EuroTermBank – a Terminology Resource based on Best Practice

Lina Henriksen (1), Claus Povlsen (1) and Andrejs Vasiljevs (2)

(1) Center for Sprogteknologi, University of Copenhagen
Njalsgade 80, DK-2300, KBH S, Denmark
(2) Tilde
Vienibas gatve 75a, LV1004, Riga, Latvia
E-mail: claus@cst.dk, lina@cst.dk and andrejs@tilde.lv

Abstract

The new EU member countries face the problems of terminology resource fragmentation and lack of coordination in terminology development in general. The EuroTermBank project aims at contributing to improve the terminology infrastructure of the new EU countries and the project will result in a centralized online terminology bank - interlinked to other terminology banks and resources - for languages of the new EU member countries. The main focus of this paper is on a description of how to identify best practice within terminology work seen from a broad perspective. Surveys of real life terminology work have been conducted and these surveys have resulted in identification of scenario specific best practice descriptions of terminology work. Furthermore, this paper will present an outline of the specific criteria that have been used for selection of existing term resources to be included in the EuroTermBank database.

1. Introduction

Access to consistent and broad-coverage terminology resources is a precondition for fast and efficient communication across countries. While it can be claimed that access to such terminology resources to a certain extent does exist for old EU member states, this is by far not the case for the new EU member countries.

One of the initiatives taken in order to remedy this unbalanced situation is the project: Collection of Pan-European Terminology Resources through Cooperation of Terminology Institutions[1] (in short EuroTermBank). This project is supported by the EU eContent programme which aims to facilitate the production, use and distribution of European digital content and to promote linguistic and cultural diversity on the global networks.

The main goal of the EuroTermBank project is to contribute to improvement of the terminology infrastructure in the new EU member countries[2]. This aim will be accomplished by establishing terminology networks and by collection and harmonization of existing terminology resources resulting in an implementation of a centralized online term base.

Selection principles defined within the project context will ensure that the pool of existing terminology resources collected in EuroTermBank will meet quality criteria reflecting the needs and demands of the users. Specification of the term base is being prepared with a view to international data exchange standards facilitating implementation of exchange mechanisms for term data from other EU terminology resources.

2. Project Outline

The overall project plan contains a number of tasks. First, an inventory of international standards and best practices in terminology work and term management in involved new EU member countries was established and recommendations for best methodology were prepared. With a view to these recommendations and conducted surveys of user needs and requirements, specification of the system and database platform was created. This specification contains a description of the overall architecture and design, data categories and structure, system functional specification and interface description.

After the implementation phase including pilot trial and standard software evaluation methodology, the final step in the project plan will be the validation phase where agreement between specified and implemented system functionality is evaluated and ensured.

The project will result in a centralized web-based terminology bank for languages of the new EU member countries interlinked to other terminology banks and resources.

3. Identification of Best Practice

One aim of the EuroTermBank project is to identify best practice within most areas of terminology work from use of terminology tools and classification systems to concept analysis and term management.

As a first step towards this goal a report was prepared describing relevant existing national and international standards together with a survey of ‘real-life’ terminology work as it is conducted in the new as well as the old EU member countries. Among the terminology resources that have been investigated are for example the state regulated or coordinated terminology collections of the new EU member countries and the IATE terminology cooperation of the old EU countries.

International and national standards have been used as a starting point for development of best practice. However, standards are very general and describe recommendations in a vacuum disconnected from specific goals and preferences and also disconnected from the set of conditions that apply in a given context. By conditions we refer to the premises or state of things that cannot (or only with much difficulty) be changed. For example a condition might be that all language professionals of a particular organization do not have access to the internet or to terminology tools. Therefore it has been necessary not only to investigate how terminology work is actually carried out in different settings, but also to investigate the

[1] For more information about the EuroTermBank project see http://www.eurotermbank.com.

[2] The countries initially addressed in EurotermBank project are: Estonia, Hungary, Latvia, Lithuania and Poland.

conditions and goals of the particular terminology settings.

