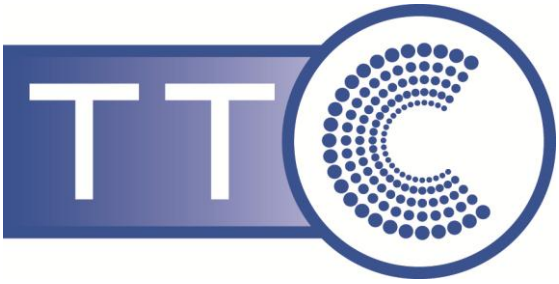
TTC Annual Public Report 2010



Terminology Extraction, Translation Tools and Comparable Corpora

www.ttc-project.eu

Project duration: 1st of January 2010 to 31st of December 2012 (36 months)

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n°248005.

Introduction

The need for linguistic resources in any natural language application is undeniable. The problem is especially difficult for translation applications because of cross-linguistic divergences and mismatches that arise from the perspective of the lexicon. Lexicons and terminologies play indeed a central role in any machine translation tool, regardless of the theoretical foundations upon which the machine translation (MT) tool is based (e.g. statistical machine translation or rule-based machine translation, example-based translation, etc). Terminologies may be extracted from parallel corpora, i.e. from previously translated texts, but such corpora are scarce. Previously translated data is still sparse and only available for some pairs of languages and few specific domains, such as Europarl (Koehn, 2005). Thus, no parallel corpora are available for most of specialized domains, especially for emerging domains (such as renewable energy).

The TTC project develops methods and tools for automatic extraction of terminologies from comparable corpora, i.e. from "sets of texts in different languages that are not translations of each other" (Bowker and Pearson, 2002, p.93). Moreover, TTC aims at developing tools for gathering and

managing comparable corpora and terminologies: a focused web crawler, a tool to handle comparable corpora using the existing open source UIMA (Unstructured Information Management Architecture1) framework, and a terminology management tool. At the end of the TTC project, a platform will be set up to regroup those tools and to compile and manage comparable corpora using standards (TMF, TBX) and UIMA framework. The open terminology platform will support tasks such as terminology storage, search, editing and export and reuse Eurotermbank[2]. Translation of technical documents for aerospace and IT domain will be performed by end-users of the consortium using CAT and MT techniques to assess impact of the TTC project outputs.

Main Goal

The main goals of the projects are as follows:

1. Using comparable corpora

The TTC will be using comparable corpora and take advantage from the huge amount of textual multilingual data available on the web ("Web as a Corpus" approach, Kilgarriff and Grefenstette; 2003). As corpora compiled from the Web can be inconsistent and too variable, the project develops a topical web crawler to gather comparable corpora from domain-specific Web portals or using query-based crawling technologies with several types of conditional analysis. The objective is to guarantee the monolingual comparability (i.e. the comparability between pairs of texts in the same language), as well as the interlingual comparability of the compiled corpora by identifying the most relevant criteria and measures, such as structural, modal or lexical criteria. Besides, the elaborated typologies should be easily adaptable to other languages.

2. Using a minimum of linguistic knowledge for candidate term extraction

Term candidates and their relevant context partners (e.g. collocations) will be extracted from corpora using available monolingual term extractors. However, the TTC platform plans to handle under-resourced European languages, i.e. languages with less available tools and resources. Therefore, the scientific objective is the assessment of the minimal amount of language-specific linguistic knowledge which is needed to identify term candidates and extract terminologies. Moreover, as word translation is not always possible, we plan to extract MWT unit translations as well as to pinpoint to terminological gaps, i.e. terms that do not exist in the target language and therefore need to be paraphrased. The extraction of MWTs and complex terms are indeed crucial when dealing with specialized domains.

3. Defining and combining different strategies for term alignment

TTC's objective concerning term alignment consists in improving methods for term alignment from comparable corpora, especially for specialized domains and MWTs. For this purpose, we intend to define, combine and evaluate lexical, contextual and corpora strategies. More specifically, we

[1] www.research.ibm.com/UIMA/

[2] www.eurotermbank.com

explore compositional methods (use of synonyms and variants, computation of interlingua representations), introduce depth in the context vectors used to express lexical contexts, and improve the adequacy of specialized comparable corpora with other textual resources such as parallel corpora or general language corpora (Morin et Daille, 2009; Daille, 2007) .

