NEALT PROCEEDINGS SERIES
VOL. 12

Proceedings of the NODALIDA 2011 workshop

CHAT 2011: Creation, Harmonization and Application of Terminology Resources

May 11, 2011
Riga, Latvia

Editors
Tatiana Gornostay and Andrejs Vasiļjevs

NORTHERN EUROPEAN ASSOCIATION FOR LANGUAGE TECHNOLOGY

Proceedings of the NODALIDA 2011 workshop

CHAT 2011: Creation, Harmonization and Application of Terminology Resources

NEALT Proceedings Series, Vol. 12

ISSN 1736-6305

Published by
Northern European Association for Language
Technology (NEALT)
http://omilia.uio.no/nealt

Electronically published at
Tartu University Library (Estonia)
http://dspace.utlib.ee/dspace/handle/10062/16956

Contents

Preface

The workshop on creation, harmonization and application of terminology resources, CHAT 2011, was held on May 11, 2011 at the University of Latvia, in Riga, Latvia. It was co-located with the 18th Nordic Conference of Computational Linguistics, NODALIDA 2011. The workshop focused on fostering the cooperation between EU projects and research and development activities in the area of terminology.

Consistent, harmonized and easily accessible terminology plays an extremely important role for ensuring true multilingualism in the European Union and throughout the world. In recent years different national and international activities have been undertaken to facilitate creation, accessibility and application of multilingual terminology resources. FP7 project TTC (*Terminology Extraction, Translation Tools and Comparable Corpora*, www.ttc-project.eu) researches novel methods how to extract multilingual terms from comparable corpora and integrate them in tools for machine translation, computer-assisted translation, and multilingual content management. Consolidation and harmonization of dispersed multilingual terminology resources is in the focus of elaboration of EuroTermBank (www.eurotermbank.com) terminology platform. FP7 Marie Curie project CLARA (clara.uib.no) has established international cooperation to involve new researchers in the terminology work and broader research on common language resources and their application.

Large scale activities are started by META-NET network (www.meta-net.eu) to create European Open Linguistic Infrastructure that will serve the needs of industry and research communities in various types of language resources. Its Baltic and Nordic branch META-NORD (CIP ICT-PSP project META-NORD, www.meta-nord.eu) is leading a work on integration of the terminology resources into this infrastructure.

We are delighted to hereby present the proceedings of CHAT 2011. Altogether, 11 papers were accepted for the presentation: 3 regular papers, 6 short papers, and 2 demonstration papers. The workshop papers cover various topics on automated approaches to terminology extraction and creation of terminology resources, compiling multilingual terminology, ensuring interoperability and harmonization of terminology resources, integrating these resources in language processing applications, distributing and sharing terminology data and others.

We are also pleased to present two invited speakers at the CHAT 2011. **Prof. Gerhard Budin** is a full professor for terminology studies and translation technologies at the University of Vienna, where he is a deputy director of the Centre for Translation Studies. He is also a director of the Institute for Corpus Linguistics and Text Technology at the Austrian Academy of Sciences and holds a UNESCO Chair for Cross-cultural, multilingual communication in the digital age. For the past 20 years he has been active in research and teaching in the fields of terminology management, specialized translation, corpus linguistics, language engineering, and philosophy of science. Multiple EU projects under his supervision deal with terminology resource development, cross-cultural eLearning, linguistic research infrastructures, eHumanities, translators' training, etc. At CHAT 2011 Prof. Gerhard Budin gave an invited talk on "Terminology Resource Development in Global Domain Communities – Practical Experiences, Case Studies and Conclusions for Future Projects".

Prof. Emmanuel Morin received his PhD degree and qualification for being full PhD adviser in computer science from the University of Nantes, France, in 1999 and 2007, respectively. He is a member of the Natural Language Processing team of the LINA laboratory (Laboratoire d'Informatique de Nantes-Atlantique, France). His research interests are multilingualism and multimodality and more specifically multilingual text mining, bilingual terminology extraction and on-line handwriting

recognition and categorization. He has published a number of scientific papers in the international journals and conference proceedings, including ACL, IJCNLP, ICDAR, ECML, ECIR, etc. At CHAT 2011 Prof. Emmanuel Morin gave an invited talk on "Bilingual Terminology Extraction from Comparable Corpora".

The organization of CHAT 2011 is a joint effort of several institutions, projects and their representatives. We would like to thank all Programme Committee members for fruitful collaboration during the preparation of the workshop and their time and attention during the review process. We would like to express our special gratitude to the workshop Organizing Committee – our colleagues from Tilde (Latvia), Norwegian school of Economics and Business Administration (Norway), the FP7 TTC project, the FP7 CLARA project, and the CIP ICT-PSP META-NORD project.

We hope that you will find these proceedings interesting, comprehensive and useful for your further research within the development of terminology resources and services of the future.

Tatiana Gornostay
Programme Committee Chair
CHAT 2011

Andrejs Vasiļjevs
Local Chair
CHAT 2011

Committees

PROGRAMME COMMITTEE & REVIEWERS

Tatiana Gornostay (Chair), Tilde, Latvia
Gisle Andersen, Norwegian school of Economics and Business Administration, Norway
Larisa Belyaeva, Herzen University, Russia
Béatrice Daille, University of Nantes, France
Patrick Drouin, University of Montreal, Canada
Judit Freixa, Universitat Pompeu Fabra, Spain
Marie-Paule Jacques, Stendhal University, France
Barbara Inge Karsch, BIKTerminology, ISO/TC 37 delegate, USA
Marita Kristiansen, Norwegian school of Economics and Business Administration, Norway
Inguna Skadiņa, TILDE, Institute of Mathematics and Computer Science, University of Latvia, Latvia
Koichi Takeuchi, Okayama University, Japan
Rita Temmerman, Erasmushogeschool Brussel, Belgium
Hanne Erdman Thomsen, Copenhagen Business School, Denmark
Andrejs Vasiļjevs, Tilde, Latvia

ORGANIZING COMMITTEE

Andrejs Vasiļjevs (Local Chair), Tilde, Latvia
Inguna Skadiņa (NODALIDA Local Chair), Tilde, Institute of Mathematics and Computer Science, University of Latvia, Latvia
Tatiana Gornostay, Tilde, Latvia
Gisle Andersen, Norwegian school of Economics and Business Administration, Norway
Marita Kristiansen, Norwegian school of Economics and Business Administration, Norway
Béatrice Daille, University of Nantes, France

WORKSHOP ORGANIZERS

Tilde, Latvia
Norwegian school of Economics and Business Administration, Norway
TTC project (FP7)
CLARA project (FP7)
META-NORD project (CIP ICT-PSP)

Workshop programme

CHAT 2011: Creation, Harmonization and Application of Terminology Resources

May 11, 2011

MORNING SESSION

9:00-9:30 **Opening –** *Welcome and workshop presentation*

9:30-11:00 **Invited presentations**

9:30-10:15	Prof. Gerhard Budin "*Terminology Resource Development in Global Domain Communities – Practical Experiences, Case Studies and Conclusions for Future Projects"*
10:15-11:00	Prof. Emmanuel Morin "*Bilingual Terminology Extraction from Comparable Corpora"*

11:00-11:20 Coffee break

11:20-13:00 **Paper Presentations**

11:20-11:40	Andrejs Vasiļjevs, Tatiana Gornostay and Inguna Skadiņa "*From Terminology Database to Platform for Terminology Services"*
11:40-12:00	Tine Lassen, Bodil Nistrup Madsen and Hanne Erdman Thomsen "*Automatic Knowledge Extraction and Knowledge Structuring for a National Term Bank"*
12:00-12:20	Fabien Poulard, Béatrice Daille, Christine Jacquin, Laura Monceaux, Emmanuel Morin and Helena Blancafort "*Comparability measurement for terminology extraction"*
12:20-12:40	Magnus Merkel and Henrik Nilsson "*Getting to terms with terminology at Swedish public agencies"*
12:40-13:00	Jody Foo *"Exploring termhood using language models"*

13:00-14:00 Lunch

AFTERNOON SESSION

14:00-15:20 **Paper Presentations**

14:00-14:20	Georg Löckinger "*User-Oriented Data Modelling in Terminography: State-of-the-Art Research on the Needs of Special Language Translators"*
14:20-14:40	Anna Savina and Elena Yagunova "*The Experimental Study of Terminology Collocations: Calculations and Experiments with Informants"*
14:40-15:00	Lorna Balkan, Taina Jääskeläinen, Christina Frentzou and Chryssa Kappi "*European Language Social Science Thesaurus (ELSST): issues in designing a multilingual tool for social science researchers"*
15:00-15:20	Dimitrios Kokkinakis "*Evaluating the Coverage of three Controlled Health Vocabularies with Focus on Findings, Signs & Symptoms"*

15:20-15:40 Coffee break

15:40-16:10 **Demonstration Presentations**

15:40-15:55	Tegau Andrews, Gruffudd Prys and Dewi Bryn Jones *"The Maes T System and its use in the Welsh-Medium Higher Education Terminology Project"*
15:55-16:10	Balázs Kis and Peter Reynolds *"Web based Terminology Management System and the Translation Market"*

16:10-16:50 **Discussion Session** – *Terminology resources and services of the future*

16:50-17:00 **Closing**

Invited presentations

Terminology Resource Development in Global Domain Communities – Practical Experiences, Case Studies and Conclusions for Future Projects
Prof. Gerhard Budin

Bilingual Terminology Extraction from Comparable Corpora
Prof. Emmanuel Morin

Regular papers

Comparability measurement for terminology extraction

Fabien Poulard and **Béatrice Daille**
Christine Jacquin and **Laura Monceaux**
Emmanuel Morin
Université de Nantes – LINA / UMR CNRS 6241
first.last@univ-nantes.fr

Helena Blancafort
Syllabs
blancafort@syllabs.com

Abstract

In this paper we describe recent work carried out in the context of the TTC project[1] towards the automatic construction of comparable corpora for multilingual terminology extraction. We focus on the communicative intention as the variable of discourse analysis that is best suited to select Web documents valuable for terminology applications and propose a classifier based on language independent features to automatically cluster crawled documents sharing the same communicative intention. The results of our experiments indicate the need to consider more sophisticated features.

1 Introduction

The notion of comparability for a corpus is still under construction. Comparable corpora are pairs (or more) of monolingual corpora which are not necessarily translations of each others but share some characteristics (domain, genre, topic...). The degree of comparability is perceived as the amount of these common characteristics: on one extremity, we find parallel corpora and on the other extremity the independent corpora wich have nothing in common (Prochasson, 2010). The choice of the common characteristics which define the content of corpus depends on its application task. For multilingual terminology extraction, the monolingual corpora must share an important part of the vocabulary in translated forms (Déjean and Gaussier, 2002). Documents domain (including the sub-domain and the topic), genre, audience, language register, communicative intentions are also characteristics of interest.

The TTC project (Terminology extraction Translation tools and Comparable corpora) aims at leveraging machine translation tools (MT tools), computer-assisted translation tools (CAT tools) and multilingual content management tools by automatically generating bilingual terminologies from comparable corpora in five European languages (English, French, German, Spanish and one under-resourced language, Latvian), as well as in Chinese and Russian. One key objective of the project is to automate methods for building comparable corpora in specialized domains from the Web. We focus on the lexical quality of the documents as we want to select documents embedding a rich terminology.

In this paper, we report our work regarding the development of a system to automaticaly classify crawled Web documents according to several characteristics in order to ensure the monolingual comparability of automaticaly compiled corpora.

First, we present various methods used to categorize Web documents according to their genre, their discourse type or their communicative intention. Then, we present a corpus we built for this study composed of documents in seven languages from five different families, as well as the terminology we observed within. Thereafter we discuss our proposition of a classifier for communicative intentions based on language independent features. We finally discuss the results of our experiments and conclude.

2 Categorizing Web Documents

Genre is one of the various variables of discourse analysis together with domain, register, document typology, document structure, etc. It is a "social type of communicative actions, characterized by a socially recognized communicative purpose and common aspect of form" (Crowston and Williams, 2000). Kessler et al. (1997) argue that the categorization of documents should not be trained on genres as atomic entities given their heigh volatility. Instead they propose a classification of gen-

[1] http://www.ttc-project.eu/

res as "generic facets" to distinguish "a class of texts that answers to certain practical interests, and which is associated with a characteristic set of computable structural or linguistic properties".

The genre is not the only characteristic to be considered to ensure monolingual comparability. The type of discourse (link between authors and audience, (Nakao et al., 2010; Ke and Zweigenbaum, 2009)) and the communicative intention may also be taken into consideration.

2.1 Webgenres

Deciding the genre of a Web document is a difficult task whether it must be done manually or automaticaly because the directory of webgenres is dynamic. Some genres are borrowed from traditional media, others derive from the formers, others again are emerging but are not yet well defined, others finally are spontaneous and have never been observed before. This evolutivity and the number of webgenres differenciates them from their traditional counterparts (Sharoff, 2011).

The attempts of automatic categorization of document in genre modelize the documents as "bags of words" (Dhillon et al., 2003) or combine dimension reduction (discriminative analysis, principal component analysis) and clustering (Poudat and Cleuziou, 2003) or classification (Cleuziou and Poudat, 2008). There has been several attemps to extend genre categorization to Web documents (Meyer-zu Eissen and Stein, 2004; Chaker and Habib, 2007; Dong et al., 2008; Mason, 2009; Waltinger et al., 2009). They usually combine various documents features with categorization algorithms based on machine learning techniques (support vector machines, clustering, neural networks...). Chaker and Habib (2007) group these features in four categories: metadata elements (URL, description, keywords...), presentation features (various HTML tags, links, images...), surface features (text statistics, function words, closed-class genre specific words, punctuation marks...) and structural features (parts-of-speech (POS), Tense of verbs...).

Experiments from Meyer-zu Eissen and Stein (2004) show that 70 % of the documents are assigned a correct genre.

2.2 Discourse

Goeuriot et al. (2008) have experimented the categorization of documents according to their type of discourse. They distinguished *scientific discourse* from *popular scientific discourse*. In the former, experts of a domain write for the same experts while in the latter experts or non experts write for non experts.

They propose a stylistic analysis on three levels implying deep linguistic analysis:

- The *structural* level consists of external criteria regarding the structure of the document and quantitative data (number of sentences and global size) ;
- The *modal* level consists of internal criteria caracterizing the position of the author in his writing. They considered allocutive [2] and elocutive modalities[3] inspired from Charaudeau (1992) ;
- The *lexical* level consists of internal criteria such as the presence of specific lexical units (specialized vocabulary, numbers, measure units), bibliographic elements, particular characters (brackets, other alphabet, symbols) and of quantitative data (size of the words, punctuation).

They obtain an average recall[4] of 87 % and an average precision[5] of 90 % for French documents and quite similar results for Russian (75 % recall and 87 % precision). The results on Japanese are lower with 46 % precision and 60 % recall.

2.3 Communicative intention

For Shepherd et al. (2004), the evolution of webgenres is also guided by the functional dimension of documents: browsing, emailing, searching, chatting, interacting, shopping, collaborating, etc. These communicative intentions may have a greater stability even if for annotators "the boundary between look'n'feel and communicative intentions is fuzzy" (Sharoff, 2011). Dong et al. (2008) consider the functionality of a Web document as part of its genre with its form and content. They associate for these three dimensions a particular kind of feature: stemmed terms for the content, HTML tags structuring the content (headings, tables, bullets...) for the form and HTML tags with

[2] Marks of the adressee presence.

[3] Marks of the author presence.

[4] Recall is a measure of completeness. It corresponds to the fraction of correct instances among all instances that actually belong to the relevant subset

[5] Precision is a measure of exactness. It corresponds to the correct instances among those that the algorithm believes to belong to the relevant subset.

content (applet, link, form...) for the functionality.

Sharoff (2011) experimented the classification of documents from the British National Corpus (BNC) according to their communicative intention (discussion, instruction, propaganda, recreation, regulation and reporting). He obtained an average precision of 83 % and an average recall of 80 %.

3 Corpus compilation

We built a multilingual corpus composed of German, English, Spanish, French, Latvian, Russian and Chinese Web documents. We present below our methodology to compile and annotate this corpus and its characteristics.

3.1 Crawled corpora

To compile the corpus, we used the first version of Babouk (de Groc, 2011), a focused web crawler (Chakrabarti et al., 1999) developed in the context of TTC to gather domain-specific corpora. To initialize the crawling, Babouk takes a list of seeds (terms or URLs) as input. During the first iteration of the crawling process, the given seeds are expanded to a large terminology using the BootCaT procedure (Baroni and Bernardini, 2004). Then, the generated lexicon is weighted automatically to build a thematic filter that is used by the categorizer in a second step to compute the relevance of webpages and filter non relevant documents. As a result, Babouk outputs a corpus consisiting of the retrived HTML files and two additional files for each HTML file:

- A Dublin Core[6] metadata file characterizing each crawled document retained for the corpus. It contains the file of the page, the seeds used for the crawling, the publisher, its original format as a mime-type, its geographic coverage, the language it is published in, the source url and the date of publication.
- A text file containing the plain text extracted from the corresponding web page.

To ensure the comparability of the corpus, we applied the same procedure to crawl the data using parallel term seeds (translation of seeds from English) in the domain of *wind energy*, a domain that is specific enough and for which corpora can be found on the web. Wind energy is one of the domains we deal with in TTC, as it is a new emerging domain for which little terminology resources exist. Other properties that may play a role in monolingual comparability, such as web genre, language register, authorship, communicative intentions and audience, are to be determined in a second step.

[6] http://dublincore.org/

3.2 Inter-annotator Campain for the Annotation in English

In addition to the files and metadata produced by the crawler, we annotated other document features whose values are detailed in Table 1:

- the webpage type (consistent with the set of web page values from Montesi and Navarrete (2008)) ;
- the communicative intentions (Sharoff, 2004; Sharoff et al., 2007) ;
- the authorship (Sharoff, 2004) ;
- the audience (Sharoff, 2004) ;
- and the language register (Goeuriot et al., 2008).

