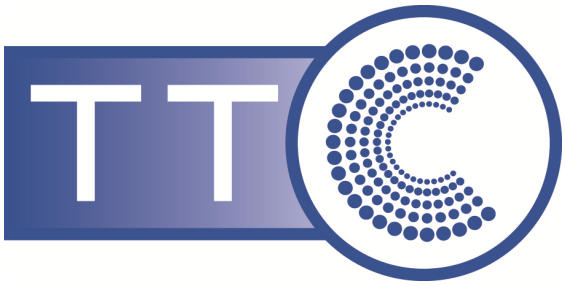
# TTC Annual Public Report 2012



*Terminology Extraction, Translation Tools and Comparable Corpora*

www.ttc-project.eu

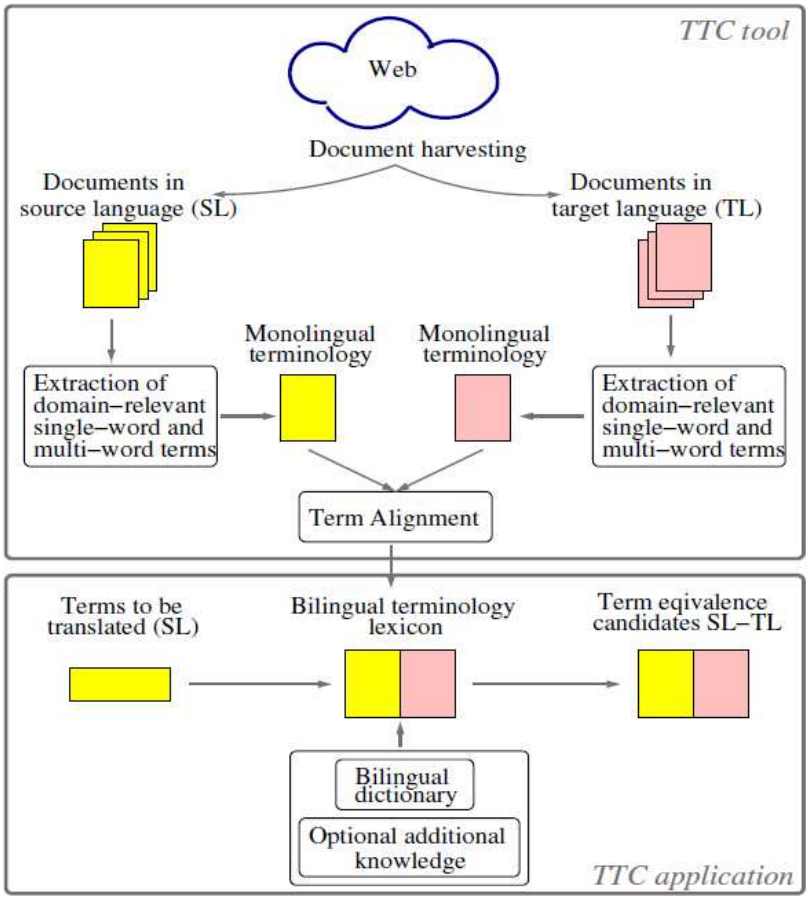
Project duration: 1st of January 2010 to 31st of December 2012 (36 months)

*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 Project objectives

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. In order to solve this issue, the TTC project first focuses on the automatic acquisition of aligned bilingual terminologies for computer-assisted translation (CAT) tools and machine translation (MT) systems. 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).




More generally, the goal of the project is to improve and raise the translation efficiency through technological means : the TTC project aims at leveraging CAT tools, MT systems and multilingual content management tools by generating bilingual terminologies automatically from comparable corpora in several European languages (including under-resourced languages, such as Latvian), as well as in Chinese and Russian. Final outcomes of the TTC project aim at improving translation activities from industry documentation to multilingual content management. This goal will be achieved by:

- using comparable corpora;
- using a minimum of linguistic knowledge for candidate term extraction;
- defining and combining different strategies for term alignment;
- developing an open web-based TTC platform for use with CAT tools and MT systems;

- demonstrating the operational benefits on CAT tools and MT systems.