4. Developing an open platform for use with MT and CAT tools

The TTC consortium develops complementary tools in order to manage comparable corpora and generated terminologies based on UIMA framework and EuroTermBank. All developed and existing tools will be integrated into a demonstration platform offering all necessary Web Services. This platform will be easily integrable into various translation processing chains (including CAT translation, rule-based MT and statistical MT). French TTC partner LINA from the Université de Nantes is building a UIMA French-speaking community in the domain of Natural Language Processing. Some services have been set up and are currently available on the UIMA[3] portal. UIMA wrappers for term extraction and alignment tools as well as UIMA collection management tools will be developed during the project (Hernandez et al., 2010).

5. Demonstrating the operational benefits on MT tools and CAT tools

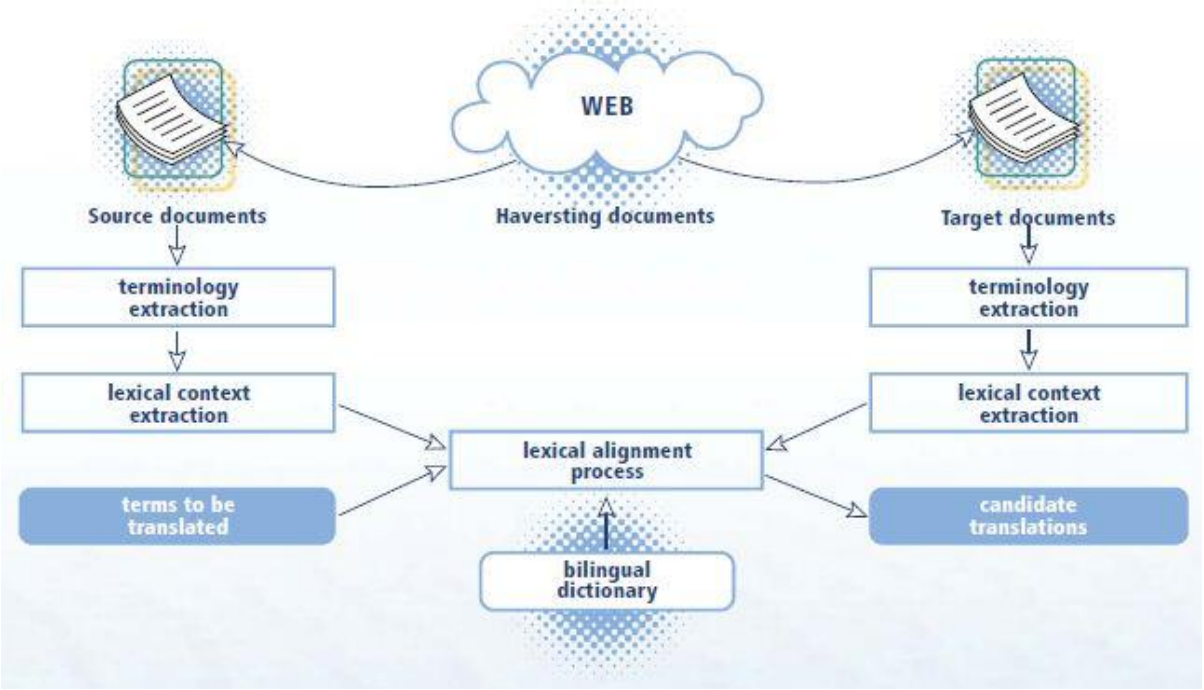
The main purpose of the TTC project is to propose solutions for translation in specialized domains and for languages with scarce linguistic resources. Besides, one of the objectives is that the proposed solutions can be rapidly used in an operational environment. As a consequence, the benefits of the developed solutions will be assessed regarding speed to generate the terminologies and quality of translation, and also regarding the amount of invested linguistic knowledge – so as to reduce the gaps in language coverage by requiring as less knowledge as possible.

