Terminology Resources and Terminology Work Benefit from Cloud Services

Tatiana Gornostay and Andrejs Vasiļjevs
Tilde
Vienibas gatve 75a, Riga, Latvia
E-mail: tatiana.gornostay@tilde.lv, andrejs@tilde.com

Abstract

This paper presents the concept of the innovative platform TaaS "Terminology as a Service". TaaS brings the benefits of cloud services to the user, in order to foster the creation of terminology resources and to maintain their up-to-datedness by integrating automated data extraction and user-supported clean-up of raw terminological data and sharing user-validated terminology. The platform is based on cutting-edge technologies, provides single-access-point terminology services, and facilitates the establishment of emerging trends beyond conventional praxis and static models in terminology work. A cloud-based, user-oriented, collaborative, portable, interoperable, and multilingual platform offers such terminology services as terminology project creation and sharing, data collection for translation lookup, user document upload and management, terminology extraction customisation and execution, raw terminological data management, validated terminological data export and reuse, and other terminology services.

Keywords: terminology work, terminology resource, cloud service

1. Welcome to the Cloud: Beyond the Conventional Terminology Work

Cloud computing technology is rapidly becoming a predominant computing paradigm applied for a large variety of data processing tasks and applications. Among its key benefits are platform and application independence, up-to-datedness, and low service costs for the user (Muegge 2012).

In the language industry, the cloud-based offerings for translation services appeared at the end of the previous decade with the launch of pioneering solutions for translation management (for example, Lingotek, Lionbridge Freeway, and Google Translator Toolkit).

To execute a translation project or to build a machine translation system in the cloud is no longer a new concept in language work[1]. It is not necessary to buy "heavy" and expensive single-company-made solutions anymore. Cloud offerings allow to "construct" and run a scalable workflow out of flexible and interoperable services, without any consumption of local computing resources.

However, conventional praxis and static models might still hinder productivity and overall progress, for example, terminology work and the creation of terminology resources suffer from time-consuming manual processing and out-of-date work patterns. The language of the profession is developing so rapidly that the conventional approach to terminology work and the creation of terminology resources no longer fits user requirements.

There is an urgent need to find a more efficient approach in terminology work as language workers acknowledge that terminology is a backbone of professional communication. Consistent terminology ensures the accuracy of created or otherwise managed content. It also ensures the quality of documentation within its life cycle, the satisfaction of users, and the consistency of brands.

Obviously, the time has come to facilitate the establishment of emerging trends in the creation of terminology resources, their management, and their utilisation in various applications, reflecting the latest developments in corpus-based terminology (see, for example, L'Homme 2004) and terminology database design (Vasiljevs et al., 2011).

Though a number of terminology work supportive tools currently exist, there is no single tool that could cover all the major steps within a term life cycle from identification to translation and further exploitation in other language applications – the so-called portable solution. Existing and/or available tools are not adjusted to new trends in terminology work, for example, few tools integrate facilities for corpus work, most tools have limited language coverage, few tools have sharing facilities and are adherent to ISO standards, no tool is based on cloud computing, etc.

This paper presents the platform TaaS[2] that is being created within the EU FP7 project TaaS[3]. It pioneers a new fashion in terminology work: an automated approach to terminology identification applying linguistic intelligence, translation lookup using major terminology resources, terminology at users' disposal, Web data, and, finally, a professional who validates the result.

[1] See, for example, the LetsMT platform at letsmt.com (Vasiljevs et al. 2010).

[2] The Beta version of the platform is available at http://demo.taas-project.eu.

[3] The public project website is available at http://taas-project.eu.

2. Overview of the TaaS Platform

TaaS is based on cloud computing technology that rapidly becomes one of the predominant computing paradigms applied for a large variety of data processing tasks and applications.

To automate various tasks in terminology work, the TaaS platform provides a set of interoperable cloud-based services (see Figure 1). The services are integrated into several workflows. These workflows automate the identification of term candidates in user-uploaded monolingual documents and the lookup of translation equivalents for identified and extracted monolingual term candidates. Translation equivalents are retrieved from online terminology term banks, multilingual terminology automatically extracted from comparable and parallel resources on the Web (in online and cashed scenarios), as well as terminology collections created by the platform's users.

