Table of Contents

Tangled Web of Concept Relations. Concept relations for ISO 1087-1 and ISO 704, A. Nuopponen .... 1
Web Content Classification with Topic and Sentiment Analysis, S. Liu [et al.] ................................. 11
A Combined Taxonomic-Frames Approach to Terminological Verb Collocations, D. Hitcheva.............. 20
Document Retrieval for Large Scale Content Analysis using Contextualized Dictionaries, G. Wiedemann [et al.] ....................................................................................................................................... 29
Leipzig Corpus Miner - A Text Mining Infrastructure for Qualitative Data Analysis, A. Niekler [et al.] .......................... 39
Combine EU Terminology with Communication and Ontology Research, R. Maslias ......................... 49
Extracting People's Hobby and Interest Information from Social Media Content, T. Forss [et al.] ....... 58
Service model for semi-automatic generation of multilingual terminology resources, A. Vasiljevs [et al.] .................................................................................................................................................. 67
Theme Classification of Arabic Text: A Statistical Approach, L. Fodil [et al.] ........................................ 77
Information Categories in a Learner's Glossary Entry, B. Alexiev ......................................................... 87
Terminology Management Systems and its Role in the Development of the Language Industry in Colombia, J. Diaz vasquez [et al.] .................................................................................................................................................. 97
Concept System Analysis of the ThirdPlaceLearning Theory, T. Orel [et al.] ..................................... 106
TBX goes TEI ? Implementing a TBX extension for the Text Encoding Initiative guidelines, L. Romary .................................................................................................................................................. 116
A Model and Method to Terminologize Existing Domain Ontologies, D. Gromann ................................. 127
Towards modeling Arabic lexicons compliant LMF in OWL-DL, M. Lhioui [et al.] ............................. 137
TBX-Min: A Simplified TBX-Based Approach to Representing Bilingual Glossaries, A. Lommel [et al.] .................................................................................................................................................. 147
A standard TMF modeling for Arabic patents, C. Ammar [et al.] ........................................................ 157
TermWise: Leveraging Big Data for Terminological Support in Legal Translation, K. Heylen [et al.] .... 157
Segueing from a Data Category Registry to a Data Concept Registry, S. Wright [et al.] ..................... 167
Exploring the Use and Usefulness of KRCs in Translation: Towards a Protocol, E. Planas [et al.] .... 177
TBX between termbases and ontologies, D. Reineke ................................................................................ 188
Tangled Web of Concept Relations. Concept relations for ISO 1087-1 and ISO 704

Anita Nuopponen

University of Vaasa, Finland
{anita.nuopponen}@uwasa.fi

Abstract. The paper discusses factors that are relevant when constructing a typology of concept relations for terminology work by focusing especially on ISO 704:2009 Terminology work - Principles and methods and ISO 1087-1:2000 Terminology work - Vocabulary - Part 1: Theory and application standards and their future revisions. At first prerequisites for a concept relation typology are discussed generally. The standards are then scrutinized as to how they introduce, define and classify concept relation types, and modifications are suggested. A concept relation typology is presented as an example of a comprehensive, generalizable and extendable typology.

Keywords: concept relation, conceptual relation, concept system, generic relation, associative relation, ISO 1087-1, ISO 704.

1 Introduction

The purpose of ISO 704:2009 Terminology work - Principles and methods is to standardize the elements which are essential for terminology work, to provide “a common framework of thinking and to explain how this thinking should be implemented” by practitioners and others involved in terminology work and terminology management. The ISO 1087-1 Terminology work - Vocabulary - Part 1: Theory and application has as its task to define the basic concepts in ISO/TC 37 standards. Both emphasize the meaning of concept relations and concept systems in terminology work. Concept systems in the standards have been scrutinized in [1] while the focus of this paper mainly lies on concept relations.

There is a growing number of researchers from terminology science and related fields (e.g. ontology research) interested in concept relations. However, the standards present a limited range of relation types, and there are some shortcomings in their definitions. This paper addresses these issues and suggests some modifications and possibilities to enhance the typology in future revisions. At first, some general prerequisites for a concept relation typology are taken up with reference to the research in concept relations. The term typology is chosen here instead of (generic) concept system in order to keep apart the object (concept relations and systems) and the metalanguage, which would easily collide in this case.
2 Prerequisites for a Concept Relation Typology

On the one hand, when comparing existing concept relation typologies, several factors influencing them can be distinguished. On the other hand, recent studies show the need for more developed and usable concept relation typologies. In the following, the observations based on these are formulated as prerequisites to be considered or aspects to be aware of when building a typology of concept relations.

1) The **theoretical background** influences the typology and terms utilized. Existing typologies can be traced e.g. to standardization, philosophy, classification studies, semantics, lexicology, ontology work etc. In terminology work and research, various modifications and combinations of these background typologies have been made. Instead of concept relation some authors use the term semantic relation.

2) A typology has usually a **target group**. The target group of the ISO standards primarily consists of field specialists and terminologists who participate in terminology work. Data modelling may utilize terminological concept relations as shown e.g. by [2], but information and data modelling are so far explicitly excluded from the ISO 704:2009. Widening the target group would mean that the different backgrounds have to be taken into consideration and an integrated methodology to be created. This would also include a shared typology of concept relations.

3) A typology is a tool made for a certain **purpose** [3]. Finding out and structuring concept systems of the field is emphasized as an important working method for terminology work. In addition to this, concept relations have a role in information dissemination, which is however lost, if the structural information is omitted from the final product as a data category [4,5]. Information on concept relations in terminology databases can help language professionals (e.g. translators, technical writers), subject-field learners, or even subject field specialists to familiarize themselves with language and concepts of the field [5]. A new generation of concept- and knowledge-oriented terminological databases (e.g. ecolexicon.ugr.es, www.coreon.com) is under development as shown by recent conference papers. More detailed typologies of concept relations are needed to enable navigation through concept relations (see e.g. [6,7]).

4) A typology must be **operationalizable** for the purpose. For instance, Wüster [8] created an extensive typology of relations, only part of which ended up in terminology work. At that time the only terminological products were printed glossaries and standards, why his typology remained a theoretical construct. A more simplified typology was adopted to the practical work, which is reflected in the ISO standards today. Costa and Roche emphasized in their paper at the LSP 2013 [9] that one of the main reasons for rethinking of the ISO conceptual principles is that they “cannot be operationalized” when considering e.g. computational representation of the conceptual system in computer aided translation, (multilingual) specialized dictionaries and content management systems, semantic search engine, knowledge mapping, e-Learning, etc.[9] They did not extend their discussion to concept relation part of the standards, but it is clear that the requirements for concept relation typologies for these purposes differ from those for traditional manual terminology work. The question is, if the scope of ISO 704 and 1087 should be widened to cover also the terminological needs of these purposes, which brings us again to the question of the target group.
5) The typologies of relations are utilized in various ways. Previously, terminology work has been solely a manual effort, which is reflected in standards and textbooks. Both in manual analysis and computer aided extraction, generic and partitive relations are relatively easy to discover. Associative relations, however, make a vast class covering all other possible relations. In manual analysis, the vague set gives a freedom to include any relevant concepts to the vocabulary or database. In computer-aided terminology extraction, however, a set of predefined lexical relation markers may be needed depending on the approach taken.

6) Depending on the purpose, the typology can be domain-dependent or independent. It is challenging to achieve a generalizability because many relation types are more or less domain-dependent [5,10]. Also domain-independency requires great adjustments when applying the typology to new domains [5]. On the other hand, the nature of the concept determines which relations are potentially activated, or seen from the opposite direction: the relation type determines what kind of concepts are involved [6].

7) It takes much effort to analyze more complicated relations than the basic generic and partitive relations. This is one of the reasons for why the manual terminology work is satisfied with a small amount of concept relations as [5] note. This is partly due to the lack of research - or rather lack of operationalization of the results - since as noted by [5] some of the more complicated relations such as causal and instrumental relations etc. have been covered. This has been basic research to find out how various types of relations appear in definitions and texts [5] or theoretical top-down classification [e.g. 11]. The basic problem is how to make the knowledge operational for practical purposes such as terminology work and standardization [cf. 9]. Some results also have more direct use for automatic extraction of terminological information. On the other hand, when systematically compiled and presented glossaries of a restricted thematic field have been scrutinized, generic and partitive relations are far from being the only relation types utilized to link the concepts together (see e.g. 12 on transactional relations).

On the basis of the discussion above the following prerequisites for creating a typology of concept relations could be summarized: theoretical background has to be considered, target group(s) and purpose(s) defined, it must be ensured that the typology is operational for the purpose and fits for its usage. For some purposes (e.g. ISO 704) a domain-independent and generalizable typology is needed than for the analysis of a certain field. The same goes for the continuum complexity-simplicity. In addition to these, there are formal requirements for a concept relation typology, such as unambiguousness and consistency as well as extensibility and flexibility. In the following, it is mostly these formal requirements function as criteria when the concept relations and their typologies in ISO standards are scrutinized.

3 Concept relations in ISO standards

While the standards define the basic concepts of terminology, the concept of concept relation does not get a definition. Instead of saying what the concept relations
“are”, ISO 704: 2009 (p. 8) states: “Concepts do not exist as isolated units of knowledge but always in relation to each other. Our thought processes constantly create and refine the relations between concepts, whether these relations are formally acknowledged or not.” Under the heading “Types of concept relations”, the reader would expect a further discussion on the nature of concept relations and criteria for subdividing them as well as on how the concept relations relate to object relations (ontical relations). Instead, the standard makes remarks on organizing concepts into concept systems and aspects that have to be kept in mind as to concept fields. Concept systems and concept fields are defined and introduced ten pages later in the document. Furthermore, the example on mice and computer mouse not being parts of the same subject field does not either belong to the topic of the chapter.

In general, the standard does not keep apart concept relations from concept systems. Concept relations are sometimes explained and exemplified by talking explicitly about concept systems as above, or by using the term generic or partitive relation when clearly talking about the respective concept system (e.g. “in a generic relation there may be several ways of subdividing a concept into subordinate concepts depending on the criteria of subdivision or type of characteristic chosen”; p. 11). The examples for generic relation and partitive relation exemplify the respective concept systems with detailed explanations and instructions for what to observe when structuring this kind of concept systems. There is a chapter for concept systems later on in the standard, which is now missing relevant content, because most of it has been already spread throughout the document [12].

### 3.1 Main relation types

Both standards distinguish between hierarchical and associative relations as the main division (see Fig. 1). The definitions of these two coordinate concepts take totally different approaches. The first one does not tell what the distinguishing characteristic is and how hierarchical relation can be distinguished from its coordinate concept. Instead, two different criteria for subdivision can be detected between the lines: ability to build hierarchies and type of associative/thematic connection.

![Figure 1. The main types of concept relations in ISO 1087-1:2000.](image-url)
Despite the asymmetry, this dichotomy has established itself in the principles of terminology work. However, if we want to define what concept relation is or to expand the amount of concept relations in the typology, this main division causes problems. Because hierarchical relations are restricted to generic and partitive by the standards, all the remaining relations are non-hierarchical by definition.

3.2 Hierarchical relations

As mentioned above, ISO 1087-1:2000 (p. 4) restricts hierarchical relations only to generic and partitive relations by. ISO 704:2009 (p. 8) makes it stipulative by adding “In this International Standard...”. The category hierarchical relation seems somewhat superfluous, because most what is said about it in ISO 704:2009, is said about generic relations and generic concept systems. The metaphorical similarity between generic and partitive relations is that they are able to form hierarchies, which could, however, be applied even to further types of concept relations. After all, there are other relation types that fill (even better) the requirements for what generally is understood by ‘hierarchy’ e.g. in systems theory, organization theory, ecology etc. According to 704:2009 (p. 8), “in a hierarchical relation, concepts are organized into levels of superordinate and subordinate concepts. For there to be a hierarchy, there must be at least one subordinate concept below a superordinate concept.”

What is common to generic and partitive concept systems is actually that they are nested hierarchies, i.e. superordinate concepts in a way or another “contain” or “consist of” the subordinate concepts. A generic superordinate concept contains the extension of its subordinate concepts; a partitive superordinate concept refers to a whole while its subordinate concept refers to a part in the whole. Also concept systems based on locational relations could be regarded as nested hierarchies, e.g. computer disc contains folders, they contain files and files contain data. The same goes for material component relations: butter contains butterfat that contains fatty acid. Examples of not nested hierarchies are military hierarchies and ecosystem’s food chains [17]. In them the entities on the higher level do not contain or consist of the entities on lower level. The hierarchical relation between them and the corresponding concepts is based on some other criteria than containment. This type of concept relation appears in e.g. [13], where it is called rank relation. These tree relations mentioned above cannot be, however, included in neither of the main relation types in the standards because of the restrictions made in the definitions.

3.3 Types of generic and partitive relations

According to the standards, both generic and partitive relations are relations between super- and subordinate concepts in respective concept systems. E.g. ISO 704:2009 (p. 9) defines generic relation as a relation that “exists between two concepts when the intension of the subordinate concept includes the intension of the superordinate concept plus at least one additional delimiting characteristic”. Even though ISO 704:2009 mentions coordinate concepts and horizontal series (e.g.
“Partitive relations, like generic relations, can be expressed as vertical and horizontal series”, p. 15), the relation typology does not cover the relation between the coordinate concepts or other types of relations in the concept systems. This has been solved e.g. in [11,13] by assigning the terms generic relation and partitive relation to wider concepts which cover also the relations between co-ordinate concepts (see Fig. 2). Respectively partitive concept relations could be partitive superordination and co-ordination. In ISO 704:2009 the concept systems formed by these relations are called generic and partitive concept systems. Thus it is motivated to call all the relations in them generic respective partitive relations.

Because ISO 704:2009 does not develop further the theoretical background of the concept relation typology, some problems appear in the definitions. The standard says forinstance that “...if the same concept is viewed as a comprehensive concept in a partitive relation, the individual concept can be subdivided into its parts” (p. 16). Here a distinction between the object and concept levels should be made clearer – it is not the concept that is subdivided into parts but the object that the concept represents (c.f. ISO 704:2009: 2). The definition in ISO 704:2009 also says that “A partitive relation is said to exist when the superordinate concept represents a whole, while the subordinate concepts represent parts of that whole. The parts come together to form the whole.” The “whole” here refers to the object that is being devided and not to the concept. Parts of the concepts are its characteristics. As to the typology of partitive relations, further types could be distinguished as has been done e.g. in Fig. 3, where they are divided into compound, partition, and set relations.

3.4 Associative relations

Similarly to partitive relations, the definitions of the associative relations do not keep clearly apart the object and concept level, e.g. “Some associative relations exist when dependence is established between concepts with respect to their proximity in space or time.” (ISO 704:2009: 17) Again, it is not the concepts that have a spatial or temporal contact but the objects. As associative relations ISO 1087-1:2000 mentions sequential, causal and temporal relations while ISO 704:2009 does not give any typology for associative relations but plenty of examples. Instead of isolated examples, the standard could present some kind of classification or a more developed typology – eventually as an annex. There is a need for one when we look at the new developments of terminological data bases. Forinstance, León Araúz et al. [6, p. 32] say that terminological knowledge bases are restricted to these basic relations, “whereas conceptual dynamism can only be fully reflected through non-hierarchical ones”, which relate to “movement, action and change, which are directly linked to human experience and perceptually salient conceptual features”.

5 Concept Relation Typology

Even though it is a challenge to compile a typology, there are some existing ones to start with e.g. the one introduced in [11], which is taken here as an example. The ty-
The main division is made into generic (syn. logical) and ontological relations, where ontological are divided into contiguity (in space or time) and influence (causal, developmental, activity, origin and interactional) relations. Influence relations have a causal component and are overlapping with each other in some degree. A distinction between causal and purely temporal concept relations is made, and purely causal concept relations are separated from other relations which include causal components.

As mentioned earlier, generic concept relations are divided into four subtypes which are those between concepts in higher and lower, lower and higher, or same level of abstraction as well as between concepts in other positions on different levels of abstraction in the same concept system (see Fig. 2). The typology takes also another approach to generic relations, and compares the intension and extension of concepts (see Fig. 2). These distinctions are useful when analysing concepts and comparing e.g. different languages or on different fields. The following types of relations are presented in [11,13]:

- intensional relation (based on similarity and differences in concept characteristics): intensional identity (concepts have same intension i.e. same characteristics); intensional inclusion (intensionally wider/narrower concept, both have same characteristics, one of them has additionally one or more): intensional overlapping (concepts have a set of same characteristics, both have one or more additional characteristics); intensional disjunction (concepts do not have any common characteristics);

- extensional relation (based on similarity and differences in concept extensions i.e. subordinate concepts or objects): extensional identity (concepts have the same ex-
tension); extensional inclusion (both have same extension, one has one or more subordinate concepts/objects in addition); extensional overlapping (concepts share a set of subordinate concepts/objects, both have one or more in addition); extensional disjunction (two concepts do not share any subordinate concepts/objects).