The impact on statistical MT will be evaluated as well. The work outlined in this project will leverage statistical MT performance in two ways. First, it will reduce OOV (out-of-vocabulary words, i.e. SWTs and MWTs which do not appear at all in the parallel data on which the statistical MT system is trained). Second, it will automatically determine additional translation candidates for SWT and MWT which occur either infrequently in the parallel corpora or which occur with a sense that that is only correct in the domain of the parallel corpus and not in the domain of the translation to be performed.

Overall strategy

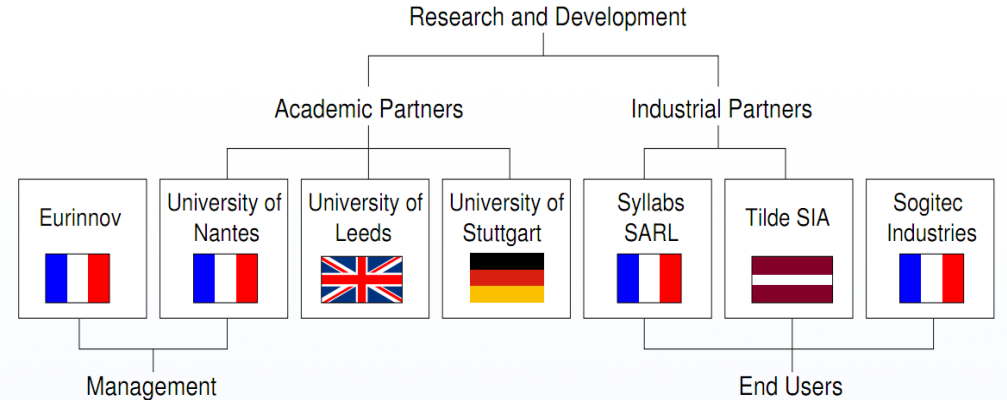
The TTC will develop a platform for automatically generating specialized monolingual and multilingual terminologies using comparable corpora and relying on existing open technologies (UIMA) and standards (TMF, TBX). The next figure illustrates the architecture of the expected multilingual terminology mining chain. Generated terminologies will be plugged into existing MT and CAT tools in order to improve their performance, especially for specialized domains and/or under-resourced languages.

[3] www.uima-fr.org



At the end of the TTC project, a platform will be set up to compile and manage comparable corpora using standards (TMF, TBX) and the existing open source UIMA framework. An evaluation and a validation of this work will be done by the consortium on CAT tools and Machine Translation tools. Translation of technical documents for aerospace and IT domain will be done using CAT and MT techniques to assess impact of the TTC project outputs.

Consortium



Summary of Activities

During the first 10 months of the project, we worked out on the following topics:

Requirements and definitions

The TTC specifications aim at:

- specifying the functional architecture with available tools and tools to be developed;
- setting up the guidelines for format definition in TTC;
- defining usage and technical scenarios taking into account the needs of real users identified in a survey run among localization professionals as well as the existing resources.

The outlined specifications might be modified and completed if necessary during the development of the tools and the setting up of the TTC platform.

1. Functional architecture

The functional architecture of the platform is driven by the term extraction and term alignment tools. Both of them will benefit from linguistic treatments: the POS tagging and the lemmatization linguistic analysis. POS tagging and lemmatization require that the word tokenization treatment. Moreover, the tokenization applies only on the textual content of documents. This dependency relation leads us to propose functional 4-steps: Text Pre-Processing, Linguistic Analysis, Term Extraction and Term Alignment.

2. Workshop with end users

Potential users of the TTC tools are translation agencies and companies which often have to produce translations for a specific domain, e.g. engineering, pharmaceuticals, energy, etc.

The users, e.g. translators and technical writers, can differ with respect to their needs concerning the extracted bilingual lexicon:

basic users: prefer to get a comparatively small amount of information, in a simple entry structure; mainly just equivalents;

advanced users: prefer to get all information relevant for the translation, including examples, metadata, etc.;

MT specialists: need specific system-adapted output.