Before annotating documents in the various languages, an annotation campaign was organized on a common language (English). The various annotators annotated in three phases the same 120 texts in English. After each phase, the results were analyzed and the annotation guide (Monceaux et al., 2011) was updated to improve the annotation. We measured the inter-annotator agreement (IAA) with the Kappa measure (Fleiss and others, 1971) to evaluate the reliability of the annotations.

Table 2 synthesizes the IAA rates obtained by the end of the campaign. While the agreement is moderate or fair for most of the annotations, no sufficient interannotator agreement could be reached on the author audience characteristic. In consequence, this characteristic has not been annotated in the final annotation process. It has to be noted that we do not obtain excellent agreement for the various annotations which gives an idea of the difficulty of the task.

3.3 Corpus characteristics

The webpage type, communicative intentions, authorship and language register features have been manually assigned to around 200 texts for seven languages (German, English, Spanish, French,

Feature	Values
Webpage type	academic article, news article, adverts, legal text, expert report, report, guides, FAQs entries, catalog, glossary entries, announcement, encyclopedia entries, not text, blog entries, threads, homepages, reviews, warning, editorial, schedule, abstract, others
Communicative Intentions	information, discussion, instruction, list of something, regulation, promotion, reporting, unknown
Authorship	single author, multiple co-authors, corporate, unknown
Register	formal, informal

Table 1: Document features and their values as they are annotated on the corpus.

Annotation	Kappa	Interpretation
Web page type	0.472	Moderate agreement
Communicative Intentions	0.501	Moderate agreement
Authorship	0.513	Moderate agreement
Register	0.345	Fair agreement
Author Audience	0.097	Poor agreement

Table 2: Inter-annotator agreement for the annotated features measured with the Kappa measure and their interpretation.

Language	No. documents	No. words
German	200	285 286
English	210	209 150
Spanish	214	226 458
French	200	504 114
Latvian	225	388 098
Russian	193	318 966
Chinese	210	NA
	1 452	1 948 735

Table 3: Characteristics of the corpus.

Latvian, Russian and Chinese). These texts constitute our gold standard corpus.

Table 3 presents the main features of the corpus: the number of documents and the number of words for each language. This corpus is composed of almost two million words in seven languages. The texts have all been converted into utf-8 for convenience. Every document is stored in the corpus as an HTML file, a text file and an XML file containing the metadata and the annotations.

4 Corpus Analysis

After the corpus annotation task, we started to analyze the terminology that we extracted from the corpora. We observed a correlation between the kind of terminology and the communicative intentions.

The richest terminologies were found in the documents with informative, promotive and regulative intentions, each one whith a specific type of terminology. Informative documents i.e. show a rich technical terminology: *rotor bobiné*, *circuit rotorique* or even *multiplicateur de type planétaire épicycloïdal* for French, and *vertical axis turbines*, *Horizontal Axis Wind Turbines (HAWT)* or *Diffuser-Augmented Wind Turbines (DAWT)* for English.

The terminology of documents aiming at promotion make reference to products, such as named entities (name of products such as *Product Model:BF-H-500*), their constitutive element (*glass fiber reinforced plastic*) and their localizations (*parc éolien de Teterchen*).

As expected, documents aiming at regulation embed a legal terminology with terms such as *unacceptable harm*, *bienes inmuebles*, *impactos ambientales* and *planeamiento urbanístico*.

The documents with other communicative intentions show less numerous terms. Still, we found some terms in documents aiming at discussion, namely documents discussing the pros and contras of the installation of wind generators : *nuisances sonores* (*noise*) or *bruit mécanique* (*mecanical noise*).

Unfortunately, the various communicative intentions are not equally present and reachable on the Web as shows the Figure 1 representing their distribution among our corpora. Hence, discussion, information, reporting, promotion and list of something are the principal communicative intentions found in the corpus while regulation is mostly invisible. Therefore communicative intentions may be interesting features to choose documents relevant for terminology applications. They both allow the selection of documents with a rich terminology and enable to differentiate several kinds of terminology.

5 Classifying Web Documents Using Language Independent Features

We believe that the monolingual comparability of a corpus can be achieved by controling the domain and the communicative intention of the documents it is composed of. As we discussed in the previous section, it is possible to crawl documents belonging to the same domain. However, we do not have tools to predict the communicative intention of a document.

We face two main challenges to build a classifier for communicative intention in our context:

- we work with a relatively wide specialized domain with few resources and no scientific journals ;
- we must handle several distant languages with the same method and therefore are limited to features without any linguistic anchor.

5.1 Proposition

We propose to use supervised learning to predict the communicative intention of a document. Given the distant languages we deal with, we need a language independent method and therefore only use very shallow text features for the classification. Among the features experimented in the litterature, we selected the URL, the page layout, char ngrams and some other quantitative features.

We represent the URL as a bag of words by splitting it in sequences using special characters as delimiters (/, ., _, #, &...). The extracted sequences are normalized using unicode. For each document we obtain a vector of booleans indicating if any of the collected words is present in the URL of the document.

The page layout of the documents is constrained by the HTML tags. We compute the distribution, in terms of frequencies, of such tags. Preliminary experiments shown that it is preferable to only consider structuring tags (p, h1, ul, li...).

We also use bags of character ngrams. Like for URL we build vectors of booleans indicating if the associated ngram is present in the document. The best discrimination is offered by ngrams composed of four characters.

Finally, we used quantitative features such as the size of the document, the number of words[7] and their average size, the distribution of these words according to the unicode category of the characters they are composed of, the number and average size of sentences, ...

5.2 Experiments and results

We experimented two supervised learning approaches: a clustering one (k-Means) and a categorization one (SVM).

Using k-Means, we want documents to form cluster for each communicative intention. Therefore we compute a centroid for each communicative intention, using training data. Then communicative intention values are associated to documents depending on the centroid they are the closest to.

On the other side, SVM (Support Vector Machines) computes hyperplanes where the density of documents for each communicative intention is the highest while maximizing the margin between documents of different communicative intentions. Then communicative intention values are associated to documents depending on the hyperplane they belong to.

We experimented both learning algorithms with our language independent features. It results that the choice of the method has virtually no impact on the result and therefore we only present the results obtained with SVM in Table 4. A classifier is built for each language and evaluated with micro-precision, micro-recall and micro-f-score that is

[7] As we refuse the use of language specific tools, we consider as a word a sequence of characters sharing the same unicode category.

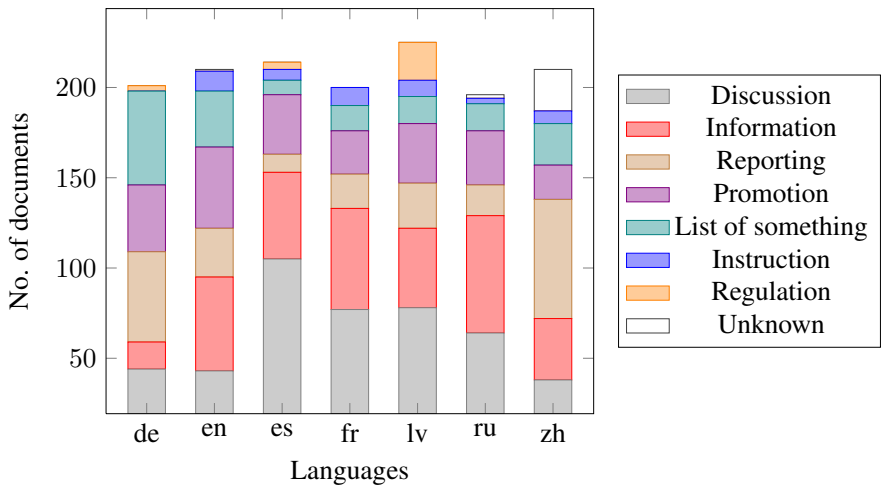

Figure 1: Distribution of the communicative intentions in terms of number of documents for each language composing our corpus.

Language	Precision	Recall	F1-score
English	25,2 %	25,8 %	13,3 %
French	39,8 %	39,8 %	24,9 %
German	6,8 %	25,8 %	10,8 %
Spanish	39,4 %	50,4 %	34,8 %
Latvian	52,2 %	41,0 %	30,6 %
Russian	32,2 %	33,4 %	20,8 %
Chinese	47,5 %	36,4 %	24,1 %

Table 4: Results obtained with SVM for each language.

the computation of these measures on the contingency table including all classes (all communicative intentions). As the various communicative intentions are not equaly distributed in the corpora, we run the evaluation with a 3-folds stratified cross-validation which preserve the same distribution of the communicative intentions among the various folds.

All the results are low which may indicate that the communicative intention is not language independent. The variations of the results between the languages mainly reflects the distribution of the communicative intentions among the documents as well as the lack of homogeneity between each monolingual corpus.

6 Conclusion

For comparable corpora extracted from the Web using a crawler for terminology oriented applications, it is important to categorize the documents with regards to terminology, named entities... Communicative intentions may be interesting features as they may allow to differentiate lexical items. Hence, informative documents should contain specific domain terminology, documents with promotion intentions should contain brand names, and regulative documents the legal terms.

In order to classify documents according to their communicative intention, in this paper we run an experiment with language independent features that seem relevant to other categorization tasks such as webgenre or discourse type. To classify documents written in seven languages belonging to five different families, we used features based on the URL, the page layout and characters ngrams. The experiments showed that these language independent features are not sufficient to distinguish communicative intentions.

More sophisticated features, including deeper linguistic features, should be considered and would require linguistic preprocessing. The best results on web genres classification make use of part-of-speech tagging while for discourse classifications very subtile features such as modality marks are used. Sharoff (2011) obtained better results in classifying English and Russian documents according to their communicative intentions using deeper linguistic features.

Another consideration is that maybe our hypothesis that the classification should be placed between the crawl process and the terminology extraction is not valid after all. Terminology may be necessary to predict the communicative intention

and not the other way around.

Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement no°248005.

We thank all the annotators (Ana Laguna, Maya Abboud, Somara Seng, Tatiana Gornostay, Iveta Keiša and others) for the time they have spent on the construction of the corpus as well as Jérôme Rocheteau for the development of the handy annotation tool OUSIA[8].

References

Marco Baroni and Silvia Bernardini, 2004. *BootCaT: Bootstrapping corpora and terms from the web*, volume 4.

Jebari Chaker and Ounelli Habib. 2007. Genre categorization of web pages. In *Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007)*, pages 455–464, Oct.

Soumen Chakrabarti, Martin Van den Berg, and Byron Dom. 1999. Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks*, 31(11-16):1623–1640.

Patrick Charaudeau. 1992. *Grammaire du sens et de l'expression*. Hachette.

Guillaume Cleuziou and Céline Poudat. 2008. Classification de textes en domaines et en genres en combinant morphosyntaxe et lexique. In *Actes de TALN*, number 1, pages 9–13. ATALA.

Kevin Crowston and Marie Williams. 2000. Reproduced and emergent genres of communication on the world-wide web. *The Information Society*, 16(3):201–216.

Clément de Groc. 2011. Babouk - exploration orientée du web pour la constitution de corpus et de terminologies. In *22es Journées francophones d'Ingénierie des Connaissances (IC'2011)*, May.

Inderjit S. Dhillon, Subramanyam Mallela, and Rahul Kumar. 2003. A divisive information-theoretic feature clustering algorithm for text classification. *Journal of Machine Learning Research*, 3(7-8):1265–1287, Oct.

Lei Dong, Carolyn Watters, Jack Duffy, and Michael Shepherd, 2008. *An Examination of Genre Attributes for Web Page Classification*, pages 133–143. IEEE Computer Society, Jan.

[8] `http://code.google.com/p/dublin-core-ousia/`

Hervé Déjean and Éric Gaussier. 2002. Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica, Alignement lexical dans les corpus multilingues*, pages 1–22.

J.L. Fleiss et al. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Lorraine Goeuriot, Natalia Grabar, and Béatrice Daille. 2008. Characterization of scientific and popular science discourse in french, japanese and russian. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008)*, number 1, pages 2933–2937. European Language Resources Association (ELRA).

Guiyao Ke and Pierre Zweigenbaum, 2009. *Catégorisation automatique de pages web chinoises*, pages 203–228. ARIA.

Brett Kessler, Geoffrey Nunberg, and Hinrich Schütze. 1997. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*.

Jane E. Mason. 2009. *An n-gram based approach to the automatic classification of web pages by genre*. Ph.D. thesis, Dalhousie University.

Sven Meyer-zu Eissen and Benno Stein. 2004. Genre classification of web pages. In S. Biundo, T. Fruhwirth, and G.Editors Palm, editors, *Advances in Artificial Intelligence (KI 2004)*, pages 256–269. Springer, Berlin Hedelberg New York.

Laura Monceaux, Christine Jacquin, and Béatrice Daille. 2011. Guidelines for monolingual annotation.

Michela Montesi and Trilce Navarrete. 2008. Classifying web genres in context: A case study documenting the web genres used by a software engineer. *Inf. Process. Manage.*, 44:1410–1430, July.

Yukie Nakao, Lorraine Goeuriot, and Béatrice Daille. 2010. Multilingual modalities for specialized languages. *Terminology*, 16(1):51–76, May.

Céline Poudat and Guillaume Cleuziou. 2003. Genre and domain processing in an information retrieval perspective. In *Proceedings of the 3rd International Conference on Web Engineering (ICWE 2003)*, pages 399–402. Springer-Verlag Berlin Heidelberg.

Emmanuel Prochasson. 2010. Alignement multilingue en corpus comparables spécialisés.

Serge Sharoff, Bogdan Babych, and Anthony Hartley. 2007. "Irrefragable answers" using comparable corpora to retrieve translation equivalents. *Language Resources and Evaluation*, 43(1):15–25.

Serge Sharoff. 2004. Analysing similarities and differences between corpora. In *Proceedings of the 7th Conference of Language Technologies (Jezikovne Tehnologije)*, volume 83.

Serge Sharoff, 2011. *Chapter 7 - In the Garden and in the Jungle*, volume 42. Springer Netherlands.

Michael Shepherd, Carolyn Watters, and Alistair Kennedy. 2004. Cybergenre : Automatic identification of home pages on the web. *Journal of Web Engineering*, 3(3-4):236–251.

Ulli Waltinger, Alexander Mehler, and Armin Wegner. 2009. A two-level approach to web genre classification. In Joaquim Filipe and JoséEditors Cordeiro, editors, *Proceedings of the 5th International Conference on Web Information Systems and Technologies (WEBIST 2009)*, pages 689–692. INSTICC Press.

European Language Social Science Thesaurus (ELSST): issues in designing a multilingual tool for social science researchers

Lorna Balkan
UK Data Archive
University of Essex
Colchester, UK
balka@essex.ac.uk

Taina Jääskeläinen
FSD Finnish Social Science
Data Archive
University of Tampere, Finland
taina.jaaskelainen@uta.fi

Christina Frentzou
EKKE National Centre for
Social Research
Athens, Greece
cfredzu@ekke.gr

Chryssa Kappi
EKKE National Centre for
Social Research
Athens, Greece
ckappi@ekke.gr

Abstract

This paper describes the methodology used to produce the European Language Social Science Thesaurus (ELSST), which has been in development for over decade, supported by a succession of EU-funded projects. Currently available in nine languages, ELSST aims to improve access to comparable social science and humanities data across geography and time. Its design is such, however, that it lends itself both as an information retrieval tool and as a terminological tool more generally.

1 Introduction

Access to good quality data in the social sciences is essential for social and economic policy makers and researchers, and in the European context, this includes in particular access to comparable data across geography and time. The Council of European Social Science Data Archives (CESSDA) operates a data portal which gives access to the data collections of its member states with the aid of a purpose-built multilingual thesaurus. This thesaurus, the European Language Social Science Thesaurus (ELSST), which has been developed over the last ten years and which currently contains nine languages[1], permits users to search for comparable data across different populations using a search term in their own language. There are currently over 3,000 terms for the majority of languages in the thesaurus. This paper explores some of the issues involved in its design and development.

2 Background

Development of ELSST has proceeded under three successive EU-funded projects, namely: Language Independent Metadata Browsing of European Resources (LIMBER), 2000-2003 (Miller and Matthews, 2001); Multilingual Access to the Data Infrastructure of the European Research Areas (MADIERA), 2003-2005; and Council of European Social Sciences (CESSDA)-Preparatory Phase Project (PPP), 2008-2010.

ELSST was initially derived from Humanities and Social Science Electronic Thesaurus (HASSET)[2], the English monolingual thesaurus created by the UK Data Archive, the social science data archive at the University of Essex. Higher level terms from the main HASSET hierarchies were selected in order to arrive at a broader-level, more 'Euroversal' thesaurus, which, it was hoped, would avoid any language or cultural bias. This first phase of ELSST as described in

[1] Lithuanian terms are due to be added to ELSST in spring 2011.

[2] HASSET [5] was originally based on the 1977 UNESCO Thesaurus, ISBN 92-3-101469-2.

Balkan et al. (2002) was confined to English, French, German and Spanish.

In the second phase of ELSST, under MADIERA, four new languages Danish, Finnish, Greek and Norwegian were added[3] and a new methodology introduced. Prior to finding multilingual equivalents to terms, hierarchies were reviewed by a multilingual and multicultural team, and subject experts consulted. Definitions were added to terms where necessary, in order to eliminate further the language and cultural bias inherited from HASSET.

In the latest phase of ELSST, under CESSDA-PPP, a number of hierarchies were amended and enlarged. Earlier translation work had revealed particular difficulties with certain hierarchies, especially education, labour, employment, social welfare and social structure, due mainly to the different systems found in different countries. One solution adopted was to align ELSST terms with international classification systems to deal with these problems.

During CESSDA-PPP maintenance and management procedures were also created, as well as a thesaurus management system.

3 Creating a multilingual thesaurus: the challenges

The first challenge for ELSST lies in the diversity of languages it contains. The second phase of ELSST included the introduction of Finnish and Greek, neither of which belong to the same family as the original ELSST languages (i.e. Romance and Germanic). Finnish in particular is less related to, and has fewer cognates with, the other ELSST languages. While this sometimes makes it more difficult to find Finnish equivalent terms, it avoids the temptation of employing 'false friends', as reported in Jääskeläinen (2006).