## Consortium

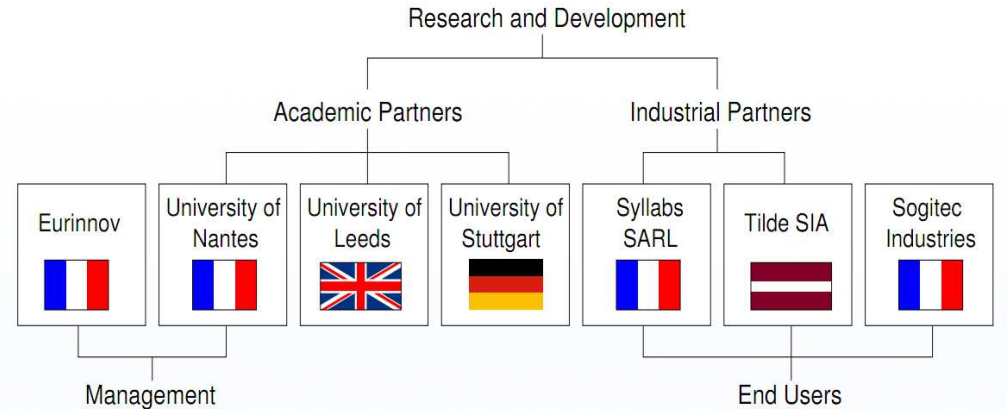



***Figure 1. TTC Consortium***

# 2 Expected results and targeted audience

The TTC project develops generic methods and tools for the automatic extraction and alignment of terminologies including adaptors to domains and languages, in order to break the lexical acquisition bottleneck in both statistical and rule-based MT. It also develops and adapts tools for gathering and managing comparable corpora, collected from the web, and managing terminologies. In particular, a topical web crawler and open terminology platform (MyEuroTermBank[1]) have been developed.

The TTC platform allows 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, translate 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.

The 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.

The target audience of the project are translators, terminologists, interpreters, CAT and MT tool developers.

[1] http://www.eurotermbank.com/

# 3 Summary of TTC activities during the period from November 15, 2011 till November 15, 2012

The TTC project can be seen according to the multilingual terminology extraction chain presented in Figure 2 below.

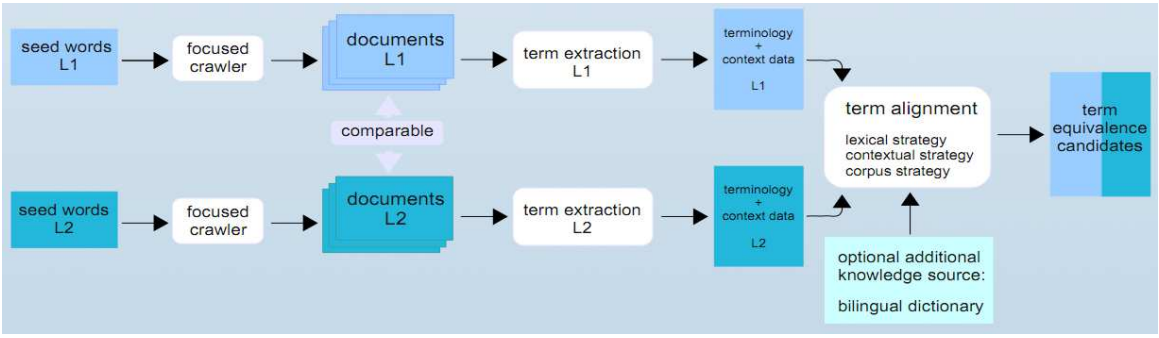



***Figure 2. TTC Multilingual Terminology Extraction Chain***

Specifications of the tools and corpora compilation methods have been addressed during the previous periods of the project. In this last year of the project, we had to accomplish the following tasks:

1. Conclude the development of our tools and platform demonstration;
2. Perform the evaluations to prove the performance and interest of our work for different targeted groups.