Input to TTC tools

The extraction of a domain-specific terminology requires considerable amounts of domain-specific textual data. Such documents are available in companies, but data privacy could prohibit their usage

for term extraction. On the other hand, on the Web, there are a lot of documents, however, the terminology used in the Internet documents in the Internet need not always be reliable and trustworthy. Nevertheless, the TTC tool will provide a webcrawler which enables the user to create a domain-specific corpus based on the documents found on the Web.

Output of TTC tools

There are different requirements on the TTC output depending on the types of users of the tool. A basic user needs a lexicon with equivalences in source and target language. The tool should however not display more than five equivalent candidates (since it is confusing; or the users do not have time to read and inspect all possible translations). Additionally, the output should not contain more than one example of the context in which the target language phrase can be used. The displayed example has to be representative showing the most common usage of a term.

On the other hand, the advanced users are interested in more information about the terminologies:

Part-of-speech tags.

Term origin. Companies often use different terms for the same concept. Moreover, within companies, different terminology is often used in different projects. Thus, the terminology lexicon created out of documents provided by a company can lead to a false usage of the extracted terms. A label denoting the origin of a term could be helpful to choose the correct term.

Corpora can also be built by crawling the Web and gathering documents in a specific domain. Web crawling is an automatic process which also harvests inappropriate documents. Having information about the document from which a specific term is extracted, the user can inspect the original document, e.g. a web page, in order to estimate the adequacy of a term in a specific domain.

Confidence values. The term equivalents are computed automatically using statistical methods. Metadata which show the probability of the equivalents being translations of each other can be useful for the user in order to choose correct equivalents.

Term variants. Terms can vary with respect to their orthography, wording and morpho-syntactic structure. Different realisations of a term can be identified by the TTC tools allowing grouping of related terms. Such term groups provide the user with information about typical realisations of a term. The user can then choose the realisation which is most appropriate in a specific context.

There are additional types of information which are considered to be useful for the potential users of the TTC tool, but which cannot be provided by the tool.

Abbreviations. Abbreviations are very common in domain-specific documents, and, particularly, in customer feedback and daily work within companies. However, it is not foreseen for the TTC tool to extract a broad range of types of abbreviations.

Definitions. Definitions will not be included in the TTC output, but rather examples of the contexts in which a term is used.

Overall, the discussions with the user experts, as well as the presentations of their own experience with terminology extraction showed that TTC focuses on relevant tasks and on a real need. Given the widely diverging needs of occasional users vs. advanced users, it will be necessary to test the TTC tools with both these user groups. Most of the external participants to the workshop expressed their interest in being continuously informed about the work of TTC. They will be kept up to date by means of mailings at appropriate points in time.

Compilation of comparable corpora

The objective is to build a specialized language corpus from the web and to guarantee the comparability of the extracted texts:

1. Crawler

The crawler objective is to gather document from the web to automatically build corpus. We need to have specific corpus focused on one specific subject such as renewable energy or mobile technology for our project. The crawler is working in parallel and is doing smart filtering such language or webspam. The crawler is currently a naïve focused (thematic) crawler that start with a list of seed terms (given by user) that are processed by search engine queries that leads to URLs). The software in queue outlinks if page is in specialized thematic; this function is done by a webpage categorizer / thematic filter. The crawling process stops when there is no more relevant documents in queue or when the crawler reaches the user limits defined such as the number of documents or the maximum depth of research… Metadata such as date, author, are automatically created for each document.

To improve the quality of the documents returned by the crawler, a lexicon filter has been implemented. This expand lexicon from seed terms, weight the lexicon using the web and define a threshold by using the web. This allow to weight terms by representativeness, for example in astronomy, “moon” is not (or less) representative (because this term is used in other domain) than “quasar”. The next step is the definition of the most appropriate threshold for defining relevant document in relation with original seed terms and lexicon.