A fundamental problem for multilingual thesauri, or for any multilingual language resources, is not only linguistic variation between languages but the fact that different languages have different ways of classifying the world. One language may choose to lexicalise a concept that is lacking in another. Often this is due to cultural differences. For example, Greek has no word for 'house husbands'. Even within the same language (e.g. German), there may be differences in concepts/lexicalisations due to differences in cultural systems such as education and legal systems which may differ between countries and regions. A multilingual thesaurus has to take account of these problems.

Another challenge for ELSST is due to its subject domain, i.e. social sciences. Social science vocabulary has a certain amount of 'hard' terms, i.e. terms which can be precisely defined (e.g. geographical regions), but in the main consists of 'soft' terms, which are much vaguer in scope and which share some overlap with general language. Social science vocabulary thus contrasts with the terminology of the physical sciences, which have a greater proportion of 'hard' terms. Moreover, the meaning of social science terms may vary not just across geographical or cultural boundaries, but across time. An example is 'old age', which means something different today than it did 100 or even 50 years ago[4].

4 Structure and function of a multilingual thesaurus

A thesaurus addresses the problem of vagueness of meaning, in that it is a controlled vocabulary. It consists of a hierarchical arrangement of ('preferred') terms, which express concepts. Terms are intended to express one and only one concept. The relationships between terms are explicitly marked. The hierarchical relationship is the Broader Term (BT) relationship and its inverse Narrower Term (NT). Non-hierarchical relationships include the Used For (UF) relationship, typically synonyms or near synonyms, or antonyms, lexical variants, etc; and the Related Term (RT) relationship, which expresses a looser association to the main 'preferred' term than the BT relationship.

Thesaurus relationships serve several purposes. First, together with the terms they link, they provide a roadmap to the conceptual space of the domain. This can be useful to information seekers who wish to get an overview of the domain or subdomain(s). Second, relationships such as BTs, NTs and RTs can suggest alternative search terms for those using the thesaurus as an information retrieval tool, allowing them to widen or narrow their search. Third, while the relationships between terms in a thesaurus are made explicit, the meanings of the individual terms are frequently only implied, either from their UFs, or from their place in the thesaurus.

[3] Swedish was also added to ELSST at this stage, though not under EU funding.

[4] This is attributable to the nature of the adjective 'old' which is comparative, rather than absolute in value.

Thus ‘courts’ in general language may have several meanings, but its position as an NT to ‘administration of justice’ in ELSST narrows its meaning to legal courts. The definition of a term may also be made precise through the use of a Scope Note (SN). Thus ‘bills’ in general language can have at least two meanings - ‘printed or written statements of the money owed for goods or services’, or ‘proposals for legislation which, if adopted by Parliament, become statutes’. In ELSST, only the second meaning is possible, as the term is assigned an explicit scope note to this effect.

The less ambiguous a term, the more precise it is as an information retrieval tool. For example, if researchers use the ELSST term ‘bills’ to search a database, they will know that the list of documents retrieved will be about legal bills and not any other kind. Contrast this with a free text search, where searching for a term equates to searching for a string, not the concept behind the string, and where the search term ‘bills’ will retrieve instances of any use of the word ‘bills’ not all of which will be relevant.

A third type of non-hierarchical relationship, the equivalence relationship, is found only in multilingual thesauri. This is the relationship which links a term to its foreign language equivalent(s) in the thesaurus. Note that in ELSST, equivalence relationships are always defined relative to the English source term. Given the different ways in which different languages lexicalise concepts, the equivalence relationship may be quite complex, ranging from complete equivalence (where two terms express exactly the same concept) to non-equivalence (where there is no equivalent concept at all in one of the two languages). Five different levels of multilingual equivalence are defined in ELSST, based on Guidelines for Multilingual Thesauri of the International Federation of Library Associations and Institutions:

1. Exact equivalence: source language (SL) and target language (TL) terms refer to the same concept.

2. Inexact or near equivalence: SL and TL terms are generally regarded as expressing the same general concept but the meanings of the terms in SL and TL are not exactly identical. Often the differences are more cultural than semantic, i.e. there is a difference in connotation or appreciation.

3. Partial equivalence: SL and TL terms are generally regarded as referring to the same concept, but one of the terms strictly denotes a slightly broader or narrower concept.

4. One-to-many equivalence: to express the meaning of the preferred term in the SL, two or more preferred terms are needed in the other language.

5. Non-equivalence: No existing term with an equivalent meaning is available in the TL for a concept in the SL, for cultural or linguistic reasons.

It should be noted that ELSST does not aspire to represent all social science concepts, merely those relevant to the existing data collections of the participating archives. Similarly, no formal logic underpins the relations between these concepts - relations such as subtype-supertype or part-whole determine the positions of the concepts in a hierarchy but do not completely define them. Thus, to use Sowa’s (Sowa 1999) terminology, ELSST can be described as a ‘terminological ontology’ rather than a formal ontology.

5 Bridging lexical and conceptual differences across languages

A central problem for multilingual thesauri construction is how to deal with these different types of equivalence relationships between concepts.

Inexact or near equivalence is treated as exact equivalence in ELSST. This is no different in essence to the relationship between a preferred term and its synonyms or near synonyms in the monolingual thesaurus.

Partial equivalence has received different treatments in ELSST. In some cases a BT or NT can be chosen instead. For example, the English term ‘paramedical personnel’ which means persons who work in ambulances and who are trained in first aid, emergency care etc, is mapped to the Finnish term ‘ensihoitohenkilöstö’. The Finnish term is broader in scope, covering, in addition to persons working in ambulances, also those working in emergency care units. In other cases where the meaning diver-

gences are due to culture-specific reasons, and where international classification schemes exist, efforts have been made to import them into ELSST. This is particularly the case for terminologies referring to systems, such as the education, legal or health care system. For example, the International Standard Classification of Education 1997 (ISCED97) was consulted for terms for educational systems and levels. While they offered useful generic terms to describe concepts (e.g. lower secondary schools) they do not necessarily correspond to terms that information seekers would use to search for documents. They thus need to be augmented with country or region-specific UFs (e.g. 'yläkoulut' in Finnish[5] and 'collèges' in French).

An example of single-to-multiple equivalence is the translation of the term 'housewives' into Finnish. The concept of housewives can only be represented by two different concepts in Finnish: 1) 'Kotiäidit' (literally translated 'stay-at-home-mothers' and 2) Kotirouvat (literally 'stay-at-home-ladies'). There is no neutral equivalent of housewives. The two Finnish terms have their own connotations: the first refers to wives staying at home to take care of children (implied by 'mothers') and the second, now becoming old-fashioned, that the family is well-off (implied by 'ladies'). Working class families would not normally have a 'kotirouva'. In ELSST, the equivalence was handled by creating a synthetic term, KOTIÄIDIT JA KOTIROUVAT, which consists of Kotiäidit and Kotirouvat conjoined by 'JA' ('and').

For cases of non-equivalence between languages, several strategies are possible including:

(1) disallow a concept if it does not exist in one or more of the thesaurus languages;
(2) allow the definition of a concept to exist in the thesaurus, without lexicalising it;
(3) adopt a loan word or some other artificial construct as its equivalent.

Strategy (1) is overly restrictive and not an option in ELSST. Similarly (2) is excluded since the structure of each language hierarchy (excluding the number of UFs, which can vary according to language) is identical in ELSST, and every preferred term has to have an equivalent in each of the other languages. Strategy (3) is adopted in ELSST. For example, the concept of 'travelling people' has no equivalence in Finnish, so is mapped to the English term 'travelling people'. From the information retrieval point of view, this is adequate, because a searcher will not be able to find Finnish data about 'travelling people' anyway, since the concept does not exist in Finland.

A novel approach to equivalence problems in ELSST is to adopt a special kind of scope note called the Translation scope note. Thus the case of the difference between 'paramedical personnel' in English and 'ensihoitohenkilöstö in Finnish is explained with the translation scope note both in English: 'The Finnish term covers all personnel with emergency care training working in ambulances or emergency care units', and in Finnish: ' Englantilainen termi kattaa vain ambulansseissa työskentelevät' (the Finnish SN says 'The English term covers only those working in ambulances').

6 Conclusion

Some of the challenges encountered in constructing ELSST stem from the fact that it was derived from an existing monolingual thesaurus, rather than being constructed from scratch (a preferable, but costlier option). The biggest problem is the lack of definitions associated with source terms. It has been necessary to add many more scope notes to the English source terms in ELSST before equivalence relationships could be established.

Another problem is that although discussing and amending English terms and hierarchies in a multilingual and multicultural terms in advance of seeking their multilingual equivalents helps to reduce language and cultural bias, this is not enough for hierarchies describing systems. In this case, there is no alternative to starting from scratch, preferably using international standard classifications and existing thesauri.

Ultimately, it is impossible to eliminate all concept mismatches due to the inherent differences in the way that different languages lexicalise concepts. However, for the information seeker, partial equivalence will in most cases still retrieve relevant data, which is the main purpose of a thesaurus. It is hoped that by adding scope notes, including translation scope notes, these different levels of equivalence will be better understood by the users of the thesaurus, thus enhancing the usefulness of ELSST both as an information tool and as a terminological aid.

[5] This work is currently ongoing and these terms are not yet available on the publicly available version of ELSST.

ELSST is currently available for the general public to view at the following web page: http://elsst.esds.ac.uk/login.aspx. It is envisioned that publicly funded bodies such as university libraries will in future be able to obtain a licence for ELSST, which will allow them to use the thesaurus as an indexing and search tool in their local systems. Anyone wishing more information on ELSST should contact Sharon Bolton at sharonb@essex.ac.uk.

Acknowledgments

We would like to thank all CESSDA archive colleagues who have participated in the construction of ELSST.

References

Lorna Balkan, Ken Miller, Birgit Austin, Anne Etheridge, Myriam Garcia Bernabé and Pamela Miller. 2002. 'ELSST: a broad-based Multilingual Thesaurus for the Social Sciences', proceedings of Third International Conference on Language Resources and Evaluation, Las Palmas.

CESSDA-PPP project, http://www.cessda.org/project/

Council of European Social Sciences (CESSDA) http://www.cessda.org/

International Federation of Library Associations and Institutions Working group on Guidelines for Multilingual Thesauri. 2009. *Guidelines for Multilingual Thesauri*, IFLA Professional Reports 115, The Hague, ISBN 978-90-77897-35-5.

Humanities and Social Science Electronic Thesaurus (HASSET) http://www.esds.ac.uk/search/hassetSearch.asp

Taina Jääskeläinen. 2006. 'Meeting the challenge of a multilingual thesaurus'. Presented at the conference Multilingual Thesauri in Social Sciences, Helsinki.

Ken Miller and Brian Matthews. 2001. Having the right connections: the LIMBER project, *Journal of Digital Information*, 1(8).

Multilingual Access to the Data Infrastructure of the European Research Areas (Madiera) project, http://www.dataarchive.ac.uk/about/projects/past?id=1633

John F. Sowa. 1999. *Knowledge Representation: Logical, Philosophical and Computational Foundations.* Brooks Cole Publishing, Co., Pacific Grove, CA.

Standard Classification of Education 1997 (ISCED97),http://www.unesco.org/education/information/nfsunesco/doc/isced_1997.htm

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.

Regular short papers

Automatic Knowledge Extraction and Knowledge Structuring for a National Term Bank

Tine Lassen
Copenhagen Business School,
Denmark
tla.isv@cbs.dk

Bodil Nistrup Madsen
Copenhagen Business School,
Denmark
bnm.isv@cbs.dk

Hanne Erdman Thomsen
Copenhagen Business School,
Denmark
het.isv@cbs.dk

Abstract

This paper gives an introduction to the plans and ongoing work in a project, the aim of which is to develop methods for automatic knowledge extraction and automatic construction and updating of ontologies. The project also aims at developing methods for automatic merging of terminological data from various existing sources, as well as methods for target group oriented knowledge dissemination. In this paper, we mainly focus on the plans for automatic knowledge extraction and knowledge structuring that will result in ontologies for a national term bank.

1 Introduction

If a term bank does not contain a sufficient number of terms, users will not feel encouraged to use it, and on the other hand, users will be frustrated if a term bank contains a large amount of terms with only little or poor quality information. Therefore it is necessary to use automatic procedures in order to extract and systematize information about terms, and the high quality that can be obtained by hand crafting the contents and the large volume that can be obtained by reusing terminology data from existing sources of varying quality must somehow be combined. One way of increasing the amount of terms in a term bank is to extract terms and information about terms automatically from texts. Another method is to merge terminology from different sources, such as other term banks or existing term lists. However, this approach will often lead to problems, since the term bank will typically contain many entries connected to the same term, but with varying formulation of the definitions and/or different translations. In order to clarify and distinguish the meanings of domain specific concepts, these must be described by means of characteristics and relations to other concepts, i.e. in the form of domain specific ontologies (or concept systems). On the basis of such ontologies, it is possible to develop consistent definitions that further the understanding and correct use of terms. Terminology work that includes development of ontologies is, however, a very labor-intensive task, and therefore most term banks do not include ontologies.

This paper describes our plans for automatic extraction of terms and information about terms as well as the automatic construction of ontologies on the basis of the extracted information. At present we have developed a prototype for retrieving relevant texts. We will describe this briefly in section 3.1.

Another goal of the project is to develop methods for automatic merging of terminological data from various existing sources; a problem that existing term banks have not solved adequately. The project also aims at developing methods for automatic construction of ontologies on the basis of definitions from the various data sources and methods for automatic merging of entries based on the merging of these ontologies.

Finally the project aims at developing methods for target group oriented knowledge dissemination. Many other term banks only offer restricted possibilities for setting up user specific search and presentation profiles.

As an introduction to the description of the current project we present some central concepts related to terminological ontologies.

2 Central concepts related to terminological ontologies

The backbone of terminological concept modelling is constituted by characteristics modelled by formal feature specifications, i.e. attribute-value

pairs. The use of feature specifications is subject to principles and constraints described in detail by Madsen, Thomsen, & Vikner (2004). Subdivision criteria, which have been used for many years in terminology work, were formalised by introducing dimensions and dimension specifications. A dimension of a concept is an attribute occurring in a (non-inherited) feature specification of one or more of its subordinate concepts. A dimension specification consists of a dimension and the values associated with the corresponding attribute in the feature specifications of the subordinate concepts: DIMENSION: [value1| value2| ...].

3 Subprojects

The current term bank project consists of three main subprojects: 1) Knowledge acquisition, 2) Knowledge structuring and 3) Knowledge dissemination. Figure 1 gives an overview of the project and its three subprojects as well as the processes involved. In subproject 1) Knowledge acquisition methods for a) automatic knowledge extraction and b) automatic merging and quality assurance of data are to be developed. Below, the three subprojects are briefly described.

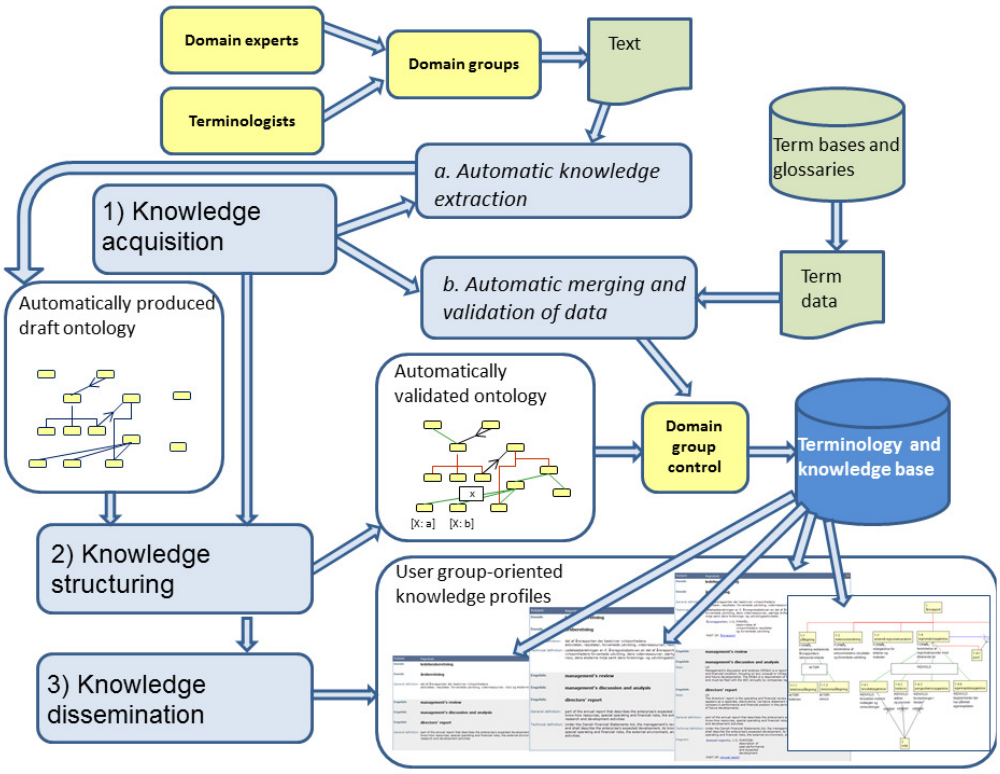

Figure 1 Outline of the project and its subprojects

3.1 Knowledge acquisition

The primary aim of the subproject 'Knowledge acquisition' is to develop new advanced models of and methods for automatic extraction of concepts and information about concepts. We develop a prototype which, on the basis of an existing domain-specific text corpus or domain texts automatically collected from the Internet, can automatically extract terms and relations and produce a draft version of a terminological ontology. The draft ontologies will contain subdivision criteria and characteristics as formal feature specifications on concepts.

One of the main ideas in this subproject is to investigate how to put together and make use of groups of domain experts, who together with terminologists in so-called domain groups (cf. figure 1) contribute to the collection of knowledge as well as to conceptual clarification.

Tools for knowledge extraction will be implemented and integrated into an interactive interface where domain experts can upload texts into a text corpus, and methods to automatically analyze these texts with respect to their (estimated) level of explicit knowledge, term density and

other LSP features (cf. e.g. Barrière, 2006 and Halskov, Braasch, Haltrup Hansen & Olsen, 2010) will be investigated.