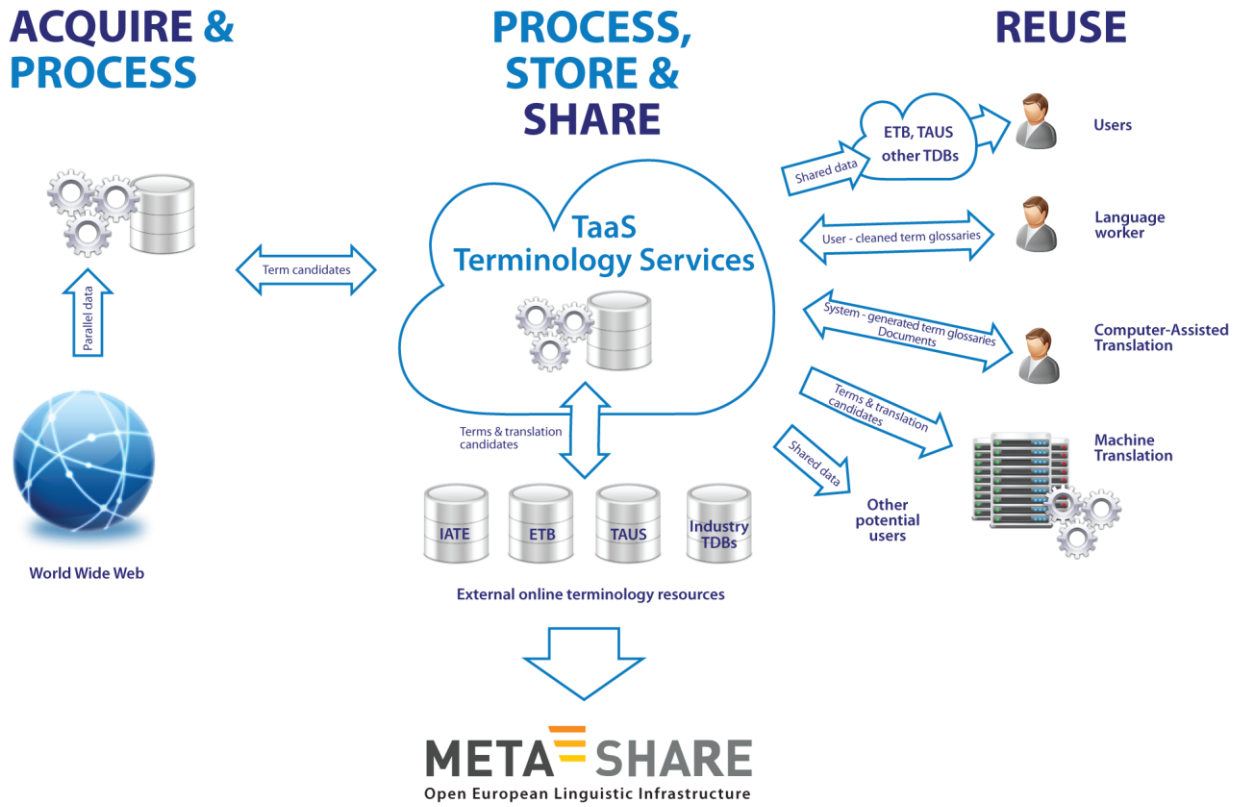

Figure 1. TaaS concept

TaaS platform also provides facilities for involvement of users in the clean-up and enrichment of raw terminological data, automated sharing and synchronisation of the resulted terminology in various use cases by language workers and language processing applications, such as computer-assisted translation tools, machine translation systems, terminology management and terminology lookup platforms, and others. User involvement is motivated by reciprocity principle, that is, users process their documents and clean-up the resulting raw terminological data, then share it and thus give access to other users and/or contribute to existing term banks.

The TaaS platform is user-oriented: we have conducted a profound user needs analysis and revealed the most needed and required functionalities that the platform should support (Gornostay et al., 2013).

The TaaS platform is collaborative: a language worker is no longer alone in his/her task. Individual autonomous work leads to errors in term usage and affects not only translation productivity and overall costs but also influences further phases of content life cycle, for example, failures in product technical support, client request processing, marketing etc.

TaaS also foresees the necessity for interoperable working environments and their capacity for inclusion into World Wide Web technology, supporting the evolution of the Internet and an emerging Web 3.0 technology.

Therefore, it is compulsory to implement standards that can be used to exchange terminological data between different applications and systems, for example, updated XML-based standards that allow for interoperability with the Linked Open Data community. Thus terminological

data will be an important part of the Semantic Web and will be accessible not only by typical terminological applications.

