From Terminology Database to Platform for Terminology Services

Andrejs Vasiļjevs
Tilde
Riga, Latvia
andrejs@tilde.lv

Tatiana Gornostay
Tilde
Riga, Latvia
tatiana.gornostay@tilde.lv

Inguna Skadiņa
Tilde
Riga, Latvia
inguna.skadina@tilde.lv

Abstract

The paper describes an emerging trend for the next generation of terminology platforms. These platforms will serve not only as a source of semantically rich consolidated multilingual terminological data but will also provide a variety of online terminological services becoming part of a multifaceted global cloud-based service infrastructure. As an example demonstrating this trend we describe the development of terminology services for the EuroTermBank database.

1 Introduction

In the development of large terminology databases or term banks we can distinguish several generations.

First term banks, including EURODICAUTOM, Termium, TEAM, LEXIS, were mostly term-oriented. The terminological data was structured around a term as a lexical unit assigning all possible meanings to a particular term.

The second generation of term banks started to implement a concept-oriented approach, where the concept is in the center of terminological data organization. Here a lexical unit term is subordinated to a concept-based entry defined by a definition, illustration or nomenclature code. Facilities for representing hierarchical relationships between concepts were provided. The Danish multidisciplinary term bank DANTERM, the Norwegian term bank on oil terminology NoTe, and the medical term bank on virology SURVIT are examples of these second generation term banks.

According to the categorization suggested by (Nkwenti-Azeh, 1993) the so called third generation of term banks are knowledge-oriented. Terminology is viewed as a problem-oriented, specialized knowledge representation, and a terminology database can be seen as an expert system for terminology. The ontology-based ECDC Core Terminology Server (Vasiljevs et al., 2008) and frame-based terminological data organization researched in the PuertoTerm project (Faber et al., 2005) are examples of the third generation term banks.

In our view, recent developments mark an emerging trend for the next generation of terminology platforms. These platforms will serve not only as a source of semantically rich consolidated multilingual terminological data but will also provide a variety of online terminology services becoming part of a multifaceted global cloud-based service infrastructure.

In this paper we describe the development of several terminology services for the EuroTermBank database as an example to demonstrate the above mentioned trend. At its core, still remaining a classical concept-oriented terminology database, EuroTermBank is being expanded with different online services to enable new models of terminology sharing and usage. The second section gives a brief overview of the EuroTermBank portal. The third section focuses on terminology sharing services for terminological data owners. The fourth, fifth and sixth sections describe terminology services for users of CAT and authoring environments, for users of MT systems and for European linguistic infrastructure respectively.

2 EuroTermBank overview

EuroTermBank[1] is a centralized online terminology database for languages of new EU member countries interlinked to other terminology resources (Rirdance and Vasiļjevs, 2006). The EuroTermBank portal was designed with the goal to collect, harmonize and disseminate dispersed terminology resources through an online terminology data bank. The EuroTermBank project was launched in December 2006 by 8 partners from 7 European Union countries – Germany, Denmark, Latvia, Lithuania, Estonia, Poland and Hungary.

[1] www.eurotermbank.com

EuroTermBank enables searching within approximately 600,000 terminology entries containing more than 2 million terms in 27 languages and coming from about 100 terminology collections. The portal serves basic terminology needs of a user by providing a single access point to distributed terminology resources and implementing query schemes suitable for particular usage scenarios.

Currently, EuroTermBank provides federated access to 5 interlinked external term banks, the major of them being IATE, the interinstitutional terminology database of the EU (Rummel and Ball, 2001). The specific functions of the EuroTermBank portal include user authentication, term search, data editing, administration, user feedback, and communication facilities with external databases as well as data import and export. An analysis of user needs through focus interviews and surveys as well as collaboration with other EU language technology RTD projects identified an increasing need to extend functionality of EuroTermBank with a number of terminology services for both human and machine users.