2. Comparability

The objective is to build a specialized-language comparable corpus. The first assumption we investigate is that monolingual comparability guarantees multilingual comparability. So that we need to identify monolingual features to automatically assign these features to texts. The monolingual criteria studied in language specific purpose are the medium (written, spoken, mixed-medium), the domain and sub-domain, the publication date, authorship (acknowledge subject-field expert), the subject, the type of document (research paper, report, instructional texts, advertisement, newspaper

article…), the type of discourse (texts written by experts to other experts, texts written by experts to non-experts), the language variety (English: UK, USA... or French: Fr, Ca, Be...), the production (native, non-native), the register (formal, informal), the authorship (single, corporate, co-authors, unknown…), the communicative intentions (information, instruction, propaganda/promotion, reporting, discussion (public, academic, communication), unknown…) and the audience (public, informed, professionals…). The Dublin core XML tags are used. Some tags are assigned automatically by the crawler: the Dublin Core features such as: creator, title, subject, contributor, date, description, publisher, type, format, coverage, rights, relation, source, language, identifier. Other features (document type, communicative intentions, register, author-audience, authorship) will be manually assigned to the corpus. OUSIA, a tool for manually annotating Dublin Core properties (elements + terms) Metadata in Dublin Core XML format and separated from their associated resource have been developed. It is possible to defined Dublin Core property values given by the means of a parameter file.

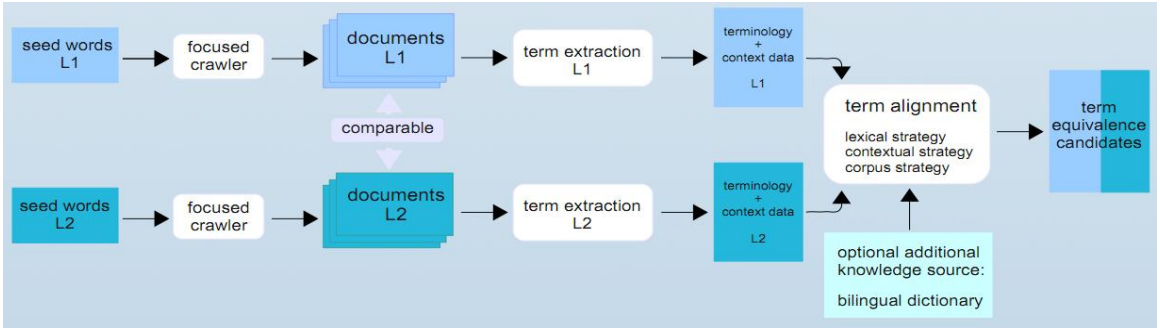
Term extraction

Basic idea about term identification is to define patterns of pos tags which are then used to identify terms in tagged and lemmatized text. The work focused focus on patterns based on nouns (adjective-noun, noun-noun, noun-preposition-noun, compounds...) including variants verb-object pairs are not computed as terms, but as phraseological examples.

Methods to improve the quality of extracted items currently under investigation are base on:

- statistical association measures;
- comparison with items extracted from non-domain-specific corpora (newspaper, text);
- ambiguities are difficult to find/resolve when extracting term candidates on monolingual data: multilingual approach might help to reduce the number of term candidates with incorrect pp-attachment.

The TTC project can be seen according to the following multilingual terminology chain:



Dissemination

Dissemination Strategy

Main goals of TTC dissemination are to make the project visible, raise awareness of the project results as swiftly as possible, reach and attract the following relevant target groups:

- **scientific community** (researchers in the area of (automated) translation, terminology, specialized communication and specialized language, intercultural communication and natural language processing),
- **end users** (freelance and staff translators, terminologists, technical writers, editors),
- **industry** (translation project managers in the localization and translation departments, translation industry managers).

To attain to successful dissemination and use of the project results, a fine-grained dissemination planning from the very beginning of the project and iterative monitoring of dissemination activities are performed. A multi-level dissemination structure is ensured at:

- **internal** (partners' organisations),
- **national** (managed by individual project partners to inform relevant audience in each specific country),
- **regional** (to encompass a broader audience in Eastern and Western Europe, Baltic countries, other regions inside EU),
- **global** (mostly spread by the Internet and other media, e.g. social networks, to reach audience on a global level) dissemination levels.