And finally, TaaS is multilingual: by now it supports terminology work for 25 languages. One of the TaaS advantages over other existing terminology extraction tools is its capability for processing languages with rich morphology. This makes extraction results better in terms of quality, for example, inflective forms are normalised, that is, canonical forms of term candidates are given to the user (see also Schmitz and Gornostay 2013).

3. TaaS Terminology Services

In the following subsections we describe terminology services provided by the TaaS platform and its user environment elements.

3.1 Search of Terminology in Various Sources

To perform most of his/her work in TaaS, the user has to sign up for the services. However, for its unregistered users, TaaS provides the service for terminology search in two sources – TaaS database, which consists of TaaS users' terminology collections made public by its users, and EuroTermBank, which is the largest European online term bank, providing access to more than 2 million standardised terms from more than 100 national terminology resources in 27 languages[4]. For advanced search, you have to select your source and target language, domain (a.k.a. subject field), and the source to be searched in (see Figure 2).

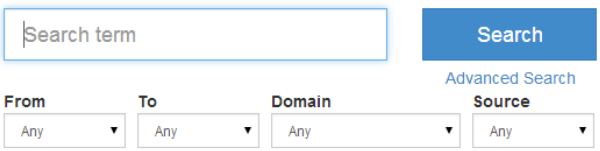

Figure 2. TaaS search form

3.2 Creation and Sharing of Terminology Projects

A registered user gains access to all TaaS services. To start his/her work, the user has to create a new project indicating the source and target languages and the domain the user works in (it is relevant to user document(s) domain). The user might also want to specify optional properties, such as product, customer, project description, and the business unit (in case of a corporate user) (see Figure 3).

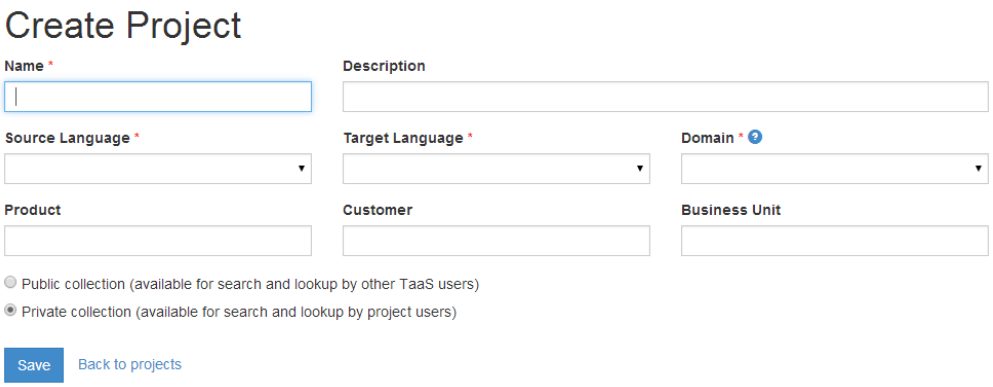

Figure 3. Creation of a new project in TaaS

TaaS also provides a default project with project properties already set for demonstration purposes. Finally, the user has to set the status of his/her project – private or public. If the status of a project is public, the user's approved terminology will be available for search and lookup by other TaaS users; if the status of a project is private, the user's approved terminology will be available only to project users.

The user can start his/her work with TaaS by using the default project or creating a new project. In both cases, the user is an administrator of his/her project.

TaaS provides facilities for project sharing among users if they work in a team. This functionality that typically involves an interchange of non-confidential, non-competing, and non-differentiating terminology across various actors is highly rated by users. Recent surveys (see, for example, Gornostay 2010; Gornostay et al. 2013) have shown that up to 60% of terminology resource users would share their resources with the community. The concept of sharing, unfortunately, is not present in the current management of major terminology databases and term banks. Instead of providing the opportunity for users to contribute their data, major term banks typically keep to the traditional one-way communication of their high-quality pre-selected terminological data.

[4] www.eurotermbank.com

To share his/her project with other users, the user has to add their e-mails and assign their roles. There are three available roles to a new user of the shared project: administrator, with full access rights; editor, with limited access to editing rights; and reader, with limited access to reading rights. One project can have more than one administrator; however, the owner of the project (the user, who has created the project) must consider assigning the administrator's role to other users of his/her project as these users will get full access, including the right to delete the project and its terminology collection. The Administrator's role is usually assigned to the project manager in the translation team, who adds documents to the project, and these are later processed by a terminologist, translator(s), editor(s), and other translation team members (see Figure 4).