**Fig. 3.** Ontological concept relations [1,11,13,14]

In the Fig. 3, the original hierarchy of ontological concept relations [1994] has been flattened on both ends to make the typology of ontological concept relations more operationalizable. The typology allows a wide variety of very specific relations and relation types to be subordinated to the relevant category. New relation types can be added; e.g. when reviewing the typology for this paper, a new relation called *nexus relation* was added (Fig. 3). It is based on connection between objects that are not parts and wholes in relation to each other, nor attachments or locations. An example could be mobile device–Internet.

Further modifications have been made here to add more flexibility to the lower level relation types. Instead of listing subtypes of relations, only examples of possible “concept roles” or “relation participants” have been listed. The complete typology includes also relationships between the two concepts also from the opposite direction, e.g. sender–receiver and receiver–sender relation. Sometimes it may be important to separate this type of information. Additionally, parallel or simultaneous relations are often involved, e.g. when we deal with concepts referring to multiple, alternative or alternating senders or receivers, which reminds the generic coordination and
sometimes overlaps with it when e.g. a classification of concepts for various senders or receivers is made in a transmission concept system.

6 Discussion

There is a need to revise concept relation typologies for terminology work. New applications for the principles of terminology work and fast pace of digitalization of “everything” emphasizes the need for more developed terminological tools. The concept relation typology presented in ISO standards is restricted to a few core relation types, and their definitions and treatment are not quite unambiguous or consistent. They may be operationalizable for manual terminology work but also there a more extended typology could be fruitful. As shown here, even the most basic relation types are quite complicated when we take a closer look at them.

Domain-dependency poses challenges to create a typology that is general enough to fit for concept analysis of various fields. On the one hand, fields may have their own frequent relation types on the micro level, which cannot be easily generalized or operationalized in other fields. On the other hand, a too abstract relation type causes also problems by being too vague and difficult to locate in specific fields. Especially when working with corpora and automatically extracting concept relations, it may be a daunting task to trace back to the concept relations and concept systems. Same relations may be expressed in myriads of ways, e.g. “is a” is only one way to express generic relation, and on the other hand, the same expression may refer to several other relation types (e.g. “can be divided into”).

The typology presented in Chapter 5 has been tested over the years in a multitude of fields and remodeled according to the problems encountered. However, several of the more complicated relation types still need more detailed analysis in order to make the typology to work properly as an analysis tool. Also when we consider needs for other than terminologists and manual terminology work, more testing and adaptation needs to be done.

Many of these questions and challenges are to be considered when remodeling and enhancing a concept relation typology for ISO standards - or for any other purpose: which theoretical background and terminology to lean on, for which target group, purpose and context the typology is meant, how to operationalize the new typology, how to handle the complexity and domain-(in)dependency of relations, how to ensure comprehensiveness, flexibility and extendibility of the typology, etc. Above all the typology must also be unambiguous and consistent even though the relations encountered may feel like a tangled web.

References

Web Content Classification with Topic and Sentiment Analysis

Shuhua Liu, Thomas Forss and Kaj-Mikael Björk

Department of Business Administration and Analytics
Arcada University of Applied Sciences
Jan-Magnus Jansson aukio 1 00560 Helsinki Finland
{shuhua.liu, thomas.forss, kaj-mikael.bjork}@arcada.fi

Abstract. Automatic classification of web content has been studied extensively, using different learning methods and tools, investigating different datasets to serve different purposes. Most of the studies have made use of content and structural features of web pages. In this study we present a new approach for automatically classifying web pages into pre-defined topic categories. We apply text summarization and sentiment analysis techniques to extract topic and sentiment indicators of web pages. We then build classifiers based on the extracted topic and sentiment features. Our results offer valuable insights and inputs to the development of web detection systems.

Keywords: web content classification, sentiment analysis, text summarization, online safety solutions

1 Introduction

Web content classification, also known as web content categorization, is the process of assigning one or more predefined category labels to a web page. It is often formulated as a supervised learning problem where classifiers are built through training and validating using a set of labeled data. The classifiers can then be applied to label new web pages, or in other words, to detect if a new webpage falls into certain predefined categories.

Automatic classification of web pages has been studied extensively, using different learning methods and tools, investigating different datasets to serve different purposes [13]. Chakrabarti et al [3] studied hypertext categorization using hyperlinks. Chen and Dumais [4, 7] explored the use of hierarchical structure for classifying a large, heterogeneous collection of web content. They applied SVM classifiers in the context of hierarchical classification and found small advantages in accuracy for hierarchical models over flat (non-hierarchical) models. They also found the same accuracy using a sequential Boolean decision rule and a multiplicative decision rule, with much more efficiency.
Our research concerns the development of classification systems for online safety and security solutions. Our work is motivated by the fact that certain groups of web pages such as those carry hate and violence content have proved to be much harder to classify with good accuracy when both content and structural features are already taken into consideration. There is a need for better detection systems that utilize enriched features coupled with good classification methods for identifying excessively offensive and harmful websites.

Hate and violence web pages often carry strong negative sentiment while their topics may vary a lot. Based on this observation, in this study we explored the effectiveness of combined topic and sentiment features for improving automatic classification of web content. We first apply text summarization and sentiment analysis techniques to extract topic and sentiment indicators of web pages. We then build classifiers based on the extracted topic and sentiment features. Large amount of experiments and analysis were carried out. Our results offer valuable insights and inputs to the development of web detection systems and online safety solutions.

2 Data and Methods

Our dataset is a collection of over 165,000 single labeled web pages in 20 categories. Each webpage is represented by a total of 31 attributes including full page, URL, Title and other meta-content, plus structural info and link information. The experiments reported in this paper mainly concern a number of specific web categories (described in Section 3 and Section 4).

In this study we only take into consideration the page attributes that are text-related. Our focus is on added value to web classification that can be gained from textual content analysis. Taking into account of missing entries for different attributes, we selected a subset of the content features as the raw data for our study: full page free text content, URL words, title words (TextTitle), meta-description terms (Cobra-Text, CobraMetaDescription, CobraMetaKeywords, TagTextA and TagTextMetaContent).

We should point out that, structural features and hyperlink information capture the design elements of web pages that may also serve as effective indicators of their content nature and category [2, 6]. They contain very useful information for web classification. In addition, analysis of images contained in a web page would provide another source of useful information for web classification [5]. However, these topics are studied in other projects.

Our approach to web content classification is illustrated in Figure 1. Exploring the textual information, we applied word weighting, text summarization and sentiment analysis techniques to extract topic features, content similarity features and sentiment indicators of web pages to build classifiers.
2.1 Topic Extraction

We start with extracting topics from each web page and the collections of web pages belonging to same categories. The extracted topics hopefully give a good representation of the core content of a web page or a web category.

Topic extraction is based on automatic identification of important and informative terms from a text. The goal is to select a set of words or phrases that are related to the main topics discussed in the given text. Topic extraction has been the subject of study for a long time, and there exists a large body of literature on it and many proposed methods. Hasan and Ng [20] presented a recent survey of the state of the art in key phrase extraction, and grouped topic extraction methods into two broad categories: supervised and unsupervised. Some early studies on key phrase extraction took a supervised approach and formulated the task as a classification problem [17, 22]. Later on Jiang et al. [18] proposed a pairwise ranking approach, to learn a ranker that rate two candidate key phrases and has been shown to significantly outperform the classification methods. These supervised methods make use of statistical features, structural features, syntactic features of the corpus, as well as external resource based features (e.g. Wikipedia). Unsupervised approach on the other hand applied graph-based ranking methods [21, 23, 24], clustering methods [21, 25, 26] (Grineva et al., 2009; Liu et al., 2009b; Liu et al., 2010) and language modeling methods [19].

The various different methods, with their pros and cons, have brought continuous development in the field. However, as pointed out in Liu et al [26] and Hasan and Ng [20], topic extraction is still a task far from being accomplished when we look at the state-of-the-art performance level. To make further improvements it is more important to incorporate background knowledge than solely focus on algorithmic development.
For our study, we choose to make use of results from text summarization research. Text summarization tools have the capability to distill the most important content from text documents. However, most of the text summarization systems are concerned with sentence extraction targeted for human users. To help web content classification, we believe simple term extraction could be a sufficiently effective and more efficient approach, as topic extraction is just one of the many middle-steps towards facilitating open domain text classification, we want to keep it generic, simple and efficient. So we applied the time-tested tf-idf weighting method to extract topic terms from web pages [16]. Terms in this study are still limited to individual words (experiments with n-grams in our later work). We will compare the effect of using n-grams and language models in another article.

For each webpage, we make use of its different content attributes as input for term weighting. Applying different compression rate, we obtained different sets of topic words (e.g. top 50, top 100, top 20%, 35%, 50%, 100%).

The content of a web category is obtained through summarization of all the web pages in the same category. For each web page collection, we apply the Centroid method of the MEAD summarization tool to make summaries of the document collection [14, 15]. Through this we try to extract topics that are a good representation of a specific web category. MEAD is applied here instead of simply tf-idf weighting to facilitate processing of large collection of web pages and reducing redundancy. MEAD offers a benchmarking text summarization method. Given a document or a collection of documents to be summarized, it creates a cluster and all sentences in the cluster are represented using tf-idf weighted vector space model. A pseudo sentence, which is the average of all the sentences in the cluster, is then calculated. This pseudo sentence is regarded as the centroid of the document (cluster). A centroid represents a set of the most important/informative words of the whole cluster, thus can be regarded as the best representation of the entire document collection.

### 2.2 Extracting Sentiment Features

Sentiment analysis is the process of automatic extraction and assessment of sentiment-related information from text. Sentiment analysis has been applied widely in extracting opinions from product reviews, discovering affective dimension of the social web [8, 11].

Sentiment analysis methods generally fall into two categories: (1) the lexical approach - unsupervised, use direct indicators of sentiment, i.e. sentiment bearing words; (2) the learning approach - supervised, classification based algorithms, exploit indirect indicators of sentiment that can reflect genre or topic specific sentiment patterns. Performance of supervised methods and unsupervised methods vary depending on text types [12].
SentiStrength [11, 12] takes a lexical approach to sentiment analysis, making use of a combination of sentiment lexical resources, semantic rules, heuristic rules and additional rules. It contains an EmotionLookupTable of 2310 sentiment words and wordstems taken from Linguistic Inquiry and Word Count (LIWC) program [9], the General Inquirer list of sentiment terms [10] and ad-hoc additions made during testing of the system. The SentiStrength algorithm has been tested on several social web data sets such as MySpace, Twitter, YouTube, Digg, Runners World, BBC Forums. It was found to be robust enough to be applied to a wide variety of social web contexts.

While most opinion mining algorithms attempt to identify the polarity of sentiment in text - positive, negative or neutral, SentiStrength gives sentiment measurement on both positive and negative direction with the strength of sentiment expressed on different scales. To help web content classification, we use sentiment features to get a grasp of the sentiment tone of a web page. This is different from the sentiment of opinions concerning a specific entity, as in traditional opinion mining literature.

As a starting point, we apply unsupervised method to the original SentiStrength system [11, 12]. Sentiment features are extracted by using the key topic terms extracted from the topic extraction process as input to SentiStrength. This gives sentiment strength value for each web page in the range of -5 to +5, with -5 indicating strong negative sentiment and +5 indicating strong positive sentiment. We found that negative sentiment strength value a better discriminator of web content than positive sentiment strength value at least for the three web categories Hate, Violence and Racism. Thus, in our first set of experiments we only uses negative sentiment strength value as data for learning and prediction. Corresponding to the six sets of topic words for each web page, six sentiment features are obtained.

2.3 Extracting Topic Similarity Features

We use topic similarity to measure the content similarity between a web page and a web category. Topic similarity is implemented as the cosine similarity between topic terms of a web page and topic terms of each web category. Topic terms are a set of top-weighted individual words. We set a length for the topic vectors based on testing of several options. A set of similarity features is extracted for each web page, considering different compression rates.

3 Sentiment based Classifier for Detecting Hate, Violence and Racism Web Pages

Three datasets are sampled from the full database. The datasets contain training data with balanced positive and negative examples for the three web categories: Hate, Violence and Racism. Each dataset makes maximal use of positive examples available, resulting in a dataset of 3635 web pages for Violence, 9040 for Hate and 6155 for
Racism. Features for learning include a number of negative sentiment strength values of each web page, based on different sets of topic terms.

We built classification model using NaiveBayes (NB) method with cross validation, as three binary classifiers: $c = 1$, belong to the category, (Violence, Hate, Jew-Racism), $c = 0$ (not belong to the category). NB Classifier is a simple but highly effective text classification algorithm that has been shown to perform very well on language data. It uses the joint probabilities of features and categories to estimate the probabilities of categories given a document. Support Vector Machines (SVM) is another most commonly used algorithms in classification and foundation for building highly effective classifiers to achieve impressive accuracy in text classification. We experimented with both NB and SVM methods, found that they achieved similar results, while SVM training takes much longer time in training.

We tested with different combination of the sentiment features. The best results show good precision and recall levels for all three categories.

<table>
<thead>
<tr>
<th>Model Performance</th>
<th>Category</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hate</td>
<td>71.38%</td>
<td>77.16%</td>
<td></td>
</tr>
<tr>
<td>Racism</td>
<td>63.29%</td>
<td>72.79%</td>
<td></td>
</tr>
<tr>
<td>Violence</td>
<td>81.91%</td>
<td>73.92%</td>
<td></td>
</tr>
</tbody>
</table>

4 Combining Topic Similarity and Sentiment Analysis in Web Content Classification

Following our first batch experiments, we extend our study from 3 to 8 web categories (dataset size 3282, 3479, 5105, 400, 4667, 5438, 1919, 3432). We first developed classifiers based on topic similarity features, and the results were very disappointing for most categories, low on both precision and recall measures in general. We thus conclude that topic similarity based classifiers alone do not perform well.

Next we seek to improve the classification performance through combined use of topic similarity features and sentiment features. The results are very encouraging and the classification performance is significantly improved for most categories.

4.1 Extracting New Sentiment Features

In this second round of experiments we made use of combined metadata of web pages as raw data, extracted topic terms and then the sentiment features again for each web page. We tried different ways to customize the SentiStrength algorithm: (1) Counts of the amount of positives and negative sentiment words in a web page; (2) Sum of word
sentiment value weighted by word frequency, normalized on total word counts, value between -5 and 5; (3) update the EmotionLookupTable. We found only few novel terms comparing with the original EmotionLookupTable, so we didn’t pursue it further as the effect would be minor.

We tested new sentiment feature based NB classifiers for a few web categories. They do not necessarily perform better than the earlier sentiment based classifier. The performance varies from category to category, some slightly better, some not.

4.2 Classification using Combined Features

Next, we built NaïveBayes classification models (with cross validation) for eight web categories, using combined topic similarity features and sentiment features. The model performances are significantly improved for almost all categories when compared with solely sentiment based or solely topic similarity based classifiers, as is shown in Table 2. Recall levels are especially good, except the Violence category, which has a bit lower recall level but very good precision level.

Table 2. Classifiers making use of combined sentiment and topic similarity features

<table>
<thead>
<tr>
<th>Category</th>
<th>Precision</th>
<th>Recall</th>
<th>Category</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cults</td>
<td>75.8%</td>
<td>90.55%</td>
<td>RacismWh</td>
<td>98.26%</td>
<td>96.30%</td>
</tr>
<tr>
<td>Occults</td>
<td>87.08%</td>
<td>91.84%</td>
<td>RacistGr</td>
<td>69.96%</td>
<td>91.82%</td>
</tr>
<tr>
<td>Violence</td>
<td>93.69%</td>
<td>82.75%</td>
<td>JewRel</td>
<td>64.43%</td>
<td>96.28%</td>
</tr>
<tr>
<td>Unknown</td>
<td>89.59%</td>
<td>93.31%</td>
<td>Religion</td>
<td>67.01%</td>
<td>92.81%</td>
</tr>
</tbody>
</table>

5 Conclusions and Future Work

In this study we set out to build sentiment aware web content detection systems. We developed different models for automatically classifying web pages into pre-defined topic categories. Word weighting, text summarization and sentiment analysis techniques are applied to extract topic and sentiment indicators of web pages. Large amount of experiments were carried out and classifiers are built based on topic similarity and sentiment features. Our results indicate that sentiment based classifiers bring much added value in the classification of Violence, Hate and Racism webpages. Topic similarity based classifiers solely do not perform well, but when topic similarity and sentiment features are combined, the classification model performance is significantly improved for most of the eight selected web categories.

Our future work would include the incorporation of LDA topic models [1] and its variations, n-grams, word ontology, domain knowledge and structural features. We will also look into new topic similarity measures. We believe there is still much room
for improvements and some of these methods will hopefully help to enhance the classification performance to a new level. Our goal will be on improving precision and reducing false positives.

