# User-centred Views on Terminology Extraction Tools: Usage Scenarios and Integration into MT and CAT Tools

Helena Blancafort, Ulrich Heid, Tatiana Gornostay, Claude Méchoulam, Béatrice Daille, Serge Sharoff

# User-centred Views on Terminology Extraction Tools: Usage Scenarios and Integration into MT and CAT Tools

**Helena Blancafort[(1)], Ulrich Heid[(2)], Tatiana Gornostay[(3)], Claude Méchoulam[(4)], Béatrice Daille[(5)], Serge Sharoff[(6)]**

[(1)] Syllabs, [(2)] IMS - Universität Stuttgart, [(3)] Tilde, [(4)] Sogitec, [(5)] LINA - Université de Nantes, [(6)] CTS- University of Leeds

blancafort@syllabs.com; heid@ims.uni-stuttgart.de; tatiana.gornostay@tilde.lv; beatrice.daille@univ-nantes.fr; cmechoulam@sogitec.fr, s.sharoff@leeds.ac.uk

This paper presents usage scenarios of the platform being developed within the TTC project (*Terminology Extraction, Translation Tools and Comparable Corpora)* along with the first feedback from potential users. The TTC project aims at leveraging translation tools, computer-assisted translation tools, and terminology management tools by automatically generating bilingual terminologies from comparable corpora in several languages of the European Union (English, French, German, Latvian and Spanish), as well as in Chinese and Russian. The TTC platform includes a web crawler and a corpora management tool, as well as tools for monolingual term extraction and bilingual terminology alignment, online terminology management, and terminology export into CAT tools and MT systems.

Overall, the paper focuses on the language activities to be carried out with the TTC tools, issues with respect to the availability of required language resources and linguistic knowledge, and different user profiles and needs. Regarding potential user needs, we discuss the results of an online questionnaire-based survey on terminology and corpora issues conducted in the translation and localization industry to reveal user needs. Furthermore, we present the envisaged usage scenarios as well as first feedback from potential users. The expected TTC input and outputs are also outlined. Finally, as it seems clear that the amount of available data and resources will not be the same for all languages, we discuss technical solutions to achieve language coverage: the TTC tools will offer different approaches depending on the amount and type of linguistic knowledge available.

## 1 Introduction

Computational applications in translation and technical documentation suffer from the terminology bottleneck. This applies to both CAT (computer-assisted translation) and MT (machine translation): there is a lack of bilingual term-related resources, especially for new or rapidly developing domains of knowledge for which no parallel corpora exist. The TTC project *Terminology Extraction, Translation Tools and Comparable Corpora*[1] explores term extraction from comparable corpora, i.e. from texts of the same domain (and possibly genre) in different languages which need not be translations of each other. TTC develops techniques for the extraction of monolingual term candidates and their contexts for Chinese, English, French, German, Latvian, Russian and Spanish. In a second step, monolingual data are aligned to identify equivalence candidates: TTC explores different

[1] www.ttc-project.eu. 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.

symbolic and statistical procedures for this purpose. The outputs of the TTC tool chain are single word terms, multiword terms and term variants, as well as contextual data.

The TTC tools will be provided as a standalone package which can be downloaded and run on a computer, as well as a web service. They will include tools for corpus crawling and corpus management, for monolingual term candidate extraction and for term alignment. An integration with an open terminology platform based on EuroTermBank[2], and selected CAT and MT tools will be provided.

In this paper, we present possible usage scenarios for the TTC tools, with a focus on the following:

- language activities to be carried out with the help of the TTC tools;
- issues with respect to the availability of language resources;
- different user profiles and needs.

We also discuss resource-related parameters of the usage scenarios. We address issues such as the quality and trustworthiness of resources, their relevance to a specific domain, as well as legal issues concerning the use of data on a web service.

Furthermore, the focus is put on the input and output of the TTC tools. As the output format of terminologies, we plan to use TBX-based (Term Base eXchange) format, an adapted version of the open XML-based standard for exchanging structured terminological data that has been approved as an international standard by ISO.

To better identify user needs in the translation and localization industry, we first conducted a survey among language professionals and then organized a workshop with experts in the domain to discuss the defined practical usage and technical scenarios and the expected input and output of the tools. We present the results of the survey as well as those of the workshop.