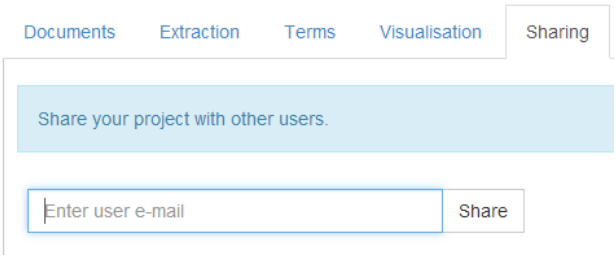

Figure 4. Project sharing

3.3 Upload and Management of Users' Documents

The main usage scenario for the TaaS services is when the user uploads his/her document(s) under the created project, in order to then execute the terminology processing. TaaS supports user document upload in more than 10 formats including the most widely used MS Word, Excel, and Power Point formats as well as the Portable Document Format (PDF), the XML Localisation Interchange File Format (XLIFF), and others. At the present time, two more formats are being added: FrameMaker (MIF) and InDesign (INX) formats to support technical writing as well. The open Beta version has certain limitations in terms of file and project size.

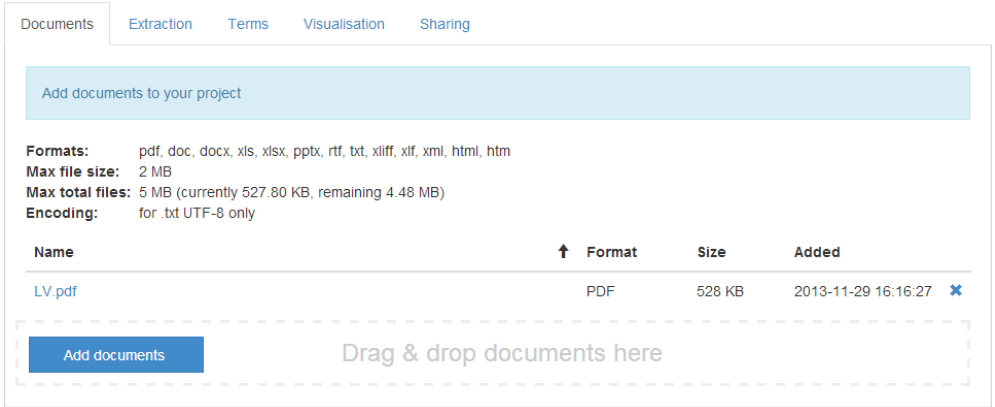

Figure 5. Document upload

3.4 Extraction of Monolingual Term Candidates from User-Uploaded Documents

The terminology extraction service performs automatic extraction of monolingual term candidates from user-uploaded documents using generic or language specific terminology extraction techniques.

The user can customise the terminology extraction process. He/she can select one or more available (on the platform) terminology extraction tools for term candidate identification in user-uploaded documents. There are two term identification tools integrated into TaaS at the moment. These are the Tilde Wrapper System for CollTerm (TWSC) [5] that includes language specific patterns and morphological analysis and Kilgray Term Extractor that applies generic statistical approach to all supported languages. It is recommended to select the first tool; however, the statistical tools might also be of help in certain cases, for example, when linguistic processing produces insufficient results.

3.5 Retrieval of translation equivalents

The platform provides the service for automatic retrieval of translation equivalents (for the extracted monolingual term candidates) in user-defined target language from different public and industry terminology databases.

The following terminology resources are available for translation equivalent lookup for term candidates identified in user-uploaded documents:

- TaaS public collections shared by other TaaS users;
- Terminology collections owned by the user;
- EuroTermBank;
- Inter-Active Terminology for Europe (IATE), an inter-institutional terminology database of the European Union[6];
- TAUS Data that stores shared translation memories;
- TaaS database of raw bilingual terminological data automatically extracted from original and translated texts (a.k.a. comparable and parallel corpora) on the Web.