References


A Combined Taxonomic-Frames Approach to Terminological Verb Collocations

Dobromira Hitcheva

University of Architecture, Civil Engineering and Geodesy - Sofia

Abstract. This paper proposes a combined taxonomic-frames approach to terminological verb collocations considered within the broad category of terminological units. The latest approaches to terminology are briefly reviewed. Special attention is paid to the methodology of frame-based terminology, emphasizing its appropriateness for analyzing that type of collocations. The concept designated by the term in the collocation is assumed to be the Patient in a dynamic scene whereas the verb designates a dynamic action/process affecting it with a consequent Result. To perform the frame semantic analysis of the scenes, it is deemed necessary to first categorize those event structures into conceptual groups according to the type of Agent and Result. Such a preliminary taxonomic analysis is required to make the overall analytical procedure more explicit and applicable in a subsequent terminological contrastive analysis involved in multilingual terminological projects. The analytical procedures are exemplified with the term ‘steel’ and its verb collocators.

Keywords: terminological verb collocations; terminological units; frame-based terminology; taxonomic-frames approach;

1 Introduction

Terminological collocations have recently attracted the attention of terminographers as necessary information categories to be used in the entries of modern translation-oriented terminological collections. Of particular relevance for technical translation are terminological collocations in which the semantically autonomous element, i.e. the base, is a nominal head term and the element selected depending on the base, i.e. the collocator is a verb. Those collocations, referred to as terminological verb collocations, have been studied by modern computational and corpus linguistic methods with the aim of organizing and presenting them in specialized dictionaries. However, as compared with general language collocations, the number of studies concerning terminological collocations is considerably smaller focusing mainly on collocation extraction techniques. One reason for that situation is the fact that most terminologists somehow refrain from acknowledging the status of terminological collocations, and especially terminological verb collocations, as terminological units. Another reason for the still scarce investigations into specialised collocations is the lack of a methodology based on sound theoretical underpinnings which can be appropriate for their analysis.

This paper aims to present a proposal for a combined taxonomic-frames approach to terminological verb collocations. To achieve that goal, I will first argue for considering terminological verb collocations as legitimate objects of terminological study, i.e. terminological units. Then I will discuss the recent trends in terminology theory and pay special attention to the frame-based approach to describing terminological relations in a special domain which is assumed to be appropriate for analyzing terminological verb collocations. With respect to the latter, I will adduce arguments to justify a proposal for a taxonomic analysis preceding the analysis of the cognitive frames underlying those collocations. That preliminary analysis is deemed necessary to first categorize the event structures into conceptual groups thus making the analytical procedure more compatible with a possible terminological contrastive analysis which requires identification of systematic sets. The head term ‘steel’ and the terminological verb collocations it forms will be used to exemplify the proposed methodology.

2 Terminological Verb Collocations as Terminological Units

Cabré [1] defines terminological units as being simultaneously language units, cognitive elements and vehicles of communication. As regards the formal aspect of terminological units, she distinguishes between simple and complex terms, the latter involving terminological phrases such as flat-glass water gauge,
which could be termed ‘terminological nominal collocations’. Cabré also identifies an intermediate formal category between terminological phrases and free syntactic formations which she refers to as “collocations” exemplified by what can logically be defined as ‘terminological verb collocations’, namely, *reboot the computer, propose an amendment*, etc. However, in her opinion these collocations cannot be categorized as terminological phrases because “they do not correspond to established concepts”. From the point of view of the most recent developments in computational terminology and cognitive terminology theories, at least two arguments can be provided against that assumption.

One strong argument for recognizing terminological verb collocations of the type Verb + Term / Term + Verb as terminological units is of purely pragmatic character. That type of collocations has definitely attracted the attention of modern computational terminologists and any attempt at citing publications on the issue of terminological verb collocation extraction would end up in a long list. It will suffice here to refer to most modern term extractors which enable users to extract automatically from text corpora terminological verb collocation bigrams. The heightened interest in those syntagmatic structures can be accounted for by the increasing need to compile collocationally rich entries in dictionaries specially made for technical translators. This need can be compared with the need people once felt for creating units of measuring length which led to introducing the meter, the foot, the centimetre, the inch, etc. They are all necessary measurement units for doing a large number of human activities and nobody has ever asked the question of which one is more or less important or whether one or another should be deleted from the list of length measuring units.

Another argument in favour of classifying terminological verb collocations as terminological units is of theoretical nature. In [2] we argue for recognizing that type of collocations as ‘terminological knowledge units’, a term introduced comparatively recently to denote terminological data extracted by semi-automatic techniques and used to discover semantic term relations, construct definitions and acquire domain knowledge. Our assumption is based on the four levels of knowledge organization in relation to their referents in the real or abstract world within a conceptual framework, postulated by Dahlberg [3]: (a) knowledge elements understood as conceptual characteristics (Level 1); (b) knowledge units equated with concepts (Level 2); (c) larger knowledge units equated with concept combinations (Level 3); and (d) knowledge systems equated with domain-specific terminological systems (Level 4). In this framework terminological collocations in general and terminological verb collocations, in particular, can be interpreted as linguistic exponents of knowledge units belonging to Level 3, i.e. concept combinations, which can be subjected to some categorization/classification (e.g. *steel: melted, cast, rolled – stages in steel production*). Thus, it is possible to identify conceptual groups of lower-level events, expressed by terminological verb collocations, which are subordinate to a higher-level event associated with a basic concept and expressed by a basic nominal term and a generalised predicate. I will develop that assumption further in the following sections. Since the methodology I will propose is a slightly modified version of the frame-based approach to terminology, I will first discuss some of the latest trends in Terminology Theory which have given rise to that approach.

### 3 Recent Trends in Terminology Theory

Faber [4] in her critical overview of terminology theories concludes that “the study of specialized language is undergoing a cognitive shift” which is expressed in laying greater emphasis on conceptual structures underlying special language text. She discusses some novel approaches to terminology against the background of the classical *General Theory of Terminology* (henceforth GTT) developed by the Vienna School of Terminology. The brief review of that theory states a number of limitations which can be summarized as follows:

- The multidimensional nature of specialized knowledge concepts is not taken into account;
- Concepts are viewed as entities referring to the objective extralinguistic world and the terms are considered to be only their abstract symbols, i.e. labels thus making a sharp distinction between general language words and terms;
- The GTT is focused on eliminating ambiguity in scientific communication by standardizing terminological units thereby excluding terminological variation which inevitably occurs in special language use;
- Since the GTT disregards completely the syntax and pragmatics of specialized language, it is not applicable to technical translation;
- Being exclusively synchronic, the GTT does not account for the diachronic dimensions of terms which are relevant for describing the semantic methods of term formation such as metaphor and metonymy;

Faber then discusses the new approaches proposed since 1990s which “paved the way to integrating Terminology into a wider social, communicative and linguistic context”. Among these are the Communicative Theory of Terminology, the Sociocognitive Terminology and the Frame-Based Terminology, the latter being her own proposal and the one that I will try to apply, with certain modifications, to the analysis of terminological verb collocations. Following Faber [4], I will present brief accounts of the former two approaches to Terminology with the view to pointing out the major differences between the GTT and those theories.

The Communicative Theory of Terminology (henceforth CTT) is proposed by Cabré [1], [5]. As has already been mentioned, this theoretical approach highlights the complex multidimensional nature of terminological units which can be seen from three different perspectives: communicative, linguistic and cognitive. Instead of taking the previously constructed concepts, as is the case with the GTT, Cabré takes the terminological units to be the central object of terminological study thus showing preference for a linguistic rather than semiotic approach to terminology. Conceived as indivisible combinations of form and content, terminological units do not only designate but they also mean with all the cognitive consequences of their meaning. Therefore, they behave like general language words, their specific character proceeding from a number of cognitive, syntactic and pragmatic constraints imposed by their functioning in specialized discourse.

Cabré proposes possible ways of accessing terminological units in order to be analyzed by making use of the metaphor “the Theory of Doors”. A terminological unit is compared to a polyhedron, a solid figure having many faces, since it also has three dimensions, viz. cognitive, linguistic and communicative. Each dimension can be a door through which one can get access to terminological units. What is important to note is that the choice of door does not exclude the other two perspectives which remain in the background. The units are usually approached through the language door but always within the framework of specialized communication.

Commenting on the CTT, Faber [4] asserts that it is “a viable, working theory of terminology” which “has led to a valuable body of research on different aspects of Terminology such as conceptual relations, terminological variation, term extraction and the application of different linguistic models to terminology”. At the same time she indicates some shortcomings of the theory which I will sum up as follows:
- The CTT does not adopt any specific linguistic model;
- The conceptual representations and the knowledge structure of the specialised discourse in which terminological units function are insufficiently clarified;
- The organization of the concepts in categories remains totally vague;
- The CTT offers no clear explanation of the components of terminological meaning as it is taken to derive from the general meaning of a lexical unit by selecting semantic features according to its realization in specialized discourse;
- The semantic analysis of terminological units is reduced to semantic decomposition which is typical of Lexical Semantics and which has lately been criticized as regards the nature of word meaning;

Among the cognitive-based theories of terminology to which Faber’s own theory belongs, she chooses to comment on Sociocognitive Terminology (henceforth SCT) proposed by Temmerman [6]. The principles of that theory are opposed to the basic principles of the GTT which are claimed to be “unrealistic and incapable of describing or explaining specialized language as it is actually used in communicative situations such as specialized translation” [4]. The GTT premises criticized by Temmerman are:
- The central role of concepts in relation to their designations;
- Concepts and categories have clear-cut boundaries;
- The monosemic reference, i.e. one-to-one correspondence between terms and concepts;
- Terminological definitions should be only intensional;
- Terminology should be approached only synchronically;

In contrast, Temmerman claims that
- Language is relevant for the conception of categories and should not be separated from concepts;
Only few categories are clear-cut, i.e. can be clearly delineated;
- Synonymy and polysemy are functional in specialised discourse
- Terminological definitions may have different content and form depending on category types;
- A diachronic approach to terminology is unavoidable.

This theory is in line with the CTT since it is descriptive rather than prescriptive and the starting point of terminological analysis is terms and not concepts. Distinctive features of the SCT are its emphasis on conceptual organization and its focus on category structure in terms of cognitive linguistics as well as its combined synchronic and diachronic approach to term analysis.

Unlike the GTT according to which concept systems are organized hierarchically, i.e. hyperonymically (is a …) and meronymically (part of …), sociocognitive categories are prototypical. The latter notion is understood as representations of conceptual structures in the form of idealized cognitive models making use of degrees of typicality to configure the conceptual categories or domains. The conceptual map thus obtained takes the form of concentric circles. Concepts are placed intuitively closer or farther away from the prototypical center represented by the most typical representative of the respective category.

Despite its advantages over other terminology theories, such as its focus on the study of terms in a diachronic perspective which contributes significantly to the cognitive analysis of terminological metaphorisation, SCT has its disadvantages formulated by Faber [4] in the following way:

- The prototypes and idealized cognitive models are totally unconstrained since they are based on an open-ended inventory of conceptual relations;
- These models seem to be largely based on the intuition of the modeller;
- Neither prototypes nor idealized cognitive models provide theoretical grounds for analyzing syntagmatic terminological data, i.e. the syntactic behaviour of terms.

In my opinion, it is this latter disadvantage that served as a major stimulus for Faber to propose a new approach to the study of terms which I will discuss in the next section.

4 Theoretical and Methodological Premises of Frame-Based Terminology

Frame-Based Terminology (henceforth FBT), proposed by Faber [4], shares many premises with the two theories discussed above such as the unnecessary distinction between words and terms for modern computerised terminology and the need to study the behaviour of terms in their natural textual environment. In view of providing useful data for translation purposes, the new theoretical approach takes into account not only the compound nominal forms by which most special language units are represented but also their syntactic valence or combinatorial capacity. As Faber rightly contends “The understanding of a terminology-rich text requires knowledge of the domain, the concepts within it, the propositional relations within the text, as well as the conceptual relations between concepts within the domain”. All the elements listed are targeted in her proposal. Below I will briefly present the theoretical and methodological premises of the Frame-Based Terminology which I consider applicable, with certain modifications, for the analysis of terminological verb collocations.

FBT is based on Frame Semantics, a theory developed by the cognitive linguist Fillmore which relates linguistic meaning to encyclopaedic knowledge. Within this theoretical framework a word is assumed to activate a frame of semantic knowledge relating to the specific concept it refers to. Before embarking on the methodological implications of that theory for the proposed model of conceptual category design of a specialized domain, Faber dwells on the conceptual organization of the latter. Pointing out the crucial role of the concept ‘domain’ in Terminology and Linguistics, she criticizes the GTT for not proposing conceptual representations “with explanatory adequacy from a psychological perspective”. Special domains are essential to the CTT and SCT. However, according to Faber, they have not seriously considered how to elaborate, design, and organize such structures but have simply regarded them as “a product of the terminologist’s intuition” later subjected to expert validation. Some other models of conceptual domain structures are mentioned which I will not discuss here except for one particular proposal made by Goldberg [7] who assumes that the “the world is carved up into discretely classified event types”. That view seems to have considerably influenced Faber in shaping up her proposal.

As previously mentioned FBT uses the basic assumptions of Frame Semantics to construct a model of a special domain structure. Frames, being a type of cognitive structuring device based on experience, provide
not only background knowledge of words (terms respectively) and their use in discourse but also make explicit their semantics and syntactic behaviour. In other words, by making use of that device, it is possible to describe the conceptual relations of a given special concept belonging to a special domain as well as the combinatorial potential of the term which designates it. According to Faber, FBT focuses on:

(a) conceptual organization;
(b) the multidimensional nature of terminological units;
(c) the extraction of semantic and syntactic information through the use of multilingual corpora;

The basic assumption of the FBT is that “conceptual networks are based on an underlying domain event which generates templates for the actions and processes that take place in the specialized field as well as the entities that participate in them” [4]. The methodological implications of that assumption can be summarized as follows:

- Special domains are described through the events that generally take place in them;
- Each special domain has its own event template;
- Generic/base-level categories are configured in a prototypical/basic domain event representing action-environment interface;
- This basic domain event provides a frame for the organisation of more specific concepts;
- The specific concepts within each category are linked by both hierarchical and non-hierarchical relations;
- Each subdomain within the event is characterized by a template with a prototypical set of conceptual relations;
- Terminological definitions are regarded as mini-knowledge representations or frames;
- The definitions are based on data extracted by corpus analysis thus providing paradigmatic and syntagmatic information about the term;

Faber exemplifies the methodology proposed with the Coastal Engineering Event which is taken to be the basic event in the Coastal Engineering domain [4]. Within that event she focuses on a typical process, namely, erosion which conforms to the process template. From that inference and the following figure representing the Coastal Engineering Event it becomes clear that a template in this view should be understood as a general conceptual category representing, in Fillmore’s terms [8], either an argument or the predicate in a generic semantic frame. Thus, in the particular event discussed the Agent template includes what I would call two basic sub-templates (natural agent and human agent) which, in turn, involve a number of specific concepts. The same applies to the other basic templates specified for the particular event as Process template, Patient template and Result template, respectively. It is essential to emphasize the relationships established between the templates, e.g. the causal relation, which according to Goldberg [7] can be interpreted as basic argument constructions realized linguistically by basic clausal expressions (X causes Y). These basic argument structures represent dynamic scenes, i.e. experientially grounded gestalts. Therefore, the conceptual representation of a special domain within the context of the Frame-Based Terminology has the advantage of representing conceptual relations in a dynamic rather than static fashion. Information about the combinatorial potential of a particular concept, e.g. erosion, in a particular language is extracted by analyzing corpus data, i.e. concordances which provide the frames and serve as definitional models.

5 Taxonomic-Frames Categorization of Terminological Verb Collocations

In this section I will attempt to use the methodological premises of the Frame-Based Terminology to analyze terminological verb collocations, exemplified with verb collocations of the term steel, and justify my proposal for certain modifications in that approach to suit the specific object of study.

Terminological verb collocations are a type of collocations defined as syntagmatic relations of words. There are two major approaches to collocations, viz. frequency-based and phraseological. The former assumes the frequency of occurrence as the main criterion for defining a word combination as a collocation, whereas the latter is semantically oriented and distinguishes between a semantically autonomous element called the base and an element selected depending on the base termed collocator [2]. When analyzing a terminological verb collocation, the term, represented by a noun or a nominal phrase according to most terminologists [9, 10], is taken to be the base and the respective verb is the collocator.
In section 2 above I argued that terminological verb collocations may be treated as knowledge units larger than individual concepts which can be equated with concept combinations conducive to categorization. Within the Frame-Based Terminology context each collocation of that type can be analyzed semantically by assuming that the term (base) is the Patient in a dynamic scene/frame whereas the verb designates a dynamic action/process affecting it with a consequent Result. What is lexicalized, i.e. explicated on the surface structure, are the Patient and the Action/Process, the action being performed by a human agent and the process induced by a natural agent.

