TTC Annual Public Report 2011

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.

Summary

1 Introduction

1.1 Project objectives

In general, the TTC project aims at leveraging computer-assisted translation (CAT) tools, machine translation (MT) systems, and multilingual content management tools by generating bilingual terminologies automatically from comparable corpora in several European languages (including under-resourced languages), as well as in Chinese and Russian.

The need for linguistic resources (terminologies, lexicons, translation memories, etc.) is overwhelming in any natural language processing application, but 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 MT system, regardless of the theoretical foundations upon which a concrete system is based (e.g. statistical, rule-based, example-based or other MT strategy). CAT tools heavily use terminologies and translation memories to assist human translators in the translation process, e.g. to create and maintain terminologies. Another functionality is the generation of rough translations produced by a dictionary-based translation approach. Besides, automatic translation of terminologies from one controlled vocabulary into another is essential to the integration and use of diverse informatics systems: bilingual terminologies for several languages are adaptive and interoperable solutions for managing multilingual content and communication.

The TTC project aims at:

- using comparable corpora;
- using a minimum of linguistic knowledge for candidate term extraction;
- defining and combining different strategies for term alignment;
- developing an open platform for use with CAT tools and MT systems;
- demonstrating the operational benefits on CAT tools and MT systems.

The goal of this project is to improve and rise translation efficiency through technological means. Final outcomes of the TTC project aim at improving translation activities from industry documentation to multilingual content management.

1.2 Project description

The TTC project focuses on the automatic acquisition of aligned bilingual terminologies for CAT and MT. To do this, the important steps of the project are the automatic extraction of monolingual terminologies and the bilingual alignment of the extracted terminologies from a large set of multilingual corpora (see Figure 1).

Such terminologies could be extracted from parallel corpora, i.e. from previously translated texts, but these corpora are scarce and available only for some pairs of languages and few specific domains. Thus, no parallel corpora are available for most of specialized domains, especially for emerging domains (such as renewable energy). As a consequence, TTC develops methods and tools for automatic extraction of terminologies from comparable corpora, i.e. from corpora corresponding to a same domain, but not necessary being a translation from each other, as well as tools for gathering

and managing comparable corpora (topical web crawler) and managing terminologies (open terminology platform).

By 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 MT systems. Translation of technical documents for the aerospace and IT domains will be done using CAT tools and MT systems to assess the impact of the TTC project outputs.

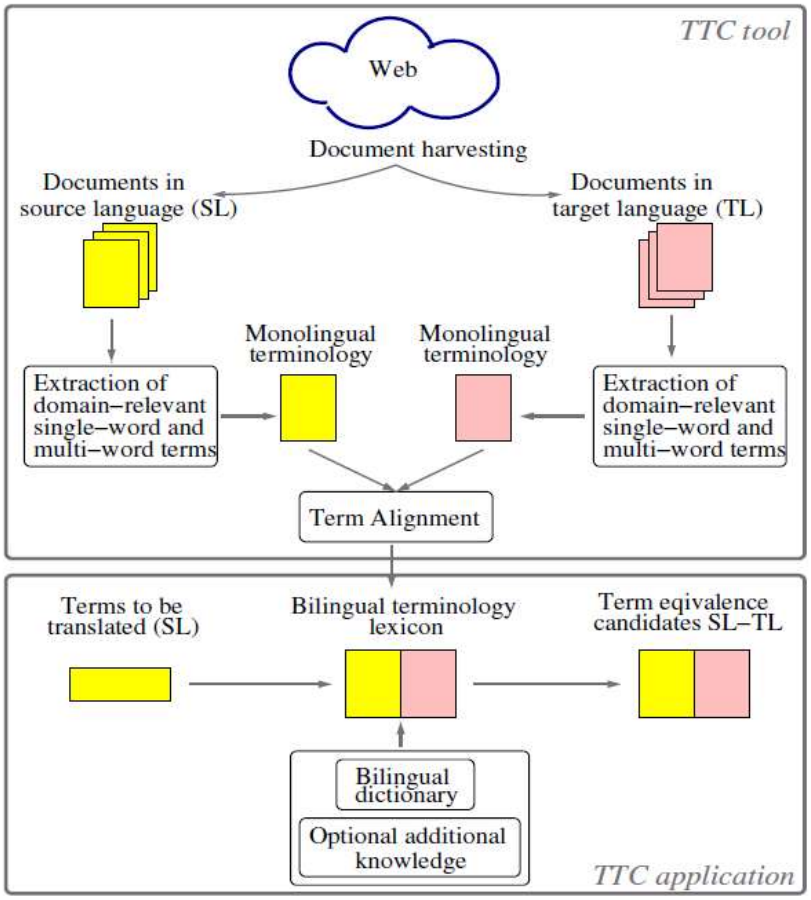

Figure 1. TTC Terminology Extraction Chain

1.3 Expected results and impacts

The TTC project develops generic methods and tools for automatic extraction of terminologies and alignment algorithms including adaptors to domains and languages, in order to break the lexical acquisition bottleneck in both statistical and rule-based MT. It will also develop or adapt tools for gathering and managing comparable corpora, collected from the web, and managing terminologies. In particular, a topical web crawler and open terminology platform will be developed.