3 Terminology sharing services for terminological data owners

The sharing of terminological and translation data is part of general process of transition towards more open and cost-efficient translation and localization business models, reducing the overhead of intermediary suppliers with little or no value added. Our survey shows that about 40% of terminology users are willing to share their resources (Gornostay, 2010).

Terminology sharing typically involves sharing of non-confidential, non-competing and non-differentiating terminology across various actors – individuals along with companies and language service providers, often with the goal to consolidate and promote accessibility to multilingual terminology per vertical industries (Rirdance, 2007). Terminology sharing involves returns from streamlined industry terminology, by ensuring the reuse of existing terminology assets. For those who share their terminology, it is a way of promoting and disseminating one's well-established terminology, possibly even to the level of de facto industry standard terminology.

Industry players have a number of benefits from terminology sharing. It helps them to develop and enhance industry terminology, particularly for minor languages (i.e. languages which have proportionally fewer terminology resources, for example, Slovenian, Latvian, Hungarian), in a cost-efficient way, resulting in the improved quality and user experience for localized products:

- sharing stimulates the harmonization and unification of industry terminology, usage of common terms for common concepts across different products and vendors, enhancing overall user experience and shorter learning curve;
- through terminology sharing vendors can distinguish their specific terms – terms that are associated with particular features and concepts differentiating a vendor's products from the products of the competition;
- sharing strengthens a vendor's market position by boosting user involvement in the particular brand and products, and nurturing the growth of communities around particular products;
- sharing enhances the public availability of language resources thus supporting the research and development of language technologies, particularly for minor languages.

However, the concept of sharing is not really present in major term banks. Instead of providing the opportunity for users to contribute their own resources or share their findings over social networks, term banks typically keep to the traditional one-way communication of their high-quality preselected resources.

A significant development in the area of sharing of linguistic resources is TAUS Data Association[2] that positions itself as "a super cloud for the global translation industry, helping to improve translation quality, automation and fuel business innovation". Although mostly oriented towards sharing translation memories, it does involve the sharing of terminology resources as well.

EuroTermBank provides an individual service for larger industry players. This service is used by Microsoft to share their multilingual terminological data. Microsoft is among pioneers in the industry data sharing on public online repositories, expanding EuroTermBank with more than 20 000 information and communication technology terms in 26 languages. Online facilities to enable every interested user to share terminological data by creating public terminology collections are currently being developed. Users will

[2] www.tausdata.org

also be able to create private online terminology collections accessible only to persons authorized by the data provider.

4 Terminology services for users of CAT and authoring environments

Another requirement identified by the user needs analysis is an integrated access to terminology resources from translation environments. Typically, translators spend about 30% of total translation time on terminology research. Therefore, it is of vital importance to ensure that they can use all the required terminology resources in the right format and in a convenient environment. Increasingly, terminology research is done using sources that are available on the Internet. Currently, translators spend a lot of time inefficiently, searching and processing information from multiple online sources, copy-pasting or changing the format to the one that they require in their work environment. Spending time on technical aspects instead of focusing on true terminology research results in cost inefficiencies and reduced translation quality.

Faced with difficulties in accessing the terms they need and participating in collaborative activities to create new terms, many translators create their own terminology resources. They typically store these terms in spreadsheets or other proprietary formats that are not efficiently connected to a multitude of translation environments that they might use. Moreover, these resources are not shared with other translators and potential users. This results in redundant work or even reduced translation quality and does not bring additional value to the creator of such custom terminology.

A further step in the direction of meeting user expectations and providing the required terminology resources to their users in a most efficient way involves integration of content delivery in the production environments of terminology users. To increase the efficiency and quality of translation, translators need an easy access to multiple terminology databases, facilities to enable collaborative efforts in creation of new terms, productivity tools to get necessary terms right from translation environment (Lengyel and Vasiljevs, 2008). There have been several efforts to provide reasonable solutions to support translators accessing multilingual terminology resources. For example, Quest tool brings consolidated terminology content closer to its user and is used internally by translators in the DG for Translation of the European Commission.