## 3.1 Single Word Term (SWT) and Multi Word Term (MWT) extraction

During the previous periods, TTC partners have worked on the identification of term candidates in comparable corpora including their variants in the individual texts of the languages handled in the project. The Year 2012 was dedicated to the evaluation of TTC tools for monolingual term extraction. Our experiments show that the various methods for monolingual term extraction implemented in the TTC platform are suitable for practical use. The results obtained for German were published and presented at the LREC 2012 conference. The general methodology of evaluation that makes use of reference term lists at the TKE 2012 conference.

## 3.2 Term alignment

This work corresponds to the step after the candidate term extraction, and the goal is to translate the monolingual terms using the comparable corpora. We developed the TeaBoat[2] tool, a fully statistical approach to alignment; we then evaluate different strategies to deal with SWTs and MWTs in term alignment. The main problem in evaluating the obtained results was that the comparison between languages and tools would require having comparable corpora of the same quality and the same dictionaries. However, these two requirements are probably impossible to meet in a study based on the web. Nevertheless, our work is probably the first that automatically builds bilingual terminology from specialized documents crawled from the web in several languages and new approaches proposed in this work give encouraging results, such as the neoclassical approach for neoclassical SWTs which was presented at CICLING 2012 conference, the ICA approach for SWTs which was presented at BUCC 2012 workshop or the Teaboat approach for mixed terms. Regarding

[2] http://www.ttc-project.eu/images/stories/Teaboat_Software.pdf

the alignment of MWTs, the compositional approach exploiting a bilingual dictionary gives goodresults in terms of precision, and a new method exploiting both the comparable corpora and the bilingual dictionary looks promising as it has been demonstrated during the user workshop.

## 3.3 Development of complementary tools and integration in the TTC platform

The goal of this part of the project was to develop three additional tools to manage corpora based on existing standard and normalization: TermSuite, MyEuroTermBank and the TTC platform – a web platform as a real size demonstrator of the TTC project results.
During the year 2012:

- even if TermSuite was functional at the end of the previous period, new improvements have been made in order to raise its performance both for the monolingual and the bilingual terminology extraction and with respect to ergonomics;
- Open terminology platform – MyEuroTermBank – has been developed and implemented as an integrated part of the EuroTermBank platform and is being integrated in the whole web-based TTC platform: it is a web-based environment for storing and managing terminological data and terminology collections for users and user groups. Its features are as follows:
    - Terminology import
    - Terminology storage
    - Terminology search
    - Terminology export
    - User management
    - Client authentication (basic authentication over SSL tunnel)
- The EuroTermBank API can be accessed using the SOAP protocol. MyEuroTermBank will interact with the open source tool for handling collections of comparable corpora – TermSuite[3] - and the TTC platform. The generated terminology collection(s) will be imported into MyEuroTermBank.
- 2 versions of the TTC Platform have been released. The second version permits to upgrade the platform in order to connect most of the TTC tools inside the TTC platform and to offer a usable framework.

Our work is almost completed but the interface of the TTC Platform will be improved in these last months of 2012.

## 3.4 Evaluation on impact of TTC Tools

After developing the tools, we worked on the best strategy to evaluate them in the most accurate way. The objective is to ensure an improvement of translation efficiency and thus to confirm the interest for our target audience: translators, terminologists, interpreters, CAT and MT tool developers.

We decide to conduct four evaluations of TTC Tools:

- An evaluation conducted on rule-based machine translation in the domain of wind energy by the University of Leeds and SYLLABS (ongoing).

[3] http://code.google.com/p/ttc-project/

- An evaluation of the impact of TTC in CAT in aerospace and IT domains performed by performed by industrial partners of the consortium (TILDE and Sogitec) (ongoing).

- An evaluation of the impact of TTC output on statistical MT conducted by the University of Stuttgart and two industrial partners TILDE and SYLLABS (ongoing).