Corpus texts will also be collected from the Internet by application of text classification algorithms. At present, in our prototype, we apply a bootstrapping algorithm, cf. BootCat (Baroni & Bernardini, 2004), where first, a small number of exemplary texts from the given domain are analyzed by applying selected statistic scores, and as a result a set of domain specific wordings or term candidates is produced. we apply co-occurrence scores, e.g. Pointwise Mutual Information (Church & Hanks, 1993) and Dice coefficient (Smadja, 1993), as well as 'termhood' scores, e.g. Log Odds Ratio (cf. e.g. Everitt, 1992) and weirdness (Ahmad et al., 1999), on n-grams, and produce a set of domain specific terms and other types of domain specific language usage that can either be the union or the intersection of the sets of term candidates produced by applying each statistic score. This set is then used as search terms, and a new collection of domain texts retrieved. The analysis and search process is iterated a number of times, until a satisfactory corpus is compiled. The definition of 'a satisfactory corpus' is still being investigated.

Another aim of this subproject is to develop methods for converting and combining terminology data from various existing sources. Two very complex types of problems exist in this process. The first type of problems that are likely to be encountered pertains to form: The data are likely to have different structures and be stored in different formats. The second type of problems pertains to content: The data may be of varying quality, and entries from the various resources may contain information about the same concept, but be associated with different sets of synonyms and with slightly varying definitions, or the other way round, have overlapping form but be associated with different concepts. Therefore, the aim of subproject 1) is also to do research in automatic ontology construction on the basis of existing term collections, and to develop methods for merging and quality assurance of term data from different sources.

3.2 Knowledge structuring

The aim of the subproject 'Knowledge structuring' is to develop methods and a prototype that may be used for automatic validation and dynamic expansion of the draft ontologies that result from the automatic knowledge extraction.

As mentioned above in section 3.1, the draft terminological ontologies will contain subdivision criteria and characteristics as formal feature specifications on concepts. This information can be used in the automatic validation of the draft ontologies: For example, if the draft ontology contains two given concepts that have been placed in a direct type relation, but where the feature specifications imply that a concept should in fact exist between them, the system can introduce a dummy concept in order to make the ontology valid. Afterwards, a domain expert must re-validate the ontology and fill in actual concepts in place of the introduced dummy concepts.

The validation process will require changes to be made in the ontology, and for this process to be performed automatically, we will develop techniques for automatic classification of concepts into ontologies with type relations based on the feature specifications that have been identified for a given concept.

Prior research distinguishes between characteristic features and conceptual relations (Madsen, Thomsen & Vikner, 2004). In the knowledge acquisition prototype, which will be developed during project subpart 1, no distinction will be made between attributes and relations per se, but all associative relations will be recorded as attribute-value pairs. For any given concept, a given characteristic feature may either be represented as a feature specification or as a relation to another concept. In a small terminology project, concepts outside the narrow domain will typically not be included in the ontology, but only exist as values of feature specifications, but if these concepts are relevant to the description of the domain, they may be included as concepts in the ontology. The project will develop new theories for distinguishing between characteristics and related concepts based on how central the values are in the given domain.

Other problems that the project will treat are multiple values and hierarchically typed values:

The knowledge acquisition prototype will potentially describe concepts with more than one (identical) relation to other concepts. However, some relations exist that can only occur once in connection with a given concept; for instance, no concept can have more than one instance of the relation HAS_LENGTH. This corresponds to the principle that a concept can have at most one value for a given attribute. Therefore, in order to facilitate ontology validation, we will develop methods for distinguishing between relations that

can only occur once, and relations that can occur several times in connection with a concept.

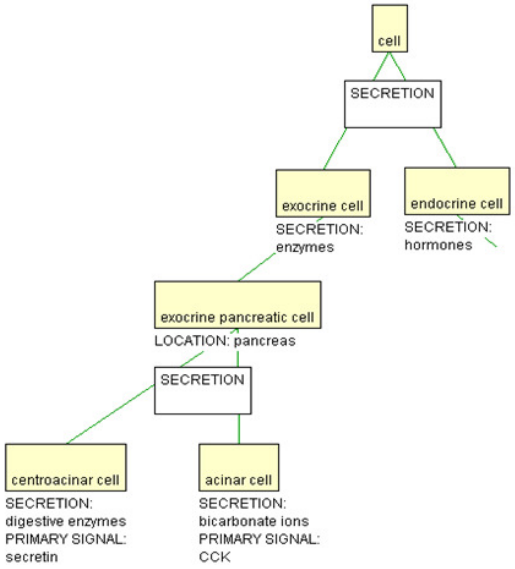

Figure 2 Excerpt of a cell ontology

In the ontology excerpt shown in figure 2, the concept *cell* is subdivided into *exocrine cell* and *endocrine cell*, based on the subdividing criterion SECRETION. The concept *centroacinar cell* inherits the feature [SECRETION:enzymes] from *exocrine cell*, but is already specified with the feature [SECRETION:digestive enzymes]. In this case, it can be argued that the value is a specialization of the inherited value, and therefore there is no conflict. To handle this, we suggest to apply a type hierarchy of values. This approach builds on the methods implemented in e.g. the Lexical Knowledge Base system (LKB) (Copestake, 1992) for use in lexical semantics.

3.3 Knowledge dissemination

The subproject 'Knowledge dissemination' will focus on presentation of data in the term bank. Traditionally, terminology and lexicography have been separate research fields with different approaches to compilation and presentation of data. However modern technology offers unlimited opportunities to meet the needs of several target groups in one database by offering the possibility of choosing between different presentations. The overall objectives of this subproject are to discuss and specify the extent to which the traditional lexicographical and terminological methods may be fruitfully combined, allowing the presentation of concepts in one single database thereby contributing added value for a defined user group, and how a combination of the two research fields may create further opportunities towards developing principles for target-group oriented knowledge transfer.

4 Conclusions

A distinctive feature of our approach includes the automatic extraction of concepts and associative relations, which can be formalised as feature specifications. The ontologies will be based on the principles for terminological ontologies as described above. No other methods or systems exist for automatic construction and consistency checking of terminological ontologies that comprise subdivision criteria and dimension specifications, which are crucial in the development of such ontologies.

References

Ahmad, K., L. Gillam and L. Tostevin. 1999. University of surrey participation in TREC8: Weirdness-indexing for logical document extrapolation and retrieval (WILDER). In: *The Eighth Text REtrieval Conference (TREC-8)*.

Baroni, M. and S. Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the web. *Proceedings of LREC 2004*.

Barrière, C. (2006). TerminoWeb: A Software Environment for Term Study in Rich Contexts. *International Conference on Terminology, Standardisation and Technology Transfer (TSTT 2006).* Beijing.

Church, K. W. and P. Hanks. 1990.Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Copestake, A. 1992. *The Representation of Lexical Semantic Information.* Doctoral dissertation, University of Sussex.

Everitt, B. 1992. *The Analysis of Contingency Tables.* Chapman & Hall/CRC Monographs on Statistics & Applied Probability), 2nd edition.

Halskov, J., Braasch, A., Haltrup Hansen, D., and S. Olsen. 2010. Quality indicators of LSP texts – selection and measurements. How to measure the terminological usefulness of a document from a particular domain in the task of compiling an LSP corpus. *Proceedings from LREC.* Malta.

Madsen, B. N., Thomsen, H. E., and C. Vikner. 2004. Principles of a system for terminological concept modelling. *Proceedings of the 4th International Conference on Language Resources and Evaluation*, *Vol. I*, pp. 15-18. Lisbon.

Smadja, F. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143-177.

Evaluating the Coverage of three Controlled Health Vocabularies with Focus on Findings, Signs & Symptoms

Dimitrios Kokkinakis
Dep. of Swedish (Språkbanken) & Centre for Language Technology (CLT)
University of Gothenburg, Sweden
dimitrios.kokkinakis@svenska.gu.se

Abstract

The medical domain is blessed with a magnitude of terminological resources of various characteristics, sizes, structure, depth and breadth of descriptive power, granularity etc. In this domain a particularly interesting and difficult entity type are signs, symptoms and findings which to a large extend are expressed in a periphrastic manner, sometimes by the use of figurative or metaphorical language, or contextualized using a wealth of vague variant expressions. We hypothesize therefore that no major official terminology source alone can accommodate for the variation and complexity present in real text data, such as electronic medical records, notes or health related documents. In this paper we evaluate the content of the three largest medical control vocabularies available for Swedish on extracted reference symptom lists and initiate a discussion on how we should proceed in order to accommodate for increased coverage on similar genres.

1 Introduction

The medical domain is blessed with a magnitude of terminological resources of various characteristics, sizes, structure, depth and breadth of descriptive power, granularity etc. This paper deals with a first attempt to investigate, understand and in the future harmonize large medical terminological resources with focus on a particular interest and difficult to describe type of terms, namely *signs*, *symptoms*, *findings* and other *symptom-based phenotypes*. We hypothesize that no major official terminological source alone can accommodate for the variation and complexity for such terms present in real text data. Preliminary experiments indicate that to a great extend signs, symptoms and findings are expressed in a periphrastic manner, sometimes by the use of figurative or metaphorical language, or contextualized using a wealth of vague variant expressions. However these characteristics seem to vary depending on the type of data examined. In this paper we evaluate the content of the three largest medical control vocabularies available for Swedish on extracted reference symptom lists and initiate a discussion on how we should proceed in order to accommodate for increased coverage.

The followed approach can be seen as exploratory in which we believe to yield insights into the nature of symptom contextualization in order to be able to enhance our knowledge of communicative events in various healthcare settings. This study is initiated in the context of a recently started project, entitled *Interpretation and understanding of functional symptoms in primary health care*. The main research goal of the project is to study health care interactions with patients suffering from *Functional Somatic Syndromes* (FSS). Relevant research has showed that the care actions taken within primary health care are unsuccessful in the purpose to reduce the patients' suffering. The project's hypothesis is that the interaction in patient/care provider encounters is dysfunctional because of diverging perspectives and interpretation frames. This is resulting in lack of understanding and explanation of the patients' symptoms, leading to unsatisfaction and frustration among patients as well as care providers. One of the project's strand of research activities is on investigating how symptom mentions are expressed and how successful automated means are for capturing symptom descriptions both on collected written (patient records) and transcribed material (patient/nurse and patient/doctor encounters).

2 Background

The medical domain is particularly well endowed with sources of terminology, but there is also a large body of work with emphasis on methods for building required terminological knowledge bases automatically or semi-automatically from

textual sources. This is guided by the assumption that even though substantial term lists are available, automated methods have the benefit of being able to discover new variant terms, acronyms etc. and add them to existing lists (*cf.* Grishman *et al.*, 2002; Krauthammer & Nenadic, 2004; Tsujii & Ananiadou, 2005). Consequently, evaluation of terminologies in various subdomains has shown that there is a long way to go in order to achieve complete coverage. For instance, Langlotz & Caldwell (2005) discuss that no lexicon achieved greater than 50% completeness for any test set of imaging terms and that no single lexicon was sufficiently complete to allow comprehensive indexing, search, and retrieval of radiology report information.

Our work is also inspired to a certain degree by Unified Medical Language System (UMLS®; Kohler, 2008) since it would have been desirable in the future to have such comprehensive platform for e.g. Swedish. UMLS facilitates the development of computer systems that behave as if they *understand* the meaning of the language of biomedicine and health. The main purpose of the UMLS is to facilitate conversion of terms from one controlled medical vocabulary to another. UMLS consist of three knowledge sources, the *Metathesaurus*®, the *Semantic Network*, and the *SPECIALIST Lexicon*. The Metathesaurus forms the base of the UMLS and comprises several million concept names, all of which stem from the over 100 incorporated controlled vocabularies and classification systems. Some examples of the incorporated controlled vocabularies are ICD-10, MeSH, SNOMED CT, DSM-IV, LOINC and the Gene Ontology.

3 Controlled vocabularies (for Swedish)

3.1 Symptoms vs. Signs

In general terms, a symptom is a manifestation of a disease, indicating the nature of the disease, which is noticed by the patient; in this respect symptoms are *subjective* by nature. This is usually contrasted to signs which are observed by a medical practitioner and are thus *objective* measures by nature. Sometimes the context is important in order to distinguish one from the other, while often the distinction is blurred.

3.2 MeSH, SNOMED CT & KSH97/ICD-10

The Medical Subject Headings (MeSH) under the hierarchy C (*Disorders*) incorporates the subhierarchy C23 (*Pathological Conditions, Signs and Symptoms*) which includes abnormal anatomical or physiological conditions and objective or subjective manifestations of disease, not classified as disease or syndromes. The Swedish MeSH (edition 2006) includes 880 term entries in C23 which we also use in the current study, examples include *smärta* 'pain', *svullnad* 'edema' and *nysning* 'sneezing'.

The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) is a systematically organized computer processable collection of medical terminology covering most areas of clinical information. A relevant top level hierarchy in SNOMED CT is *finding*. The Swedish version of SNOMED CT (first release of April 2010) includes 32 911 findings, such as *brännande känsla* 'burning feeling' (90673000), *undernärd* 'malnourished' (248325000) and *kronisk hosta* 'chronic cough' (68154008).

Finally, the International Statistical Classification of Diseases and Related Health Problems (ICD) contains a listing of chapters one of which, Chapter XVIII, *Symptoms, signs and abnormal clinical and laboratory findings*, is relevant for this study. XVIII contains 532 terms, examples include *onormal hjärtrytm* 'abnormal heart rhythm' (R00), *dysuri* 'dysuria' (R30.0) and *dåliga matvanor* 'unhealthy nutrition habits' (R63.3). The Swedish translation of ICD is based on the Classification of Diseases 1997 (KSH97) and a systematic list that was released in September 1996. KSH97 (ICD-10) was recently replaced by ICD-10-SE (January 2011). In this study we use the older version.

4 Material and Method

There are several health related portals on the internet that provide a rather thorough description of diseases, their symptoms, etiology, treatment etc. The data sources of the symptoms' encoding used for the empirical evaluation were extracted from three popular health portals. The first site is intended for professional users, i.e. medical doctors <http://www.praktiskmedicin.com> the second and third are intended for laymen <http://www.netdoktor.se> & <http://www.1177.se>.

Fifteen randomly selected disease description pages were visited from each portal (Appendix A1). The symptoms' discussion parts for each disease was transferred to an external file, tokenized and automatically annotated with the three terminologies. The total number of manually identified symptoms was 552 (475 unique).

5 Evaluation

For the evaluation of the existing terminologies we chose a pragmatic approach as previously outlined, since available gold standards for such evaluation do not exist. A quantitative and qualitative analyses of the results are shown in table 1. Qualitative analysis in this context implies a thorough, manual examination of each annotated symptom description mention. A process that allows us to get a clearer picture on how symptom descriptions are formulated in text, what the limitations of the terminologies are and whether there is a need of harmonization of the terminologies and the gains we can expect. Moreover, it became apparent that enhancement with other mechanisms, such as extensive inclusion of variant forms (if available) or links to laymen vocabularies is necessary in order to enable a highly accurate and sufficient coverage of the textual content.

5.1 A Reference List

For ease of evaluation we chose to manually produce three reference lists, one from each site, a part of the accumulative term list is given in-Appendix A2. The quality of the controlled vocabularies, with respect to coverage, was evaluated in terms of i) the number of exact matches of text mentions; ii) the number of exact matches of text mentions after semiautomatic enhancement of the terminologies with various transformed variants (*cf.* Kokkinakis, 2009); iii) partial matches after the vocabulary enhancements and iv) the number of non-match after the vocabulary enhancements. The average symptom is 2,46 words long. Table 1 summarizes the results.

	ND	1177	PM
SNOMED	31,9%	34,4%	**34,9%**
MeSH	22,4%	24%	**26,8%**
ICD-10	3,2%	**4,3**%	4,1%
SNOMED+	38,1%	51,2%	45.6%
MeSH+	29,4%	32,8%	34,4%
ICD-10+	5,1%	6%	4,9%
No match SNOMED+	6,2%	19,2	13,9%
No match MeSH+	9,5%	13,6%	15%
No match ICD-10+	85,2%	85,4%	86,8%
Partial SNOMED+	55,6%	29,6%	40,3%
Partial MeSH+	60,9%	53,6%	50,5%
Partial ICD-10+	9,7%	8,6%	8,3%

Table 1: Evaluation results based on three samples (ND: *NetDoktor* and PM: *Praktisk Medicin* and 1177: *1177.se*) without/with vocabulary extensions (variations) the latter designated by the *plus* sign.

In the table above *Partial* implies that the obtained annotation is not complete. Sometime partial matching is sufficient in order to grasp the meaning of a text sequence such as in the example *hörselnedsättning på ena örat* 'hearing loss in one ear' in which both *hörselnedsättning* (C23.888.592.763.393.341) and *örat* (A01.456. 313;A09.246) have been recognised by MeSH but not the whole composite term. In other cases partial annotation is insufficient to capture the proper meaning such as in the case of *rasslande ljud i bröstkorgen* 'rattling sound coming from the chest cavity' in which only *bröstkorgen* could be matched.

6 Discussion

The initial findings of this study suggest that in combination the three resources have the potential to adequately represent a large number of the terms required to describe symptoms. All three together provide substantially more exact matches than any individual vocabulary in the set, although SNOMED CT gives the far better results. This is a natural consequence since its content is far more extensive and nuanced than both MeSH and ICD-10 together.

A problem faced with our approach is the fact that it is hard to determine whether potential missed terms (i.e. unmatched) were truly "absent" from the vocabularies or there might have been synonyms/variants in the resources that could not be identified despite the use of a large number of generated variant forms and near synonyms. Another important issues is the difficulty, in some cases, to differentiate between findings/symptoms and disorders/diseases. Although there is a separation in the three resources, sometimes fuzzy, as indicated in the MeSH-SNOMED distinction, in which a number of findings according to SNOMED were labeled with other hierachies in MeSH, such as *irritabilitet* 'irritable mood' which is found with the label "F01.470.047.110" which belongs to the *Psychiatry and Psychology* hierarchy; or *högt blodtryck* 'high blood pressure' which is found with the label "C14.907.489" which belongs to the *Cardiovascular Diseases* subhierarchy. However, these cases were marked as correct.