If we are interested in representing the verbs that go with a basic term, let us say, steel in a kind of a learner’s glossary designed for technical translators as proposed by Alexiev [11], we can extract automatically the necessary terminological data by submitting the appropriate text to a term extractor and using the concordances and bigrams it yields. In the case of steel I retrieved the following verbs/collocators to the term/base (arranged alphabetically):

<table>
<thead>
<tr>
<th>BASE</th>
<th>COLLOCATOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>anneal</td>
<td>embrittles</td>
</tr>
<tr>
<td>bolt</td>
<td>extrude</td>
</tr>
<tr>
<td>braze</td>
<td>forge</td>
</tr>
<tr>
<td>cast</td>
<td>oxidizes</td>
</tr>
<tr>
<td>corrodes</td>
<td>quench</td>
</tr>
<tr>
<td>cracks</td>
<td>rivet</td>
</tr>
<tr>
<td>draw</td>
<td>roll</td>
</tr>
</tbody>
</table>

In a knowledge-oriented terminological collection [11] the terminological collocations these verbs form with the term steel should be represented in different categories. In terms of the frames approach, those categories can be specified according to the type of Agent and the type of Result since steel is a building material on which human agents act and which is subjected to natural forces. According to the type of Agent (human or natural) a terminologist with a linguistic background can easily classify the verbs and the collocations, respectively, into two groups, all the more that in the knowledge contexts extracted the processes caused by a Natural Agent are expressed by verbs which in the corpus commonly occur in a grammatically marked form, viz. 3rd person singular, present simple (-s). Therefore, as a first step towards categorization we can represent two generic frames for the verb collocations of steel, which according to the type of Agent can be termed the Process Frame and the Action Frame within the general Steel Processing Event:

**STEEL PROCESSING**

<table>
<thead>
<tr>
<th>Natural Agent</th>
<th>Process</th>
<th>Patient</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weather</td>
<td>Corrodes</td>
<td>Steel</td>
<td>Steel quality deterioration</td>
</tr>
<tr>
<td></td>
<td>Cracks</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Embrittles</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Oxidizes</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rusts</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Splits</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

However, even for the laymen it seems obvious that a second step towards categorizing all other terminological verb collocations in one Action Frame would be an act of very imprecise categorization. A need is intuitively felt for further subdivision of the actions carried out in relation to the material concept ‘steel’. A more precise categorization can be performed only with the help of an expert. By applying an integrated top-down and bottom-up approach as recommended by Faber [4], i.e. extracting terminological data from a textual corpus and from reference sources including an expert’s opinion I collected the following definitional information on the verbs in question:

---

1 I am grateful to Assoc. Prof. Dr. Eng. Roumiana Zaharieva from the University of Architecture, Civil Engineering and Geodesy-Sofia for her expert assistance in categorizing the verbs which collocate with the basic concept steel.
**Heat Treating** - a group of industrial and metalworking processes used to alter the physical, and sometimes chemical, properties of a material
- *Quench* - To cool (hot metal) by thrusting into water or other liquid
- *Temper* - To harden or strengthen (metal or glass) by application of heat or by heating and cooling
- *Anneal* - To heat (glass, earthenware, metals, etc.) to remove or prevent internal stress

**Steel Shaping/Forming** - changing the shape of a piece of metal
- *Cast* - To form (liquid metal, for example) into a particular shape by pouring into a mold
- *Roll* - To pass metal stock through one or more pairs of rolls to reduce the thickness
- *Extrude* - To produce (moulded sections of plastic, metal, etc) by ejection under pressure through a suitably shaped nozzle or die
- *Forge* - To form (metal, for example) by heating in a furnace and beating or hammering into shape
- *Stamp* - To form or cut out by application of a mold, form, or die
- *Draw* - To flatten, stretch, or mold (metal) by hammering or die stamping

**Joining** - Bringing two separate materials together through some type of forming. Joining is one of the main ways metals can be formed.
- *Weld* - To join pieces of metal together by heating the edges until they begin to melt and then pressing them together
- *Braze* - To solder (two pieces of metal) together using a hard solder with a high melting point
- *Solder* - To join together two or more metal items by melting and flowing a filler metal (solder) into the joint, the filler metal having a lower melting point than the adjoining metal
- *Rivet* - To hammer the headless end of so as to form a head and fasten something.
- *Bolt* - To secure or lock with or as if with a bolt

As can be seen from the data collected, within the Action Frame we can identify three subcategories: Heat Treating, Steel Shaping and Joining. These subcategories can be represented in tabular form to illustrate the semantic frames to which each terminological verb collocation from the generic Action Frame belongs:

### STEEL PROCESSING

<table>
<thead>
<tr>
<th>Action Frame 1 – Heat Treating</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Human Agent</strong></td>
</tr>
<tr>
<td>Metallurgist</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Action Frame 2 – Steel Shaping</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Human Agent</strong></td>
</tr>
<tr>
<td>Metallurgist</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Action Frame 3 – Joining</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Human Agent</strong></td>
</tr>
<tr>
<td>Builder</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
Given that comparatively precise categorization, two questions logically arise to which I will try to give reasonable answers. The first question is whether the approach proposed can be considered a frames one or a different one. My answer to that question is that the approach is frame-based but with certain modification involving taxonomic elements. The categorization tables presented above show clearly that the procedural steps in my proposal, which can be termed Taxonomic-Frames Approach, should start with a preliminary expert-based taxonomy of the basic event frames lexicalized by the respective terminological verb collocations thereby arriving at scientifically accurate categorization schemes. It is important to note that the concept ‘taxonomy’ in this context is understood in a broad sense, viz. as the science of classification. Cruse [12] defines taxonomic hierarchies as “essentially classificatory systems” which “reflect the way speakers of a language categorize the world of experience”. Taxonomic hierarchies are characterized by well-developed levels. Therefore, if we want to subsume all events associated with the concept steel described above, we can designate that higher-level event Steel Processing (as denoted above the tables) thus making it possible to represent, if necessary, the categorization scheme in the form of the classical tree-diagram typical of classification representations.

The second question we have to answer is what the particular application of the Taxonomic-Frames Approach can be in Terminology Work and especially in Knowledge-Oriented Terminography. Alexiev [11], in his extended model of a learner’s glossary entry, has shown the applicability of such an approach in creating the microstructure of a translation-oriented, knowledge-based terminographic collection. Another argument for proposing the combined Taxonomic-Frames Approach to terminological verb collocations is that in my opinion the analytical procedure is explicit and applicable in a subsequent terminological contrastive analysis involved in multilingual terminological projects. Such a contrastive analysis requires, as claimed by Alexiev [11], preliminary identification of systematic conceptual groups and the corresponding term sets. In this respect I will provide only one example with the collocations within the Joining Action Frame 3 above, some of which are translated into Bulgarian by adding a gloss (brief explanation) to the verb collocate, e.g.:

1. Braze (EN) – Споявам с твърд припой/Solder with hard solder (literal translation of the BG translation equivalent);
2. Solder (EN) – Споявам с мек припой/Solder with soft solder (literal translation of the BG translation equivalent);

6 Conclusion

A proposal has been made for a combined taxonomic-frames approach to terminological verb collocations. Arguments have been adduced for recognizing terminological verb collocations as terminological units. Faber’s Frame-Based Terminology has been discussed in comparison with two other recent terminology theories, namely, Cabré’s Communicative Theory of Terminology and Temmerman’s Sociocognitive Terminology. I have claimed that Frame-Based Terminology offers theoretical and methodological premises which are appropriate for the analysis of terminological verb collocations. However, I have opted for a combined taxonomic-frames approach involving a taxonomic analysis preceding the analysis of the cognitive frames underlying those collocations, which is deemed necessary to first categorize the event structures into conceptual groups thus making the analytical procedure more compatible with a possible terminological contrastive analysis requiring identification of systematically grouped conceptual and term items.

The approach I propose can handle translation problems because a preliminary categorization of a special domain of knowledge undoubtedly helps in systematizing the object of a subsequent terminological contrastive analysis, in our case terminological verb collocations, in this way assisting the translator in making justifiable translation decisions in case of terminological gaps or long terminological expressions.

References

Document Retrieval for Large Scale Content Analysis using Contextualized Dictionaries

Gregor Wiedemann and Andreas Niekler

University of Leipzig
Augustusplatz 10
04109 Leipzig, Germany

gregor.wiedemann@uni-leipzig.de, aniekler@informatik.uni-leipzig.de

Abstract. This paper presents a procedure to retrieve subsets of relevant documents from large text collections for Content Analysis, e.g. in social sciences. Document retrieval for this purpose needs to take account of the fact that analysts often cannot describe their research objective with a small set of key terms, especially when dealing with theoretical or rather abstract research interests. Instead, it is much easier to define a set of paradigmatic documents which reflect topics of interest as well as targeted manner of speech. Thus, in contrast to classic information retrieval tasks we employ manually compiled collections of reference documents to compose large queries of several hundred key terms, called dictionaries. We extract dictionaries via Topic Models and also use co-occurrence data from reference collections. Evaluations show that the procedure improves retrieval results for this purpose compared to alternative methods of key term extraction as well as neglecting co-occurrence data.

1 Introduction

Due to the vastly growing availability of (retro-)digitized large scale text corpora computer-assisted Content Analysis (CA) is of increasing interest for various disciplines and applications ranging from social sciences to business intelligence. When exploring large corpora analysts are confronted with the problem of selecting relevant documents for qualitative investigation and further quantitative analysis. Standard tasks in Information Retrieval (IR) usually rely on small sets of concrete key terms for querying a collection. Highly abstract research interests in CA often cannot describe research objectives by such small queries. For example, analysis on the war in Iraq certainly can query for Iraq AND war. But, what would be a reasonable query for documents containing neoliberal justifications of politics?

To meet special requirements of CA retrieval tasks we propose a procedure where a query is not (primarily) based on single terms, but on a set of reference documents. Compared to the problem of determining concrete key terms for a query it is rather easy for analysts to manually compile a collection of ‘paradigmatic’ documents which reflect topics or manner of speech matching their research objective. Retrieval with such a reference collection is then performed in three steps:
1. Extraction of a set of key terms from the reference collection, called dictionary. Terms in the dictionary are ranked by weight to reflect difference in their importance for describing an analysis objective.
2. Extraction of co-occurrence data from the reference collection as well as from an additional generic corpus representative of a given language.
3. Scoring relevancy of each document on the basis of dictionary and co-occurrence data to create a ranked list of documents.

**Related work:** Using documents as queries is a consequent idea within the Vector Space Model (VSM) of IR where key term queries are modeled as document vectors for comparison with documents in the target collection [14]. Our approach extends the standard VSM approach by additionally employing of shared meanings of topic defining terms captured by co-occurrence data. Co-occurrence data has been used in standard IR tasks for term weighting as well as query expansion with mixed results (e.g. see [12, 3, 13, 18]). These applications differ from our approach as they want to deal with unequal importance of terms in a single query due to term correlations in natural language. The method presented in this paper does not weight semantically dependent query terms by co-occurrence information globally. Instead, in CA analysts are often interested in certain aspects of meaning of specific terms. Following the distributional semantics hypothesis [9] meaning may be captured by contexts better than just by isolated terms. Therefore, we score relevancy based on similarity of individual contexts of single query terms in sentences of the target documents compared to observed contexts from the reference collection.

The paper is organized in 5 sections. After having clarified our motivation the next section presents our approach of semi-supervised dictionary extraction with the help of statistical Topic Models. The third part explains how to utilize ranked dictionaries together with co-occurrence data for document retrieval. Section 4 briefly introduces an example application of this procedure on a target collection of German newspaper articles, followed by an evaluation of the approach.

## 2 Dictionary Extraction with Topic Models

The generation and usage of dictionaries is an important part of quantitative CA procedures [10]. Dictionaries provide the basis of code books and category systems within CA studies. As we want to exploit dictionaries for document retrieval we suggest a procedure of semi-supervised term extraction for dictionary creation. For this we apply a statistical Topic Model on a collection of paradigmatic documents. Reference documents should be selected carefully by the analyst in consideration of representing domain knowledge or specific language use of interest. The resulting list may be compared to domain term extraction based on reference corpus comparison [7] or tf/idf weighting of words. These calculate ‘keyness’ of terms isolated from each other. In contrast to those the Topic Model based approach takes account of the fact that terms do not occur independently of each other.
**Statistical Topic Models:** Topic Models are a set of statistical models for unsupervised extraction of latent semantic structures from document collections. ([4] first introduced *Latent Dirichlet Allocation*). They generate results where underlying latent variables with $K$ dimensions are extracted from a document collection. Those variables represent distributions over words $\phi_{k} = p(w|z_k), (k = 1, \ldots, K)$ representing their alignment to a semantically coherent group which may be interpreted as topic. Words $w$ with a high probability $p(w|z_k)$ in a topic $k$ represent its determining terms and allow for interpretation of the meaning of an underlying thematic composition.

**Extracting Dictionaries from Topic Models:** Distributions $p(w|z_k)$ represent a simplification of the collection content as a composition of topics. We can assume that highly probable words play an important role in the semantics of the whole corpus. Therefore, those words may be used to compile a dictionary of key terms within that collection. We define the weight of a term in the dictionary by the sum of its probability values in each topic.

In comparison to term frequency counts in a collection the probability weight of a term in a corpus represents a word’s contribution to a certain context. Even if a context has relatively low evidence in the data because of a low frequency a term can have high probability $p(w_n|z_k)$ within a topic. We don’t want to overly bias the ranks in the dictionary with very improbable topics and their words—e.g. the high probability of the top terms in topics of low probability would be ranked almost equal to the top words in highly probable topics. Therefore, we need to normalize the term probabilities $p(w|z_k)$ according to either their topic’s probability $p(z_k)$ or their term frequency $tf(w_n)$ within the corpus. We decide to normalize each terms probability in a certain topic with its term frequency, but to use log frequency to dampen the effect of high frequency terms. The final weight of a term in the dictionary is determined by

$$tw_n = \log(tf(w_n)) \sum_{k=1}^{K} p(w_n|z_k)$$  \hspace{1cm} (1)

where $K$ is the number of topics, $tf(\cdot)$ the term frequency, and $w_n$ the dictionary term. Descended sorting of term weights $tw$ results in a list of ranked words which can be cut to a certain length $N$.

**Analyst supervision:** Within a Topic Model usually topics with undesired content can be identified. A few topics normally capture rather syntactic information of the collection representing co-occurring functional words in a corpus [1]. Other topics, although capturing semantic structure, may be considered as irrelevant by the analyst with a view to her/his research interest. In contrast to other methods of key term extraction the Topic Model approach allows to exclude those unwanted semantic clusters. Before calculation of the weights of terms one has to identify these topics which do not represent meaningful structures and to exclude them from the set of $K$ topics. This is an important step for the analyst to exercise influence on the so far unsupervised dictionary creation process and a clear advantage over other methods of key term extraction.
3 Retrieval with dictionaries

Dictionaries can be employed as filters in IR systems reducing general collections to sub collections containing sets of documents of interest for further analysis. Using a dictionary of ranked terms for IR might be translated in the standard Vector Space Model (VSM) approach in combination with ‘term boosting’. In this approach prior knowledge of unequal importance of terms is incorporated into query processing via factors for each term. A VSM-based scoring function can be computed for a document $d$ and a dictionary as query $q$ as follows:

$$\text{score}_{\text{VSM}}(q, d) = \sum_{w \in q} \text{tf}(w, d) \cdot \text{boost}(w) \cdot \text{norm}(d) \quad (2)$$

Usually IR weightings consider inverse document frequency of a term as a relevant factor. As the dictionary ranking is derived from Topic Models, information comparable to document frequency has indirectly already been taken into account. We skip the $\text{idf}$ factor for each term in favor of our own dictionary weight. Rank information from the dictionary can be translated into a boosting factor for the scoring function. We suggest a factor ranging between 0 and 1

$$\text{boost}(w) = \frac{1}{\sqrt{\text{rank}(w)}} \quad (3)$$

for each term $w$ which reflects that the most prominent terms in a dictionary of $N$ terms are of high relevancy for the retrieval process while terms located nearer to the end of the list are of more equal importance.