In the following we will describe the conditions and goals that have been identified during the project and by an example demonstrate how different sets of conditions and goals influence the outline of best practice within the new EU states involved in the EuroTermBank project.

3.1. Goals and Conditions

Goals and conditions identified in the above mentioned survey were collected and project partners prepared in cooperation with terminology resource owners an assessment of the influence of each condition and the importance of each goal by assigning scores to them. The aim of allocating scores was i.a. to identify sets of goals and conditions that typically co-exist as a first step towards establishment of a number of fixed scenarios with best practice descriptions for each terminology task.

The tables below show goals and conditions considered as having a profound impact on terminology methodologies.

Goal	Explanation
High quality in general terms	Terminology work is based on sound research principles; consistent, non-ambiguous, broadly accepted etc.
Harmonization	In many contexts an inherent part of 'high quality'
Exchangeability	Exchange of data between term resources using standard approved exchange methodology
Availability	Terminology available to external users
Speed and up-to-dateness	Speed of terminology work and data that are always up-to-date

Table 1: Goals

Condition	Explanation
Terminology tools	Users may have access or no access to these tools. Terminology tools in this context include corpus/term extraction tools, but not the term base itself
Type of language professionals	May include or exclude terminologists and domain experts
Financial situation	Satisfactory or unsatisfactory
Language(s) in terminology resource	Mono-, bi- or multi-lingual
Domain coverage	Broad or focused
Purpose: translation, coordination, regulation	Some organizations have translation as their main focus. Some also have coordinating and regulatory obligations

Table 2: Conditions

3.2. Scenarios

The scenarios that were identified are based on the distinction between international, national and local terminology settings.

The international scenario is concerned with coordination and management of multilingual terminology work in a well-organized infrastructure and primarily concerns approval/dismissal and harmonization of terms.

The main activities in a national scenario are similar to those of an international scenario though one main distinctive element is that terminology work at the national level usually is mono- or bilingual. Another difference is that organizations belonging in a national framework in some countries have regulatory obligations as well.

The local scenario covers organizations that do not belong in an international or national framework and concerns terminology work that involves translation/creation of documents and often coinages of new terms. Characteristic features in a local framework are that terminology work usually is limited to one or a few closely related domains, that harmonization does often not play a significant role and that restricted budgets and tight time frames are more likely than in national or international frameworks.

These three scenarios represent schematic frameworks of terminology work. Requirements, aims and circumstances can differ some, even within one framework. Therefore best practice described for one scenario may also in some cases be applicable for an organization that would in this context belong in another scenario. Besides, some factors, not mentioned in the above goals and conditions, also play a role as for example the nature and size of the particular organization. The impact of these factors has however been assessed as less measurable and clear.

The below tables show typical goals and conditions in the international, national and local scenarios.

International	National	Local
High quality in general terms	High quality in general terms	Tight time frames coexist with - and put limitations on requirements for - high quality
Harmonization is high priority	Harmonization is high priority	Harmonization is not a priority
Exchangeability is high priority	Exchangeability is high priority/is sometimes not a priority (recommended as high priority)	Exchangeability is often not a priority (recommended as high priority)
Availability is high priority	Availability is high priority	Availability is not a priority

Table 3: Goals in the international, national and local scenarios

International	National	Local
Access to terminology tools	Access/no access to terminology tools	No access to terminology tools
All types of language professionals represented	All types of language professionals represented	Terminologists often not part of terminology developer team
Adequate financial support	Adequate financial support	Often a tight budget
Multilingual	Mono- or bilingual	Usually bi- or multilingual
Broad domain coverage	Broad domain coverage	Focused domain coverage
Coordination (translation)	Coordination (regulation, translation)	Usually translation

Table 4: Conditions in the international, national and local scenarios

Terminology work involves many types of activities and most of these activities have been dealt with in the EuroTermBank project with a view to extracting best practice. For this paper *data structure* of a terminology data base has been selected to demonstrate how best practice differs from one scenario to another.