The TTC platform will allow to create thematic corpora given some clues (such as terms or documents on a specific domain), extract monolingual terminology from such corpora, create a comparable corpus in the target language from a corpus in the source language, align bilingual terminologies, choose the tools to apply for terminology extraction, expand a given corpus, and export monolingual or bilingual terminologies in order to use them easily in automatic and semi-automatic translation tools.

Impact of translation tools and methods resulting from the TTC project will be evaluated in four domains of application: CAT tools, MT systems, multilingual content management and terminology management tools.

1.4 Consortium

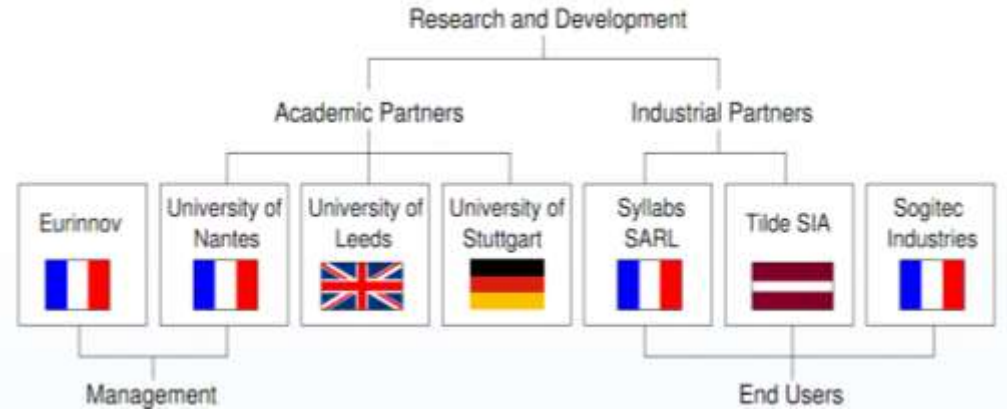

Figure 2. TTC Consortium

2 Summary of TTC activities during the period from November 15, 2010 till November 15, 2011

The TTC project can be seen according to the multilingual terminology extraction chain presented in Figure 3 below.

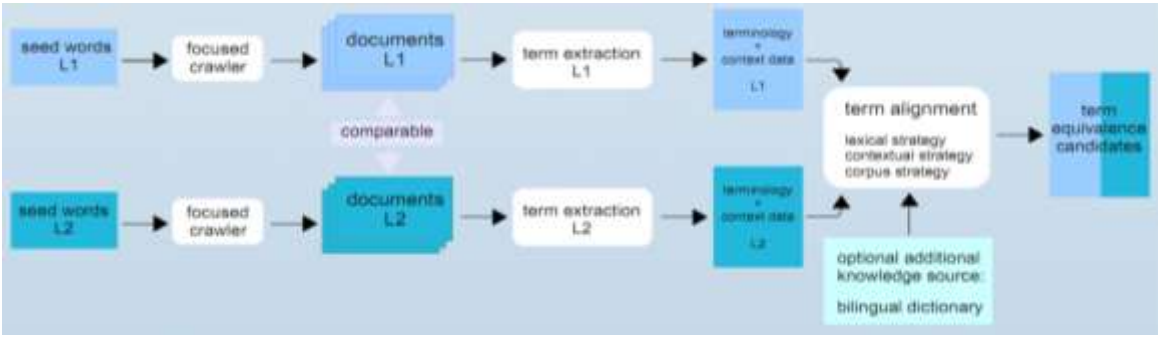

Figure 3. TTC Multilingual Terminology Extraction Chain

2.1 Corpora: principles, methods and development for the compilation of corpora

Working with corpora during the reporting period, we succeeded in conceiving and producing a state of the art web crawler that allowed TTC partners to gather and compile topical monolingual corpora. On the material of such corpora, the consortium researched monolingual and interlingual comparability using new concepts and focusing on automatic classification and categorization.

2.2 Single Word Term (SWT) and Multi Word Term (MWT) candidate extraction

TTC partners worked on identifying term candidates in comparable corpora including their significant context partners (e.g. collocations) in the individual texts of the languages handled in the project.

During 2011, a first set of tools for SWT and MWT candidate extraction was released based on the work done on the automatic identification and term variation, in particular. The remained work is related to the evaluation of the tools and collected corpora.