The structure of this paper is as follows: firstly, we provide an overview of the project (section 2). Secondly, we sum up the results obtained from the TTC survey on terminology and corpora issues in the industry (section 3). Then, we present the envisaged usage scenarios as well as first feedback from potential users (section 4). Next, expected TTC input and outputs are outlined (section 5). Finally, we discuss technical solutions to achieve language coverage: the TTC tools in fact will offer different approaches depending on the amount and type of linguistic knowledge available (section 6). We conclude in section 7.

## 2 Project Overview and Functionality of the TTC tools

The TTC project aims at leveraging translation tools, computer-assisted translation tools, and terminology management tools by automatically generating bilingual terminologies from comparable corpora in several languages of the European Union (English, French, German, Latvian and Spanish), as well as in Chinese and Russian. This three-year project started in January 2010. The main terminology extraction components of the TTC tools are monolingual term candidate search and term alignment, i.e. the identification of bilingual equivalent candidates. The first step identifies single word terms (SWTs), compounds and multi-word terms (MWTs). It relies on generic (e.g. statistical) and language-specific processes which are based on different amounts of linguistic knowledge of each of the languages covered. The second step, i.e. term alignment, uses statistical and generic linguistic devices to group items from the first step into equivalent pairs.

[2] www.eurotermbank.eu

Alongside this term identification function, the TTC tool chain serves a second main purpose, which in typical usage scenarios will in fact precede term identification, namely harvesting texts from where terminology can be extracted. This component is a focused web crawler which starts from a small set of seed words of a domain and downloads texts from the Internet which are relevant to the domain. Users may use both components (first harvesting, then term identification) whenever no appropriate own texts are available, or they may provide their own texts as an input to term identification. The first case may be typical of a situation where a translator has to work on a technical sub-domain which is not known to him or her in detail, or on an emerging domain, for which no or only few terminological resources are available, and where a fully-fledged and commonly accepted terminology has not yet been created. In the project work, we take wind energy as an example of such a domain. TTC deals also with the aerospace and the IT domain.

A second dimension where the TTC tools will be useful concerns languages for which only few computational linguistic resources are available (in particular lexicons and processing components). For such languages, TTC prepares statistical processing components which support the bootstrapping of the linguistic knowledge needed for term candidate extraction.

## 3 User Needs Survey

An online questionnaire-based survey[3] about terminology and corpus practices among language professionals was conducted to identify needs in the translation and localization industry. A questionnaire was designed to know more about the practical use of terminology management tools as well as about the use of computer-aided translation and machine translation tools. 139 professionals participated in the survey. The online survey consisted of more than 40 questions about practices in the translation and localization industry.

The call for participants was published via mailing lists (e.g. LinguistList), thematic groups on professional networks on LinkedIn (e.g. Trad Online), professional online forums, Tilde's localization departments and personal contacts of professional translators. Respondents came from 31 countries and consisted of in-house translators, freelance translators, terminologists as well as language and translation teachers. The participants cover a wide range of topics and text types, but most of them produce translations in a specialized domain (technical, software and legal translation).

The main objectives of the survey were to summarize trends in current translation and localization projects, to understand current terminology management practices, to reveal user needs, and to obtain a view on user expectations. Furthermore, the survey included questions about the use of MT tools and other Natural Language Processing (NLP) applications, such as corpora concordancers. Previous user need surveys concentrated more on terminology management and CAT tools. A survey[4] of the Localization Industry Standards Association (LISA)[5] conducted in 2004, focused on trends in terminology management within the localization industry. The results showed that the majority of companies within the localization industry performed terminology management systematically (75%). However, the level of sophistication varied widely and many companies used spreadsheets as a primary tool for terminology collection, storage,

[3] "Calling Professionals: Help Us to Understand Your Needs!", Syllabs, Tilde, 2010, http://www.visidati.lv/aptauja/345124026.

[4] "Terminology Management Practices and Trends: LISA Terminology Management Survey", 2005, LISA, http://www.lisa.org/index.php?eID=tx_nawsecuredl&u=8762&file=fileadmin/filestore/terminology_report_2005.pdf&t=1269012084&hash=d1be2a4b4daf6ad053270425fbdb2e47.