Although the consolidation of terminology in EuroTermBank provides single access point to a variety of terms, still an extra effort is required from the user to switch from translation environment to terminology webpage, specify a search query, select a result and go back to the translation tool and type the term there.

EuroTermBank integration services provide the solution where access to online terminology databases is supported directly from the most widely used translation environments, such as SDL Trados and MemoQ, as well as authoring applications that are commonly used in the translation process, such as Microsoft Word. These services provide terminology integration component for instant access from text editing environment to web-based terminological data by invoking web service based queries.

External terminology database API enables third party software manufactures to provide their users with direct access to the content of terminology database. This is especially useful in the translation usage scenario since such a solution will deliver well-targeted content from a terminology database to productivity environments used routinely by translators and other language workers. Target clients of terminology integration component are translation service providers (freelance translators, translation agencies, localization service providers), translation service consumers (using outsourced and / or in-house services), providers of web-based CAT (computer-assisted translation) tools, students, etc. Freelance translators and in-house translators are foreseen to be major target user groups for the tool.

Furthermore, about 90% of respondents use Google for terminology research. Nevertheless, the survey results show users' interest and necessity for additional terminology tools especially for Microsoft Word. Besides, Microsoft Word integrates with SDL Trados and thus bridges the gap to the user of CAT tools. The goal is to provide access to online terminology content with a single keyboard shortcut, even without opening a browser window. The component for the integration of terminology portal in authoring systems should meet such requirements as easy download, quick setup, low usage of computer resources, integrated representation of terminological data inside authoring system, intuitive use of the tool, no hidden or complicated features. A terminology database should be able to perform

analysis of textual segments to identify terms and provide respective terminological entries.

A layer of connectivity tools was developed for terminology research in specific work environments, such as plug-ins for use with Microsoft Word and MemoQ (Gornostay et al, 2010). For example, in Microsoft Word terminological content is provided inside Word environment in a special terminology pane easily evocable by a single keyboard shortcut. The Microsoft Word integration mechanism automatically detects the source language, filters terminology by domain and language, identifies terms in a segment / sentence and researches the EuroTermBank internal and external resources for the identified terms. It should be mentioned that the function of identifying terms in a segment or sentence and then searching the EuroTermBank resources for them is highly appreciated by end users. The tool identifies terms and shows them hyperlinked in the topmost part of the pane. Moreover, the user can change the language and domain settings, and the tool updates the relevant links in specified languages or domains.

The developed tool was tested and evaluated by end users before its release (internal beta testing). General results of the internal beta testing showed that 70% of respondents consider the tool as a useful or very useful for their translation needs.

Quest is a similar tool that brings consolidated terminology content closer to its user. This metasearch interface which translators can use to query several databases simultaneously is used internally by translators in the Directorate-General for Translation of the European Commission and was developed with a view to centralizing, simplifying and speeding up terminology searches. A Quest search can be launched by pressing a button in Microsoft Word. Translators can select the source and target language pair, and one of three available profiles determining which databases they wish to search. However, this tool is not made available to the general public.

Obviously, the connectivity could also be provided and supported from the side of translation tools. Although a number of translation tools already provide basic integration with terminology web searches, for instance, a user can define a number of term banks to be queried, the nature of these features is such that they will necessarily be general and not adapted to specifics of each term bank, thus possibly making the results of these searches quite useless.

5 Terminology services for users of MT systems

This section overviews terminology services for users of MT (machine translation) systems provided by Open terminology platform being developed within the TTC project (Terminology Extraction, Translation Tools and Comparable Corpora)[3]. Open Terminology Platform (OTP) will be integrated with EuroTermBank and will be interlinked to EuroTermBank as an external database.

Open Terminology Platform will provide support for terminology work for different categories of language workers (translators, terminologists, translation / terminology team managers, technical writers, and researchers in relevant areas) who use MT in their translation workflow[4]. It is motivated by the analysis of current patterns in terminology usage in the translation and localization industry identified in the survey performed within TTC (Blancafort and Gornostay, 2010; Gornostay, 2010; Vasiljevs et al., 2010). More than 65% of respondents use online terminology databases and about 80% of respondents are interested in storing and working with / processing their terminology online. More than 30% of respondents use MT in their translation workflow and 66% of respondents are interested in new terminology management solutions.