While an absent synonym can be remedied by simply adding a surface form, a missing concept represents a more significant absence but we could not identify such cases *cf.* the discussion by Wasserman & Wang (2003). There were a small number of lexical ambiguities (homographs) such as the phrase *sena skeden* litt: 'late stages' for which the SNOMED returned an an-

notation for *sena* ‘tendon’ (body structure) and *skeden* ‘the spoon’ (physical object); obviously both annotations are wrong in this context. Although the sample is not spontaneous language a number of metaphoric and figurative language expressions could still be found, such as *brännande smärtor* ‘burning paint, *bubblig i magen* ‘bubbly in the stomach’, *månansikte* ‘moon-face’, *buffelpuckel* ‘buffalo hump’, *motorisk klumpighet* ‘motor clumsiness’ and *produktiv hosta* ‘productive cough’. Finally, an issue that needs attention is various types of coordinations that need to be resolved in order to increase coverage, such as *minnes- och koncentrationsstörning* ‘memory and concentration disturbance’ and *fingrarnas ytter- och mellanleder* ‘fingers outer and middle joints’ and which may be resolved as *minnesstörning* & *koncentrationsstörning and fingrarnas ytterleder* & *fingrarnas mellanleder*.

7 Conclusions

Term matching in new subdomains of medicine is likely to identify further omissions highlighting the importance of a responsive updating process (Brown & Odusanya 2001). In the near future we intend to make detail analyses of other types of data, patient records and transcribed data, which will shed more light to whether controlled vocabularies can capture the patients' contextualization of symptoms, which is the main focus of this initiated activity. For future work we also intend to investigate whether partial or uncaptured symptom/finding-like terms are parts of disease/disorder descriptions. There might be other sources of lexical/terminological knowledge that might have been useful such as the International Classification of Functioning, Disability and Health (ICF) that we haven’t yet investigated. We anticipate that transcribed data will impose other source of problems due to the nature of how spoken language is transformed into written form. It might be fairly cumbersome to capture patients' perceptions of health-related problems in a simple straightforward manner. A general language, near synonym dictionary should also be worth to investigate since there are numerous cases that could be captured by such resources such as *smärta* ‘pain’, *ont* ‘hurt’, *värk* ‘pain’, and enhance controlled vocabularies in order to achieve better matching. In the same spirit Zeng & Tse (2006) discuss the development of consumer health vocabularies that would reflect the different ways consumers express and think about their health is necessary for extending research on various types of information-based tools. Such resources would be also beneficial as a complement to controlled vocabularies, and particularly for health information retrieval and understanding applications. The results should serve as a useful model, both for distributed input to the enhancement of controlled vocabularies and for devising new and better means for achieving better coverage.

Acknowledgments

This work is supported by the Gothenburg Centre for Person-Centred Care (GPCC) & the CLT.

References

Brown PJ and Odusanya L. 2001. Does size matter?--Evaluation of value added content of two decades of successive coding schemes in secondary care. *Proc AMIA Symp*. 71–75.

Grishman R., Huttunen S. and Yangarber R. 2002. Information Extraction for Enhanced Access to Disease Outbreak Reports. *J Biomed Inf – Special issue: Sublanguage*. Vol. 35(4): 236-246.

Kohler M. 2008. *Unified Medical Language System for Information Extraction*. VDM Verlag.

Kokkinakis D. 2009. Lexical granularity for automatic indexing and means to achieve it – the case of Swedish MeSH®. In *Information Retrieval in Biomedicine: NLP for Knowledge Integration*. Prince V. & Roche M. (eds). Pp:11-37. IGI Global.

Krauthammer M. and Nenadic G. 2004. Term identification in the biomedical literature. *J Biomed Inf*. 37(6):512-26.

KSH97/ICD-10. 2002. Klassifikation av sjukdomar & hälsoproblem 1997; Rev 2002/04. Socialstyrelsen. <http://www.socialstyrelsen.se/Lists/Artikelkatalog/Attachments/10871/2002-4-2_200242.pdf> .

Langlotz CP. and Caldwell SA. 2005. Completeness of existing lexicons for representing radiology report information. *J Digit Imag.* 15 Suppl 1:201-5.

Tsujii J. and S. Ananiadou. 2005. Thesaurus or Logical Ontology, Which One Do We Need for Text Mining? *J. Lang Res & Eval*. Pp:77-90. Vol. 39:1.

UMLS <www.nlm.nih.gov/pubs/factsheets/umls.htm>

Wasserman H. and Wang J. 2003. An Applied Evaluation of SNOMED CT as a Clinical Vocabulary for the Computerized Diagnosis and Problem List. *AMIA Annu Symp Proc*. 699–703.

Zeng QT. and Tse T. 2006. Exploring and Developing Consumer Health Vocabularies. *J Am Med Inform Assoc.* 13:24-29.

Appendix A1

< http://www.1177.se//Fakta-och-rad/Sjukdomar>
Astma </Astma/>
Blindtarmsinflammation </Blindtarmsinflammation/>
Blodpropp i benet </Blodpropp-i-benet/>
Bältros </Baltros/>
Gallsten </Gallsten/>
Havandeskapsförgiftning </Havandeskapsforgiftning/>
Hjärtsvikt </Hjartsvikt/>
Klamydia </Klamydia/>
Laktosintolerans </Laktosintolerans/>
Ménières sjukdom </Menieres-sjukdom/>
Näthinneavlossning </Nathinneavlossning/>
Påssjuka </Passjuka/>
Rabies </Rabies/>
Ulcerös kolit </Ulceros-kolit/>
Urinvägsinfektion </Urinvagsinfektion/>
<http://www.praktiskmedicin.com/>
Akut lymfatisk leukemi <sjukdom.asp?sjukdid=897>
Analfissurer <sjukdom.asp?sjukdid=309>
Bronkit. Luftrörskatarr < sjukdom.asp?sjukdid=10>
Demens < sjukdom.asp?sjukdid=90>
Diabetes ketoacidios < sjukdom.asp?sjukdid=744>
Järnbristanemi < sjukdom.asp?sjukdid=900>
Kol. Emfysem. Kroniskt Obstruktiv Lungsjukdom < sjukdom.asp?sjukdid=14>
Lungödem < sjukdom.asp?sjukdid=147>
Njursten < sjukdom.asp?sjukdid=469>
Osteoporosis < sjukdom.asp?sjukdid=98>
Polyneuropati < sjukdom.asp?sjukdid=369>
Prostatacancer < sjukdom.asp?sjukdid=670>
Psoriasis < sjukdom.asp?sjukdid=234>
Soleksem < sjukdom.asp?sjukdid=239>
TBE-infektion < sjukdom.asp?sjukdid=1158>
<http ://www.netdoktor.se/>
ADHD <adhd/?_PageId=113320>
Artros <artros/?_PageId=162>
Bihåleinflammation <forkylning-infektion/?_PageId=505>
Cushings syndrom <hud-har/?_PageId=524>
Diskbråck <smarta/?_PageId=360>
Enterohemorragisk E. Coli (EHEC) <mage-tarm/?_PageId=550>
Fönstertittarsjuka (claudicatio intermittens) <hjart-karl/?_PageId=107115>
Genital Herpes <sex-relationer/?_PageId=432>
Hemorrojder <mage-tarm/?_PageId=583>
Irriterad tjocktarm (Colon Irritabile/IBS) <mage-tarm/?_PageId=509>
Kolera <mage-tarm/?_PageId=622>
Multipel skleros (MS) <neurologi/?_PageId=652>
RS-virus <barn/?_PageId=713>
Skrumplever (levercirrhos) <mage-tarm/?_PageId=630>
Vinterkräksjukan <mage-tarm/?_PageId=694>

Appendix A2

Reference list (top occurrences)
8 feber
6 diarré
5 trötthet
5 kräkningar
4 ångest
3 trött
3 sveda
3 smärta
3 magsmärtor
3 förvirring
2 ökad törst
2 yrsel
2 vätskeförlusten
2 viktminskning
2 tryck på ryggmärgen
2 tinnitus
2 smärtor
2 oro
2 ont i magen
2 nedsatt vibrationssinne
2 muskelsvaghet
2 medvetandesänkning
2 lätt feber
2 kramper
2 koncentrationssvårigheter
2 kallsvett
2 impotens
2 hög feber
2 hematuri
2 gaser i magen
2 förstoppning
2 dålig aptit
2 dyspné
2 depression
2 blåskatarr
2 blekhet
2 benskörhet
1 övergående ospecifik feber
1 överaktivitet
1 ömt över gallblåsan
1 ömma öronspottkörtlar
1 ömhet runt naveln
1 ökad trötthet
1 ökad hårväxt
1 ögonvitan blir gul
1 ögat kännas torrt
1 ögat bli rött
1 ögat blir känsligt för ljus
1 ödem
1 ängslan
1 återkommande trötthet
1 åldrandet
1 åderbråck i matstrupen
1 ytsensibilitet
1 vätska samlas i kroppen
…

Exploring termhood using language models

Jody Foo
Linköping University
Linköping, Sweden
jody.foo@liu.se

Abstract

Term extraction metrics are mostly based on frequency counts. This can be a problem when trying to extract previously unseen multi-word terms. This paper explores whether smoothed language models can be used instead. Although a simplistic use of language models is examined in this paper, the results indicate that with more refinement, smoothed language models may be used instead of unsmoothed frequency-count based termhood metrics.

1 Background

Terminology work is the process of creating, harmonizing and standardizing term banks. The process involves the use of human terminologists and domain experts to a high degree, which can be costly for even small sized (e.g. 300 terms) term banks.

Automatic term extraction (ATE) or automatic term recognition (ATR) is a research area where methods researched that can to some degree automate the task of finding term candidates from document collections.

For the discussions in this paper we will be considering ATE used to facilitate terminology work done by terminologists. Looking from above, a workflow may be as follows.

1. Extract term candidates from corpus
2. Let domain expert process term candidates
3. Let terminologists create a term bank

This paper concerns step 1 which can be broken down into the following smaller steps.

a) Extract phrases
b) Asses termhood of phrases
c) Output term candidates

The following assumptions are used in this paper regarding the context of terminology.

- Term banks are used to reduce misunderstandings, eliminate ambiguity and raise the efficiency of communication between domain experts within the same domain, and to aid non-experts to understand domain specific texts.
- Terms represent Concepts.
- Definitions are attached to Concepts, not to terms.
- Terminologists are detectives that work together with domain experts to maintain a consistent terminology within the domain.

1.1 Term ranking concepts

Term ranking metrics can be categorized in several ways. One facet divides metrics into contrastive and non-contrastive measures. The contrastive model was introduced by Basili et al (2001) and explicitly argues that distributional differences between different document collections can be used to say something about extracted phrases.

The concept of termhood was introduced by Kageura and Umino (1996) and is defined as "*The degree to which a stable lexical unit is related to some domain-specific concepts*.". Unithood was also introduced by Kageura and Umino (1996) and is defined as "*the degree of strength or stability of syntagmatic combinations and collocations*". Both Wong and Liu (2009), and Zhang et al (2008) provide good overviews of termhood and unithood related metrics such as C-Value/NC-Value (Frantzi et al, 1998), Weirdness (Ahmad et al, 1999), Termextractor (Sclano and Velardi, 2007).

The ideal goal regarding termhood is to find a metric that correlates perfectly with the concept of termhood. Such a metric does however not yet exist and it is quite probable that constructing such a metric is a near impossible task for several reasons; one of them being that the properties of terms are difficult to capture. With

regard to actual work done by terminologists, a termhood metric is quite artificial. Also, it is important to keep in mind that a usable term ranking metric does not necessarily measure termhood – i.e. it may not be necessary to use a termhood metric to implement a useful term extraction application.

1.2 Support Vector Machines

In this paper, a *Support Vector Machine classifier* is used in an attempt to classify phrases into *term candidates* and *non-term candidates*.

The framework used is the e1071 package for R[1] (Dimitriadou et al 2009), which interfaces with libSVM, a Support Vector Machines implementation (Chang and Lin, 2001).

Support Vector Machines were introduced by Boser, et al (1992) and is a linear classifier that can use kernels to also classify non-linear data.

2 Questions

The existing research on term extraction is focused on term extraction as a once-off process using relatively large document collections. However, in reality, one may want to perform term extraction on smaller document sets containing new unseen documents from a previously processed domain. This may present a problem for frequency-count based metrics for two reasons

1. The document set may be too small for frequency based term metrics to be of use.
2. The first problem may be solved using a larger document collection is used to produce the metric values for extracted words/phrases from the smaller document collection. However, previously unseen multi-word terms cannot be assigned a score.

One way of solving problem 2, may be to use probability and perplexity scores from smoothed n-gram language models instead. The key point here is that a smoothed language model can produce a probability score for a multi-word term that uses a combination of words that has never been seen in previous document collections. Language Models have not been used in this way to the author's knowledge.

However, Patry and Langlais (2005) used language models of POS tags to improve phrase extraction beyond ordinary POS pattern extraction.

The work described in this paper is a preliminary study on using smoothed n-gram (word) language models to capture termhood.

3 Dataset

In this paper, two corpora are used 1) the British National Corpus (BNC) (BNC Consortium, 2000) and 2) English patent texts from the C04B IPC subclass (lime; magnesia; slag; cements; compositions thereof) as well as a set of domain expert validated terms from the subclass (note: the list of validated terms is not complete). See Table 1 for details of the used patent corpus.

C04B statistic	Value
Number of segments (sentences)	96,390
Number of tokens	2,395,177
Number of characters	1,2836,222
Validated terms	2,677

Table 1 C04B patent document corpus in numbers

3.1 Language models

Both the BNC corpus and C04B corpus were lemmatized using the commercial tagger Conexor Machinese Syntax[2]. The lemmatized corpora were then processed using SRI Language Modeling Toolkit, which produced one n-gram language model per corpus (two language models in total).

4 Phrase extraction and validation

The phrases from the dataset first extracted using IPhraxtor, a phrase extractor developed at Fodina Language Technology AB. A randomly sampled subset was then validated with regard to term candidates and non-term candidates.

4.1 Phrase extraction

Using IPhraxtor, noun phrases were extracted from the C04B corpus resulting in 101,191 extracted phrases. Among these phrases, 2,143 of the validated terms were found.

4.2 Term candidate validation

A sample was then extracted for manual term candidate markup. The sample was processed in Microsoft Excel where a non-domain-expert classified the phrases as either *term candidates* or *non-term candidates*. Note that the classification is between term candidates and non-term candidates; not between term and non-term. The reason is that the process we want to improve

[1] http://www.r-project.org/

[2] http://www.connexor.eu/technology/machinese/machinesesyntax/

outputs term candidates, not terms. Below are the guidelines used during the manual validation.

1 When validating the phrase as a term candidate the whole phrase must be considered, not just a part of the phrase. E.g. the phrase "*mold temperature*" may be considered a term candidate, but not "*measure mold temperature*"

2 Non-term candidates are

   a grammatically incomplete phrases, e.g. "involves passage", "improves compressive strength"

   b phrases that contain non words, misspelled words, or tokenization errors, e.g. "die(51a", "grains)"

   c phrases that are obviously general language such as idioms and general collocations, e.g. "*infinite length*", "*major role*"

   d phrases containing numbers

   e phrases starting with a verb

   f chemical formulas, e.g. H20 are not terms. Names of chemicals however, are, e.g. hydrogen oxide.

   g phrases starting with a "subjective" or referring adjective, e.g. *desired*, *intended*, *indicated*. Quantifying adjectives however, are fine, e.g. *poor*

Regarding guideline 2c, it is still a decision that depends on the validators experience and knowledge. Therefore, it is recommended that validators are domain experts in at least one field. For example the word "*accurate*" might be classified as a non-term candidate by a validator not familiar with the term "*accuracy*" in e.g. the domain of machine learning. Regarding guideline 2e; no phrases starting with a verb were intentionally extracted, but POS-tagger errors resulted in a few such phrases being included.

5 Contrastive features

The validated, extracted phrases were annotated with several features using the previously created language models. Each phrase was given a logarithmic probability value (logProb) and a perplexity value (ppl), first using the BNC language model, then the domain specific C04B language model. A probability ratio using the $logProb_{C04B}/logProb_{BNC}$ was also calculated and added. Finally, each phrase was annotated with the number of words in the feature. Each phrase also belonged to the class term candidate or non-term candidate. All values were normalized to the scale of 0-1. The features are summarized in Table 2.

Feature	Description
class	term candidate/non-term candidate
number of words	number of words in phrase
$logProb_{BNC}$	logarithmic probability of phrase in BNC language model
ppl_{BNC}	perplexity value of phrase in BNC language model
$logProb_{C04B}$	logarithmic probability of phrase in C04B language model
ppl_{C04B}	perplexity value of phrase in C04B language model
logProbRatio	the ratio between $logProb_{C04B}$ and $logProb_{BNC}$

Table 2 Features used for SVM classification

6 Looking for patterns

To understand the results of the SVM classification experiment, extracted phrases were ordered by class (term candidates first) and plotting their corresponding feature values in graphs. Figures 2-4, are examples of such graphs. In Figure 1, the precision of the ordered list is presented. This just shows how many term candidates and how many non-term candidates are in the list (# correct stops increasing where the non-term candidates begin). From Figures 2-4 it is clear that there does not seem to be any visible correlation between the language model output and the phrases classified as term candidates.