[5] www.lisa.org

exchange, etc. (35%). The main reason why respondents did not have terminology tools was that they were dissatisfied with the functionality of the then available tools and did not have enough information about the latest tool developments.

On the other hand, SDL[6] conducted two surveys[7] in 2008 to explore the trends in terminology management within businesses and the translation/localization industry. The results of these surveys showed that only 29% of business respondents already had a terminology management solution and the major internal processes for managing and sharing terminology were the following: publishing terminology in style guides (36%), using terminology lists in Microsoft Excel (33%) and using a specific terminology management tool (28%). Within the translation and localization industry, the most common methods used by translators to manage their terminology were Microsoft Excel (42%) and specific terminology management tools (31%). Another crucial issue was how terminology extraction was carried out: 84% of respondents selected terms from documents manually. Thus, the survey showed that terminology management was important for both user groups, however, the two groups were sometimes using different means of managing terminology and there was an obvious need for a uniform tool between business and translators.

The TTC survey showed that the situation in terminology management has not changed greatly since 2004 (see the LISA survey), even if terminology plays an increasingly important role in the translation process: 56% of the respondents spend 10 to 30% of their time working with terminology. Bilingual terminology research and collection is the main terminology-related activity. However, spreadsheets still remain the most often used means of storing and exchanging terminology, even though their use obviously limits the possibilities of representing and illustrating terminological data quite considerably. Budget and/or time constraints, as well as deficiencies in the functionality of existing tools, are still the main obstacles for using terminology management tools, and terminological consistency and productivity are still high priorities. Among standard formats, TBX is most used (10%). This is an important point which led the TTC project to decide that a TTC output will comply, among others, with this standard (see more in 5.2). Good news of the survey is that still 66% of the respondents are interested in new solutions and that 40% would agree to share their terminology within the online database. This is encouraging for open terminology projects.

The main task concerning terminology search is lexical work: look up of translation equivalents, provision of definitions in source language, etc. A terminology database should include information about definitions, indicate contexts, and offer information about usage (style, frequency). Lookup speed is an important issue. A terminological database should include refined filter techniques for looking up a term, e.g. by domain, category, or area. This information has been of great help to define the specifications of the open terminology tool developed under TTC that aims at better responding to user needs and at offering a user-friendly tool with rich functionalities regarding terminology look-up.

Moreover, the survey showed that 50% of the respondents collect corpora of the relevant domain, but that only 7% use automatic processing. Regarding the use of these available or collected corpora, as well as the use of concordancer and other NLP tools to manipulate them, the common strategy with TTC respondents is to manipulate corpora manually: quick reading, highlighting of terms and manual search for an equivalent is the

---

[6] www.sdl.com

[7] "Terminology: An End-to-End Perspective: SDL Research Paper", SDL, 2008, http://www.sdl.com/en/globalization-knowledge-centre/research_results/terminology-an-end-to-end-perspective.asp.

most current strategy when working with a corpus to learn about terminology (cf. Pearson 1998). Only 30% use corpus concordance tools and only 10% use NLP tools (some do not know of their existence). A precious output of the survey for the specifications of the tool to be developed to handle corpora is the wish list of functionalities that the user expects from a such a tool: rich metadata search, concordancing to show the context of a term, automatic updating of the database and automatic categorization according to defined domains, automatic saving of links, generation of frequency lists, and annotation functions. Users expect the tool to be collaborative and using a standard: TBX format with one sentence per line.

Concerning the use of CAT tools and MT systems, 74% of the respondents are using CAT tools, Trados being the most used translation software followed by Similis. Systran (software used in TTC for evaluation purposes) is the most used commercial MT software after the Language Weaver, while Google Translate is the most used free online MT software. A more detailed analysis of the TTC survey is included in (Blancafort and Gornostay, 2010), Gornostay (2010), and Vasiljevs et al. (2010). The conclusions of this survey have been taken into account when preparing the specifications of TTC tools.

## 4 Usage Scenarios: An Overview

In this section we present the usage scenarios for the TTC tools from the viewpoint of the user. To better illustrate the envisaged scenarios, the TTC tool user scenario is schematized in figure 1, below. This schema is subdivided in the automatic (upper) part and the interactive (lower) part, as the automatic TTC tools are meant to provide the lexical and/or terminological knowledge needed by the user to carry out an interactive translation task. We intend the TTC tool output (i.e. term candidates with their equivalents) as an input to both (i) interactive translation, e.g. with CAT tools, and (ii) automatic translation, with rule-based or statistical MT systems.