[5] See the ACCURAT Toolkit 3.0 at www.accurat-project.eu.
[6] http://iate.europa.eu/

3.5.1 TaaS Bilingual Terminology Extraction Workflows for Web Data

In the dynamic pace of technological developments and societal changes, new terms are coined every day by industry, translation and/or localisation agencies, collective and individual authors. Although these terms can be found in different online and offline publications, the inclusion of new terms in online public terminology databases and term banks takes months or even years, if it happens at all. As a result, terminology databases and term banks fail to provide users with extensive up-to-date multilingual terminology, especially for terms in under-resourced languages or specific domains that are poorly represented in online public terminology resources.

At the same time many new terms and their translations can be found on the Web – in multilingual websites, online documents, support pages, etc. TaaS provides four bilingual terminology extraction workflows for Web data: one workflow for terminology extraction from parallel data and three workflows from comparable data. The latter three are customized to collect terms from comparable news corpora, from multilingual Wikipedia, and from focused comparable corpora, respectively.

Web data are collected and then automatically processed. As a result, a list of bilingual raw term candidate pairs are extracted and fed into the TaaS terminology repository. During the execution of a terminology project at the translation candidate lookup step, these data are retrieved and proposed to the user for his/her validation. Thus the TaaS aligns the speed of terminology resource acquisition with the speed at which the content is created by mining new terms directly from the Web.

The data collection process is ongoing constantly feeding the TaaS repository with new terms. By April 2014, the TaaS database included more than 8 M bilingual term pairs extracted from the Web data.

3.6 Clean-up of Raw Terminological Data

TaaS provides facilities for cleaning up raw terminological data extracted automatically that is noisy and needs validation by users. The process of validation can be regarded as a three-step procedure (see, for example, Chambers 200):

- monolingual validation (deletion of "unwanted" and/or unreliable term candidates, definition of termhood, term variant identification, deduplication, deletion of "incorrect" extraction, for example, a part of a longer noun group, synonym identification etc.);
- bilingual validation (bilingual checking of term candidates and their translation candidates, defining the right translation for the source term, deletion of irrelevant and/or incorrect translations, etc.);
- and validation in context.

As soon as extraction finishes, the user can see extracted terms from his/her documents and their translation equivalents retrieved by TaaS. The user can hover over terms to get additional information, such as grammar, source, and context (see Figure 6).

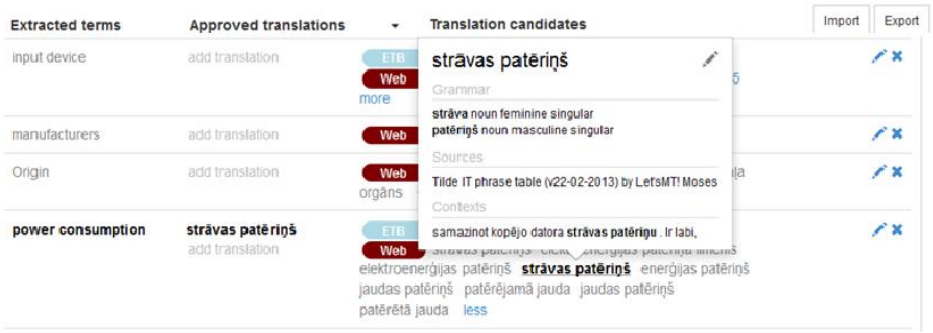

Figure 6. TaaS interface for cleaning up and validating raw terminological data

The user can approve terms with a single click and add translations him-/herself, is the right translation from proposed translation candidates is not found.

An extracted term with its translation equivalent(s) forms a terminology entry. For advanced purposes, the user might want to edit a term entry in full entry view using the term entry editor and to add additional information about terms, for example, definitions, notes, grammatical information, and usage properties, such as term type, register, administrative status, temporal qualifier, geographical usage, and frequency. The history of editing is saved and is seen in the full entry view (see Figure 7).

The user might also want to see term candidates identified by TaaS in his/her documents, and the visualisation functionality is available for this purpose (see Figure 7).

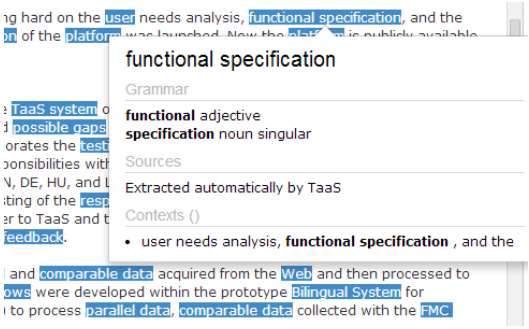

Figure 7. Visualisation of term candidates in the user's document

3.7 Export and Reuse of Validated Terminology

Validated terminological data can be exported and then used in other translation working environments.