To address the problem of document length normalization and identify relevant documents of all possible length we utilize pivoted unique normalization as introduced in [15]. Pivotal length normalization slightly lowers relevancy scores for shorter documents of a collection $D$ and consequently lifts the score for documents after a pivotal value determined by the average document length. The normalization factor for each document is computed by

$$\text{norm}(d) = \frac{1}{\sqrt{(1 - \text{slope}) \cdot \text{pivot} + \text{slope} \cdot |U_d|}} \quad (4)$$

where $U_d$ represents the set of unique terms occurring in document $d$ and $\text{pivot}$ is computed by $\text{pivot} = \frac{1}{|D|} \cdot \sum_{d \in D} |U_d|.$

When evaluation data is available, the value for slope might be optimized for each collection. Lacking a gold standard for our retrieval task we set slope to 0.7 which has proven to be a reasonable choice for retrieval optimization in various document collections [15]. Further, the $\text{tf}$ factor should reflect on the importance of an individual term relative to the average frequency of unique terms within a document: $\text{avgtf}(d) = \frac{1}{|U_d|} \cdot \sum_{t \in U_d} t f(t, d).$ Thus, the final scoring formula yields a document ranking for the entire collection:

$$\text{score}_{\text{dict}}(q, d) = \sum_{w \in q} \frac{1 + \log(\text{tf}(w, d))}{1 + \log(\text{avgtf}(d))} \cdot \text{boost}(w) \cdot \text{norm}(d) \quad (5)$$
Contextualizing dictionaries: The approach described above yields useful results when looking for documents which can be described by a larger set of key terms. When it comes to more abstract research interests, however, which aim to identify certain meanings of terms or specific language use, isolated observation of terms may not be sufficient. Fortunately the approach described above can be augmented with co-occurrence data from the reference collection to judge on relevancy of occurrence of a single key term in our target document. This helps not only to disambiguate different actual meanings of a term, but also reflects the specific usage of terms in the reference collection. Therefore we compute patterns of significant co-occurrences of the $N$ terms in our dictionary with each other resulting in an $N \times N$ matrix $C$. Co-occurrences are observed in a sentence window. Statistical significance of a co-occurrence is calculated by the Dice measure which is the fraction of the count of all sentences containing term $a$ and term $b$ over the sum of all sentences containing each single term:

$$\text{dice}(a, b) = \frac{2n_{ab}}{n_a + n_b} \quad (6)$$

We decided for this measure instead of more sophisticated co-occurrence significance tests like Log Likelihood [8] because it also reflects syntagmatic relations of terms in language relatively well [5], but, more important its values range is between 0 and 1 which makes measurements over different corpora comparable. This is useful for dealing with an unwanted effect we experienced when experimenting with co-occurrence data to improve our retrieval mechanism. Co-occurrences of terms in the sentences of a reference collection may reflect characteristics in language use of the included documents. However, certain co-occurrence patterns may reflect general regularities of language not specific to a collection of a certain domain (e.g. strong correlations between the occurrence of parents and children or multi word units like United States in one sentence). Applying co-occurrence data to IR scoring tends to favor documents where many of those common language regularities can be observed.

Instead of using the co-occurrence matrix $C$ solely based on the reference collection we ‘filter’ the co-occurrence data by subtracting a co-occurrence matrix based on a second, randomly composed reference corpus.\footnote{Suitable corpora for this purpose, such as the ones provided by the “Leipzig Corpora Collection” [2] which is carefully maintained by computational linguists, may be seen as representative of common language characteristics not specific to a certain domain or topic.} A second $N \times N$ matrix $D$ of co-occurrences is computed from such a generic corpus. The subtraction of $D$ from $C$ delivers a matrix $C'$ reflecting the divergence of co-occurrence patterns in the reference collection compared to common language: $C' = \max(C - D, 0)$.\footnote{$\max$ asserts that all negative values in $C'$ (representing terms occurring less together in sentences of the reference collection than in sentences of common language) are set to zero.} Values for common combinations of terms (United States) in $C'$ are significantly lowered while combinations specific to the reference collection remain largely constant.
Using sentence co-occurrences: To exploit co-occurrence data for IR the scoring function in eq. 5 has to be reformulated to incorporate a similarity measure between a co-occurrence vector profile of each term $w$ in the dictionary and each sentence $s$ in the to-be-scored-document $d$. Instead of using just term frequency we add information on contextual similarity:

$$tf sim(w, d) = \sum_{s \in d} \sum_{w \in s} tf(w, s) + \alpha \cdot \frac{s \cdot C'_{w}}{||s|| \cdot ||C'_w||} \tag{7}$$

The frequency of $w$ within a sentence (which usually equals 1) is incremented by the cosine similarity between sentence vector $s$ (sparse vector of length $N$ indicating occurrence of dictionary terms in $s$) and the dictionary context vector for $w$ out of $C'$. This measure rewards the relevancy score, if the target sentence and the reference term $w$ share common contexts. In case they share no common context $tf sim$ is equal to $tf$.

Because term frequency and cosine similarity differ widely in their range the influence of the similarity on the scoring needs to be controlled by a parameter $\alpha$. If $\alpha$ is set to zero $tf sim$ replicates simple term frequency counts. Values for $\alpha$ higher than 0 yield a mixing of unigram matching and contextual information for the relevancy score. Optimal values for $\alpha$ can be calculated by our evaluation method (see section 5). The context-sensitive score is computed as follows:

$$score_{context}(q,d) = \sum_{w \in q} \frac{1 + \log(tf sim(w,d))}{1 + \log(avg tf(d))} \cdot boost(w) \cdot norm(d) \tag{8}$$

4 Example

The procedure presented above is applied to a political science research task performed as part of a German research project. The project aims at analyzing the influence of neoliberal ideas on domestic politics by studying the discourse in public media. In our example application we identify documents which (potentially) contain neoliberal argumentation in a collection of 101,032 newspaper articles of the German magazine DIE ZEIT (volumes 2000–2009). This example is also used for evaluation of the method in section 5.

To retrieve documents of interest for further analysis political scientists compiled a reference corpus consisting of 36 German books and journal articles written by self-confessed neoliberal theorists (e.g. Milton Friedman). In this reference corpus sentence boundaries are detected and tokens are lemmatized. A topic model based on the Pitman-Yor Process [16] is calculated. For this process all paragraphs of books and articles in the collection were treated as single ‘documents’ for modeling. In the resulting 23 topics of this model we can identify a very large topic containing only English words (originating mostly from bibliographies in the reference collection) which have been clustered by the topic model process. Since this topic does not represent meaningful semantics for the analysis analysts could exclude it for the process of dictionary extraction. Using
the process described in Section 2 a dictionary of 500 key terms is extracted from
the reference collection. Co-occurrence patterns of these 500 terms are extracted
according to Section 3.

This contextualized dictionary is then used to query the target corpus of
DIE ZEIT newspaper articles. The 2,000 highest ranked articles are retrieved and
used as starting point for further qualitative and quantitative analysis procedures
by the analysts.

5 Evaluation

Determining a large set of key terms from a reference collection and extract-
ing its co-occurrence profiles to compose a “query” is an essential step in the
proposed retrieval mechanism to meet requirements of content analysts. Due to
this, standard approaches of IR evaluation [6] are not applicable. There are no
test collections like the TREC datasets [17] regarding such type of retrieval task.
In order to evaluate our method we therefore follow two approaches:

1. Generating a quasi-gold standard of pseudorelevant documents to show per-
formance improvements through the use of co-occurrence data as well as key
term extraction via topic models,
2. Judging on the overall validity with precision at k evaluation in our example
retrieval performed by domain experts.

Generating pseudorels: Due to the lack of proper gold standards for our spe-
cial retrieval task we create a custom evaluation set to evaluate our approach.
Therefore, we use a strategy of data fusion. This strategy merges results, namely
ranked lists of documents, of multiple retrieval systems to a set of pseudorelevant
documents. These pseudorels can be used as an automatically generated quasi-
gold standard. The approach proposed by [11] shows that merging results of
different retrieval systems by the Condorcet method generates pseudorels which
produce evaluation results of IR systems that highly correlate with the TREC
testset rankings of IR systems. Although it is hard to judge on the overall abso-
lute performance of a system, the procedure allows for relative judgment between
tested systems. We employ this strategy in a two-fold manner:

1. Optimizing parameter $\alpha$ for the best mix of unigram / co-occurrence matches
2. Deciding whether using topic models improves retrieval results over a simple
tf/idf measure for key term extraction.

For the latter a second dictionary is created on the basis of the highest 500 tf/idf
measures for each term in the reference collection. In (7) we set the influence of
the context similarity within a sentence by the parameter $\alpha$. For the evaluation
we will treat every setting for this parameter as a different retrieval system in
order to artificially create a large set of different systems. We vary the parameter
$\alpha$ by steps of 2 in a range of $[0, 30]$. Furthermore, a system is added where just
context similarities contribute to the ranking by setting $\alpha = 1$ and leaving out
Each of the 17 ‘systems’ produces a ranked list of 2000 highest-scored documents using the contextualized dictionary created by the topic model approach. Another set of retrieval systems is created in the same manner, but using the dictionary which was extracted with the tf/idf measure. Both sets together provide 34 lists of ranked retrieval results allowing for a comparison between the represented systems. For this, results of the systems need to be merged to a set of pseudorelevant documents by the following procedure taken from [11]:

1. Selecting a set of the most biased systems: As we want to compare results in two dimensions (i.e., use of co-occurrence data, ii. key term extraction procedure) we select 4 from our 34 ‘systems’: The system which neglects co-occurrence data ($\alpha = 0$) and the system which solely relies on co-occurrence data—both combined with a dictionary generated by Topic Model or tf/idf.
2. Select the top documents of each of the most biased systems as candidates for pseudorels. For each retrieved document in each system a norm weight can be computed and summed up over all systems: $n_d = \sum_{s=1}^S m_s/i_{d,s}$, with system $s$, $m_i$ number of ranks in system $s$ and $i_{d,s}$ rank of document $d$ in system $s$. Figure 1 shows the sorted values of norm weights $n_d$ for the best 100 documents in our example. The first documents rank very high in most of the tested systems. Documents with lower $n_d$ ranks are retrieved only by a few of the tested systems and, thus, are considered not to be good candidates for pseudorels. We select the top 50 documents of each system.
3. Rank the candidates for pseudorels by using the Condorcet method which relies on counting wins and losses of direct comparison of document rankings within the most biased systems.
4. The selection of pseudorels should reflect only those documents which yield a robust ranking within the Condorcet method. Figure 2 shows that roughly half of the documents of our example dataset have distinguishable values. Others produce more equal amounts of wins. Following [11] we define the top 50% of the Condorcet ranked list as pseudorels.

**Mean average precision:** With this procedure a list of 54 documents is selected as pseudorels for evaluating the example application. Performance of each of the 34 retrieval systems is measured by utilizing the mean average precision (MAP). Since we only have one result e.g. one query for each system MAP is used as $MAP = \frac{1}{R} \sum_{k=1}^R P(R[k])$, with the number of relevant documents $R$ and $P(R[k])$ as precision within the ranking of a system up to the document $R[k]$.

The best performance of all tested systems is achieved by the system which mixes unigram and co-occurrence matching with parameter $\alpha = 14$. Table 1 displays MAP values for the two retrieval results based solely on unigram or context matching and the best performing mixed approach. Furthermore, it contrasts systems based on the tf/idf dictionary with systems based on the Topic Model dictionary. The results indicate that systems based on a Topic Model approach perform indeed better than systems using tf/idf for term extraction if co-occurrence data is used. This is not surprising considering the fact that
Fig. 1. Plot of the accumulated values $n_{rd}$ of all documents from each retrieval system (first 100 ranks).

Fig. 2. Plot of the wins of the documents found by the Condorcet method (values have been sorted).

<table>
<thead>
<tr>
<th>Retrieval System</th>
<th>tf/idf</th>
<th>TM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unigram</td>
<td>0.732</td>
<td>0.714</td>
</tr>
<tr>
<td>Co-occurrence</td>
<td>0.657</td>
<td>0.723</td>
</tr>
<tr>
<td>$\alpha$-mix</td>
<td>0.823</td>
<td>0.861</td>
</tr>
<tr>
<td>$\alpha = 6$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha = 14$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. MAP evaluation of compared retrieval systems.

<table>
<thead>
<tr>
<th>Rank</th>
<th>$\alpha = 0$</th>
<th>$\alpha = 14$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-10</td>
<td>0.9</td>
<td>1.0</td>
</tr>
<tr>
<td>101-110</td>
<td>0.7</td>
<td>1.0</td>
</tr>
<tr>
<td>501-510</td>
<td>0.8</td>
<td>0.9</td>
</tr>
<tr>
<td>1001-1010</td>
<td>0.7</td>
<td>0.6</td>
</tr>
<tr>
<td>1501-1510</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>1991-2000</td>
<td>0.5</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Table 2. Precision at 10 evaluated at different ranks of the unigram and $\alpha$-mix retrieval results.

Topic Models are based on co-occurrences as well. Our dictionary extraction method preserves this information in contrast to the independence assumption underlying tf/idf. Overall, the evaluation shows that using context information outperforms the base line approach without context information for our retrieval purpose. Nonetheless, because the evaluation method uses pseudorels care has to be taken in interpreting absolute MAP values of the systems.

**Precision at k:** A second evaluation target is to test how dense the relevant documents on different ranges in the ranks are. The precision at $k$ measure can be utilized to determine the quality of the process by manually assessing the first 10 documents downwards from the ranks 1, 101, 501, 1001, 1501, 1991. Documents were marked as relevant in case a domain expert was able to annotate text snippets therein regarding arguments, topics or claims representing a discourse framed in neoliberal terminology. The results in table 2 confirm the usefulness of our approach. Density of positively evaluated results in the upper ranks is very high and gets lower towards the bottom of the list. Precision in the best performing system remains high also in lower ranks while it drops off in comparison with the system which solely exploits unigram matchings.
6 Conclusion

We presented an extension of the VSM approach of IR by exploiting automatically extracted dictionaries and co-occurrence data from manually compiled reference collections. This method has proven to produce valuable results for Content Analysis studies to extract collections specific to a certain research interest from large unspecific corpora. The use of co-occurrence data improves upon merely taking into account raw frequencies. Results could be enhanced further by creating dictionaries with the help of Topic Models. The retrieval mechanism allows domain experts, such as social scientists, to build collections for further analyses based on paradigmatic example documents or theory texts representing abstract domain knowledge.

References

Leipzig Corpus Miner – A Text Mining Infrastructure for Qualitative Data Analysis

Andreas Niekler, Gregor Wiedemann, and Gerhard Heyer

University of Leipzig
Augustusplatz 10
04109 Leipzig, Germany
{aniekler,heyer}@informatik.uni-leipzig.de, gregor.wiedemann@uni-leipzig.de

Abstract. This paper presents the “Leipzig Corpus Miner”—a technical infrastructure for supporting qualitative and quantitative content analysis. The infrastructure aims at the integration of “close reading” procedures on individual documents with procedures of “distant reading”, e.g. lexical characteristics of large document collections. Therefore information retrieval systems, lexicometric statistics and machine learning procedures are combined in a coherent framework which enables qualitative data analysts to make use of state-of-the-art Natural Language Processing techniques on very large document collections. Applicability of the framework ranges from social sciences to media studies and market research. As an example we introduce the usage of the framework in a political science study on post-democracy and neoliberalism.

1 Introduction

For several years humanists, social scientists and media analysts working with text as primary data have been opening up to large scale text analysis procedures. This development, sometimes referred to by the term “Digital Humanities”, is fueled by increased availability of digital text and algorithms for identification of (semantic) structures in unstructured data. Nonetheless, the complexity of handling ‘Text Mining’ (TM) procedures as well as problems in managing of ‘big data’ prevents those approaches from being used by a wider audience lacking a computer science background.

To facilitate the handling of large document collections and make use of algorithmic Text Mining procedures for content analysis (CA) we built the “Leipzig Corpus Miner” (LCM). With CA we refer to a broad set of methods and corresponding methodologies for analyzing textual data common in various scientific disciplines. These may include classic (quantitative) content analysis [12] as well as rather qualitative approaches like discourse analysis [13], grounded theory [9] or qualitative content analysis [16]. In this respect, techniques integrated into the LCM do not replicate analysis procedures of these methods exactly. Rather they offer a set of tools which enable analysts to support certain steps of an
applied method and to extend the size of collections under investigation to a
degree which could not be handled manually.

In contrast to most computer-assisted CA studies which employ only a single
or very few TM procedures [22] LCM allows for application of multiple proce-
dures which may be integrated systematically into complex analysis workflows.
Thus, results of single processes are not restricted to be interpreted in an iso-
lated manner. Beyond that, they may be used as input data for further mining
processes. For example, a list of key terms automatically extracted from a sub-
set of a reference corpus may be utilized to retrieve documents of interest in a
second target corpus. Semantic topics automatically retrieved from Topic Mod-
els [2] may be applied to disambiguate homonymous term usage and thus, help
defining concepts under investigation for individual CA studies.

The systematic integration of TM procedures in a coherent user-friendly app-
lication enables content analysts to employ complex algorithms while simul-
taneously retaining control over major parameters or configurations of analysis
process chains. Utilizing those procedures on large document collections may
improve quality of qualitative data analysis especially in terms of reliability and
reproducibility. It thereby gives analysts without deeper knowledge of founda-
tions of natural language processing (NLP) the ability to develop best practices
for computer-assisted large scale text analysis. Because NLP experts usually lack
background knowledge about requirements and methodologies in the humanities
and social sciences such best practices could not be developed by either discipline
alone. Providing a common interface on text analysis is the major advantage of
using a framework like the LCM.

2 Architecture

The LCM is a combination of different technologies which provide a qualitative
data analysis environment accessible by an interface targeted towards domain
experts unfamiliar with NLP. Analysts are put in a position to work on their data
with more methodical rather than technical understanding of the algorithms.
Applied technologies behind the user interface need to support analysts in tasks
such as data storage, retrieval, processing and presentation. A schematic overview
of the architecture and workflow is given in figure 1.