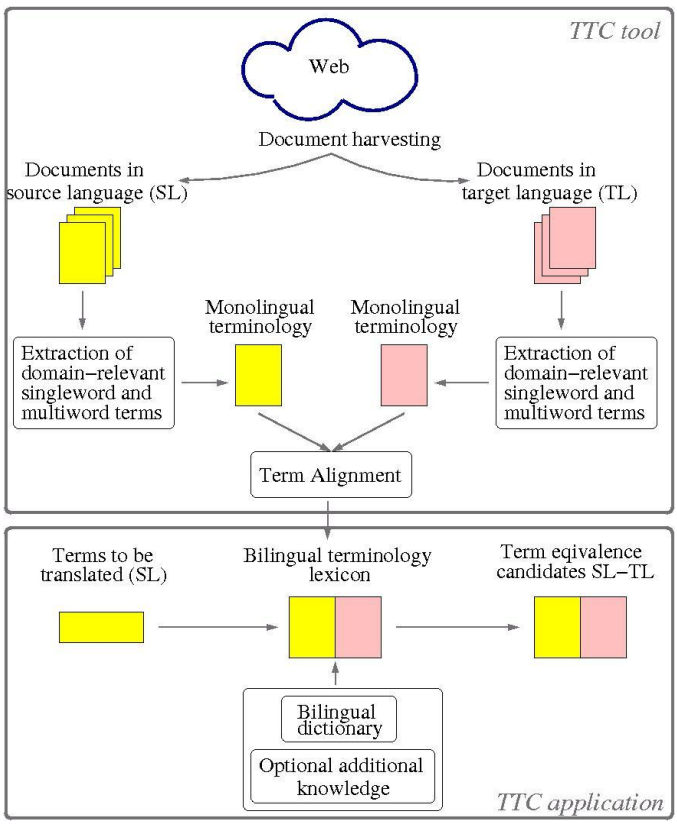


Figure 1: Multilingual Terminology Mining Chain

From the user perspective, we see three main categories of scenarios which need to be addressed. One category concerns the types of language activities to be carried out with the help of the TTC tools, the other concerns the situation of users with respect to the availability of language resources as the input for the TTC tools. This second category will be discussed in section 5.1. A third category deals with the profile of users and suggests three types of users depending on their level of expertise with respect to the use of translation tools.

The main objective of TTC is to support interactive translation work, especially the following tasks:

- translation of texts in a new domain, requiring the provision of larger amounts of terminology (Bowker and Pearson 2002);
- translation of texts in a known domain, but with new terminology;
- correction of translations, possibly with terminological checking.

Next to translation, we could also envisage the TTC tools to be useful for technical writers.

# 5 TTC Input and Output

## 5.1 TTC Input

Many translators will not start from scratch when having to do terminology work for a given domain. They may have available one or more of the following types of data:

1) the texts to be translated (source language (SL));

2) other texts which have already been translated, or which are variants of the texts to be translated;

3) translation memories created in previous translation work;

4) monolingual termbank data;

5) bilingual termbank data for the language pair to be translated;

6) bilingual termbank data for a language pair where only one of the languages is part of the language pair to be translated.

This material may be used in several ways as an input to the TTC tools:

a) SL texts may provide seed words for crawling more texts, as part of a comparable corpus; the same applies obviously for target language (TL) texts;

b) translation memories may also provide seed words, as well as a basic stock of equivalence candidates; this is also true of variant texts and texts which have been already translated and which are closely related with the domain under study (e.g. material from earlier translation work). Discussions during the workshop concluded that the usefulness of translation memories would need to be further analysed in experiments in the TTC setup.

c) lexical data, both monolingual and bilingual can be used both as seeds for crawling and in a basic stock of terminology.