2.3 Term alignment

This work corresponds to the step after the candidate term extraction, and the goal is to improve existing methods for SWT and MWT alignment. The research is based on different strategies: lexical (interlingua representation, synonyms, variants, etc.), contextual (cognates, linguistic features for context vectors, etc.), and corpora (data mining, corpus extensions, etc.). We delivered a program able to detect neoclassical MWT for English / French / German (TTC public deliverable D4.1). We investigated the problem of MWT / SWT translations. In addition, we worked on strategies to improve comparability of the corpora so that the translation could be optimal.

2.4 Development of complementary tools and integration in the TTC platform

This task aims at developing tools to be integrated into the TTC online platform. The platform is already available in its preliminary version, functional online for beta testing by the partners. By August 2011, the platform had worked for monolingual terminology extraction for English, French, German, Russian, Spanish, and Chinese. The integration still continues.

2.5 Evaluation on impact on Computer Aided Translation (CAT) tools

The next task aims at investigating the impact of TTC tools on CAT which is a specific translation activity. During 2011, partners especially focused on producing the test corpora. Initial work was done about evaluation (methodology and planning).

2.6 Evaluation of impact on Machine Translation (MT) tools

This work focuses on automated and human evaluation of MT quality which can be achieved by enhancing MT dictionaries with automatically extracted terminology. We already produced the baseline MT output as well as human translations. We also collected parallel texts for TTC languages and aligned them on the sentence level.

2.7 Dissemination

TTC dissemination activities comprise participation in conferences with publications, oral / demo / poster presentations, and communication such as the project website and workshops. A summary of TTC dissemination activities performed during the second project reporting period are presented in the present report (see section 9).

2.8 Administrative, financial and legal management

The work done during November 15, 2010 – November 15, 2011 was essentially oriented toward the reporting period and review requirements. The integration of new partners was abandoned due to administrative and organization issues regarding Hildesheim. Other partners also approached the consortium, however, neither new contributions nor expertise were required to perform project tasks, thus it was decided to keep the consortium and the project as defined originally.

2.9 Scientific and technical management

The S&T management pursued the management of the distribution lists and the extranet platform for data sharing. The consortium meetings were organized and the minute was produced. In addition, the work performed was also to follow up the project progress and documentation delivery.

3 Corpora: principles, methods and development for the compilation of corpora

3.1 Monolingual comparability. Handling the problem of intertextuality

The objective of this task is to build specialized language corpora from the web in TTC languages. This is done in two steps: first, the texts are extracted from the web with a tool developed for the project: a crawler that guarantees domain comparability. Another tool classifies the documents according to discourse features in order to exploit only documents that are likely to contain terminology.

3.2 Interlingual comparability

We originally planned to:

- define metrics for intralingual and interlingual comparability for genres, domains and topics;
- measure intralingual distance between corpora and documents within corpora in the same language;
- define methods for distance measurement in feature spaces and machine learning;
- measure interlingual distance between corpora and documents in different languages;
- define methods for dictionaries creation and existing MT to map feature spaces between languages;
- validate the scale by independent annotation.

We analysed the solutions to represent the corpus graphically to investigate the possibility of interlingual comparability by using distance concept for terms.

3.3 Focused web crawling for corpus compilation

This task focuses on the development of a tool to gather domain-specific comparable corpora from the web. All crawled data will be available on the TTC web platform and access will be given to all partners to compile a corpus for the domains for which TTC aims to build terminologies: wind energy and mobile technology. By the end of the task the tool will be evaluated and the domain-specific corpora will be distributed to the general public. The corpora will be enriched with linguistic annotation (pos-tagging, lemmatization) performed with TTC Term suite and TreeTagger.

3.4 TTC public deliverable D2.2: Analysis of typologies to measure intertextuality. Evaluation on several monolingual corpora

For comparable corpora extracted from the Web using a crawler for terminology oriented applications, it is important to categorize the documents with regards to its terminology and itsnamed entities. Communicative intentions are interesting features as they allow to differentiate documents that contain specific domain terminology from documents that contains brand names, or from regulative documents that contain legal terms.

In order to classify documents according to their communicative intention, in this paper we run an experiments with language independent features that seem relevant to other categorization tasks such as web genre or discourse type. To classify documents written in seven languages belonging to five different families, we used language independent features that provide a rough classification of the document.

3.5 TTC public deliverable D2.3: Open-source tools for measuring the composition of a corpus within a language and across languages

In D2.3, we investigated the ways of ensuring that corpora collected from the web in different languages are comparable, and we implemented a software suite that enable the collected corpora to be checked and filtered for comparability before being used for subsequent tasks, such as multilingual term extraction. This entailed the devising and testing procedures for monolingual and bilingual comparability assessment and enhancement.

4 Single Word Term (SWT) and Multi Word Term (MWT) candidate extraction

4.1 Term identification and annotation. Single word terms, multi-word terms

This work deals with monolingual term candidate extraction from the text. It thereby provides data for each language, which are to be aligned into equivalent pairs in a subsequent step of the project using different new strategies under investigation.