Storage and processing of large amounts of text data are key tasks within the
proposed environment. To process data we use UIMA, an architecture to identify
structures in unstructured textual data [7]. Within this architecture data read-
ers, data writers and data processing classes can be chained together in order
to define standardized workflows for datasets. In UIMA data processing is done
by so called “annotators” which (usually) add additional information to the an-
alyzed documents. UIMA employs “stand off” markup for annotations which in
contrast to inline markup does not alter the original text. Therefore, annota-
tions can be stored in a separate data structure [6]. Annotations could include
sentence information, token separations or classification labels. But annotators
can also be used to aggregate single text documents for further processing such as topic modeling. We defined UIMA processes for the following tasks:

- **Initial processing of XML sources:** To initially process raw sets of text data we defined a process which identifies areas like titles and paragraphs, separates sentences, applies tokenization and POS-Tagging, detects entities and multi word units (MWU) and performs entity resolution. *OpenNLP*[^1], the *Stanford NER* tool [8], a Wikipedia based multi word unit detection and the *JRC-Names* resource [20] for entity resolution are utilized within this tool chain. We trained custom models for sentence detection and tokenization with resources provided by the *Leipzig Corpora Collection* (LCC) [1].

- **Classification, Corpus Linguistics (CL) and Machine Learning (ML):**
  The UIMA based *ClearTK* [18] framework is applied to provide convenient access to classification and machine learning libraries to support certain analysis steps. We also implemented inference algorithms for different Topic Models and corpus linguistic analyses based on ClearTK.

Annotations made by processing chains, e.g. tokens, sentences, entities or classification labels, can vary throughout different analysis approaches. We therefore decided to employ a schema free data storage which gives us the flexibility to delete or add annotation structures to stored text documents. For this, we use MongoDB[^2]—a NoSQL database implementation which stores datasets as JavaScript Object Notation (JSON) structures. It supports the distribution of very large datasets to different shards and machines. This is a very important feature as we want to open the system for new large corpora. Every imported corpus can be accessed separately within the architecture.

Alongside with data storage the user-friendly accessibility of corpora by different indexes is an additional requirement. As we want to provide indexes accessible by a convenient query language based on metadata fields and full-text of documents we use the *Solr*[^3] enterprise search server. For this a set of metadata fields and full-text contents for indexing each corpus can be defined. This information is provided to Solr which uses annotations made by UIMA directly to populate its indexes. As well as MongoDB we can distribute the Solr indexes to a cloud and thus are able to index and store new large corpora. Indexes are used for full-text search, faceted search and search on meta data fields.

Access to the datasets and indexes is facilitated by a web-based front end (fig. 2). This web application, based on the *YAML* CSS framework[^4], accesses stored corpora and datasets through Java Servlets[^5] hosted by an Application Server. The front end to back end communication is managed by *AJAX* calls via the *jQuery*[^6] javascript framework. Presentations of results, graphs and charts are

[^1]: http://opennlp.apache.org
[^2]: http://www.mongodb.org
[^3]: http://lucene.apache.org/solr
[^4]: http://www.yaml.de
[^6]: http://jquery.com
implemented in D3[4], a visualization library to create, manipulate and animate SVG Objects.

One central objective of the LCM is to enable analysts to perform Text Mining tasks without explicit guidance by NLP experts. For that reason we implemented a middleware (UIMAWS), a webservice to start, stop and manage UIMA processes for certain tasks. The analysis tasks are described in detail in section 3 of this paper. The UIMAWS is deployed to a dedicated server and is able to manage multiple instances of UIMA processes. The webservice communicates information on the progress of running processes. Results of completed tasks will be written to the database and visualized by the front end. The UIMA processes always operate on a finite set of documents which are referenced by their ID within MongoDB. Those identifications can be retrieved by querying the Solr index, a custom information retrieval algorithm or manually compiled lists of documents. Starting and managing of UIMA processes can be done in the front end as shown in figure 3.
3 Analysis capabilities

The LCM integrates several procedures for retrieving, annotating and mining textual data. Flexibility in combining these tools lends support to various analysis interests ranging from quantitative corpus linguistics to qualitative reconstructivist methodologies. We briefly introduce the inbuilt technologies and give an example of a workflow in a concrete research environment in section 4.

– Information retrieval: Assuming the availability of a large document collection, e.g. complete volumes of a daily newspaper over several decades, a common need is to identify documents of interest for certain research questions. The LCM provides two ways of IR for this task. First, a full text index allows for key term search over the entire collection (see section 2). But, for CA relevancy often cannot be tied to a handful of key terms. Thus, we provide a second retrieval method based on contextualized dictionaries which can be retrieved from a reference corpus of paradigmatic documents covering a certain research interest. A list of several hundred key terms, called ‘dictionary’, is extracted automatically from this reference corpus and co-occurrence patterns of these terms are measured. Relevancy in our target collection is then defined through occurrence of dictionary terms in their desired co-occurrence contexts from the reference collection. This method allows not only for retrieving coherent document sets on a single subject-matter (e.g. ‘Iraq war’), but rather for distinguishing modes of speech over various topics (e.g. military cadence in politics or sports). Document collections retrieved by either IR system can be stored for further procedures.

– Lexicometrics: The LCM has implemented computation and visualization of basic corpus linguistic measures on stored collections. It allows for fre-
quency analysis, co-occurrence analysis (figure 4) and automatic extraction of key terms. As documents contain a time-stamp in their metadata, visualizations can aggregate these measurements over days, weeks, months or years. Researchers may choose if they want to have absolute or relative frequencies displayed. For co-occurrence analysis LCM allows selecting different significance measures to compute term relations (raw counts, Tanimoto, Dice, Mutual-Information, Log-Likelihood)[5].

- **Topic models**: For analysis of topical structures in large text collections Topic Models have been shown to be useful in recent studies [17]. Topic Models are statistical models which infer probability distributions over latent variables, assumed to represent topics, in text collections as well as in single documents. So far the LCM has implemented models of Latent Dirichlet Allocation (LDA) [3], Hierarchical Dirichlet Process (HDP), Pitman-Yor Process (HPY) [21] and Online LDA [11], a very fast inference algorithm for the LDA model [3]. Results from Topic Models can be seen as an unsupervised clustering which gives analysts the opportunity to identify and describe thematic structures on a macro level as well as to select subsets of documents from a collection for further refinement of the analysis or for a “close reading” process. Distributions of topics can be visualized over time (figure 5).

- **Classification**: Supervised learning from annotated text to assist coding of documents or parts of documents promises to be one major innovation to CA applications [19]. The LCM allows for manual annotation of complete documents or snippets of documents with category labels (figure 6). The analyst may initially develop a hierarchical category system and / or refine it during the process of annotation. Annotated text parts are used as training examples for automatic classification processes which output category labels for unseen analysis units (e.g. sentences, paragraphs or documents). For that the LCM integrated several classification approaches like SVM, Naive Bayes and Semantically Smoothed Naive Bayes [23]. Feature engineering can be performed by the analyst up to a certain degree by configuring the classification process (e.g. restriction to certain POS-types for training). An iterated process of manual labeling and evaluation of (best) automatically retrieved labels may replicate “Hybrid Active Learning” [15].

## 4 Example use case

The LCM was initially built to facilitate a CA study on political theory. Within the German “ePol-project”\(^7\) political scientists conduct large scale text analyses with the help of NLP researchers. The project aims to identify changes in discursive patterns of policy justifications in public media. By identifying certain patterns and measuring their quantities the project strives to verify or reject central hypotheses about a phenomenon called “Post-democracy”. One central

\(^7\) [http://www.epol-project.de](http://www.epol-project.de)
Fig. 4. The LCM provides various data visualizations on sub collections, e.g. co-occurrence graphs.

Fig. 5. ... or topic model distributions over time.

Fig. 6. The UI provides assistance for manual annotation of document parts which subsequently may be used as training examples for classification.
question is: Has there been an “economization” of justifications in some or all policy fields during the last decades? [14]. To answer this question a corpus consisting of 3.5 million newspaper articles from 1946 to 2012 is investigated.8

Fig. 7. Schema of a Content Analysis workflow realized within the LCM.

The LCM was built with the requirement to support analyses of and inquiries into these rather abstract questions. Figure 7 shows a simplified workflow of the analysis process which is realized with the help of the aforementioned procedures.

1. A subset of relevant documents for the analysis is retrieved from the complete corpus of 3.5 million documents. To identify “neoliberal” themes and modes of speech a reference corpus consisting of 36 works from neoliberal theorists was compiled. A dictionary of 500 automatically extracted key terms and their sentence co-occurrences were computed. With our IR system we select 10,000 (potentially) relevant documents from the initial corpus.

2. A topic model computation over this subset identifies thematic clusters which help to describe its content from a macro perspective. The topic model result allows to distinguish of policy fields mentioned within the articles. It also enables analysts to remove documents from the selected subset which contain topics irrelevant to the specific purpose of this analysis (e.g. articles mainly concerned with foreign policy are not considered relevant for this study).

3. Best ranked documents from step 1 of each topic (step 2) are investigated manually by the political scientists to identify argumentative structures in different policy fields. A hierarchical category system is derived from common, distinctive structures. Relevant sections of the documents were then annotated with category labels.

4. An automatic classification procedure is invoked on the unlabeled data to identify more text parts containing argumentative structures. The NLP group supports the analysts by identifying discriminating linguistic features for this task.

5. Text snippets identified in the previous step (supposedly containing arguments of interest) are presented to the analysts ranked by certainty of label assignments. Analysts can verify or reject the results manually. In this active learning paradigm we calculate internal precision / recall measures while the analysts are evaluating the process qualitatively. If those ongoing evaluations show satisfactory results, the process of creating training data is concluded.

8 The corpus consists of complete volumes of DIE ZEIT, TAZ, Süddeutsche Zeitung and a representative sample of Frankfurter Allgemeine Zeitung.
6. The classification procedure is run on the entire collection under investigation. Results can be described qualitatively and quantitatively (e.g. proportions of categories over time slices).

5 Future Work

The LCM supports manual Content Analysis (CA) via basic corpus linguistic procedures as well as supervised state-of-the-art Text Mining techniques. In a next step we will incorporate a user rights management system which allows for access restrictions to different corpora in the LCM. In addition, work-flows will be further enhanced to combine intermediate results of different procedures. For example, a centralized management of (semi)automatically generated dictionaries could be useful to exploit controlled lists of key terms throughout different steps of the entire process chain (e.g. computing co-occurrences just for key terms; using occurrence of terms in dictionaries as additional feature for classification tasks). To support more inductive research approaches, we plan to integrate unsupervised data-driven analysis procedures into the LCM. These could help analysts especially between steps 2 and 3 of our proposed process chain to identify categories of interest in unknown data. This includes the identification of stable as opposed to volatile semantic concepts in certain topics over time [10].

In classification and annotation processes we enable analysts to define categories and training data for those categories. Within this process many evaluation measures, frequently used by CA scholars, could be applied. In future developments we will incorporate more of these measures regarding cooperative annotation and categorization tasks. This will be a further step towards integration with common social science research methods away from a narrow computer linguistic oriented work-flow. Its our hope that the LCM may contribute to improve the comparability, reliability and validity of social science research standards.

References

15. Lughofer, E.: Hybrid active learning (hal) for reducing the annotation efforts of operators in classification systems. Pattern Recognition 45(2), 884–896 (2012)
Combining EU Terminology with Communication and Ontology Research

Rodolfo Maslias

Terminology Coordination Unit, European Parliament, Luxembourg

Abstract

The Terminology Coordination Unit of the European Parliament has an original approach among the Coordination Services of the 10 EU Institutions. It has opened up EU terminology to the science and industry aspects of the field.

This presentation will focus on

the access to EU terminology that TermCoord offers through a public website and free tools, through the Interinstitutional Terminology Portal EurTerm and its expert collaborators, through traineeship opportunities, study visits, open seminars, the publication of studies, university courses, terminology projects based on the interactive IATE template, participation in major international terminology networks, etc.

the constant improvement of the content and functionality of the EU database IATE; including free access to its content, cross-referencing features, as well as TermCoord's cooperation with several universities in order to explore possibilities for adding an ontological dimension to its structure of more than 100 domains.

Key words: Terminology, Cooperation, Communication, Access, IATE, Universities, Portal, Ontologies

1. Communicating for Terminology

1st Level: Internal Communication and Awareness Building

Coordinating Terminology Management in a Multilingual Public Institution

The European Union issues legislation in 24 languages and works with 552 language combinations. Among the tools used for translation on an interinstitutional level, the EU regrouped all separate databases in an interactive database permitting permanent input by the some 5000 translators of the participating Institutions. The management of this database requires a very complex system of rules and guidelines (i.e. a handbook for best terminology practices in referencing and creating definitions, legal notice addressing copyright, ownership and responsibility issues) to ensure linguistic quality and consistency between the legislative texts in all languages. In the European Parliament, which gradually obtained more and more power in the legislative process until it became the final legislator under the Lisbon Treaty, the input of terminology is created mainly from the thousands of amendments of 766 MEPs, pertaining to 7 political groups and 23 committees dealing with all policy fields. The input was organized in each translation unit in a different way and according to different approaches depending on the various usage conventions and cooperation methods of each language community and each country. In 2008, the European Parliament decided to create a horizontal unit to coordinate, harmonize, support and assist the terminology research and storage of the findings into IATE and also to cooperate with the units and the other Institutions in the huge effort of cleaning and updating a database containing millions of terms coming from different databases, massively imported without any filtering, some of them being obsolete and outdated and creating duplicates with newly inserted terms. They also needed completion with the new languages that joined the family of the EU official languages in 2003. To achieve this coordination, the new unit needed to raise awareness on the inherent role of terminology in
translation, especially of a legislation governing 28 countries and permitting each citizen to go to court on the basis of the translated version. It needed to organize a group of translators dealing with terminology in the units; today this group is the Terminology Network and numbers more than one hundred translators in the 23 language units with at least two per unit. Special skills and knowledge are needed for the "terminologists", who have to be appointed as such by their superiors among the translators to accomplish the terminology tasks.

These tasks had to be inserted into the workflow and accounting system. Training was needed, tailored for the different target groups, advanced administration of the database for the terminologists, input instructions for the translators, as well as basic training for the newcomers and for the average 60 translation unit trainees who work at the Parliament for a period of 3 months. These trainees dedicate two weeks of their traineeship to field or language specific terminology projects, both for educational and professional experience reasons and to cover needs of lacking human resources because of increasing budgetary constraints combined with an extreme workload in the translation units. Since 2012, TermCoord has been able to offer terminologists the possibility of an on-line ECQA Terminology Management Certification in cooperation with TermNet and covered by the training budget of the EP. TermCoord staff is also invited to provide training to other Directorates in the EP, interpreters and web-editors, as well as to other Institutions like to the translators of the European Investment Bank and the Court of Justice.

Different IT tools were needed such as term extractors, a macro to ease preliminary storage of terms during translation, a search engine to browse the hundreds of links to specialized glossaries imported, a webpage concentrating the links to all reference documents of all EU Institutions. Some of these resources (DocHound and GlossaryLinks) have also been made public on the website of the unit and are used also by the external contractors translating 30% of Parliament’s texts.

All these efforts needed a strong communication policy based mainly on an internal website in the intranet, making part of an internal portal of the Directorate General of Translation. Awareness campaigns on different slogans like "recycle your terms", "in termino qualitas" or "tailored training" and "let's IATE, your database is what you make of it", as well as a series of terminology seminars under the generic title "Terminology in the changing World of Translation" and on topics such as "multilingual legislation process", "terminology and CAT tools", "e-lexicography and terminology in the media" that have been very well attended and evaluated by translators of the European Parliament and other Institutions.

2nd Level: Interinstitutional Communication and Information Sharing

Cooperating and Administrating a Common Interactive Database and a Terminology Portal

The internal and interactive part of IATE needs a tight and continuous cooperation between all the participating Institutions. This cooperation concerns:

a) The content, consolidation, merging, constant updating of methods and guidelines, common or coordinated projects in specific fields; also decisions on import of collections, and referencing criteria to be followed by all EU translators

b) Administrational aspects such as rights and roles attributed to different groups of users, transfer of ownership to another Institution for updating or completion and interinstitutional validation

c) Technical aspects such as prioritizing improvements to functionalities made by the permanent external developers team established in the Translation Center in Luxembourg, testing and bug management.

All these decisions are taken by the IATE Management Group, accountable to an Interinstitutional Directors’ body. With a presence in the Management Group since 2008, TermCoord brought about several changes increasing its efficiency by suggesting the creation of several working groups and task forces dealing with the
data entry and interface improvement, with the updating of the handbook and the integration of IATE in the new CAT tools. It also suggested the creation of an interinstitutional platform for the sharing of information and providing common resources to all EU translators, as well as the conversion of IATE in a modern web 2.0 version. Both of these initiatives have been endorsed by the interinstitutional hierarchy and a specific budget was allocated.

