- Vučković, K., Agić, Ž., Tadić, M. Sentence Classification and Clause Detection for Croatian // *The 7th International Conference Formal Approaches to South Slavic and Balkan Languages (FASSBL7), Dubrovnik, Croatia, 4-6 October 2010.*
- Vučković, K., Bekavac, B., Dovedan, Z. Improved Parser for Simple Cro Sentences // *NooJ2010 Conference, Komotini, Greece, 27-29 May 2010.*
- Skadiņa, I., Aker, A., Giouli, V., Tufis, D., Gaizauskas, R., Mieriņa M., Mastropavlos, N. A Collection of Comparable Corpora for Under-resourced Languages. // The Fourth International Conference Baltic HLT 2010, Riga, Latvia, 7-8 October 2010.

5.2.4. Organized events

- Workshop on Methods for the automatic acquisition of Language Resources and their evaluation methods (May 23, 2010) was organized by FLaReNet, ACCURAT, PANACEA and TTC in LREC2010 conference
- The Fourth International Conference Human Language Technologies –Baltic Perspective (Baltic HLT 2010) in Riga, 7-8 October, was supported by ACCURAT, LetsMT! and CLARIN projects.

6. COLLABORATION

ACCURAT cooperates with national and international research activities in related areas. The project has close collaboration with FP7 projects TTC, PANACEA and EuroMatrixPlus, ICT PSP projects LetsMT! and EASTIN-CL, .ACCURAT also collaborates with FLaReNet and CLARIN projects. The project plans to sign a collaboration agreement with the META-NET Network of Excellence.

The ACCURAT project and its partners have participated in various activities of knowledge sharing within their domains of expertise. A joint Workshop on Methods for the automatic acquisition of Language Resources and their evaluation methods was organized by FLaReNet, ACCURAT,

PANACEA and TTC during the LREC2010 conference.

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**