TTC term extraction techniques rely on low-level annotated corpora (sentence boundaries, word classes and lemmas are annotated). We use patterns to extract term candidates: nouns, adjectives, and combinations of nouns and adjectives, noun sequences, etc. In a second step, frequency-based filtering against texts which are not domain-specific is used to identify those items of each of the patterns which are specific for the text (collection) analysed and thus can be considered to be term candidates.

During the reporting period, extraction patterns for all TTC languages, as well as data for the frequency-based filtering were collected, the tools for carrying out the above steps were implemented, integrated into the UIMA tool chain and made available. The techniques and the tools were delivered to the consortium for project purposes.

4.2 Term variation

Accurate terminology extraction from corpora needs to handle terminology variation. Terminology variation in texts is now a well-known phenomenon, whose amount is estimated from 15% to 35%, depending on the languages and on the characteristics of the textual documents such as the domain, genre, and discourse. The aim of this task is to define a multilingual typology of variations and to propose various methods to detect and classify them.

4.3 Morphological processing of term candidates

This task concerns all those components and aspects of term candidate extraction which involve the morphological processing of language data. This includes the work on inflection forms and word

formation, in particular for Germanic languages (in TTC: German). Furthermore, this part of the project includes the work on neoclassical word formation: scientific terminology contains many terms which are derived from Greek or Latin. These items have a similar form in most languages; tools need to be developed that provide a high-precision match of such items during term alignment.

The tool corresponding to this work is needed (i) to provide users of interactive terminology tools with standardized citation forms and correct data about the inflection of the terms, and (ii) to provide analyses that support a correct term alignment of Germanic compounds and neoclassical ones.

During the reporting period, the work on the finalization of the involved inflection guesser for inflection forms of terms and the specification of algorithms for the provision of citation forms was carried out. Furthermore, tools for the identification and alignment of neoclassical terms were developed, as well as a splitter for German compounds (based on (Koehn/Knight, 2003)) which takes into account word classes.

4.4 Set of tools for monolingual term candidate extraction: single and multi-word terms, and context properties, e.g. collocations

We conceived and produced a set of tools for extraction purposes of monolingual term candidates.

The TTC approach relies on a shallow pre-processing of the corpora from where terms are to be extracted: sentence boundaries, and the word class and lemmas of each word form found in the text are annotated. To provide this annotation, both knowledge-poor (mainly learning-based) and knowledge-rich strategies (relying on standard corpus linguistic tools and their lexical resources) are applied, in separate strands of the tool building work.

Term extraction relies on these annotations and consists of two subsequent steps, namely (i) the extraction of all text occurrences of a given pattern (e.g. nouns, noun + adjective groups, noun + preposition + noun groups) and (ii) filtering of the extracted items by comparison of their (relative) frequency in a specialized text with their (relative) frequency in a domain-independent text.

The assumption underlying this procedure is that items that are much more frequent in a specialized text than in a non-domain-specific text tend to be terms.

5 Term alignment

5.1 Lexical strategies (interlingua representation for compound terms, synonyms, variants)

We explored compositional methods for aligning term lists produced from comparable corpora using available dictionaries and term banks of SWTs and MWTs, as well as other lexical data collected during the term extraction phase such as variants and collocations (WP3). We investigated:

- the use of synonyms and variants to increase the coverage of the bilingual dictionary;
- the computation of an interlingua representation of an SWT or MWT to solve the problem of mapping some MWTs in one language to SWTs in another language, or to MWTs of different

morpho-syntactic structures. The work done about this strategy is strongly linked with that investigated for contextual and corpora strategies.

5.2 Contextual strategies: cognates, linguistic features for context vectors, identifying the best context

The purpose of this task is to propose new approaches to automatically extract bilingual lexicon from comparable corpora (i.e. documents that deals with a similar subject without being translations of each other). The traditional approach to compile bilingual lexicon from comparable corpora consists in implementing the assumption of the linguist Firth "You shall know a word by the company it keeps" (called the standard approach). We seek to propose new strategies for this task, either by improving the standard approach or by proposing new methods.

We worked on the standard approach improvements, starting from the intuition that each nearest lexical unit (nlu) contributes to the characterization of a lexical unit to be translated, our proposition aims at providing an algorithm that gives a better precision while ensuring higher stability with respect to the number of nlu.

We also investigated alignment of neoclassical terms and internationalisms and especially on a knowledge-poor method based on string distance (Levenshtein ratio), which is language-independent. We also developed a tool to align neoclassical terms (English, French, German): neoclassical terms are decomposed into their basic elements, which are translated separately.

5.3 Corpora Strategies

The aim of this task is to compile a comparable corpus using a parallel corpus as input. Parallel data are indeed difficult to find and constitute very small resource for specialized domains. Size limitations are a hurdle for the extraction and alignment of terminologies. Therefore, here we plan to compile comparable corpora as an extension of an existing parallel corpus. One of the keystones of this task is how to guarantee a satisfactory comparability of the extended corpus.