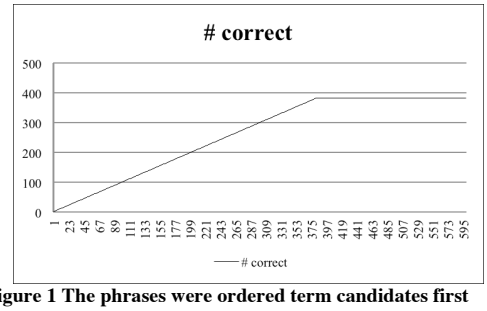

Figure 1 The phrases were ordered term candidates first

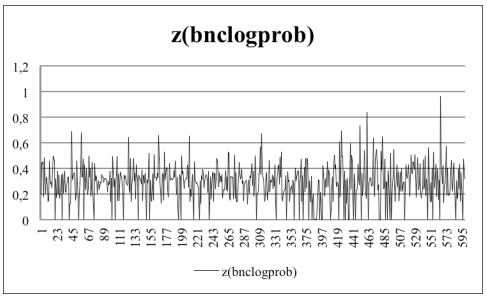

Figure 2 Probability values from the BNC language model for phrases ordered term candidates first

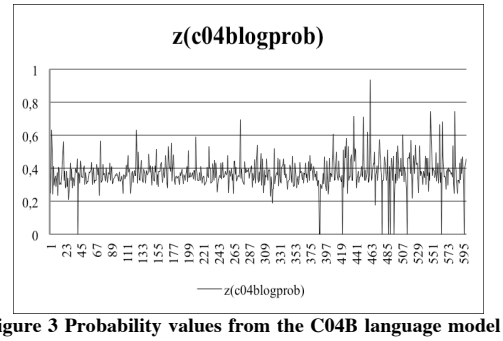

Figure 3 Probability values from the C04B language model for phrases ordered term candidates first

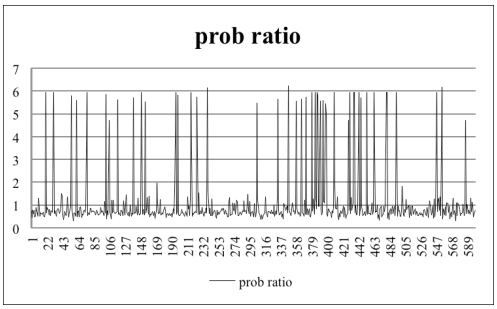

Figure 4 Probability ratio values for phrases ordered term candidates first

7 SVM classification results

A simple SVM experiment was conducted using the 1800 classified phrases. First a model was trained using 1200 of the phrases. Then the model was used to predict the class of the 600 phrases that were held back during training. The model used, predicted term candidates with a precision of 66.4% and a recall of 88.0%. Considering that the test partition contained 368 term candidate phrases, i.e. 61.3% of the test data were term candidates, the result of the classification is not much better than using the extracted phrases as they are.

8 Discussion and future work

Though the results from the classification experiment are not that strong, they were also the result of a rather simplistic use of language model provided features. The frequency count based metrics described in current research are still much more refined, as using the raw probability and perplexity values can be compared to using raw phrase frequency counts. Therefore, the author believes that there is more to gain from a language model approach. A higher level of refinement however is needed.

For example, a next step could be to consider phrases of different word length separately, as phrases containing more words have a lower probability in an n-gram language model by nature.

References

Ahmad, K., Gillam, L., & Tostevin, L. (1999). Weirdness Indexing for Logical Document Extrapolation and Retrieval. In *Proceedings of the Terminology and Artificial Intelligence Conference (TIA 2001)*.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *COLT '92 Proceedings of the fifth annual workshop on Computational learning theory* (p. 144-152).

Chang, C., & Lin, C. (2001). *LIBSVM: a library for support vector machines*. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm (accessed 2011-04-25)

BNC Consortium. (2000). *The British National Corpus, version 2 (BNC World)* (2nd ed.). Oxford University Computing Services.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., & Weingessel, A. (2009). *Package "e1071."* Software available at http://cran.r-project.org/web/packages/e1071/index.html (accessed 2011-04-25)

Frantzi, K. T., Ananiadou, S., & Tsujii, J. (1998). The C-value/NC-value Method of Automatic Recognition for Multi-word Terms. In *ECDL '98* In *Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries* (p. 585-604).

Kageura, K. (1996). Methods of Automatic Term Recognition. *Terminology*, *3*(2), 259-289(31).

Patry, A. & Langlais, P. (2005) Corpus-Based Terminology Extraction. In *Proceedings of the 7th International Conference on Terminology and Knowledge Engineering*, pp. 313-321

Sclano, F., & Velardi, P. (2007). TermExtractor: a Web Application to Learn the Common Terminology of Interest Groups and Research Communities. In *Proceedings of the 9th Conference on Terminology and Artificial Intelligence*.

Wong, W., & Liu, W. (2009). Determination of unithood and termhood for term recognition. In Song, M., & Brook Wu, Y. (Eds.), *Handbook of Research on Text and Web Mining Technologies. (pp. 500-529)*. IGI Global.

Zhang, Z., Iria, J., Brewster, C., & Ciravegna, F. (2008). A Comparative Evaluation of Term Recognition Algorithms. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*.

Getting to terms with terminology at Swedish public agencies

Magnus Merkel
Linköping University/Fodina
Linköping, Sweden
magnus.merkel@
{liu|fodina}.se

Henrik Nilsson
Terminologicentrum (TNC)
Stockholm, Sweden
henrik.nilsson@tnc.se

Abstract

This paper describes on-going work aimed at assisting public agencies in Sweden to conform to the new Swedish Language Act (passed in 2009). The Language Act highlights terminology as a key factor for a public agency, as well as a responsibility for a public agency to ensure that its terminology is made available, used and developed. Term-O-Stat is an action program to help public agencies to improve their terminological efforts. Term-O-Stat is divided into four distinct steps: 1) term inventory, 2) term classification, 3) conceptual analysis and term choice, and 4) term implementation. We describe the four steps and also experiences from the realization of step 1 and 2 at the Swedish Social Insurance Agency.

1 Introduction

A Language Act was passed in Sweden in July 2009. It contains a clause which clearly emphasizes that public agencies have a responsibility in making sure that Swedish terminology within their specific domain is "available, used and developed" (SFS 2009:600).

This means that there now is a clear legal incentive for public agencies to address terminological issues for their particular subject field. The specific terms that are currently being used within a public agency have often been developed over many years, and it is not uncommon that there is evidence of inconsistent term usage. For instance, a close scrutiny will usually reveal that one concept is denoted by a number of different terms on a website or in other public documents. Inconsistent term usage makes communication within a public government agency more difficult, and, furthermore, it can also make communication with citizens inefficient and confusing.

But, getting to terms with inconsistent and confusing term usage need not be that complicated. First of all, it is essential to investigate the actual term usage. Such an investigation can help to bring order in what terminology an organization is actually using. It is also important to try to specify the actual areas of responsibility for a public agency. The tax authorities have their specific responsibility to handle and maintain terminology for the area of taxation, and this differs from e.g. the Swedish Social Insurance Agency (Försäkringskassan). The latter will have to take the main responsibility for social insurance terminology. However, as terms from different areas are often used across public agencies, it is necessary to clarify who "owns" what terminology and also that terminologies will be shared across public agencies.

An effort to have public terminologies spread in Sweden was made by Terminologicentrum TNC (Swedish Centre for Terminology) in 2009 when they launched "Rikstermbanken" (www.rikstermbanken.se). Rikstermbanken is Sweden's national termbank on the web and holds over 77,000 term records spanning over a variety of domains. More than 150 organizations (most of them public agencies) have contributed to Rikstermbanken.

In many public agencies fragmented terminological resources are kept in Excel files or binders, but very few public agencies have systematically built a concept-oriented term database, and even fewer have integrated it in their writing environment.

2 Term-O-Stat

An action program called "Term-O-Stat" was launched in 2009 as an attempt to assist the public sector in Sweden to comply with the new Language Act, specifically directed to terminological issues. Term-O-Stat is constituted by the following four steps:

1. Term inventory
2. Term classification
3. Conceptual analysis and term choice
4. Term implementation

In short, step 1 concerns the collection and automatic analysis of documents in order to find what the actual term usage looks like. In step 2 and 3, the term candidates found in step 1 are processed further by their classification in subdomains, and the corresponding concepts are analysed and defined. In step 4, the results from steps 1–3 are implemented in a term database and a writing tool in order to integrate the established terminology into the normal writing and publishing workflow.

The overall program has been introduced to Swedish public agencies via seminars and on site visits.

In the following subsections the Term-O-Stat steps are explained in more detail.

2.1 Step 1: Term Inventory

In step 1, a document collection is analysed automatically, although some of the work involves manual inspection. The different phases of step 1 are the following:

1. Collection of suitable documents
2. Conversion from Word, Excel, Power Point, HTML and PDF to text
3. Grammatical analysis
4. Term extraction
5. Import to database
6. Filtering
7. Linguistic validation 1
8. Generation of synonym clusters
9. Linguistic validation 2
10. Cross-reference to internal linguistic resources (wordlists etc.)
11. Cross-reference to Rikstermbanken
12. Export to Excel sheet

The first phase involves a discussion with the agency in order to decide a suitable set of documents to use as input. This could result in all documents on the external website being selected, e.g. brochures, regulations, press releases, etc. The documentation volume can vary considerably between different agencies; for smaller agencies there may only be a couple of hundred thousand words, and for large agencies there may be several million words. The different file formats are then converted into plain text and sent to a grammatical analysis component. The grammatical and lexical analysis is made by Connexor's Machinese Syntax (Tapanainen and Järvinen (1997)). The analysis gives parts-of-speech information together with information on baseforms, morphological features as well as syntactic functions. After this step, term candidates are extracted; mainly noun phrases and verbs are extracted, but also syntactic function such as subject and object relations are utilized. The term candidates are then imported into a database and filtered (using stop word lists and syntactic patterns). All contexts for each term candidate are stored in the database and presented in an application (called TermViewer) where a linguist can validate the term candidates in context.

When the term candidates have been validated a search is made for synonyms among the candidates. This means that some term candidates are found to be possible synonyms to each other and therefore clustered together in a synonym set. The synonym clustering is made by string comparisons (for example the candidates "oral kirurgi" and "oralkirurgi" are clustered) and also with the use of Swedish synonym lexicons. The synonym clusters are generated automatically but inspected by a linguist for validation.

If the agency has internal word lists or lexicons, these are cross-referenced in order to find term candidates. A similar lookup is also made in Rikstermbanken and a reference is made for each term candidate that is also found there. At this point the data in the database is exported to an Excel sheet and presented to the agency. See fig 1.

C	D	E	G	H	I	J	
KonceptID	**POS**	**Term**	**Frek**	**RTB**	**FKEN**	**FKBK**	**Exempel**
#00042&	subst	oral kirurgi	5	oral kirurgi (Status:			Till–exempel kan åtgärd 491 bara debiteras av specialist
#00042&	subst	oralkirurgi	4				Om den som utför behandlingen är specialist eller utbilda
#00043&	adj	begärd	191				Du ändrar eller avbryter **begärd** föräldrapenning genom a **Begärd** ersättning Kontoförande bank är inte skyldig att pröva behörigheten Försäkringskassan har därför in en särskild skrivelse **beg** Försäkr Privat vårdgivare **Begärt**
#00043&	adj	beställd	6				En--del av en förklaring kan vara att fokus i slutrapporten Även beslut i massärenden utgör emellertid myndighetsu
#00044&	subst	F-skattesedel	2				Kopia på registreringsbevis från Skatteverket (**F-skattese**
#00044&	subst	Fskattsedel	4				Med innehav av F-skattsedel jämställs en skriftlig uppgift
#00044&	subst	F-skattsedel	116		corporate tax certificate		Det gäller även--om du är egen företagare med **F-skattse** **F-Skattsedel** **F-skattsedeln** utfärdas endast för ett år i taget . **F-skattsedelns** från och med datum är före eller samma

Fig. 1 Output data from step 1. Three synonym clusters with term candidates are shown with frequency data, cross-references and sample context.

2.2 Step 2: Term Classification

In step 2 of Term-O-Stat, the term candidates found in step 1 are used as input. Terminologists inspect all term candidates and classify them into different groups:

1. Terms specific to the public agency
2. Terms common in the public sector
3. ”General” terms
4. Non-terms
5. Names

Group 1 is the most important category for any given public agency. Terms classified as belonging to that category are terms that constitute “their own terminology” and thus the agency’s area of responsibility. The second group contains terms that are not unique for the agency but which are also relevant for other public agencies. Groups 3–4 will contain term candidates that are deemed not important enough to investigate further. These candidates can be terms of a more general character and words that superficially look like terms but belong to general language. It has proven useful to separate names in a special category (group 5). The classification of the term candidates is made by using the database produced in step 1 and by using the GUI interface TermViewer, which allows several users to access the database simultaneously. Terms from group 1 are further subclassified; the public agency in question may have its own classification system that can be applied here.

2.3 Step 3: Conceptual Analysis and Term Choice

In step 3, the terminologists continue to work with the terms in group 1 (terms specific to the public agency), together with experts from the public agency. The work here is more of traditional terminology work where concept clusters are analyzed and described in concept systems and then defined. The main difference from traditional terminology work procedures lies in that the starting point is a fuller inventory and categorized terms (in synonym clusters) that emanate from a large amount of public agency documents. The objective is to come to a consensus about the concepts, how they should be defined and what terms should be used to denote them.

An important aspect of step 3 is also to work with term status. If a concept is denoted by several terms, one term may be “recommended”, another “admitted” while three other terms could be classified as “deprecated”. This is very important and useful information when the terminology is to be “put to use”. The status indication of a term is a prerequisite for the integration of the terminology into the existing writing tools of the authoring and publishing environment (see step 4). Although agency-specific terms (group 1) are in focus in step 3, terms from group 2 could also become important, and this would entail the co-operation with other public agencies.

2.4 Step 4; Term Implementation

In step 4, the objective is to integrate the results from the earlier steps into the authoring and publishing environment. This is actually one way of complying with the Swedish Language Act that states that the terminology within the subject domain of a public agency also should be used, e.g. in the documents and website created by the agency.

It is not enough to publish a web page on the intranet listing all terms alphabetically to promote “usage”. It is of course better than nothing, but the optimal solution would be if the terminology could be integrated and embedded in the writing tools (word processors, presentation or web authoring tools, etc.) and used in the manner of ordinary spellcheckers and grammar checkers in applications like Microsoft Office.

Remembering terminology suggestions and detailed writing recommendations is extremely difficult for authors. Many public agencies in Sweden conduct regular training on writing and the use of proper terminology, but it is still hard to spread information on newly changed terminology and new policies. If it were possible to make changes to a central language server such changes could be made available directly through a language checking plug-in programme for the standard word processor.

One example where a central language server combined with language checking plug-in clients for various applications is acrolinx IQ. In acrolinx IQ it is possible to check documents for terminology, spelling as well as grammar and style rules. In other words, from a terminological point of view one can

- store and administer terms in an integrated term database
- highlight terms in documents that are admitted by the term database
- mark deprecated terms when such are used and propose a recommended term instead
- manage different term sets for different text types, users and domains within an organization

- extract new term candidates from existing documentation

The language checking can be performed by the author from the plug-in client or applied as a batch checking process on a set of documents.

3 Term-O-Stat so far …

Term-O-Stat has been active for approximately one year and during that time three open seminars have been held where more than 25 public agencies have attended. Step 1 and 2 have been implemented at Försäkringskassan (Swedish Social Insurance Agency). At Försäkringskassan, around 2,000 documents were processed in step 1, resulting in 17,000 term candidates that were fed into step 2. In step 2, the term candidates were distributed over the category groups in the following way:

- Terms specific to the public agency (2,628)
- Terms common in the public sector (2,320)
- ”General” terms (6,235)
- Non-terms (4,618)
- Names (726)

The first group, with terms specific to Försäkringskassan’s area of responsibility, was divided into eleven subareas, e.g. administration, housing, dental care, immigration, disease, etc.

4 Conclusions

The first Term-O-Stat project showed that the agency-specific terminology is much more complex than one could have expected, and also that there may be a considerably higher degree of inconsistency in how terms are used in practice. By using existing term lists in the inspection, it is possible to compare these to the actual usage in the analyzed document set. At Försäkringskassan, it was discovered that a number of terms specified in a rather small termbank were not used at all in any of the documents on the external website. This does not have to mean that they are unimportant, instead it may reflect the fact that the termbank focused on concepts that are not used in external communication.

So far, we have only dealt with monolingual term extraction. Bilingual and multilingual material exists but usually makes up only a fraction of the information published in Swedish. If parallel texts were available it would be possible to do bilingual term extraction and find terminological inconsistencies in both the source and target texts.

Automatic term extraction methods, filtering techniques, database technology and the performance of modern computers open up new exciting possibilities for making terminology projects much more efficient. On the other hand, terminology work requires access to domain experts, in this case experts at the public agency that has the in-depth knowledge of their subject area. The participation of the public agency representatives will be necessary for the following activities:

- Search and select documents that form the input data.
- Assist in thee categorization and clustering of term candidates (classification systems, clustering criteria).
- Participate in concept analysis, definition writing and term selection.
- Assist in the publishing of the material internally and externally,
- Train users in using new tools in the internal authoring environment.

The exact time that is required for these activities varies from agency to agency. A successful end result will to a large extent depend on how much time the agency can devote to the project, especially to step 3.

By combining automatic methods from language technology with manual validation and categorization, Term-O-Stat has shown that it is possible to get an overview of terminology usage that would have been practically impossible to acquire using only traditional terminological methods.

References

Foo, Jody & Merkel, Magnus. 2010. Computer aided term bank creation and standardization: Building standardized term banks through automated term extraction and advanced editing tools. In Marcel Thelen & Frieda Steurs (red.), Terminology in Everyday Life, (pp. 163–180). John Benjamins Publishing Company. Amsterdam.

Tapanainen, Pasi & Järvinen, Timo. 1997. A non-projective dependency parser. In Proceedings of the 5th Conference on Applied Natural Language-Processing, pp. 64–71, Washington, D.C., April. Association for Computational Linguistics.

Språklag (Language Act), Svensk författningssamling (SFS 2009:600). http://www.riksdagen.se/webbnav/index.aspx?nid=3911&bet=2009:600.

The Experimental Study of Terminology Collocations: Calculations and Experiments with Informants

Elena Yagunova
Saint-Petersburg State University
Saint-Petersburg, Russia
iagounova.elena@gmail.com

Anna Savina
Saint-Petersburg State University
Saint-Petersburg, Russia
anja.savina@gmail.com

Abstract

This paper presents an experimental solution of the problem of the nature of the terminology collocations and possibility of their ranging, which depends on the degree of coherence of these collocations. Within this paper the combination of two different approaches – calculation and experiments with informants – is proposed to the study of the terminology collocations. The proposed approach is particularly relevant for those scientific areas, where still there isn't precise terminology.