Specific functions of OTP relevant to such usage scenario will be terminology import, editing and export into formats compliant with several MT systems. Users will be able to import their terminology collections into OTP and store them online. A widely-accepted term exchange standard format – TermBase exchange (TBX) – will be used to enable exchange of terminological data. TBX framework defined by ISO 30042: 2008[5] is designed to support various types of processes involving terminological data, including analysis, descriptive representation, dissemination, and intercharge (exchange), in various computer environments. The primary purpose of TBX is for standardized interchange of terminological data. To maximize interoperability of the actual terminological data, TBX also provides a default set of data categories that are commonly used in terminology databases. How-

[3] www.ttc-project.eu

[4] One of OTP's usage scenarios evaluated and demonstrated within the project.

[5] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=45797

ever, subsets or supersets of the default set of data categories can be used within the TBX framework to support specific user requirements.

Moreover, OTP users will be able to edit their proprietary terminological data (terms themselves and their corresponding data fields), as well as add / delete individual terms or terminology collections. OTP will also support export into formats compliant with MT software. Within the TTC project evaluation experiments will be performed with the rule-based SYSTRAN system[6] and statistical MT systems based on Moses toolkit (Koehn et. al., 2007), for example, English-German, English-French, English-Latvian statistical MT system and some other language pairs.

Open Terminology Platform is an ongoing development of TTC, it is currently being tested by the project consortium, and will be delivered by June, 2012.

6 Terminology services for European linguistic infrastructure

It is expected that terminology resources and respective services will play an increasingly important role in the European infrastructure for language resources and services that is under construction by EU co-funded CLARIN and META-NET initiatives.

In 2006 CLARIN (Common Language Resources and Technology Infrastructure) initiative came up with the concept of a language resource infrastructure. The aim of CLARIN[7] is to make language resources and technologies available and readily usable for the European researchers in Humanities and Social Sciences through the integrated and interoperable research infrastructure of language resources and technologies (Váradi et al., 2008).

The idea of an infrastructure of language resources and technologies is also among the aims of META-NET Network of Excellence[8]. One of the META-NET goals is to create an open distributed facility META-SHARE for the sharing and exchange of language resources. META-SHARE will be a sustainable network of repositories of language data, tools and related web services documented with high-quality metadata, aggregated in central inventories allowing for uniform search and access to resources.

Three recently initiated ICT Policy Support Programme projects CESAR, META4U and META-NORD will contribute to META-NET aims by assembling, linking across languages, and making widely available language resources. These initiatives will help to build and operate broad, non-commercial, community-driven, inter-connected repositories and exchange facilities of META-SHARE.

Terminology resources are among core datasets of META-SHARE. Thus the META-NORD project will consolidate distributed terminology resources across languages and domains to extend the open linguistic infrastructure with multilingual terminology resources. The EuroTermBank platform will be integrated into the open linguistic infrastructure by adapting it to relevant data access and sharing specifications. The sharing of terminological data will also be based on TBX mentioned above.

Terminology coverage in EuroTermBank for some languages (for example, Latvian, Lithuanian, Polish, Hungarian) is much stronger than for some others which have limited terminology resources integrated. Therefore META-NORD will approach holders of terminology resources in European countries, especially in Nordic countries, facilitating the sharing of their data collections through cross-linking and federation of distributed terminology service. In addition, mechanisms for consolidated multilingual representation of monolingual and bilingual terminology entries will be elaborated. META-NORD has a tight collaboration with CESAR and META4YOU projects to identify and consolidate matching resources and ensure pan-European language coverage and critical volume for the key resources.

Conclusions

The evolutionary development of EuroTermBank from the database of consolidated multilingual terminology to a platform for multifaceted online terminology services reflects a growing trend in the development of terminology management systems.