3.3. Best Practice with respect to Data Structure

Irrespective of terminology scenario it is as a principal rule recommended to observe the basic data modeling principles as described in ISO 12200:1999 and 12620:1999. This will ensure exchangeability and facilitate recognition and comprehension of data categories for new or outside users. Principles of these ISO standards i.a. involve that the term entries:

- are concept oriented
- contain a rather broad selection of data categories that permits the necessary level of detail (data categories and the contents of these should reflect each other precisely)
- permit full descriptions of each term (NOT: *main term* with descriptions and *synonyms* with no possibility of descriptions)

3.3.1. Local Scenario

In a local scenario some conditions and goals that will have significance for the design of a data structure are for example *tight time frames,* that it is usually *translation oriented*, that *exchangeability* should be high priority and that terminology work is usually restricted to *few domains*. These criteria speak in favour of a highly customized and only moderately exhaustive data structure where data categories are consistent with the requirements of the particular application area and have a translation related focus.

A focus on translation requirements implies coverage of more than one language. It must therefore be considered whether it is necessary with descriptive concept related information as *definition* or *explanation* for each language or only for one language. If the term collection is multilingual a *definition* for each language is usually necessary. If the term collection is only bilingual it may not be necessary.

A focus on translation requirements also indicates inclusion of data categories permitting sufficient information about the use of a term, for example different types of grammar information, context information and collocation information. Some translation settings may also require grammar information for each word of a term. Furthermore, it is often considered very important to document the degree of equivalence between terms of different languages. Data categories that could be relevant in this respect are for example *false friend, directionality* and *transfer comment.*

The below data structure containing four levels reflects a multilingual terminology setting permitting for example concept descriptive information for each language and grammar information for each word. In multilingual as well as bilingual terminology settings it can however be considered to omit the word level and locate grammar information at the term level instead. In some bilingual terminology settings it can also be considered to have a *definition* for only one language. Consequently, the data structure in a bilingual framework may include only 2 levels, namely concept and term levels.

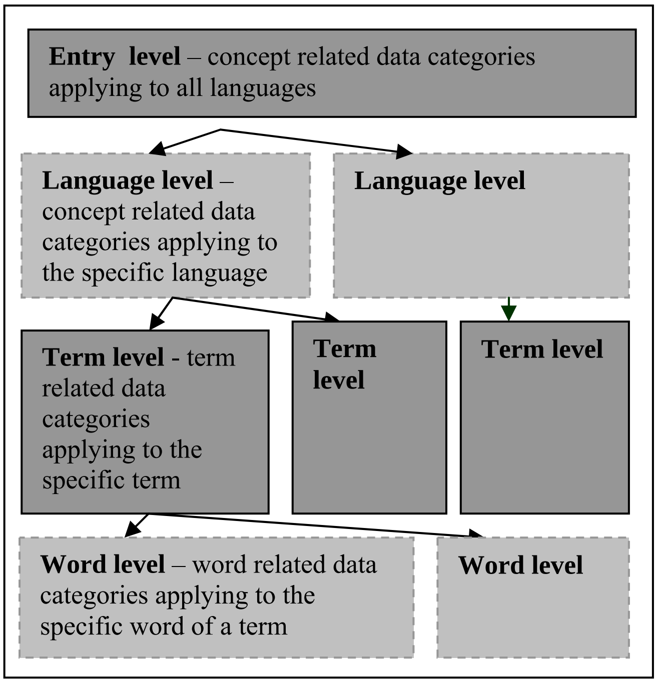
Entry level – concept related data categories applying to all languages
Language level – concept related data categories applying to the specific language
Language level
Term level - term related data categories applying to the specific term
Term level
Term level
Word level – word related data categories applying to the specific word of a term
Word level

Figure 1: Data structure

3.3.2. National Scenario

Conditions and goals influencing the design of a data structure in the national scenario are *adequate financial support*, *exchangeability, broad domain coverage* and *high quality* in general terms. Besides, a national term collection is aimed at terminology *coordination and regulation* rather than at translation. These criteria point towards a data structure that permits an exhaustive selection of data categories covering very different user requirements and enabling users to develop entries for very different purposes and of a very high quality.