We tried this strategy to extend a parallel corpus in the wind energy domain. The goal was to bootstrap a large thematic comparable corpus from a small thematic parallel corpus. We extended crawler seed terms with extracted terms from the parallel corpora. These queries are then submitted to a general-purpose search-engine and top-N documents are retrieved (Figure 4).

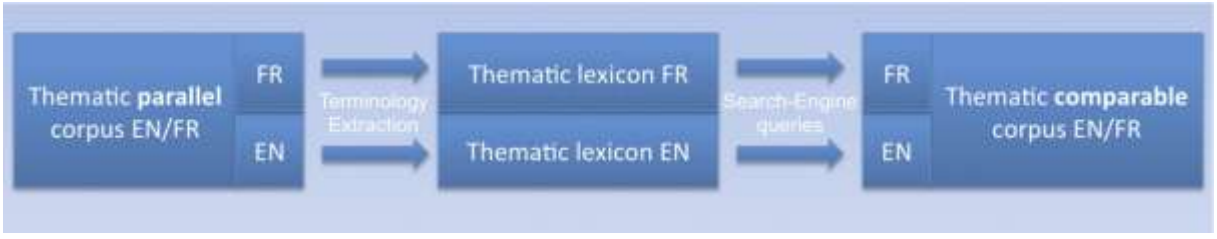

Figure 4. Comparable corpus bootstrapping experiment

The comparability measure is important for the corpora approach, it could be possible to combine corpora strategies to measure and evaluate the corpus comparability. A further strategy to evaluate the extended corpus will be to evaluate the resulting aligned terminology.

5.4 TTC public deliverable D4.1: Neo-classical MWT detection program for English / French / German

A neoclassical compound is a term that contains at least one Greek or Latin root, e.g. the term *cardiology* consists of the roots *cardio* and *logy*. A method has been developed in order to align neoclassical compounds between any pair of the languages: English, French and German (see Figure 5). The resulting alignments can help in improving the translation tools from one language to another. In order to do this, neoclassical compounds are automatically extracted for source and target languages from large texts in the two languages with the help of some predefined Greek and Latin roots. Then, translation candidates are automatically generated for each extracted neoclassical compound of the source language, this is done with the help of a bilingual dictionary and aligned list of Greek and Latin roots between the two languages (e.g. cardio=cardio, logy=logie,…etc). The generated translation candidate that is identical to a neoclassical compound extracted from the texts of the target language is considered to be the correct translation.

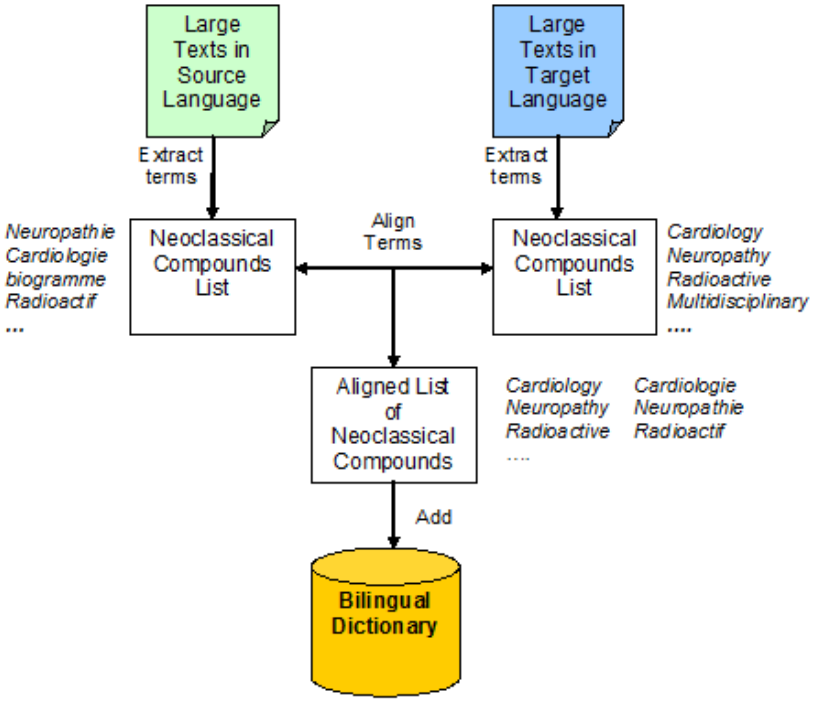

Figure 5. Neoclassical Compound Alignment in TTC

6 Development of complementary tools and integration in the TTC platform

6.1 Open-source tool to handle comparable corpora (i.e. collections) and multiple languages

This task is devoted to the design, development and documentation of a public domain application which aim is to extract terminology from bilingual comparable corpora. Terminology will be composed of monolingual candidate terms which come along with translation candidates. This

application has been released under the Apache License 2.0 and is composed of UIMA components. The work carried out consists in setting up the whole process that deals with monolingual terminology extraction for English, French, German, Russian, Spanish and Chinese.