- An evaluation through a user workshop held on October 11-12, 2012. The workshop was intended to comprise information and feedback gathering phase, as well as a discussion step. After the presentation of the TTC Tools, the goal was to give each participant a hands-on experience with the tools. Ca. 50% of the attendees were from companies and administrations, another half from academia. Participants from outside academia were from the following domains of activity: translation services (in industrial companies, in large international administrations, as well as SMEs offering translation services), deployers of language tools, developers of terminology tools and developers of MT.

While waiting for the output of the three other project evaluations, conclusions of the user workshop were very positive: users are positively impressed by the different tools, and the notion of an online terminology service is very well received. Future work on the TTC tools could include customisation and further work on alignment in order to cover all equivalence types. Expressions of interest for further cooperation came already from tool developers and deployment companies (ESTeam, Lingua & Machina and ABBYY), and a massive interest in following TTC's work on machine translation was also voiced (Systran, and Lingenio).

## 3.5 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;

- a joint workshop CHAT 2012 was organised at the conference TKE 2012, TTC was one of the co-organisers;

- the second TTC workshop with experts and end users;

- web statistics Webalizer (hits, files, pages, visits) was used for the analysis of the project website and it indicated the growing popularity of the project website according to November 2011-November 2012 data;

- the TTC Advisory Board was updated:
    - Diego Bartolomé, an expert in SMT;
    - Bianca Buschbeck, an expert in MT from SYSTRAN;
    - Kurt Eberle from Lingenio, Heidelberg, an expert in RBMT and language resources;

  - Michael Wetzel, ESTeam AB, Berlin, an expert in CAT.

- 17 scientific papers/abstracts were submitted and accepted to the following events:
  - HLT 2012: The 5th International Conference Human Language Technologies "The Baltic Perspective", October 4–5, 2012, Tartu, Estonia.
  - EURALEX 2012: The 15th International Congress, August 7-10, 2012, Oslo, Norway.
  - IVACS 2012: The 6th Inter-Varietal Applied Corpus Studies group International Conference on Corpora across Linguistics, June 21-22, 2012, Leeds, UK.
  - 2 - TALN 2012: Traitement Automatique des Langues Naturelles Conference, June 4-8, 2012, Grenoble, France.
  - IVACS 2012: The 6th Inter-Varietal Applied Corpus Studies group International Conference on Corpora across Linguistics, June 21-22, 2012, Leeds, UK.
  - TKE 2012: Terminology and Knowledge Engineering, June 19-22, 2012, Madrid, Spain.
  - EAMT 2012: The 16th Annual Conference of the European Association for Machine Translation", May 28-30, 2012, Trento, Italy.
  - CREDISLAS 2012: Workshop on Creating Cross-language Resources for Disconnected Languages and Styles co-located with LREC 2012, May 27, 2012, Istanbul, Turkey.
  - 2 - BUCC 2012: The 5th Workshop on Building and Using Comparable Corpora with special topic "Language Resources for Machine Translation in Less-Resourced Languages and Domains", May 26, 2012, Istanbul, Turkey.
  - 3 - LREC 2012: The 8th International Conference on Language Resources and Evaluation, May 23-25, 2012, Istanbul, Turkey.
  - EACL 2012: The European Chapter of the Association for Computational Linguistics, April 23-27, 2012, Avignon, France.
  - CICLing 2012: The 13th International Conference on Intelligent Text Processing and Computational Linguistics, March 12, 2012, New Delhi, India.
  - DGfS's Poster Session, March 8, 2012, Frankfurt, Germany.

- 4 invited talks were given at the following events:
  - Polytechnic Museum, November 6, 2012, Moscow, Russia.
  - EURAC, July 28-29, 2012, Bolzano, Italy.
  - CREDISLAS 2012: LREC 2012 Workshop on Creating Cross-Language Resources for Disconnected Languages and Styles, May 27, Istanbul, Turkey.
  - BUCC 2012: The 5th Workshop on Building and Using Comparable Corpora with special topic "Language Resources for Machine Translation in Less-Resourced Languages and Domains", May 26, 2012, Istanbul, Turkey.