1 Intoduction

Our research is devoted to solving one of the most important problems of collocation study: about the nature of scientific (terminology primarily) collocations and their possible classification. The report presents the result of the first stage of work within the overall project on this topic. We understand a **collocation** as a non-random combination of two or more lexical items that characterizes the language as a whole (texts of any kind) or a certain text type (or even (sub)sample of texts). In our research of language and speech we go from the text **realization**, from the available material. The material dictates the choice of certain theoretical positions and classification principles. Such research may be conducted only by using statistical measures to evaluate the degree of the non-randomness of the sequence of words. It's obvious that the list of combinations isn't completely homogeneous and requires a subsequent classification and some theoretical interpretation (Pivovarova and Yagunova, 2010; Khokhlova, 2011; Yagunova and Pivovarova, 2011).

Ample opportunities for understanding the nature of the collocations – within the lists received on the basis of statistical measures – are given by the reference to experiments with the native speaker informants. The purpose of the report is the demonstration of such kind of capabilities.

2 Material and Methods

We want to illustrate the suggested methodology based on an example of a monothematic collection of conference materials "Corpus Linguistics" for 2004-2008[1]. Volume of the collection is about 220,000 "tokens" – word usages and punctuation marks. Corpus Linguistics (especially in Russia) is the scientific area, where still there isn't precise terminology.

We used two statistical measures for bigram extraction: MI and T-score (Evert, 2004; Manning and Schütze, 1999; Stubbs, 1995). MI allows to extract terminological combinations, T-score highlights scientific clichés and those terminological combinations that characterize all texts from the collection (or most of those texts) (Pivovarova and Yagunova, 2010). A.Savina has made the program, which is very convenient for research purposes and allows to select lists of bigrams with a nuclear word on the basis of those statistical measures.

We have considered two ways to get lists of interesting (for us) collocations:

- for all collocations with a maximum value of these measures;
- for collocations with interesting (for us) nuclear word.

[1] For comparison, we used a collection of news texts lenta.ru in 2009. Thus, we tried to research collocation, describing the text of a certain functional style (type, genre).

Based on the collection of scientific texts we first of all had received lists of bigrams for each of the measures of association (MI and T-score) and sorted them by descending value of these measures. Then from each list were selected from 25 bigrams with the highest values of the measure.

After that for each nuclear word ("corpus", "word") we also consider sublists of 25 bigrams for each of the measures.

Later on the combined listings of the 50 randomized collocations are the input data for the two types of experiments with informants.

Such combined lists allowed us to estimate the degree of connectivity for terminological combinations, which was allocated on the basis of measures MI, and then compared with scientific clichés, function words (and other combinations), allocated on the basis of the measures T-score.

For this monothematic collection terms, that are common to all texts of the collection, were distinguished on the basis of both measures.

Thus, we obtained two lists of bigrams – with nuclear words and without.

As it has been already mentioned, we conducted two types of experiments with each of the lists:

- experiment 1 – the classification of bigrams;
- experiment 2 – the scaling of bigrams.

Experiment 1: 25 informants offered informants a questionnaire in which they were required to determine which of the three classes – "right", "predictable" and "others" – applies to each combination of the proposed list.

Experiment 2: 22 informants were given the task to evaluate the degree of connectedness between words – for the same lists – at a scale of 0 to 5, where "0" corresponds to the minimum, and "5" – the maximum degree of connectivity from the perspective of informants.

In the instructions we said to the informants about the domain specificity: "You see the combinations of words (bigrams) from the specialized (linguistic conference) texts, selected on the basis of statistical criteria. ..."

3 Results. Conclusions

We have obtained the classification of bigrams in a given set of classes for a holistic list (without specifying the nuclear word) based on the results of Experiment 1. Each class is divided into the core and the periphery. Collocation was considered as core, if more than 65% of informants referred it to this class and the peripheral – if the number of informants ranged from 33% to 65% (amount of these classes in table 1)[2].

Table 1. Bigram classes according to Experiment 1

bigrams	"right"		"predictable"	
classes	core	periphery	core	periphery
without nuclear words	12	12	6	27
with nuclear words	9	6	10	15

Table 2. The results of Experiment 1 (bigrams without nuclear words)

core of "right"	core of "predictable"	core of "others"
математической лингвистики [mathematical linguistics]	в качестве [as]	но и [but also]
художественной литературы [fiction]	за счет [due to]	так и [and]
русского языка [the Russian language]	(в) свою очередь [(in) turn of]	что в [that in]
корпусная лингвистика [corpus linguistics]	(в) том числе [including]	
имена собственные [proper names]	на основе [on the basis of]	
словарной статьи [vocabulary entry]	(с) точки зрения [(in) terms of]	
машинного перевода [machine translation]		
корпусной лингвистике [corpus linguistics]		
корпусной лингвистики [corpus linguistics]		
речевой деятельности		

[2] Part of peripheral collocations was attributed to the intersection of classes (unless this requirement observed with respect to two classes).

Continuation of Table 2. The results of Experiment 1 (bigrams without nuclear words)

[speech perception and speech production]		
XX века [XX century]		
корпуса текстов [text corpora]		

The term *"контекстной предсказуемости [context predictability]"* is a very clear example of differentiation between statistical (maximum of MI-score) and informant-used approaches: our informants attributed this term to the intersection between "right' and "predictable". What does it mean? Maybe this term is not wide-used, or it belongs to the other domain, but degree of intuitive connectedness and wholeness of the bigram is more important then statistical one for many terminology classification tasks.

We illustrate the capabilities of our method of analysis on a sample list of bigrams (for word forms bigrams) with the nuclear word "corpus" and "word". 42% of collocations were related to nuclear bigrams:

- the core of "right" contains scientific terms (or their components);
- the core of "predictable" contains mainly compound words;
- the core of "others" – a combination that is difficult to interpret.

Experiment 2 has allowed us to establish the degree of connectedness between the components of bigrams and define flexible boundaries between classes.

The data of Experiment 2 verifies those hypotheses on the classification that has been received as the results of Experiment 1. The list of bigrams, the connection of which is estimated by a group of informants is not less than 4 points (average of the group), fully consistent with the core of "right" (according to Experiment 1). These bigrams are allocated on the basis of MI (and sometimes T-score). Those bigrams whose connection is more than 2,8 – the core of the "predictable" (according to Experiment 1).

The core of "others" (the connection is less than 2,8), these are high-mix combinations, which could not be cut off by the correction factor to the extent of T-score.

We have obtained similar results for bigrams with the nuclear word – "corpus" or "word" – as in Experiment 1: the group "core of "right" was just terminological combinations (Table 3).

Table 3. The results of Experiment 1 (bigrams with the nuclear word)

core of "right"	core of "predictable"
аннотированный корпус [annotated corpus]	корпус является [corpus is]
национальный корпус [national corpus]	корпус содержит [corpus contains]
параллельный корпус [parallel corpus]	корпус представляет [corpus represents]
международный корпус [international corpus]	корпус позволяет [corpus allows]
представительный корпус [representative corpus]	данный корпус [this corpus]
размеченный корпус [labeled corpus]	большой корпус [large corpus]
электронный корпус [electronic corpus]	второе слово [second word]
служебное слово [function word]	первое слово [first word]
составное слово [compositum]	данное слово [this word]
	слово встретилось [word is used]

The data analysis of Experiment 2, at first, confirmed the results of Experiment 1, i.e. all bigrams of the list "core of "right" had a value of the connection at least 4. Secondly, Experiment 2 expanded our list of the most associated bigrams by terminological combinations *"главное слово" [main word], "зависимое слово" [depended word], "отдельное слово" [separate word]*, which in experiment 1 were in the intersection of groups of "right" and "predictable" bigrams.

The results of Experiments 1 and 2 allow us to install additional scales, based not only on the values of statistical measures, but also of the feeling the degree of connectivity (by native speakers), which can become explicit during the experiments.

Thus, we propose an experimental approach, which combines the methods of computing experiment, and the experiments with informants. Terminological combinations are the most connected from the viewpoint of the experiment, even when compared with such units as Multiword function words (such as, *в качестве [as], в частности [in particular], за счет [due to]* and etc).

Multiword terminology gets explicit hierarchy in terms of the degree of connectivity

within their own class. For example, from the perspective of our informants (students after corpus linguistics lectures), there are several levels of this kind of connectedness (in descending order), see Table 4 with levels of connectedness from two different perspectives.

Table 4. Examples of levels of connectedness

From the perspective of our informants	From the perspective of MI-score
художественной литературы [fiction]	контекстной предсказуемости [context predictability]
математической лингвистики [mathematical linguistics], корпусной лингвистике [corpus linguistics], имен собственных [proper names], корпусная лингвистика [corpus linguistics], имена собственные [proper names], машинного перевода [machine translation], корпусной лингвистики [corpus linguistics], корпуса текстов [text corpora]	речевой деятельности [speech perception and speech production], художественной литературы [fiction], имен собственных [proper names], корпусная лингвистика [corpus linguistics], имена собственные [proper names]
контекстной предсказуемости [context predictability], речевой деятельности [speech perception and speech production], русского языка [the Russian language]	математической лингвистики [mathematical linguistics], словарной статьи [vocabulary entry]
предметной области [(knowledge) domain]	предметной области [(knowledge) domain]
словарной статьи [vocabulary entry]	машинного перевода [machine translation],

We don't pretend to give an exhaustive description of terminology (for example, the terminology of corpus linguistics). However, the results allow us to take a fresh look at the application of terminology. It is particularly relevant to some new scientific paradigms or to interdisciplinary areas, where there is a process of formation of terminology. In our opinion, the proposed approach has a great future in the study of terminology and predicting the potential of different terminology variants.

Our goal is not only in terminology extracting. Set of terms will not be homogeneous. The advantages of this approach – combined the corpus based and informant-used methods – are in classifying of multiword terminology. This classification must have both statistical and human (informants and/or experts) basis.

We plan to compare the application of terminology in some different domains. We have got a monothematic scientific collection "Hydrogeology" (more then 200 000 "tokens"). Hydrogeology is semantically less closely related to the researchers' field of computational linguistics. We suppose also that hydrogeology has more precise terminology then corpus linguistics. The next step is the comparison of multiword hydrogeology terminology with the terminology of corpus linguistics.

References

Stefan Evert. 2004. *The Statistics of Word Cooccurences: Word Pairs and Collocations*. PhD thesis. Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart.

Chris Manning, Hinrich Schütze. 1999. Collocations. *Foundations of Statistical Natural Language Processing*. MIT Press. Cambridge, MA.

Michael Stubbs. 1995. Collocations and semantic profiles: on the case of the trouble with quantitative studies. *Functions of language* 2(1): 23-55, Benjamins.

Lidia Pivovarova, Elena Yagunova. 2010. Extraction and Classification of Terminology Collocations of the Material of Scientific Texts (preliminary observations). *Materials of II International Symposium "Terminology and Knowledge"*. Moscow.

Maria Khokhlova. 2011. The Research of Lexical-semantic Compatibility in Russian Language with Statistical Measures (on basis of corpora). AKD Saint-Petersburg.

Elena Yagunova, Lidia Pivovarova. 2011. From Collocations To Constructions. *Russian Language: Structural and Lexical-semantic approaches*. Saint-Petersburg.

User-Oriented Data Modelling in Terminography: State-of-the-Art Research on the Needs of Special Language Translators

Georg Löckinger
University of Vienna and
Austrian Academy of Sciences
Vienna, Austria
georg.loeckinger@oeaw.ac.at

Abstract

Special language translators[1] need tailor-made subject field-related information in their daily work. Yet there is a gap between their needs and the subject field-related resources available to them. In my doctoral thesis project, the central research question is whether special language translation can be made more efficient by means of an ideal translation-oriented special language dictionary. To answer this question, first a couple of postulates are put forward. On this basis, a model is built which will later be verified/falsified in an empirical test using "ProTerm", a software for terminology work and text analysis. This will show whether the implemented model can satisfy the needs of special language translators. In the present paper, I aim to give an overview of the research work done so far. In particular, I will provide a summary of 15 postulates derived from scholarly literature and my own professional experience in special language translation and terminology work. Then, I will outline a model that serves as an interface between the specific requirements expressed in the 15 postulates and the implementation using "ProTerm" (bottom-up/top-down approach). Finally, I will briefly describe the next steps in my doctoral thesis project.

1 Introduction

Special language translators have long been waiting for a reference tool that is tailor-made for their needs. In historical terms, Tiktin (1910) provides a good starting point for tracing scholarly literature on the needs of special language translators up to the present. In summary, many authors state that these requirements are known and have been partly met in some cases, but they do not seem to have been implemented consistently or to the full for the benefit of special language translators. Referring to the dream/reality dichotomy, the titles of some relevant publications point very clearly to the gap between what is needed and what exists (e.g., Hartmann, 1988; de Schryver, 2003). Due to this gap, special language translators have started to create their own terminological resources and reference tools, thus assuming the role of terminology producers over and above their original role of terminology users.

2 The Needs of Special Language Translators: 15 Postulates[2]

There are many different requirements that the translation-oriented special language dictionary has to fulfil. This is because special language translation, despite widespread belief to the contrary, is a highly complex process (e.g., Wilss, 1997). The 15 postulates listed below are used as a means to merge all those requirements; they have been derived both from scholarly literature on the practice of special language translation and from this practice itself. Depending on its nature, each postulate is assigned to one of the three requirements categories called "methodology-related", "contents-related" and "related to the presentation and linking of contents". Just as the postulates themselves, these categories complement each other and overlap at some points.

[1] Translators who deal with texts written in special language as defined in ISO 1087-1 (2000): "language used in a subject field … and characterized by the use of specific linguistic means of expression".

[2] It is well beyond the scope of this paper to give a detailed account of the rationale behind each postulate and to cite all the relevant sources. A list of references can be obtained from the author.

2.1 Methodology-Related Requirements

Postulate 1 – Systematic Terminology Work: The translation-oriented special language dictionary must have been compiled in accordance with the principles and methods of systematic terminology work, which is defined in ISO 1087-1 (2000) as "the systematic collection, description, processing and presentation of concepts … and their designations".

Postulate 2 – Description of Methodology Used: The translation-oriented special language dictionary must provide information about the methods used in the underlying lexicographical and/or terminographical process.

2.2 Contents-Related Requirements

Postulate 3 – Terms and Phraseological Units as well as Their Equivalents: The translation-oriented special language dictionary must contain terms, phraseological units and equivalents in the source and target languages.

Postulate 4 – Grammatical Information: The translation-oriented special language dictionary must provide relevant grammatical information on terms, phraseological units and their equivalents.

Postulate 5 – Definitions: The translation-oriented special language dictionary must contain definitions of the concepts described.

Postulate 6 – Contexts: The translation-oriented special language dictionary must provide authentic contexts (primarily in the target language).

Postulate 7 – Encyclopaedic Information: The translation-oriented special language dictionary must contain encyclopaedic information (subject field-related background information, e.g. information about the use of the material object in question).

Postulate 8 – Multimedia Content: The translation-oriented special language dictionary must provide multimedia content, i.e., non-textual illustrations such as figures, videos, etc.

Postulate 9 – Remarks: There must be remarks on the terminology contained in the translation-oriented special language dictionary, e.g. comments on frequent translation mistakes.

2.3 Requirements Related to the Presentation and Linking of Contents

Postulate 10 – Electronic Form: To fulfil most of the other requirements, the translation-oriented special language dictionary must be available electronically.

Postulate 11 – Systematic and Alphabetical Arrangement: The translation-oriented special language dictionary must be both systematically and alphabetically arranged to offer possible solutions to a broad range of translation-related problems.

Postulate 12 – Representation of Concept Relations: The translation-oriented special language dictionary must show concept relations that indicate how various concepts are interrelated.

Postulate 13 – Use of Text Corpora: Since authentic text corpora contain a lot of valuable information, the translation-oriented special language dictionary must both be based on such text corpora and provide direct access to them.

Postulate 14 – Additions and Modifications by the Special Language Translator: The translation-oriented special language dictionary must enable the special language translator to add to and modify it according to his/her needs.

Postulate 15 – One Single User Interface: It must be possible for the special language translator to access the translation-oriented special language dictionary via one single user interface.

3 Model of the Translation-Oriented Special Language Dictionary

The 15 postulates listed in section 2 are to be converted into an appropriate model. They represent requirements for the translation-oriented special language dictionary all of which also reflect the empirical practice of special language translation. Therefore, a model of the translation-oriented special language dictionary is derived inductively from this empirical practice.

Except for postulates 10, 14 and 15, which will become relevant only at the implementation stage, all postulates can be merged into one single model that describes the contents of the translation-oriented special language dictionary. From the meta-model in the international standard ISO 16642 (2003), which represents the highest level of abstraction, a model of the translation-oriented special language dictionary is developed at two lower levels of abstraction (a conceptual data model at the intermediate level and a specific data model at the lowest level). This follows the three-level approach that Budin and Melby (2000) adopted in the "SALT" project.

The modelling process provides a twofold link between empirical practice and theory: firstly, the model at the two levels of abstraction is derived inductively from the postulates listed in

section 2, i.e., from the empirical practice of special language translation; secondly, the model is to be transformed (back) into empirical practice by means of deduction (Budin, 1996) and put to the test in a real-life scenario. The benefit of this step-by-step method is that you can fully dedicate yourself to creating a model that is abstract and thus independent of any specific implementation that might be chosen later according to your needs (e.g., Sager, 1990).

The following subsections 3.1, 3.2 and 3.3 deal with the conceptual data model (including the model of the terminological entry) and the specific data model, respectively. The main focus is on the conceptual data model since this has already been developed to an advanced stage. For a detailed discussion of the meta-model, i.e., the highest level of abstraction, please refer to ISO 16642 (2003).

The conceptual data model is based on the terminological entry model presented by Mayer (1998) and has been modified and extended according to the requirements in my doctoral thesis project. A sketch of the conceptual data model looks as follows:

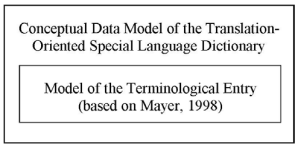

Figure 1. Sketch of the conceptual data model.