The EurTerm Terminology Portal developed by TermCoord in “Confluence” will be fed by an editorial board composed from the Terminology Coordination Departments and will contain access to internal IATE with which it will be connected, a centralized access to all interinstitutional and worldwide terminology resources and databanks and access to communication platforms on a language level (terminology wikis and forums).

Additionally, it will provide access to the intra-EU meta-search terminology tool QUEST searching in the most important national terminology databases in addition to IATE, the translation memory Euramis, and EurLex. Further features will include a blog, a calendar of events, a multimedia section for webinars or web streaming of events, as well as the search tool for the links to glossaries selected and updated by our Unit. TermCoord prepared in 2011 the model of this portal that has been agreed and will be finalized and opened to the translators in 2014 and in the future partly to the wider public.

TermCoord established and regularly hosts videoconferences with the Terminology Coordination Departments in Luxembourg and Brussels on strategy and management issues, deciding on communication issues, training and even an exchange of staff members for a period of two months to increase knowledge transfer and cooperation.

Except IATE, the unit also represents the EP in the management of the interinstitutional text related forum ELISE, in which translators from the three legislative Institutions translating texts related to the same act discuss on-line terminology in their respective language.

**3rd Level: Cooperation with international organizations**

On an international level, the Joint Inter-Agency Meeting on Computer-Assisted Translation and Terminology (JIAMCATT) regroups all international organizations and deals with all subjects related to CAT tools and Terminology. Within the scope of its main goal, which is the cooperation among all partners and the exchange of knowledge and material also in the field of terminology, these organizations try to gather all terminology resources and grant access to each other and to share their experience with terminology projects.

JIAMCATT’s newly set up Terminology Working Group (consisting of representatives of EP TermCoord, UN, European Court of Justice and European Commission) contributed substantially to enrich the terminology section of JIAMCATT’s website with the EU and other terminology resources. Many of these resources can be found on EU interinstitutional terminology portal EurTerm, which in the near future will be opened to all JIAMCATT members.

In the (most recent) 27th JIAMCATT meeting (April 2014), EP TermCoord proposed to create an interinstitutional access page to the members’ databases and portals. The idea behind this is to offer all members of JIAMCATT a common meta-search tool (restricting access to pages containing sensitive information) that will enable searches in all databases including IATE, UNterm, FAOterm, AGORA (OECD), etc.

Sharing of terminology resources and tools among international organizations was recognized as a very beneficial exercise for the quality of translation and the need for enhanced communication and networking has been stressed in order to face the spectacular rise of the importance of terminology in a globalised context.
The common platform will provide a single central access to terminology resources, tools and databases; efficient and user-friendly interinstitutional communication, linguistic knowledge and skills sharing as well as enabling communication on language level and on specific topics.

4th Level: External Communication, Promotion and Coining of Resources

Connecting with the External World of Terminology

Terminology has evolved in the last years, thanks to and because of the inherent need for multilingual communication for every public, private and academic sphere of activity in a globalised world.

Important companies worldwide possess huge terminology departments and databases; private providers or non-profit organizations offer databanks and terminology support. Also, very large professional forums on terminology have been created, each international activity includes a glossary on the relevant website, there are yearly hundreds of conferences on terminology and connected activities, very large organizations regroup hundreds of public, private and academic actors dealing with terminology.

Terminology has also become a separate discipline in the linguistic sector and the humanities. Hundreds of Universities worldwide have created terminology departments or integrated terminology courses and practice into their programs in translation and linguistics faculties. Terminology in several fields is very frequently the subject of academic theses of all levels and the Terminology Coordination Unit of the European Parliament has to select every three months among more than 1500 applications from all over the EU to cover the 6 posts for a specialized traineeship in terminology. While in the recent future, terminology specialists tried to enrich the web with terminology resources, today in the era of web 3.0 and big data, the main task is to regroup in an easily accessible way all the huge and multiple resources in all languages of the world and all fields of activity.

The European Union has the largest and most multilingual human translation production in the world and cannot stay isolated in this wonderful evolution of terminology and its adaption to the current needs of our localized and globalized world. TermCoord has started an intensive campaign and could include IATE in some of the biggest terminology search tools and webpages.

TermCoord has contributed a module on Terminology to the University of Luxembourg Master's program "Multilingual Learning in a Multicultural Environment". This consists of courses on multilingualism, terminology management, digital terminology coining and practical exercises, accompanied by theoretical courses given by Prof. Rute Costa in the next semester. The students of this module are offered internships in TermCoord. Staff members of the Unit are invited to several universities for seminars and to initiate IATE based terminology projects.
As from its creation, TermCoord has developed a wide network of external contacts mainly in the other international institutions and the academic world. Even in the first months of its existence it invited 10 prominent terminology professors and cooperated with them for the development of an original approach of terminology management in a public institution. A very important asset for the unit has been the knowledge and enthusiasm of the by now some 100 trainees from all countries with studies or specific interest in terminology, who offered an important input from the most recent developments and keep cooperating with the unit in several projects through specific accounts and discussions in the social media.

In April 2011, the unit went public with a website that attracted up until December 2013 300.000 visitors and became an important reference in the world of terminology offering EU specific resources and interesting posts on translation and terminology prepared by our staff in cooperation with trainees specializing in communication. In 2012, this external presence was extended with accounts and pages in the most important professional social media networks.

The biggest interest was shown for our pages containing the IATE term of the week, related and linking to an article on an important current event, the interviews with prominent terminologists conducted by our trainees, and the specific EU terminology resources and the different possibilities offered by the European Parliament for traineeships and study visits in our unit.

In 2012, TermCoord launched IATE projects with Universities: Interested translation and terminology departments receive a template to make exercises with their students simulating input into the database and compile lists of entries in their languages in cooperation with our unit, which after validation are inserted in IATE. These projects are still limited in the pilot phase and are intended to become in the future, projects conducted by the Terminology Coordination Departments of all Institutions cooperating in IATE.

As from the first semester of 2013, the University of Luxembourg inserted in the syllabus of its master "multilingualism in a multicultural environment" a series of courses and workshops on terminology management, multilingualism, terminology coining and web monitoring, given in three languages by members of TermCoord staff with the agreement of the European Parliament. Several universities invited the lecturers to give a part of these courses combined with a terminology project (Saarbrücken, Vienna, Thessaloniki, Bologna, Paris, Sofia, Salerno, Malaga, Vigo a.o.).

TermCoord is preparing a project to insert terminology research and ontologies in the European Master of Translation offered by the European Commission in cooperation with several Universities.

Our unit was the subject of a master thesis at the University of Vienna and provides the possibility to students and researchers to publish on its website their theses and studies on terminology or containing a chapter on terminology. Students working on terminology in their universities are allowed to apply for an unpaid traineeship in TermCoord, where they are familiarized with practical terminology work.

Several university professors were invited to our big seminars as speakers or were invited to run specific workshops for the translators and terminologists. The next seminar will take place on the 5th of June under the title "Terminology in Academy", focusing on this cooperation and ontologies.

TermCoord is member of EAFT (European Association for Terminology) and TermNet and participates in international terminology conferences autonomously or representing all EU Institutions like in the latest JIAMCATT hosted by the European Parliament in Luxembourg in 2012. Presentations are also given in Universities upon request or organized in the European Parliament in Luxembourg for groups of professors and students.
II. Extending Terminology to New Horizons: TermCoord Experiments with the Semantic Web Technologies

The discipline of Terminology is in constant evolution. It is widely considered as a stand-alone science and has been put into the spotlight by universities, companies, state bodies and international institutions worldwide. In the era of open linked data, big data and web 3.0, a highly specific subject like terminology cannot ignore neither new technology inputs from these fields nor the mutual benefits to be obtained.

In this sense, IATE Public has recently undergone several reviews and improvements. The most interesting ones include entry-to-entry links and cross-references as active hyperlinks. Moreover, a web service that permits queries has also been implemented. In particular, these changes will soon make it possible to download IATE data exports in TBX format after updating its restrictive "legal notice".

In addition, LTAC Global (www.ltacglobal.org), the host organization of the TBX Steering Committee, will cooperate with IATE in the design of a term base template that matches the content of document-specific glossaries that could then be imported into different CAT tools. With this goal in mind, the Committee will define a dialect of TBX that supports the selected fields and will provide working software in Perl or some other programming language that automatically converts between TBX and the term base format. Together with the IATE software developers, the Committee will then create an application that formats information from IATE in the selected TBX dialect.

On top of that, TermCoord has also suggested transforming IATE (linguistic resource in form of a terminological relational database) into a knowledge management resource in the form of linked data, in order to produce multilingual solutions for different working domains that are homogeneous and responsive.

The main difference between a terminological database and a knowledge base is that the former presents concepts, terms and their multilingual equivalents in isolation with no semantic interconnectivity. The latter, on the other hand, presents concepts as 3D networks with several different aspects from which you can request and obtain linguistic and cognitive information in a well-structured, logical manner.

For this reason, the IATE Management Group is looking to achieve the amelioration of IATE with the semantic web technologies. By developing a model of linguistic-terminological knowledge management, the Group wishes to define new typology for a terminologist profile within the EU Institutions. This profile is set to deal not only with concepts and terms, but also with knowledge management, multilingual webpage content indexing, and website content management.

In this context, the IATE Management Group, represented by 3 of its members (EP TermCoord, the Translation Centre and the Publications Office of the EU) will launch a project concerning the development of an ontological structure in 2 sub-domains of IATE. This will be done in cooperation with the University of Luxembourg (philosophy and linguistics), the University of Salerno (communication) and the University Sapienza of Rome (computational linguistics). This project should refer to the initiatives related to the ISA Program, such as a common metadata vocabulary or Asset Description Metadata Schema (ADMS).

The first step will be to find an interoperability solution for IATE in order to overcome the semantic heterogeneity within the Institutions of the EU, i.e. Council of the EU, European Commission, European Parliament, Court of Justice of the EU, European Court of Auditors, European Central Bank, European Investment Bank, European Economic and Social Committee, Committee of the Region and Translation Centre.
On the one hand, this common project aims at enhancing stakeholders’ access to EU policies, as well as supporting EU citizens in understanding institutional terminology. Yet on the other hand, it should improve the management of institutional terminology and communication and, above all, it should help translators in their daily tasks with non-ambiguous terms by providing direct links to reference texts.

The domain study will be conducted taking into account sociolinguistic, economic, organisational and translation aspects. This is intended to narrow the initial conceptualization of the field of expertise and to establish an initial conceptual structure prototype.

As a second step, a specialised corpus will be compiled to develop the database, i.e. linguistic (terms, contexts), semantic (semantic relations, concepts) and pragmatic information (frequency of use, communicative situation).

Since an extraction from a corpus unicum can lead to ambiguity due to the varying concepts a single expression can refer to in different contexts, a text classification will be implemented to avoid this situation.

Moreover, this facilitates linking a term to the text where it was extracted from in the ontology and in turn will improve the references and quality of the extracted terminology.

Terminology extraction (with a special focus on multi-word expressions) will be carried out following this corpus-based approach by a classification on text types for different pragmatic finalities.

The focus of this project lies on multi-word expressions since they are rarely, and often poorly, recognized by statistical extraction tools. Most terminology extraction tools that focus on frequency do not recognize multi
word expressions, because they are based on mathematical-statistical algorithms. For this reason, TermCoord will continue testing different extraction tools -from statistical, linguistic to hybrid ones- to approach this project in the most appropriate way.

To acquire knowledge from corpora, the most widely used method is based on knowledge patterns. These are logical hierarchical relationships, ontological hierarchical relationships and non-hierarchical or associative relationships. This data will be extracted from the compiled corpora in order to gain vertical, as well as horizontal relationships that will then be converted into computational linguistic resources to be used in ontologies.

Although there are different ways to develop domain ontology, either by using a linear approach or by means of graphics or semantic frameworks, if the latter is applied to terminology, it allows a clear and orderly structuring of concepts; thus, it allows observing the context in which different concepts and terms appear. Therefore, semantic frameworks will be implemented in our project and every ontology entry will be linked to a specific type of text.

Apart from the above described goals, the project also intends to examine the following aspects:

- The integration of IATE with EuroVoc and VocBench (SKOS) and its use in document and website indexing

  EuroVoc is a multilingual thesaurus which was originally built up specifically for processing the documentary information of the EU institutions and now it is a controlled set of vocabulary which can be also used outside the EU institutions. The aim of the thesaurus is to provide the information management and dissemination services with a coherent indexing tool for the effective management of their documentary resources and to enable users to carry out documentary searches using controlled vocabulary.

  VocBench is a common platform offered by the Publications Office to the EU Institutions and agencies for the maintenance/dissemination of controlled vocabularies as well as indexing/annotation of their documents.

- How to reduce redundancy and overlapping in EU terminology in order to foster a more coherent and standardised use of terminology in the European Union

- The use of the European Open Data Portal to share the resources (in TBX or TBXmin)
Cooperation of TermCoord with the Laboratory of Terminology of the University of Bologna

In the recent seminar, Terminology in Academia (termcoord.eu/terminology-academia/) organised at the European Parliament, Prof. Franco Bertaccini presented the difference between term databases and knowledge bases to the translators and terminologists of the EU Institutions.

In the context of this cooperation, a student of this University will dedicate her thesis to a sample IATE database including ontologies.

References


W3C and the Semantic Web: http://www.w3.org/2001/sw/

You can find any additional information on TermCoord’s website (http://termcoord.eu)

Presentation by Rodolfo Maslias (Head of the Terminology Coordination Unit of the European Parliament)

In collaboration with: Carolina Dunaevsky (University of Cologne) and Marie-José Palos Caravina (Translation Centre of the EU)

6.6.2014
Extracting People’s Hobby and Interest Information from Social Media Content

Thomas Forss, Shuhua Liu and Kaj-Mikael Björk

Department of Business Administration and Analytics
Arcada University of Applied Sciences
Jan-Magnus Janssonin aukio 1 00560 Helsinki Finland
{thomas.forss, shuhua.liu, kaj-mikael.bjork}@arcada.fi

Abstract. In this study we investigate how to analyze people’s social media profiles to extract hobby and interest information. We developed a baseline system that applies heuristic rules and TF-IDF term weighting method in determining the most representative terms indicating hobbies and interests. A pilot test was done to collect feedback from users concerning the perceived usefulness of the extracted tags. The baseline system was then extended to include new functionality to help set limits on the scope of relevant content, extract Named Entities, use of predefined dictionaries to identify even low-scoring hobbies and interests, and use of machine translation to handle content in multiple languages.

Keywords: social media analysis, term extraction, hobby and interest, Facebook profile, Named Entity extraction

1 Introduction

Since the beginning of the 21st century more and more users use some form of social media every day. Popular social media domains include blogs, web forums, photo and video sharing sites as well as social networking sites [1]. Popular social media networking sites that focus on sharing information with friends include but are not limited to Facebook, Twitter, LinkedIn, Google+, and Diaspora. As of January 2014 Facebook has over 1.3 Billion users spread all over the world while Twitter has about 645 Million users [2].

In parallel with the huge explosion of the usage of social media, we have an Analytics movement that strive to create competitive advantage and added value based on analyzing the huge amount of both structured and unstructured data. In Delen and Demirkan [3], the future of data, information and Analytics as a service is predicted. We will mostly probably see a significant number of new services emerging to help us analyze social media content.
Numerous research on social media analysis have been made, however there is much still to explore. Social media content analysis has been reported in several studies on shorter information snippets such as the ‘tweets’. Shamma et al [4] built a tool, Statler that looks for trending topics, level of interest, and geo-locations of tweets. Zhao et al [5] extracts keywords and organizes the keywords according to topics learned from Twitter through a context-sensitive topical Page Ranking method. Yang et al [6] studied automatic summarization of Twitter tweets through topic modeling and event detection. Much less research has been done on analyzing and/or summarizing a more complete social media presence of people, i.e. user profiles, in sites like Facebook, LinkedIn, Diaspora, and Google+.

Text summarization and keyword extraction was originally done on structured documents that have paragraphs, sentences and correct grammar. In social media the format the text is saved in depends upon how the site or platform in question stores the data and how the user chooses to write. The challenge in analyzing a person’s social media profile lies in that not all the available information is relevant to the user who is subject to the analysis, and not all the content is created by the one user. Additionally, the content can be fragmented and neither structured like normal text nor written in a grammatical way. Further, multilingual users tend to change languages between posts and can also sometimes write posts mixing several languages.

Clark & Araki [7] identifies 8 different grammatical and structural problems that can make extraction and/or summarization harder in social media content. Out of these 8 problems the following ones are relevant for this study: non-dictionary slang, punctuation omission/errors, intentional misspellings, and abbreviations. Bertoldi et al [8] extended a statistical machine translation tool with the capabilities to adjust for misspellings, combining the approach Clark & Araki [7] propose with machine translation could provide a similar result.