This trend is determined by shifting patterns of terminology usage such as data sharing and user participation in data collection, as well as rapid development of data-driven language technology applications, for example, machine translation.

The integration of terminology services in the European open language resource infrastructure provides new possibilities for usage of termino-

[6] http://www.systran.co.uk/
[7] www.clarin.eu
[8] www.meta-net.eu

logical data in all kinds of current and future natural language-based applications.

Acknowledgements

Many thanks to colleagues from the EuroTermBank Consortium (the European Union eContent Programme) organizations: Tilde (Latvia), Institute for Information Management at Cologne University of Applied Science (Germany), Centre for Language Technology at University of Copenhagen (Denmark), Institute of Lithuanian Language (Lithuania), Terminology Commission of Latvian Academy of Science (Latvia), MorphoLogic (Hungary), University of Tartu (Estonia), and State Commission of the Lithuanian Language (Lithuania).

Open Terminology Platform is being developed within the TTC project which has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 248005.

The concept of sharing of terminology resources through the European open linguistic infrastructure is being discussed within the META-NORD project which has received funding from the ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, grant agreement n° 270899.

References

Helena Blancafort and Tatiana Gornostay. 2010. Calling Professionals: Help us to Understand Your Needs! The results of a questionnaire-based online survey. Power Point presentation: http://www.ttc-project.eu/images/stories/TTC_Survey_2010.pdf.

Faber, P., Márquez Linares, C. & Vega Expósito, M., 2005. Framing Terminology: A Process-Oriented Approach. *Meta: Translators' Journal*, 50(4).

Tatiana Gornostay. 2010. Terminology management in real use. *Proceedings of the 5th International Conference Applied Linguistics in Science and Education*. Saint-Petersburg, Russia.

Tatiana Gornostay, Andrejs Vasiljevs, Signe Rirdance and Roberts Rozis. 2010. Bridging the Gap - EuroTermBank Terminology Delivered to Users' Environment. *Proceedings of the 14th Annual European Association for Machine Translation (EAMT) Conference*. Saint-Raphael, France.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris C. Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. *ACL '07 Proceedings of the 45th Annual Meeting of the ACL*: 177-180.

Lengyel Istvan and Andrejs Vasiljevs. 2008. How to get the right terms to the right people – terminology sharing and integration in translation environments. *TCWorld Conference*, Wiesbaden, November 2008.

Nkwenti-Azeh, B., 1993. New trends in terminology processing and implications for practical translation. *Proceedings of ASLIB*: 83-98.

Signe Rirdance. 2007. IP vs. Customer Satisfaction: EuroTermBank and the Business Case for Terminology Sharing. *The Globalization Insider*, LISA, 6/2007.

Signe Rirdance, Andrejs Vasiljevs (eds.). 2006. *Towards Consolidation of European Terminology Resources: Experience and Recommendations from EuroTermBank Project*, Tilde, Riga.

Rummel, D., Ball S. (2001). The IATE Project – Towards a Single Terminology Database for the EU. *Proceedings of ASLIB 2001, the 23rd International Conference on Translation and the Computer*, London.

Tamás Váradi, Steven Krauwer, Peter Wittenburg, Martin Wynne and Kimmo Koskenniemi. 2008. CLARIN: Common Language Resources and Technology Infrastructure. *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, 2008, May 28-30, Marrakech, Morocco.

Andrejs Vasiljevs, Signe Rirdance, Laszlo Balkanyi, 2008. Ontological Enrichment of Multilingual Terminology Databank. In *Proceedings of the 8th International Conference on Terminology and Knowledge Engineering TKE 2008*. Copenhagen, 2008, pp.279-289.

Andrejs Vasiljevs, Signe Rirdance, and Tatiana Gornostay. 2010. Reaching the User: Targeted Delivery of Federated Content in Multilingual Term Bank. *Proceedings of the TKE (Terminology and Knowledge Engineering) Conference 2010*: 356-374, Dublin.