This implies that the data structure should often contain 2 levels: concept and term levels (at least when the term collection is monolingual) and that data categories should represent a wide selection of information types and include term status qualifiers reflecting for example acceptability, approval or applicability of a term in a given context. An example of a term status qualifier is *normative authorization* which is assigned by an authoritative body and includes qualifiers as *standardized term, preferred term, admitted term* and *deprecated term.*

3.3.3. International Scenario

The criteria considered important in an international scenario are very similar to those considered important in a national scenario. A crucial difference is however that an international terminology cooperation is multilingual by nature. Therefore it is recommended that the data structure should usually include four levels permitting concept descriptive information for each language, translation related information types and grammar information for each word of a term (see fig. 1).

4. Evaluation of Terminology Resources

One of the major tasks of the project is identification and evaluation of a large number of terminology resources (available in participating countries) and selection of resources for possible inclusion in the EuroTermBank database.

In order to evaluate the terminology resources systematically and to make selection and prioritization for inclusion in the EuroTermBank database several criteria have been used:

- Language for Special Purposes (LSP) only - Language for General Purposes (LGP) resources are not included in the project
- Authority, reputation and expertise of the creating institution or person – whether resource is prepared by a group of experts or by an individual expert, whether specialized lexicographers have been involved etc. Data originators listed by degree of authoritativeness are:
    - legal international or national authority determined by legislation or jurisdiction
    - officially authorized harmonization-/standardization body
    - institution authorized or recognized as a subject field authority
    - formally or informally recognized subject-field authority
    - non-authoritative terminology source
- Methodological approach – observance of relevant national or international standards, completeness of entries (priority to terms with most fields populated), existence of internal/external validation mechanisms. Central quality criteria are concept orientation, subject field indications and usage notes, alphabetical indices in all languages, abbreviations and definitions.
- Availability of the data - to make use of the data, either the terminology resources must be freely accessible or the respective copyright holder should be ready to cooperate and to conclude a copyright agreement with the project consortium.
- Actuality of the data – topicality, frequency of use, date of input or revision. This criterion is closely connected with the respective subject field. For example, in some subject fields old terminology resources of new EU countries include concepts and terms related to soviet-time realities that are not of general interest today.

5. Status and Future Work

This paper provided an outline of the EuroTermBank project and focused primarily on those aspects of the project that concern best practice.

The project results described in the project deliverables and achieved during the first year are:

- Assessment of current standards and best practices that provides an overview of terminology standards, current terminology processes and best practices in the participating new EU countries
- User needs and requirements assessment that serves as a basis for system specification
- System and implementation specification
- Standard document templates and procedures to set the legal and procedural framework for the data collection, integration and exchange
- Framework for identification of existing terminology resources and the key resources identified in the participating countries

Further work is concentrated on technical development of the system specification, implementation of user interface requirements, selection and acquisition of terminology resources and establishment of cooperative relationships with institutions involved in the terminology development. A major part of the EuroTermBank resource will be public, but certain services will probably be commercial in order to cover moderate infrastructure expenses that will arise after the end of the project period.

6. References

Johnson, I., A. MacPhail (2000). IATE – Development of a Single Central Terminology Database for the Institutions and Agencies of the European Union. In *LREC Workshop on Terminology Resources and Computation,* Athens.

Piccioni, Lorenzo, Eros Zanchetta (2004). XTERM: A Flexible Standard-Compliant XML-based Termbase Management System. In *Proceedings of LREC, the IV International Conference on Language Resources and Evaluation.* Lisboa, pp. 469-473.

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

Vasiļjevs A., Skadiņš R. (2005). Eurotermbank terminology database and cooperation network. In *Proceedings of the Second Baltic Conference on Human Language Technologies,* Tallinn, pp. 347-352.

Wright, Sue Ellen (2005). A Guide to Terminological Data Categories – Extracting the Essentials from the Maze. In *Proceedings of TKE 2005, the 7th International Conference on Terminology and Knowledge Engineering*. Copenhagen, pp. 63-77.

(2005). *Final Methodology Report*, deliverable 1.2 University of Copenhagen.

(2005). *Implementation Specification,* deliverable 3.2, Tilde