- 13 oral presentations were made at the following events:
  - BAAL 2012: British Association for Applied Linguistics conference, September 6-8, 2012, Southampton, UK.
  - EURALEX 2012: The 15th International Congress, August 7-10, 2012, Oslo, Norway.
  - IVACS 2012: The 6th Inter-Varietal Applied Corpus Studies group International Conference on Corpora across Linguistics, June 21-22, 2012, Leeds, UK.
  - TKE 2012: Terminology and Knowledge Engineering, June 19-22, 2012, Madrid, Spain.
  - CHAT 2012: The Second Workshop on the Creation, Harmonization and Application of Terminology Resources, June 22, 2012, Madrid, Spain.

  - 2 - TALN 2012: Traitement Automatique des Langues Naturelles Conference, June 4-8, 2012, Grenoble, France.
  - BUCC 2012: The 5th Workshop on Building and Using Comparable Corpora with special topic "Language Resources for Machine Translation in Less-Resourced Languages and Domains", May 26, 2012, Istanbul, Turkey.
  - 3 - LREC 2012: The 8th International Conference on Language Resources and Evaluation, May 23-25, 2012, Istanbul, Turkey.
  - EACL 2012: The European Chapter of the Association for Computational Linguistics, April 23-27, 2012, Avignon, France.
  - CICLing 2012: The 13th International Conference on Intelligent Text Processing and Computational Linguistics, March 12, 2012, New Delhi, India.

- 2 demo presentation was made at the following events:
  - CHAT 2012: The Second Workshop on the Creation, Harmonization and Application of Terminology Resources, June 22, 2012, Madrid, Spain.
  - CICLing Poster Session, March 12, 2012, New Delhi, India.

- 9 poster presentations were made at the following events:
  - EAFT 2012 Terminology Summit, October 11-12, 2012, Oslo, Norway.
  - HLT 2012: 5th International Conference Human Language Technologies "The Baltic Perspective", October 4–5, 2012, Tartu, Estonia.
  - CHAT 2012: The Second Workshop on the Creation, Harmonization and Application of Terminology Resources, June 22, 2012, Madrid, Spain.
  - META-FORUM 2012: A Strategy for Multilingual Europe, June 20-21, Brussels, Belgium.
  - EAMT 2012: The 16th Annual Conference of the European Association for Machine Translation", May 28-30, 2012, Trento, Italy.
  - LREC 2012: The 8th International Conference on Language Resources and Evaluation, May 23-25, 2012, Istanbul, Turkey.
  - CICLing 2012: The 13th International Conference on Intelligent Text Processing and Computational Linguistics, March 12, 2012, New Delhi, India.
  - DGfS's Poster Session, March 8, 2012, Frankfurt, Germany.
  - The Open Day's event for FP7 R&D in Latvia, November 21, 2011, Riga, Latvia.

- 2 other dissemination materials mentioned TTC at the following events:
  - MultilingualWeb workshop, June 11-13, Dublin, Ireland.
  - International conference "Crosslingual Language Technology in service of an integrated multilingual Europe - 20 years on", May 4-5, 2012, Hamburg, Germany.

- TTC submitted papers/abstracts to the following 2012 events:
  - DGfS's Poster Session 2013, March 114, 2013, Potsdam, Germany.
  - TRALOGY II 2013 conference, January 17-8, Paris, France.

- 409 members joined the TTC LinkedIn group since the project start (cf. 192 members in November 2011 and 336 in June 2012);
- Twitter and Facebook were used for TTC dissemination;
- TTC poster and leaflet were updated during the reporting period;
- D8.4 and D8.5 deliverables were prepared, reviewed and delivered to the project coordinator;

- TTC collaborated with other EU projects: ACCURAT, LetsMT!, PANACEA, PRESEMT, META-NORD, TaaS;
- TTC resources are planned to be shared via Tilde META-SHARE node.