Export is available in the most popular formats:

- TBX (TermBase eXchanged ISO-standardised format);
- CSV (a comma-separated value format that is the top first user-required format according to the user needs survey (Gornostay et al. 2013));
- TSV (a tab-separated value format that is also widely used by the community).

During the analysis of user needs and requirements, we also proved our hypothesis that terminology, as a language resource, is central for the second large group of users – language applications (the first user group is represented by language workers). Under language applications (or machine users in other words), in the first place we consider computer-assisted translation (CAT) tools and machine translation (MT) systems. We have already performed first successful experiments on the integration of terminological data acquired within TaaS into the statistical MT system (Skadins et al. 2013; Pinnis and Skadiņš, 2012). At the time, the memoQ CAT tool[7] owned and developed by Kilgray, the TaaS project partner, is being integrated with TaaS via the TaaS Application Program Interface (API) developed in the project and available for machine users.

3 Conclusion and Future Work

In this paper we have presented the concept of the innovative platform TaaS "Terminology as a Service". At the present time, TaaS is a unique dynamic cloud-based solution that provides a wide range of terminology services.

We foresee the potential of the established platform for a wide range of user groups, both language workers (for example, lexicographers and commercial publishing houses, second language learning students, students and specialists in various domains of knowledge, and others) and language applications, or so-called machine users (for example, knowledge organisation systems in library and information science, search engines, and others).

We are also researching the ways of collaboration between terminology resources and Linked Open Data as well as the benefits that terminology might bring to the Web technology development.

4 Acknowledgements

Research within the TaaS project, leading to these results, has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement no 296312.

The TaaS platform is a result of fruitful collaborative work of the project partners – coordinator and lead developer Tilde (Latvia), research partners Cologne University of Applied Sciences (Germany) and University of Sheffield (UK), industry partners Kilgray (Hungary) and TAUS (Netherlands).

5 References

Chamgers D. (2000) Automatic Bilingual Terminology Extraction: A Practical Approach. In Proceedings of the 22nd International Conference "Translating and the Computer 22", November, 2000.

Gornostay T. (2010) Terminology Management in Real Use. In *Proceeding of the 5th International conference "Applied Linguistics in Research and Education"*, April, 2010, St.-Petersburg, Russia.

Gornostay T., Vodopiyanova O., Vasiļjevs A., Schmitz K.-D. (2013) Cloud-Based Terminology Services for Acquiring, Sharing and Reusing Multilingual Terminology for Human and Machine Users. In *Proceedings of the TRALOGY II Conference "The quest for meaning: where are our weak points and what do we need?"*, January, 2013, Paris, France.

L'Homme M.-C. (2004) *La terminologie: principes et techniques*. Montréal: Les Presses de l'Université de Montréal.

Muegge U. (2012) The Silent Revolution: Cloud-Based Translation Management System. In TCWorld, July, 2012.

Pinnis M. and Skadiņš R. (2012) MT Adaptation for Under-Resourced Domains – What Works and What Not. In *Proceedings of the 5th International Conference "Human Language Technologies – The Baltic Perspective"*. October, 2012, Tartu, Estonia.

Schmitz K.-D. and Gornostay T. (2013) Beyond the Conventional Terminology Work. In *Proceedings of the conference TCWorld 2013, the track "CHAT: Creation, Harmonization, and Application of Terminology"*. November 6-8, 2013, Wiesbaden, Germany.

Skadiņš R., Pinnis M., Gornostay T., Vasiļjevs A. (2013) Application of Online Terminology Services in Statistical Machine Translation. In *Proceedings of MT Summit XIV*, September, 2013, Nice, France.

Vasiļjevs A., Gornostay T., Skadiņš R. (2010) LetsMT! – Online Platform for Sharing Training Data and Building User Tailored Machine Translation. In *Proceedings of the 4th International Conference "Human Language Technologies – The Baltic Perspective"*, October, 2010, Riga, Latvia.

Vasiļjevs A., Gornostay T., Skadiņa I. (2011) From Terminology Database to Platform for Terminology Services. In Proceedings of the 1st workshop "CHAT: Creation, Harmonization, and Application of Terminology", May, 2011, Riga, Latvia.

[7] See the description at http://kilgray.com/products/memoq.