In the focused workshop with high-level experts with different specializations (translators, tool manufacturers, etc.) held in October 2010, more issues concerning possible inputs to the TTC tools were discussed. In fact, most companies own considerable amounts of documentation, but data privacy could prohibit their usage for term extraction. A translator may thus not be able to use already existing data to bootstrap automatic term extraction, e.g. using a client's translation memory as input for the TTC crawler to generate seeds and extend the existing corpus; or introducing a client's glossary into the open terminology platform to enrich it with TTC output. TTC is conscious of these legal issues since the beginning of the project, and thus, TTC tools, such as the web crawler, provide private user accounts. The tool stores the seeds and the crawled corpus, but does not store the corpora used as seeds.

If the user does not have or cannot use available data, he/she can use the crawler to gather domain-relevant terminology from the Internet, although this type of resource is not necessarily reliable and trustworthy. Reliability and trustworthiness of the data is a major concern of users. In the above-mentioned workshop, the following issues were discussed:

- Non-availability of company-specific terminology: company-specific terminology will likely not be available on the Web, unless it is related to advertisement; thus, resources for highly specialized in-house translation tasks will require the availability of company resources.

- Heterogeneity of crawled data: the texts crawled from the Web are of a wide variety of text types. For the domain of wind energy, for instance, we found several text types ranging from technical texts from physics or electrical engineering, to legal documents about possibilities to install windmills, as well as documents about energy pricing and remunerations; advertising for and against wind energy, including pamphlets criticizing the installation of windmills; advertisement for "green energy" produced by windmills; children's book articles (encyclopaedia-like) on windmills; articles from Wikipedia. In TTC, we address this issue by means of classifying crawled text into broad text type categories; not all categories are relevant for term extraction, and irrelevant texts can be excluded from further processing.

- Degree of specialisation: The data retrieved using an html web crawler such as Bootcat (Baroni and Bernardini, 2004) typically contains few specialized documents. A crawler for terminology extraction purposes should deal with several formats of documents. Highly specialized texts, such as reports and scientific publications, are often encoded in PDF or DOC formats. Text provision should those distinguish between scientific, popular science and legal discourse (Goeuriot et al. 2009). Moreover, the quality and the amount of data available go hand in hand with the presence of the language on the internet. Thus, for languages like Latvian, it is more difficult to find relevant domain-specific data, as less data is available on the internet.

## 5.2 TTC Output: Data Categories

The excepted output from TTC tools includes both lists of monolingual and/or aligned bilingual terminologies in a standard format as well as textual data (corpus collection gathered with the crawler).

The TTC survey has shown that around 70% of the respondents use proprietary formats (based half and half on Microsoft Word and Excel) to represent their terminological data. Few use standard formats, and if they do, they use TBX (see above, section 3). In TTC, we have opted for using TBX as an output option. In addition, we used the TBX (lite) data categories as a “checklist” with respect to the data categories that need to be provided by the TTC tools. We did so, because we anticipate more TBX users in the future, because TBX data categories are well-defined and clearly described, and, finally, because TBX compliance means standards compliance (ISO 30042).

TBX (lite) proposes some 20 data categories, along with descriptions and guidelines for their use. In table 1, we list the TBX data categories, grouping them into (i) those which the TTC tools will provide (left half of the table), and (ii) those which cannot be provided by the TTC tools (right side of the table). For part (i), we distinguish between two modes of provision: the extraction from metadata available together with the texts used for term extraction (signalled by “M”), and the extraction from the actual data, by means of the TTC tools (signalled by ”T”).

As table 1 shows, a considerable amount of TBX data categories can in principle be provided by TTC. The items marked with an asterisk (*) can however not be provided in full.

| Data categories provided by the TTC tools | Type | Data categories not provided by TTC |
|---|---|---|
| Subject field | M | Image |
| Language | T | Note |
| Term | T | External Reference |
| Source (of term) | M | Definition |
| Date of last modification* | M | Usage Status |
| Author of last modification* | T | Localization Type |
| Term type | M | Cross Reference |
| POS | T | Customer |
| Gender | T | Project |
| Context(s) | T | |
| Graphical usage* | M | |

Table 1: Data categories from TBX and their provision by TTC: under “type”, "M" indicates that the metadata provide the respective information, “T” stands for "produced by the TTC tools"; items with asterisk (*) cannot always be provided in full, depending on the source quality

The data categories provided by the TTC tools include the linguistically most relevant ones. Administrative data can however not always be provided, as not all text sources may have metadata for the text creation and release data, possible modifications, etc. Data that cannot be provided by TTC are definitions, as well as external references and pragmatic features of terms, e.g. customer or project-specificity features.