3.1 Model of the Terminological Entry (based on Mayer, 1998)

According to the current state of the art in terminographical modelling, the model of the terminological entry has to conform to the following principles: concept orientation (e.g., ISO 16642, 2003), term autonomy (e.g., Schmitz, 2001), data elementarity (e.g., ISO/PRF 26162, 2010), data granularity (e.g., Schmitz, 2001) and repeatability (e.g., ISO/PRF 26162, 2010). Also, the meta-model in ISO 16642 (2003) provides three levels that are relevant for the structuring of terminological data. These three levels are called "terminological entry", "language section" and "term section", respectively.

The data categories listed below result from the 15 postulates mentioned in section 2 and/or from the current state of the art in terminographical modelling (see, in particular, ISO 12620, 1999, and ISO's data category registry "ISOcat" available at www.isocat.org). A plus sign in superscript format "$^{+}$" indicates that the data category in question may contain data elements at one or more of the three levels mentioned above. A superscript capital letter "^{R}" denotes a data category that must be repeatable within the level at which it appears.

The terminological entry level comprises the following data categories: encyclopaedic information^{+}, multimedia content^{R}, remark^{+R}, concept position (if one single concept is described), source identifier^{+R}, administrative information^{+R}. The data categories at the language section level are the following: definition (if one single concept is described) or definition^{R} (if several quasi-equivalent concepts are described), encyclopaedic information^{+}, remark^{+R}, concept position^{R} (if several quasi-equivalent concepts are described), source identifier^{+R}, administrative information^{+R}. Finally, the term section level holds the following data categories: term/phraseological unit^{R}, grammatical information^{R}, context^{R}, encyclopaedic information^{+}, remark^{+R}, source identifier^{+R}, administrative information^{+R}.

3.2 Conceptual Data Model of the Translation-Oriented Special Language Dictionary[3]

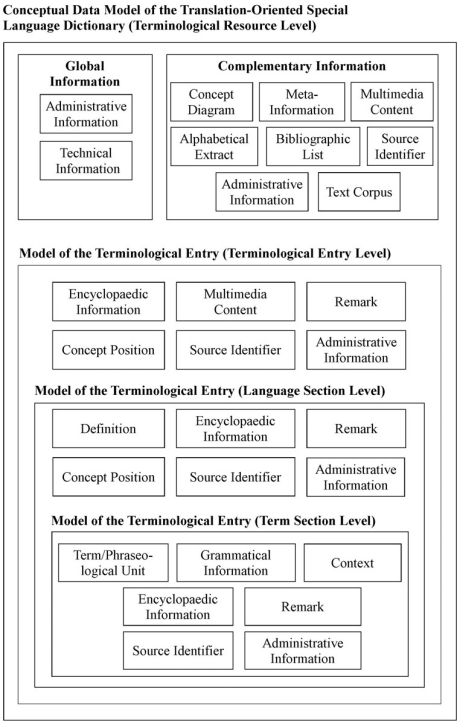

Figure 2. Detailed schematic view of the conceptual data model.

[3] Again, it is well beyond the scope of this paper to describe in detail each of the elements in the model derived from the 15 postulates.

In addition to the three levels discussed in subsection 3.1, the meta-model in ISO 16642 (2003) specifies another two containers at the terminological resource level which are called "global information" (information applying to a complete terminological resource) and "complementary information" (information shared across a terminological resource). The data categories for these two containers have again been derived from the 15 postulates listed in section 2 and/or from the current state of the art in terminographical modelling (see, in particular, ISO 12620, 1999; ISO 16642, 2003; ISO/PRF 26162, 2010; see also ISO 1951, 2007). Thus, the global information container holds technical and administrative information, whereas the complementary information container holds concept diagrams, meta-information describing the translation-oriented special language dictionary, multimedia content, alphabetical extracts (e.g., term indices), bibliographic lists, text corpora, source identifiers and administrative information.

3.3 The Specific Data Model

On the basis of the conceptual data model discussed in subsection 3.2, a specific data model is to be created that will later be implemented in an empirical test using "ProTerm". To that end, the object-oriented modelling language called "Unified Modeling Language" (UML) will be used. The UML is used in relevant international standards (e.g., ISO 16642, 2003; ISO/PRF 26162, 2010) and lends itself to data models that are implemented in relational databases. Yet in principle, UML models are independent of any specific implementation and can thus be used in various technical environments.

The UML model is work in progress, which is why it cannot be published at this stage. The current draft can be provided upon request.

4 Future Work

After refining the conceptual data model as necessary, the next step will be to build a specific data model in the form of a UML diagram that can be used for implementation in "ProTerm". An empirical test will show whether the implemented model can serve the needs of special language translators and answer the central research question. While the model is independent of any specific subject field or language combination, the subject field of terrorism, antiterrorism and counterterrorism will provide the relevant text material in the English and German languages for the empirical test.

References

Gerhard Budin. 1996. *Wissensorganisation und Terminologie: Die Komplexität und Dynamik wissenschaftlicher Informations- und Kommunikationsprozesse*. Narr, Tübingen.

Gerhard Budin and Alan Melby. 2000. Accessibility of Multilingual Terminological Resources – Current Problems and Prospects for the Future. In Antonio Zampolli et al., editors, *Proceedings of the Second International Conference on Language Resources and Evaluation*, volume II. Athens, pages 837–844.

Reinhard R. K. Hartmann. 1988. The Learner's Dictionary: Traum oder Wirklichkeit? In Karl Hyldgaard-Jensen and Arne Zettersten, editors, *Symposium on Lexicography III: Proceedings of the Third International Symposium on Lexicography, May 14–16, 1986 at the University of Copenhagen*. Niemeyer, Tübingen, pages 215–235.

ISO 1087-1. 2000. *Terminology work – Vocabulary – Part 1: Theory and application*.

ISO 12620. 1999. *Computer applications in terminology – Data categories*.

ISO 16642. 2003. *Computer applications in terminology – Terminological markup framework*.

ISO 1951. 2007. *Presentation/representation of entries in dictionaries – Requirements, recommendations and information*.

ISO/PRF 26162. 2010. *Systems to manage terminology, knowledge and content – Design, implementation and maintenance of Terminology Management Systems*.

Felix Mayer. 1998. *Eintragsmodelle für terminologische Datenbanken, Ein Beitrag zur übersetzungsorientierten Terminographie*. Narr, Tübingen.

Juan C. Sager. 1990. *A Practical Course in Terminology Processing*. Benjamins, Amsterdam.

Klaus-Dirk Schmitz. 2001. Systeme zur Terminologieverwaltung: Funktionsprinzipien, Systemtypen und Auswahlkriterien (online edition). *technische kommunikation* 23(2):34–39.

Gilles-Maurice de Schryver, 2003. Lexicographers' Dreams in the Electronic-Dictionary Age. *International Journal of Lexicography* 16(2):143–199.

H. Tiktin. 1910. Wörterbücher der Zukunft. *Germanisch-romanische Monatsschrift II*:243–253.

Wolfram Wilss. 1997. Übersetzen als wissensbasierte Tätigkeit. In Gerhard Budin and Erhard Oeser, editors, *Beiträge zur Terminologie und Wissenstechnik*, TermNet, Wien, pages 151–168.

Demo papers

The Maes T System and its use in the Welsh-Medium Higher Education Terminology Project

Tegau Andrews
Language Technologies Unit
Bangor University
t.andrews@bangor.ac.uk

Gruffudd Prys
Language Technologies Unit
Bangor University
g.prys@bangor.ac.uk

Dewi Bryn Jones
Language Technologies Unit
Bangor University
d.b.jones@bangor.ac.uk

Abstract

This paper describes the Maes T terminology development system, its use in the Welsh-Medium Higher Education Terminology Project, and the manner in which it facilitates collaboration between geographically dispersed terminologists and subject specialists.

1 Introduction

Maes T is a language neutral ISO standards-based online system for the development of terminology resources. It facilitates the creation and publication of dictionaries, whether online, on CD, on mobile phones or in hardcopy. It enables terminologists and subject specialists to collaborate online to standardize terms. The Maes T infrastructure currently supports the Welsh National Terminology Portal and the Terms for Welsh-Medium Higher Education website, which together encompass 18 terminology dictionaries. This paper discusses its use in the standardization of English-Welsh Higher Education terminology.

2 The Welsh-Medium Higher Education Terminology Project

The Welsh-Medium Higher Education Terminology (WMHET) Project is an ongoing scheme which began in September 2009, funded by the Centre for Welsh Medium Higher Education. The project aim is to standardize Welsh terminology in academic fields that lack the terminology required at university level. Using the Maes T system, online dictionaries are developed to aid students, researchers and lecturers across all Welsh universities. The use of a common system with an inbuilt step-by-step standardization process ensures a consistency in terminology development in the sector and helps coordinate the work of geographically dispersed teams of subject specialists and terminologists.

As ISO standard 15188 states (2001: 4.3.4), involving subject specialists in the standardization of terms improves terminological quality, whilst also ensuring the implantation and dissemination of those terms. As the opportunities for providing terminology training for subject specialists are limited, the interface has been carefully designed to be user-friendly and accessible to non-terminologists, with terminology standardization presented as a clear, interface-driven process. The user interface is currently bilingual (English-Welsh) to allow users to access the system in the language of their choice. Crucially this enables non-Welsh speaking subject specialists to contribute in standardization activities such as writing an English definition for a concept.

3 The Maes T Interface

For practical purposes, the design of the Maes T interface incorporates certain assumptions. It assumes that subject specialists will submit an English term for which they wish to standardize a Welsh language equivalent term. Therefore, English is considered a source language throughout. Maes T then divides the terminology standardization process into 4 stages, each identified by a separate tab in the user interface. These are:

1. Collecting Terms,
2. Defining the Concept,
3. Standardizing Terms and
4. Linguistic Information.

In the first stage, *Collecting Terms*, the following initial information is collected, as prescribed in Prys and Jones (2007):

- The source language (SL) standard term
- Source of the SL term
- Any SL synonyms which may occur, along with their sources
- Welsh language candidate terms
- Source of the Welsh candidate terms.

In the second stage, the concept may be defined in either or both the source and target languages, and a disambiguator may be added.

While term collection and concept definition are seen as two steps in standardization within Maes T, they are not considered in isolation from each other. The link between each is manifest in a shared comments box. The comment box is used by all team members as a way to register agreement or disagreement on the candidate terms or underlying concept, and this feature facilitates the development of a consensus between all parties before a term is declared standard.

In third step, *Standardizing Terms*, the terminologist records the consensus achieved regarding the normative status of each candidate term, while in the fourth step, *Linguistic Information*, the terminologist records data such as the part of speech, gender and plural form of the term. Leaving this as a last step removes the tendency, seen in previous Maes T iterations, of recording superfluous grammatical information for candidate terms that are later discarded. It also ensures that subject specialists are free to concentrate on the conceptual aspect of terminology work, leaving grammar to the terminologist.

Once each of these four steps is completed, the terminologist publishes the term online by clicking the 'Publish' button. This feature ensures that only standardized terms are released on the web. Currently, across all terminology projects, some 83,000 concepts are live online through the 'Publish' feature of the Maes T system.

4 Technical Information

Maes T utilizes current web technologies to provide the user interface and to store and disseminate terminology dictionaries. Google Web Toolkit is used to allow for a more interactive responsive web-based interface. The server side software utilizes Welsh linguistic components, such as lemmatizers, spelling and grammar checkers that are then available to Maes T.

5 Conclusion

Maes T was developed to provide a rapid, inexpensive and user-friendly way of standardizing terminology, based on principles of international standards and inclusive, consensus-based working methods. It facilitates the multi-format publication of terminology dictionaries, and has streamlined the process of inputting terms, storing definitions, discussing candidate terms and reaching conclusions, whilst archiving the decisions taken for future reference.

In the context of the WMHET Project, Maes T has given geographically dispersed academics a common platform to facilitate collaborating on mutually beneficial projects, where they can develop a common sense of 'ownership' over newly developed or standardized terms. It is hoped that this will ensure the implantation and dissemination of those terms in the Higher Education sector and beyond.

The application itself is language neutral, and its interface could be translated into other languages as necessary. Generous licensing models for developing countries and lesser-used languages are being discussed.

Reference

Bangor University. 2009. *Maes T*. <www.maes-t.com/gwt> (accessed 17/02/11).

Bangor University. 2010. *Porth Termau Cenedlaethol Cymru/ Welsh National Terminology Portal*. <www.termau.org/porth> (accessed 17/02/11).

Centre for Welsh Medium Higher Education. 2009. *Termau Addysg Uwch/ Terms for Higher Education*. <www.porth.ac.uk/termau> (accessed 17/02/11).

ISO 15188:2001. *Project management guidelines for terminology standardization*.

Prys, Delyth and Dewi Bryn Jones. 2007. *Guidelines for the Standardization of Terminology for the Welsh Assembly Government Translation Service and Welsh Language Board*. <http://www.byig-wlb.org.uk/English/publications/Publications/5338.pdf> (accessed 17/02/11).

A Web-based Terminology Management System and the Translation Market

Balázs Kis
Kilgray Translation Technologies
Gyula, Hungary
balazs.kis@kilgray.com

Peter Reynolds
Kilgray Translation Technologies
Warsaw, Poland
peter.reynolds@kilgray.com

Abstract

This paper introduces a new multilingual terminology management system that suits translators and translation organizations, as well as larger enterprises that are currently the largest purchasers of commercial terminology management systems. The paper outlines the principles behind translation-related terminology management, and demonstrates how the new terminology management tool integrates with the translation environment and the translation workflow.

1 Introduction

There is a latent conflict between terminology management and translation. Terminology management is considered as the means of standardization in order to facilitate unambiguous communication in specific fields. Therefore, terminology management systems tend to focus on creating highly structured, very detailed and elaborate data structures so that all information related to concepts and terms can be entered, looked up, and shared.

As a result, terminology management systems integrate poorly with translation environments. This is not apparent from the technical implementation since commercial examples of integration do exist. The conflict is not in the technicalities but the use case: translators have a different focus that apparently falls outside the attention of creators of legacy terminology management tools. When we created our company, and started selling our translation environment, we were surprised how scarcely terminology tools are used among translators and (smaller) translation organizations (also in Fulford-Zafra 2005).

2 The Business Problem Statement

The main problem of translators is that the time constraints of a translation task do not allow for extensive terminology research. Of all the sophistication of terminology management, translators seldom need more than fundamental dictionary functionality: they immediately need the target-language equivalents while editing the translation, and they also need to be able to add new terms with the smallest possible effort, without leaving the editing environment. This observation led us to the principles upon which we have originally built the terminology component in our translation environment.

As our business was growing, so was the size of our customers. As opposed to small or medium translation organizations, larger enterprises need sophisticated terminology management – it is a core component of their research, development, and communication infrastructure. Likewise, larger translation organizations working for enterprises need the same level of terminology management. Our company adopted the organic-growth approach both in its business and its development strategy. As a result, we have recently encountered customers who needed more than the simplistic terminology management that our translation environment offered. This prompted us to create the terminology management system we are demonstrating in this workshop.

It is not necessarily the technical implementation or the fundamental concepts of the terminology management system that represent a novelty. It is rather the approach to terminology management and the resulting integration of systems that is new in the field, although the basic principles were outlined as early as in 1980 by Martin Kay (Kay 1980).

3 Simplistic Terminology Management in a Translation Tool

The translation environment we mentioned earlier is an integrated desktop application that offers all features in a single program. Features include terminology management in a simplistic way – let us summarize them in a few words:

- Term bases follow a proprietary database format (no relational database systems are used).
- The internal structure of entries complies with legacy terminology systems: we have implemented the three-level structure of (1) concepts (entries), (2) languages (indexes), and (3) terms, with meta-data attached to each level.
- We use a restricted set of meta-data, and the entry structure cannot be modified (although the internal database representation is flexible). If one uses the translation environment without the new terminology system, this limitation still applies.
- On the translation user interface, the structure of the term base entries is hidden: users primarily see source-language terms and their target-language equivalents, both in a match list and highlighted in the source-language text. There are simple commands (clicks and key shortcuts) to insert target-language equivalents in the translation. In addition, simple commands are available to add new terms.

The translation environment has had a server component from the start. Translation memories, term bases, documents (projects), and other resources can be published on a server, and they can be used the same way as in the desktop tool – full transparency is provided.

With server-based term bases, we had also introduced a simple collaborative workflow. A term base can be moderated, which means there are two user roles: translator and terminologist (or proof-reader). Translators can add new terms to the term base, but they are not visible to other translators until the terminologist approves them.

4 The New Development and its Integration

The new terminology management system aims at overcoming the limitations outlined in Section 3. We have planned the development in several stages, of which the first stage is completed as of now.

The system eliminates the restriction of meta-data at all levels of entries (concepts, indexes, and terms). It offers a web-based collaborative editing interface. Full support for importing and exporting TBX files is also included. The system facilitates read-only end-client access, and sophisticated access permissions in general.

The real challenge of this development lies in maintaining the compatibility with the existing translation environment. We have decided to integrate the terminology management with the translation resource server as closely as possible. All functionality related to term base management is still implemented in the translation resource server. The data model required no change because it was originally designed with flexibility in mind, and it was recently scaled up to ca. 2 million entries per term base.

We have also made effort to keep the system transparently interoperable with the translation environment itself. There are two types on the translation resource server: qTerm term bases and legacy term bases. From the translation environment, both types can be queried the same way. However, qTerm term bases can be edited from the new system only. Any legacy term base can be converted into a qTerm term base.

5 Conclusion: Further Development

As we have mentioned above, only the first stage of development is completed. The real novelties in the area of collaborative terminology management and feedback/approval workflow are yet to come in the next 12 months. We plan to conduct extensive beta tests, and one of our reasons to present the tool in this workshop is to invite professionals to give us feedback as the development progresses. Our aim is to create a tool that meets the requirements of the strictest terminology management scheme, yet it retains its simplicity and value for translators and translation organizations.

References

Heather Fulford, Joaquín Granell-Zafra. 2005. Translation and Technology: a Study of UK Freelance Translators. *The Journal of Specialised Translation*, Issue 4, JoSTrans, http://www.jostrans.org/issue04/art_fulford_zafra.pdf

Kay, Martin. 1980. The Proper Place of Men and Machines in Language Translation. *Xerox report CSL-80-11*, Xerox Palo Alto Research Center.