6.2 TTC public deliverable D5.1: UIMA Type System specification for bilingual term extraction from comparable corpora[1]

This report details the UIMA-based TTC tool chain for terminology extraction for the seven languages involved in the TTC project. It presents the main concepts of UIMA that help to understand what kind of components are included in the TTC tool chain. Their behaviour is explained and illustrated through their own type systems. Such type systems express the kind of information these components require and produce. This was needed as the whole TTC type system is drawn out from its component ones which, themselves, are directly implied by their component behaviour.

6.3 TTC public deliverable D5.2: UIMA components to integrate existing partners' tools for term extraction over a given collection[2]

The TTC deliverable D5.2 completes this report as it aims at providing guidelines to embed new components into the TTC tool chain. Indeed, such integrations require this TTC type system for component interoperability.

The TTC Collection Processing Engine (CPE) for terminology extraction consists of the following components (behaviour, parameters and capabilities) detailed in this document:

- TTC Dublin Core Collector;
- TTC Text PreProcessing;
- TTC Preliminary Linguistic Analysis;
- TTC Terminology Extraction.

The report then explains how to improve existing components mostly by increasing the quality of their resources. Finally, the report presents how to extend such a CPE by components that process either document by document or the whole collection by the following means:

- improving existing mappings from TreeTagger to Multext;
- defining new mappings from TreeTagger to Multext;
- enriching the Analysis Engine called TreeTagger Multext Annotator.

6.4 Open terminology platform

The goal of this particular task is to develop an open terminology platform (OTP) for terminology storage and processing (import, search, editing, and export). The following activities were performed during the reporting period:

- finalisation of OTP user interface and interaction facilities;
- development of OTP architecture and implementation of main functionality in accordance with specifications;
- development of OTP demo with limited functionality (import, editing, export)[3];

[1] Published on the project website

[2] Published on the project website

[3] https://otp.eurotermbank.com/

- OTP internal (in-project) Beta testing initiated;
- partners' feedback analysis.

6.5 TTC platform development

The main challenge of the TTC project is to provide tools for terminology extraction and term alignment from comparable corpora crawled from the web for seven languages using both knowledge-rich and knowledge-poor tools that can be integrated into CAT tools and MT systems to enhance the translation of documents with rich terminology. The tools developed during the project will be assembled on the TTC platform to have a demonstrator showing their performance and validating the whole chain. Therefore, the platform will allow performing all the steps to extract terminologies from comparable corpora, which include the following tasks:

- crawling thematic comparable corpora;
- monolingual term candidate extraction and bilingual term alignment;
- processing the extracted terminologies via a tool for terminology management.

7 Evaluation of TTC impact on CAT tools

7.1 Preparation of test corpora of technical documentation

The work done here is a case study concerning the translation of technical documentation using CAT as well as terminology management tools. However, as a result of the SOGITEC work during this period, for the aeronautics domain, this work will address post edition associated to MT which has to be considered as a kind of CAT. TILDE is planning to work in the IT domain (mobile technologies) and evaluate TTC-generated term/translation candidate lists and TTC tools during the localization process.

This evaluation on CAT tools includes two tasks:

- comparable corpora selection and delivery, definition of the evaluation methodology;
- CAT implementation and testing, analysis of generated dictionaries, report on the evaluation of the efficiency of TTC for CAT tools (planned in 2012).

8 Evaluation of TTC impact on MT systems

8.1 Development of test corpora for evaluation of machine translation

This project part focuses on automated and human evaluation of MT quality which can be achieved by enhancing MT dictionaries with automatically extracted terminology. During the reporting period we produced the baseline MT output as well as human translations. The first work was dedicated to the production of the first collection of parallel corpora in the following domains:

- wind energy using translation and the skystream manuals;
- IT, using Open Office documentation (except for Latvian), Google help files;
- mobile technologies, using iPhone and HTC manuals.

We worked on alignment with Hunalign[4] (got clean texts), and Gargantua[5] (got noisy texts). The results revealed that accuracy thresholds were not always helpful, and human evaluation was crucial. The output of these initial tasks is aligned texts in TMX (Translation Memory eXchange) format.

8.2 TTC public deliverable D7.1: Sample texts with their translations by RBMT

The goal of this deliverable is to prepare a collection of corpora to be used for evaluation purposes. They cannot be used in the terminology extraction exercises, so that we could evaluate how an MT system performs with various parameter settings. The collection includes texts from three subject domains (wind energy, software and mobile interfaces) with translations for the language pairs of the project. The corpus is publicly available and will provide a basis for further research in MT. The corpus is unique in terms of the number of language pairs and domains covered. Other corpora commonly used in MT evaluation are based on either a limited number of language pairs or subject domains, like Europarl[6].