## 5.3 TTC Output vs. User Preferences

Not all TBX lite data categories have been checked with users in the TTC survey. However, a ranking by importance of most categories falls out from the questions about the most important tasks in terminology work and about the most important data categories in external resources and in possible own terminology creation work. The ranking obtained from the survey is given in table 2 along with marks indicating the availability (+/-) in TTC output and the way of production (M/T, as above, in table 1).

Table 2 shows that from the top important data categories, only definitions, synonyms, antonyms and (style) usage labels are not provided by the TTC tools.

| Rank | Data category | TBX | TTC | M/T |
|---|---|---|---|---|
| 1 | Translation equivalent | (+) | + | T |
| 2 | Example sentence | Context | + | T |
| 3 | Definition | Definition | - | - |
| 4 | Style, usage | Geographical | (+) | (M) |
| 5 | Frequency | - | + | T |
| 6 | Subject fields | Subject field | + | T |
| 7 | Synonyms, Antonyms | - | - | - |
| 8 | Word class | POS | + | + |

Table 2: TTC survey: data categories and their importance for the users - here confronted with what TTC can produce

## 5.4 User Profiles vs. TTC Output

The facts sketched in the two preceding subsections of this section were presented and discussed with experts participating in the above-mentioned workshop. We summarize the main outcome of the discussion.

First, it is worth mentioning that a major conclusion from the discussion about the potential users of TTC technology concerns the definition of the different user profiles. TTC will have to serve may need to serve different kinds of "clients", as far as their prior knowledge, their interest in the tools, as well as their availability for involvement into experimental work with TTC output is concerned:

- *basic users*: prefer to get a comparatively small amount of information, in a simple dictionary entry structure; mainly just equivalents;
- *advanced users*: prefer to get all information relevant for the translation, including examples, metadata, etc., which support them in selecting equivalents;
- *MT specialists*: need specific system-adapted output.

This means that there are different requirements on the TTC output depending on the types of users and on the envisaged applications. A basic user needs a lexicon with equivalences in the source and target language. The tool should typically not display more than five equivalent candidates (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 ("keep it simple!").

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

- Part-of-speech tags.
- Term origin: Labels indicating the source of a proposed item allow the user to judge reliability and trustworthiness, and to understand the context where the item plays a role.
- Reliability data and 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, form and morpho-syntactic structure (Daille, 2007). Different realisations of a term can be identified by the TTC tools allowing for a 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 were considered to be useful for the potential users of the TTC tool, but which cannot be provided (in full) by the tool.

- Abbreviations. Abbreviations are very common in domain-specific documents, and, particularly, in customer feedback and daily work within companies.

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

The discussion showed that the main data categories foreseen as TTC output are useful for advanced users. Following suggestions from the participants in the above-mentioned workshop, TTC will allow users a manual selection in terms of data categories, before importing TTC output into, e.g. CAT tools.

Furthermore, TTC has to pay special attention to the merging of existing terminology of the user and newly generated data produced by the TTC tools. This task will need human intervention: the merge cannot be carried out automatically, and the user will have to validate results and avoid removing existing correct entries and/or replacing them without validation.

This lesson learned from the discussion has an impact on the design of the TTC GUIs as well as on the specification of the tool outputs.

## 6 Technical Scenarios for Bilingual Term Extraction

In the previous sections, we have presented the expected user scenarios and input and output formats. Here we outline the technical scenarios depending on the resources available for the linguistic processing of the data. The extraction of bilingual terminologies from comparable corpora is a task that includes the processes of monolingual term identification and bilingual term alignment, as well as the compilation of comparable corpora and bilingual corpora alignment. The tool pipeline that will be set up as well as the quality of the results will depend on the resources that are available as well as on the strategies adopted for each step. The quality and type of resources may not be the same for each language pair. For under-resourced languages, it seems clear that the amount of available data and resources will be lower than for more well-resourced languages. Therefore, we will evaluate the quality of the results in connection with a comparison between slim and knowledge-rich approaches.