TTC visual identity

The visibility of TTC is ensured by a unique visual identity (logo) used for defined:

- **dissemination channels**
  - public website, partners' websites, mass media (including social and professional networks, e.g. Linkedin, Tweeter, Wikipedia), translation community portals (e.g. ProZ.com, Translators' Café, Terminómetro and others), conferences and workshops, joint events with relevant projects (e.g. joint workshops), scientific journals, TTC Advisory User group, relevant projects (both ongoing and supported beyond the FP7, e.g. ACCURAT, METRICC, PANACEA, CLARIN, T4ME), mailing lists (e.g. MT-list, Corpora-List and others), personal contacts
- **dissemination means**
  - news and event announcements, presentations (Power Point and oral), papers, posters, leaflets, press releases, contributions to discussions at meetings, face-to-face communication, public deliverables, questionnaires and surveys, demonstrations, e-mails

The basic TTC public dissemination channel and the project communication tool is the **project website** (www.ttc-project.eu). TTC website has static (general information about the project, key

objectives, consortium, etc.) and dynamic (news and events, releases, e.g. TTC public deliverables and scientific papers free to download) content.

Dissemination activities

1. Organized events

TTC project. The First TTC Workshop with End Users. October 6-7, 2010, Stuttgart, Germany.

FLaReNet, ACCURAT, PANACEA and TTC projects. Methods for the automatic acquisition of Language Resources and their evaluation methods: a joint workshop // LREC, May 23, 2010, Valletta, Malta.

2. Scientific papers accepted to conferences and **free to download**

Sharoff S. Analysing Similarities and Differences between Corpora // The 7th Conference "Language Technologies" (Jezikovne Tehnologije), October 14-15, 2010, the Institute "Jožef Stefan", Ljubljana, Slovenia.

Blancafort H., Daille B., Gornostay T., Heid U., Mechoulam C., Sharoff S. TTC: Terminology Extraction, Translation Tools and Comparable Corpora // The 14th EURALEX International Congress, July 6-10, 2010, Leeuwarden/Ljouwert, The Netherlands.

Hernandez N., Poulard F., Vernier M., Rocheteau J. Building a French-speaking community around UIMA, gathering research, education and industrial partners, mainly in Natural Language Processing and Speech Recognizing domains // New Challenges for NLP Frameworks Workshop, Language Resources and Evaluation Conference (LREC), May 22, 2010, Valletta, Malta.

Gornostay T. Terminology Management in Real Use // The 5th International Biannual Conference "Applied Linguistics in Research and Education" in Memoriam Rajmund Piotrowski (1922-2009), March 25-26, 2010, Saint-Petersburg, Russia.

3. Invited talks

Daille B. TTC: Terminology Extraction, Translation Tools and Comparable Corpora // Language Technology Days, March 22-23, 2010, Luxembourg.

4. Presentations

Sharoff S. Analysing Similarities and Differences between Corpora // The 7th Conference "Language Technologies" (Jezikovne Tehnologije), October 14-15, 2010, the Institute "Jožef Stefan", Ljubljana, Slovenia.

Daille B. TTC: Evaluation procedures of multilingual terminology acquired from comparable corpora // Methods for the automatic acquisition of Language Resources and their evaluation methods: a joint workshop, Language Resources and Evaluation Conference (LREC), May 23, 2010, Valletta, Malta.

Heid U. TTC: Terminology Extraction, Translation Tools and Comparable Corpora // Workshop on Requirements for Metadata, Language Resources and Evaluation Conference (LREC), May 2010, Valletta, Malta.

Gornostay T. Terminology Management in Real Use // The 5th International Biannual Conference "Applied Linguistics in Research and Education" (Прикладная лингвистика в науке и образовании) in Memoriam Rajmund Piotrowski (1922-2009), March 25-26, 2010, Saint-Petersburg, Russia.