9 Dissemination

The main goal of TTC dissemination is to make the project visible and raise awareness of the project results as swiftly as possible, as well as to reach and attract the following relevant main target groups: **scientific community**, **end users**, and **industry**. To attain to successful dissemination and use of the project results, analytical recording and monitoring of TTC dissemination activities is undertaken on an iterative basis.

The following TTC dissemination activities were performed during the reporting period:

- the project website was updated regularly, including news and events (e.g. conferences, workshops, and consortium meetings), as well as public deliverables and TTC publications on the releases section;
- advanced web statistics of the project website was analysed using Webalizer which provides data about:
  - hits, files, pages, visits, etc. (e.g. data for October 2011: 16096, 15077, 6071, 2935 correspondingly);
  - monthly history (e.g. the highest during the reporting period was in May and June after the CHAT 2011 workshop);
- a joint workshop CHAT 2011 was organised at the NODALIDA 2011 conference:
  - TTC was one of the organisers of a workshop co-located with the 18th Nordic Conference of Computational Linguistics NODALIDA 2011. The Workshop on Creation, Harmonization and Application of Terminology resources CHAT 2011 was held on May 11, 2011 at the University of Latvia, in Riga, Latvia;
- the TTC Advisory Board was updated:
  - Jürgen Porsiel resigned,
  - Caroline Champsaur was invited,
  - Michael Wetzel moved from TRADOS to ESTEAM and would contribute to the Board on behalf of ESTEAM,

[4] http://mokk.bme.hu/resources/hunalign Varga et al. (2005)
[5] http://gargantua.sourceforge.net/ Braune and Fraser (2010)
[6] http://www.europarl.europa.eu/

- Bianka Buschbeck from SYSTRAN and Diego Bartolome from TA with you joined;

- 16 scientific papers/abstracts were submitted to conferences:
  - IJCNLP 2011: The 5th International Joint Conference on Natural Language Processing, November 8-13, 2011, Chiang Mai, Thailand.
  - TIA 2011: the 9th International Conference on Terminology and Artificial Intelligence, November 8-10, 2011, Paris, France.
  - GSCL: German Society for Computational Linguistics and Language Technology conference, September 28-30, 2011, University of Hamburg, Germany.
  - RANLP 2011: Recent Advances in Natural Language Processing conference, September 12-14, 2011, Hissar, Bulgaria.
  - IEEE/WIC/ACM International Conference on Web Intelligence, August 22-27, 2011, Campus Scientifique de la Doua in Lyon, France.
  - Corpus Linguistics Conference: Discourse and Corpus Linguistics, July 20-22, 2011, Birmingham, UK.
  - The 4th Workshop on Building and Using Comparable Corpora (BUCC 2011), June 24, 2011, Portland, Oregon, USA.
  - TALN 2011: Traitement Automatique des Langues Naturelles Conference, June 27 – July 1, 2011, Montpellier, France (PDF).
  - The 15th Annual Conference of the European Association for Machine Translation (EAMT 2011), May 30-31, 2011, Leuven, Belgium.
  - International Conference on Computational Linguistics and Artificial Intelligence Dialog 2011, May 25-29, 2011, Moscow region, Russia.
  - IC 2011: 22es Journées francophones d'Ingénierie des Connaissances, Chambéry, May 16-20, 2011, France.
  - CHAT 2011 Workshop on Creation, Harmonization and Application of Terminology resources, May 11, 2011, Riga, Latvia.
  - TRALOGY 2011: Translation Careers and Technologies: Convergence Points for the Future!, March 3-4, 2011, Paris, France.
  - The 33rd Annual Conference of the German Linguistic Society (DGfS-CL), February 23-25, 2011, Gottingen, Germany.
- 3 invited presentations were given at:
  - Pre-GSCL-Workshop (GSCL 2011 Conference), September 27, Hamburg, Germany.
  - CHAT 2011 Workshop on Creation, Harmonization and Application of Terminology resources, May 11, 2011, Riga, Latvia.
  - The Centre for Translation Studies of the University of Leeds (Boardroom, E.C. Stoner Building), March 17, 2011, Leeds, UK.
- 13 oral presentations were made:
  - RANLP 2011: Recent Advances in Natural Language Processing conference, September 12-14, 2011, Hissar, Bulgaria.
  - Corpus Linguistics Conference: Discourse and Corpus Linguistics, July 20-22, 2011, Birmingham, UK.
  - The 4th Workshop on Building and Using Comparable Corpora (BUCC 2011), June 24, 2011, Portland, Oregon, USA.
  - The 15th Annual Conference of the European Association for Machine Translation (EAMT 2011), May 30-31, 2011, Leuven, Belgium.
  - International Conference on Computational Linguistics and Artificial Intelligence Dialog 2011, May 25-29, 2011, Moscow region, Russia.
  - CHAT 2011 Workshop on Creation, Harmonization and Application of Terminology resources, May 11, 2011, Riga, Latvia.