One of the challenges of the project is to set up a tool pipeline that is valid for all language pairs, including less resourced language as Latvian, and different language families. The pipeline should also be applicable to morphologically rich languages. Thus, one of the research questions is to assess the quantity and type of linguistic resources and tools that we need. How shallow can the resources or tools be? What is the impact on the performance (precision and recall)? How dependent are the results on the resources used? To what extent do the type and quantity of knowledge used have a direct impact on performance?

We envisage several scenarios depending on the knowledge but also on the processing approach. We have identified four basic pipelines for bilingual terminology extraction.

1. Corpora compilation: Important factors at this stage are the type and number of seeds used. Three different types of seeds can be used for corpus compilation: lists of terms, lists of URLs, or pre-existing corpora. Does the number of terms or the degree of specialization of the seeds (SWTs vs. MWTs) have an impact on the quality of the corpora output by the crawler?

2. Bilingual corpus alignment: Does the degree of comparability have an impact on the performance of term alignment? Are the results better with parallel corpora? For comparable corpora, are results better when all document features are the same? When documents are "strongly" comparable? What results do we obtain with less comparable corpora, such as corpora from a different genre?

3. Linguistic knowledge for corpora analysis: Do results improve when using more ore less linguistic knowledge, e.g. when using a part-of-speech-tagger? What is the impact of naive vs. more sophisticated tokenization (especially for Chinese as tokenization of this language is a very complex issue)? Envisaged scenarios concern the following linguistic processes: tokenization, lemmatization, PoS tagging, syntactic analysis (dependency vs. constituent parsing), morphological analysis (e.g. compound splitting) and extraction of single vs. complex terms with or without variants.

4. Bilingual term alignment strategies: What is the impact of using pre-existing monolingual and bilingual terminologies or bilingual lexica for term alignment? Does the size of the pre-existing monolingual and bilingual terminologies have an impact? Do we obtain different results when using existing translation tools to generate term candidates instead of using bilingual terminologies?

## 7 Conclusions

Overall, the discussions with the user experts at the October 2010 workshop, 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.

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.

Concerning the input to the 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 via web service. To solve this limitation, domain-specific corpora can be gathered using the TTC web crawler, although the harvested material may not necessarily be reliable and trustworthy. Because web data might not always be relevant or specialized, the web

crawler should include parameters to filter texts that should not be crawled, and a blacklist of URLs. The tool should be intuitive and easy to parameterize for all types of users.

Regarding the output of the TTC tools, there are different requirements depending on the types of users of the tool. A basic user needs a lexicon with equivalences in the source and target languages. The tool should however not display more than five equivalent candidates. 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, such as grammatical tags and information about the source.

Furthermore, advanced users and MT users should be given a possibility to validate TTC's terminology extraction output before integrating the data into translation tools. Another scenario concerns the merging of resources: existing terminologies and TTC terminology output. In both cases manual validation should lead to better results, but as translation activities are performed under tight time constraints, we will pay special attention to the question of how much manual checking could be involved in the work with TTC output.

To conclude, we can confirm that the definition of usage scenarios as well as the interaction with potential users and experts in the domain was vital for the specification and development of the tool components and the whole pipeline.

## References


*Baroni, M. and Bernardini, S*. (2004) "Bootcat: Bootstrapping corpora and terms from the web". In Proceedings of LREC 2004, Lisbon.

*Blancafort, H. and Gornostay, T*. (2010) "Calling Professionals: Help Us to Understand Your Needs!" TTC Survey 2010 results. (presentation published on the project website[8]).

*Bowker, L. and Pearson, J.* (2002) "Working with Specialized Language: A Practical Guide to Using Corpora". London/New York: Routeledge.

*Daille, B.* (2007) "Variations and application-oriented terminology engineering". In F. Ibekwe-SanJuan, A. Condamines, and M. T. Cabré Castellvì, editors, Application-Driven Terminology Engineering, volume 2 of Benjamins Current Topics, pages 163-177. John Benjamins. ISSN 1874-0081.

*Gœuriot, L., Daille B., and Morin, E.* (2009) "Compilation of specialized comparable corpus in French and Japanese". In Proceedings of ACL-IJCNLP workshop "Building and Using Comparable Corpora" (BUCC 2009), August, Singapore.

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

*Pearson, J.* (1998) "Terms in Context". Amsterdam/Philadelphia:John Benjamins.

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


[8] http://www.ttc-project.eu/images/stories/TTC_Survey_2010.pdf