5. Poster presentations

TTC: Terminology Extraction, Translation Tools and Comparable Corpora // The Open Day's Event for the 7th Framework Programme Research and Development in Latvia, November 16, the University of Latvia, Riga, Latvia.

Blancafort H., Daille B., Gornostay T., Heid U., Mechoulam C., Sharoff S. TTC: Terminology Extraction, Translation Tools and Comparable Corpora // The 14th EURALEX International Congress, July 6-10, 2010, Leeuwarden/Ljouwert, The Netherlands.

TTC: Terminology Extraction, Translation Tools and Comparable Corpora // The 14th Annual European Association for Machine Translation (EAMT) Conference, May 27-28, 2010, Saint-Raphael, France.

Hernandez N., Poulard F., Vernier M., Rocheteau J. Building a French-speaking community around UIMA, gathering research, education and industrial partners, mainly in Natural Language Processing

and Speech Recognizing domains // New Challenges for NLP Frameworks Workshop, Language Resources and Evaluation Conference (LREC), May 22, 2010, Valletta, Malta.

TTC: Terminology Extraction, Translation Tools and Comparable Corpora // EC Project Village, Language Resources and Evaluation Conference (LREC), May 17-23, 2010, Valletta, Malta.

6. Dissemination materials mentioning TTC

Heid U. Invited Talk on Ongoing Work in Stuttgart with a View to the Extraction of Specialized Terminology and Terminological variance // The 6th International Symposium on Lexicography was held on October 25-26 at the University of Maribor, Slovenia.

Sāmīte I. Defining Quality Levels for MT: Terminology Management and Quality in MT: presentation // The Localization Industry Standards Association (LISA) Forum Europe: Building Quality, Building Customers, October 11-14, Budapest, Hungary.

Gornostay T. Machine Translation System for Social Communication Support in a Multilingual Environment: scientific paper // The 13th International Conference "Russian Philology in Modern Times", October 7-9, 2010, Riga, Latvia.

Sāmīte I. Machine Translation for Small Languages and Under-Resourced Domains: presentation // Translation Automation User Society Conference, October 4-6, 2010, Portland, USA.

Vasiljevs A. Consolidation of Heterogeneous Terminology Resources: presentation // Common Language Resources and Their Applications (CLARA) Researcher Training Course Terminology and Resource Harmonization, September 13-17, 2010, Norwegian School of Economics and Business Administration, Bergen, Norway.

Vasiljevs A., Rirdance S., Gornostay T. Reaching the User: Targeted Delivery of Federated Content in Multilingual Term Bank: scientific paper // Terminology and Knowledge Engineering (TKE) Conference "Presenting Terminology and Knowledge Engineering Resources Online: Models and Challenges", August 12-13, 2010, Dublin City University, Ireland.

Sharoff S. Beyond Googleology: Assessing the Composition of the Web as a Large Corpus: invited talk // Web N-gram Workshop, Special Interest Group on Information Retrieval (SIGIR) Conference, July 19-23, 2010, UniMail, Geneva, Switzerland.

Gornostay T., Vasiljevs A., Rirdance S., Rozis R. Bridging the Gap – EuroTermBank Terminology Delivered to Users' Environment: scientific paper // The 14th Annual European Association for Machine Translation (EAMT) Conference, May 27-28, 2010, Saint-Raphael, France.

Vasiljevs A. European Industry and Academic Cooperation for Innovation and Leadership in Language Technologies: presentation // The International Forum "Baltic IT&T: eBaltics", April 21-22, Riga, Latvia.

Vasiljevs A. EU Language Technologies Research Projects in Latvia: presentation // Common Language Resources and Technology Infrastructure (CLARIN) seminar, February 26, 2010, the Institute of Mathematics and Computer Science of the University of Latvia, Riga, Latvia.

7. Other events

META-NET Media and Information Services Vision Group Meetings, September 10, 2010, Paris, France: TTC overview.

The 4th International Conference Baltic HLT "Human Language Technologies – The Baltic Perspective", October 7-8, 2010, Riga, Latvia: introductory leaflet, face-to-face communication.