- Panel Session: Rethinking Corpus-based Translation Studies in the Web Era of the Research Models in Translation Studies II Conference, April 29 – May 2, 2011, Manchester, UK.
- Software Summit 2011, April 13-15, Paris, France.
- A training workshop for translators from IAMLAPD, March 28, 2011, Centre for Translation Studies, University of Leeds.
- TRALOGY 2011: Translation Careers and Technologies: Convergence Points for the Future!, March 3-4, 2011, Paris, France.
- META-FORUM conference event "Challenges for Multilingual Europe", META-NET/META-SHARE meeting, November 17-18, 2010, Brussels.

▪ 6 demo and 8 poster presentations were made at:

- IJCNLP 2011: The 5th International Joint Conference on Natural Language Processing, November 8-13, 2011, Chiang Mai, Thailand.
- TIA 2011: the 9th International Conference on Terminology and Artificial Intelligence, November 8-10, 2011, Paris, France.
- GSCL: German Society for Computational Linguistics and Language Technology conference, September 28-30, 2011, University of Hamburg, Germany.
- IEEE/WIC/ACM International Conference on Web Intelligence, August 22-27, 2011, Campus Scientifique de la Doua in Lyon, France.
- TALN 2011: Traitement Automatique des Langues Naturelles Conference, June 27 – July 1, 2011, Montpellier, France.
- META-FORUM 2011 "Solutions for Multilingual Europe", June 27-28, Budapest, Hungary.
- The 15th Annual Conference of the European Association for Machine Translation (EAMT 2011), May 30-31, 2011, Leuven, Belgium.
- IC 2011: 22es Journées francophones d'Ingénierie des Connaissances, Chambéry, May 16-20, 2011, France.
- CHAT 2011 Workshop on Creation, Harmonization and Application of Terminology resources, May 11, 2011, Riga, Latvia.
- A training workshop for translators from IAMLAPD, March 28, 2011, Centre for Translation Studies, University of Leeds.
- The 33rd Annual Conference of the German Linguistic Society (DGfS-CL), Postersession 2011, February 23-25, 2011, Gottingen, Germany.

▪ 10 other dissemination materials mentioning TTC:

- A Case Study of Knowledge-Rich Context Extraction in Russian: scientific paper at TIA 2011: the 9th International Conference on Terminology and Artificial Intelligence, November 8-10, 2011, Paris, France.
- META-NET, related projects and industry research collaboration: oral presentation at EUROLAN 2011, August 28 – September 4, 2011, Cluj-Napoca, Romania.
- Terminology Turbocharges Your Translation! Oral presentation at Translation Forum Russia, September 23-25, 2011, Saint-Petersburg, Russia.
- TTC: Terminology Extraction, Translation Tools and Comparable Corpora: oral presentation at the lecture on language technologies for students of Baltic International Academy, September 17, 2011, Riga, Latvia.
- LOL –Langage objet dédié à la programmation linguistique: poster presentation at TALN 2011: Traitement Automatique des Langues Naturelles Conference, June 27 – July 1, 2011, Montpellier, France.
- TTC: Terminology Extraction, Translation Tools and Comparable Corpora: oral presentation at "Les Phenomenes de Figement Linguistique", June 9-10, 2011, Universite de Bourgogne.

- CHAT 2011 Workshop and Terminology Research: oral presentation to language workers (translators, editors, terminologist) at Tilde Localization department meeting, May 26, 2011.
- Extraction of Knowledge-Rich Contexts in Russian – A Study in the Automotive Domain: scientific paper at the 18th Nordic Conference of Computational Linguistics NODALIDA 2011, May 11-13, 2011 Riga, Latvia, p. 311-314.
- TTC presentation at Délégation générale à la langue française et aux langues de France (DGLFLF) (workshop on the impact of technologies on terminological resources), April 12, 2011, Ministry of Culture, Paris, France.
- TTC: Terminology Extraction, Translation Tools and Comparable Corpora: oral presentation CHAT 2011 Workshop and Terminology Research: oral presentation to the Translation Department students at the Baltic International Academy, December 17, 2010, Baltic International Academy, Riga, Latvia.
- Machine Translation in the System of Social Communication: scientific paper at the 4th International Conference "Modern Tasks of Linguistics, Language Teaching and Intercultural Communications" (APL-2010), December 3-4, 2010, Uljanovsk State Technical University, Uljanovsk, Russia.

- Term Suite demo video was published on YouTube[7,8];
- TTC midterm poster was produced and published on the project website[9];
- TTC LinkedIn group was regularly updated with new discussions:
  - currently it has 192 members (data on November 15, 2011).

[7] http://www.youtube.com/watch?v=v9jdsnEDw5c

[8] http://www.youtube.com/watch?v=Vi6yoXaFZ44

[9] http://www.ttc-project.eu/images/stories/TTC_poster_November_2011.pdf