In this paper we present our work on analyzing social media profiles to extract users’ hobby and interest information from Facebook content. We developed a baseline system that applies heuristic rules and TF-IDF term weighting method in determining the most representative terms indicating hobbies and interests. A pilot test was done to collect feedback from users concerning the perceived usefulness of the extracted tags. The baseline system was then extended to include functionalities that help limit the scope of relevant data, Named Entity recognition, and predefined dictionaries containing hobbies and interests, and possibilities to handle multiple languages.

The extensions of the system were added after the pilot test. As such the extensions while implemented have not been tested except for by the authors. We plan on running a second test with the extended system in the near future.
2 A baseline system and pilot test

2.1 A baseline system

To the best of our knowledge there are no systems available that extract or summarize content from a Facebook profile. So we started with building a baseline system. Our approach can be described as a three-step model shown in Figure 1. First we retrieve content and group it. Then we pre-process the content and lastly we extract hobbies and interests.

Keyword/key phrase extraction help select a small set of words or phrases that are central to the information in the text, which in our case is the sum total of one person’s activity on a social media site. In the simplest case, keywords can be determined using word weighting methods. TF-IDF is one of the most popularly used and robust word-weighting methods, combining the effect of term frequency with inverse document frequency to determine the importance of words [9]. This method automatically gives a low value to common words like pronouns and prepositions that are normally neither relevant to a summary nor should be counted as key phrases. On the other hand a word with a high term frequency in the text we are trying to summarize and a low inverse document frequency would give a high TF-IDF value and thus identify a word that is important but not found in many different texts.

2.2 Pilot test

A pilot test with the baseline system is done to get a hands-on experience and benchmark with the performance of such systems, mainly the usefulness of the extracted tags.

What is considered a useful tag or a bad tag is mostly subjective and as such the only way of finding out whether the extracted tags are useful is to ask the users. The pilot test was set up so that each user gets twenty extracted tags and is asked to rate...
the tags as useful, neutral or not useful. The system is limited to handling English language so only profiles with content mostly in English are considered valid.

The pilot test was conducted among international students and researchers and collected helpful feedback and observations concerning the type of useful tags we may use and system performance issues. In total we had 42 tests. Test results from non-English speakers were rejected from statistical analysis. This leaves the number of valid tests to 21. On average the percentage of useful tags was 43%, with highest at 55% (11 out of 20 tags) and lowest at 25% (5 out 20 tags).

In addition to interests and hobbies such as “cooking”, “biking” and “reading”, some other types of tags were also perceived as useful by the users. Those tags fall mostly into two categories: adjectives and names. Examples of positive adjectives would be “stylish” and “happy”. Names of people and locations were also perceived as positive. We can assume that the people are closely related or close friends if they are perceived as positive and that locations hold some special meaning for them to be perceived as positive tags. Such information could be linked to which city the person lives in, where the person grew up, or simply some place where the person recently has been.

While the system seems to work quite well on English profiles, it almost always failed to extract relevant tags from multilingual profiles. Other observations include the need for a social media specific stop word list, which could include slang and shortened words. Sentence ordering is not reliable in social media summarization. We expect sentence extraction to give a less than satisfactory result in the social media context. Keyword extraction takes into account neither sentence ordering nor length. As such, keyword extraction is perhaps more suitable to the analysis of social media content than a sentence based summary.

3 Extension of the baseline system

The pilot test showed us some obvious shortcoming in the system. As a further step, we want to be able to parse multilingual profiles so that finding test users becomes easier. We also want to look at the possibility of increase the accuracy of the extraction. So we try to extend our system further in three different areas. We try to increase the accuracy of the system by techniques that will help target relevant text portions in a profile. We implemented Named Entity Recognition to find names and locations of significance but not found through previous methods. We experiment with adding a predefined hobby and interest dictionary in case any hobbies or dictionaries receive low significance score in the extraction. Lastly we implement support for multiple languages through online translation. The extended approach is illustrated in Figure 2.
3.1 Targeting relevant text portion

Here we discuss ways to keep input data relevant, and how we can increase the significance of parts of the data to make it more relevant to the subject of the extraction.

In addition to the TF-IDF weighting of word informative-ness, Luhn [10] suggested taking into consideration the positioning of sentences in texts. A sentence at the beginning of a paragraph or at the end of a text has a higher chance of being of high importance in the text [10]. However, this is not necessarily true for texts gathered from social media sites due to the following reasons: 1) the information will be structured according to whatever database model the company uses, 2) the information will be a collection of different areas such as personal information, communication with friends, interests and other issues which do not necessarily have anything in common, and 3) the data in question can include dialects, multiple languages, intentionally misspelled words etc.

As sentence ordering cannot be usefully taken into account in analyzing Facebook content, we suggest firstly: publishing dates, and secondly: site specific counters, as new criteria for determining relevant data. By limiting publishing to recent dates we can either leave out old information from social media profiles or decrease the significance of TF-IDF values for posts older than the specified date by a pre-determined factor. By using site-specific counters, for example increasing the significance of content that has received a higher amount of "likes" on Facebook, we can increase the content relevance. In addition, we can also try to make good use of semi-structured nature of pages such as those on Facebook and target content under different interest categories such as Places, Music, Movies, Books, Events, and Groups.

Figure 2 Determining the most representative terms for hobbies and interests in the extended system.
3.2 Dictionary based methods and machine translation

Posts containing several languages or non-English languages need to be translated for the extraction to work properly. As mentioned earlier, social media content can consist of several different languages between posts but also within the same posts.

Dictionaries can be used in two ways to improve the results of our extraction. The first approach is to combine support for multiple languages through machine translation and dictionaries. Clark & Araki [7] introduces a system called Casual English Conversation System (CESC) to detect abbreviations, misspellings, punctuation omissions, non-dictionary slang, emoticons, wordplay, and censor avoidance. The approach is based on having databases/dictionaries and matching the text to the phrases in the database/dictionary and replacing the matched erroneous word with the grammatically correct version. [7]

The second way to use dictionaries is to directly extract hobbies and/or interests from social media content. This approach implies that we need dictionaries with the relevant hobbies and interests predefined. With the term weighting approach, new hobbies and interests will be hard to get identified.

The drawback of this dictionary approach is that, if a word is not in the interest/hobby dictionary, then it cannot be extracted. Unlisted hobbies and interests could never occur in an extraction based solely on this approach. The approach is easy to set up but hard to maintain as new hobbies and interests need to be added to the dictionaries to keep up to date.

To extend the capabilities of our system we use an approach that combines dictionaries for hobbies and interests with machine translation and extraction of terms with high significance. The result consists of all existing hobbies, interests, as well as words not registered in the dictionaries.

We use the freely available resource Yandex translation [11] to translate content into English. A separate stop word list is not needed for the extra languages since stop words will be translated to English and then removed by the English stop word list.

We continue by creating dictionaries for abbreviations, slang and intentionally misspelled words to supplement the system. An abbreviation is a shortened word, for instance the word “year” has an abbreviation “Yr.”. A typical example of an intentionally misspelled word would be writing “u” instead of “you”. Slang are words with the same meaning as another word but are not found in standard dictionaries, an example of an English slang word is “aggro” which often means the same as “angry” or “aggravate”. Such dictionaries can be either hand gathered or automatically gathered from the social media site if the structure of the database supports it. For instance, one approach would be to gather learning data from Facebook profiles. The information in question would be from the “like” and “group” tags.
When extracting the information in our system we first remove the abbreviation, slang and misspelled words by going through the dictionaries and match them to the profile that is being parsed. After that we translate the text into English. When we have the profile text in English we can use our interest and hobby dictionaries to supplement the TFIDF extraction. Lastly we decide how to order the extracted results. The options for ordering we have are; according to TFIDF significance, by publishing date, and/or by site-specific counters.

3.3 Named entity recognition

Named Entity Recognition (NER) is used to locate and classify names in texts [12]. While the name of the owner of a Facebook profile is not interesting when extracting hobbies and interests, there are still other named entities that could add value to the extraction. Recent work by Liu et al [13] includes semi-structured methods for finding Named Entities in Twitter “tweets”. Tweets are not structured in any way and limited training data is available. This means that in order to reliably be able to extract Named Entities from tweets a semi-supervised method is created [13].

Hasegawa et al [12] presents an unsupervised approach to how we can link several Named Entities together. The approach works in the following way; first they Tag Named Entities in the text with a state-of-the-art NER tagger, then they pair Named Entities and look for similarities among the found pairs, lastly they cluster paired Named Entities and label them [12]. Liu et al [13] has also done research on Named Entity Recognition in Twitter “tweets”.

When doing extraction of Named Entities in a Facebook profile we can take advantage of the semi-structured data. Each interaction on Facebook is linked to a person, and since we know which person we are extracting data from we can limit the Named Entities in an appropriate way. If we consider Named Entities as a possible part of the extracted words we would need to limit them so that only Named Entities that are relevant to the owner of the profile should be considered. The approach we end up with is a modified version of what Hasegawa et al [12] did for unstructured data, combined with Named Entities directly extracted from certain categories.

Since we are focusing on extracting data that is relevant to the profile owner we can directly extract Named Entities from the Facebook categories Groups and Pages. Groups consist of people that share a common interest; a user can share updates, photos and documents with other people in the group. Pages can be a place, company, institution, organization, Brand, Product, Artist, Band, Public Figure, Entertainment, Cause, or Community. Each profile is only linked to the Groups and Pages that the user him- or herself has decided to be linked to it. This means that Named Entities from these categories are linked to the user with a high probability. For the rest of the data we use the structure to pair Named Entities with the creator of the post, message or comment. Lastly we remove all unwanted Named Entities we have found in the profile from the list of Named Entities.
To be able to order the Named Entities according to relevance we can then increase the significance of the Named Entities so that they appear higher in the TF-IDF weighting. Another approach to sorting Named Entities by relevance is to find the highest weighted TF-IDF word that is linked to each Named Entity and sort them according to these TF-IDF values.

4 Conclusion and future work

In this study we investigate how to analyze people’s social media profiles to extract hobby and interest information. We developed a baseline system that applies heuristic rules and TF-IDF term weighting method in determining the most representative terms indicating hobbies and interests. Our pilot test with limited amount of English-dominant user profiles shows 43% average useful tags, with highest at 55% and lowest at 25%.

The baseline system was extended to include new functionality to help set limits on the scope of relevant content, extract Named Entities, use of predefined dictionaries to identify even low-scoring hobbies and interests, and use of machine translation to handle content in multiple languages. When dealing with social media analysis it becomes important to support multiple languages as the extractions can fail if falsely portray unrecognized words as being of higher significance than they should be. When translating social media content we also need to take into account that people not necessarily follow grammatical rules.

Key word extraction seems more useful to the analysis of social media content than a sentence based summary, due to the fact that social media content is not structured as professional texts. Combining keyword extraction with predefined dictionaries and Named Entity Recognition gives us a broader scope. Named Entity Recognition becomes effective when combining state-of-the-art tools with the semi-structured architecture of for example a Facebook profile. Predefined dictionaries supplement the extraction by including lower-scoring words that still might have a high personal significance. An alternative to keyword extraction would be to extend the baseline system further to n-gram based weighting and topic models. In addition, LDA (Latent Dirichlet Allocation) topic modeling methods [14] could help us in identifying topics embedded in texts.

References

Service model for semi-automatic generation of multilingual terminology resources

Andrejs Vasiļjevs, Mārcis Pinnis, Tatiana Gornostay

Tilde, Vienibas gatve 75a, Riga, Latvia
{andrejs|marcis.pinnis|tatiana.gornostay}@tilde.lv

Abstract. The authors present a service-based model for semi-automatic generation of multilingual terminology resources which, if performed manually, is very time consuming. In this model, the automation of individual terminology work tasks is rendered as a set of interoperable cloud-based services integrated into workflows. These services automate the identification of term candidates in user documents, and the lookup of translation equivalents in online terminology resources and on the Web by automatically extracting multilingual terminology from comparable and parallel online resources. Collaborative involvement of users contributes to the refinement and enrichment of the raw terminological data. Finally, we present the TaaS platform, which implements this service-based model, particularly focusing on the processing of Web content.

Keywords: terminology, service, language resource, multilingual, cloud computing, automation

1 Introduction

Terminology resources are among the most used language data providing lexical designations assigned to concepts, term equivalents in different languages, their definitions, usage contexts, and other data. Digital collections of terminology data are in everyday use by language workers – terminologists, translators, technical writers, and are also increasingly used in automated tools for machine translation, information extraction, semantic search, and other applications. In this paper we present a service-based approach to automate the creation and use of multilingual terminology resources, which is a very time-consuming process if performed manually.

The creation and maintenance of terminology resources is usually organised in one of two settings. The objective of descriptive terminology work is to document terms used to designate concepts of a given discipline. It usually involves manual or semi-automated analysis of documents to identify candidate terms, which are then checked in existing terminology resources for corresponding entries to create a terminology glossary. This is a typical practise in preparing terminology for translation and writing (Wright and Budin, 2001). Prescriptive terminology work is practiced by standardisers, government regulators, and nomenclature specialists to ensure consistent and unambiguous use of terminology in regulated areas as well as to facilitate precise
communication in general usage. Standardisers may also perform descriptive work as they collect data on usage prior to agreeing on standardised terms.

The traditional model for the creation of a bi-(multi-)lingual terminology resource in the descriptive scenario involves (1) collection of domain specific documents, (2) term identification and preparation of a glossary, (3) lookup for matching terminology data in prescriptive sources, (4) creation of new terminology data for glossary entries that have not been found in other sources. This work is usually carried out by an individual expert or a group, involves a great deal of manual work, and is rarely shared to the community.

In the last decade, significant progress has been achieved in the automation of these terminology work steps. Web crawling tools have been successfully adapted and applied to collect corpora for terminology needs. Baroni and Bernardini (2004) use a list of seed terms and bootstrapping approach in Web crawling, Blancafort et al. (2010) and Pinnis et al. (2012a) collect comparable corpora for term extraction. Methods for term extraction range from language independent n-gram extraction using relative frequency based termhood estimation (Delač et al., 2009; Pantel and Lin, 2001) to linguistically motivated methods based on syntactic analysis and the application of term phrase patterns (Bourigault, 1992). The combination of statistical and linguistically motivated techniques (Justeson and Katz, 1995; Dagan and Church, 1994) is the most used approach in the practical tools. The automation of term lookup is hindered by fragmentation of terminology resources in numerous databases with differing data structure and coverage. In recent years, several efforts have been invested to consolidate and harmonise terminology resources in national and international online term banks (e.g., Rikstermbanken, IATE and related lookup tool Quest, EuroTermBank and its term lookup API). But these term banks mostly incorporate terminology resulting from prescriptive work and are still limited in coverage.

Due to laborious manual work and the incompleteness of terminology data in prescriptive sources, it is still very time consuming to find and prepare the terminology data needed in practical translation work. Several surveys show that technical translators spend more than 30% of translation time on terminology work (Massion, 2007; Gornostay, 2010). Creators of terminology resources and operators of term banks struggle to cope with the need to incorporate an increasing number of new terms resulting from the rapid developments in technological, scientific and social spheres.

Although several tools are available to automate individual steps of terminology work, there is no solution that covers all major tasks for terminology creation. As a result, the major deficiencies of terminology resources are high cost and time needed to create them, insufficient coverage of terminology, particularly for the most recent concepts, poor language coverage, insufficient sharing of terminology resources, and a lack of collaborative mechanisms for involving terminology practitioners.

2 Overview of the Service Model

To facilitate the creation and usage of terminology resources and to benefit from the recent advances in computational linguistics, we propose a cloud-based service model that automates the major steps in terminology work. The automation of individual
tasks in terminology work is rendered as a set of interoperable cloud-based services integrated into workflows. These services automate identification of term candidates in user-provided monolingual documents and the lookup of translation equivalents for extracted monolingual term candidates. Translation equivalents are retrieved from online terminology term banks, automatically extracted multilingual terminology from comparable and parallel resources on the Web (in online and cached scenarios), as well as from terminology collections created by the platform’s users.

An essential element of this model is the collaborative involvement of users in the refinement and enrichment of raw terminological data, automated sharing and synchronisation of the terminology in various use cases by language workers and language processing applications (e.g., computer-assisted translation tools, machine translation systems, terminology management and terminology lookup platforms etc.).

The model is based on a reciprocity principle. Users process their documents and refine and enrich resulting terminological data, which can be shared and provided to other users and contributed to term banks.

This model is being successfully piloted in the TaaS platform (Pinnis et al., 2013) serving all 24 official EU languages and providing the following services:

- Automatic extraction of monolingual term candidates from user-uploaded source documents using terminology extraction techniques;
- Automatic retrieval of target translation equivalents for the extracted monolingual term candidates from various existing public and industry terminology resources;
- Acquisition of translation candidates from parallel or comparable Web data for terms not found in existing terminology resources using terminology extraction and bilingual terminology alignment methods;
- Facilities for refinement and enrichment of the resulting automatically extracted terminological data by the platform’s users;
- Data sharing and integration via API and export tools for sharing terminological data with major existing terminology resources and reuse it in various applications;
- Instant access to term translation equivalents and translation candidates for professional translators via CAT tools;
- Domain adaptation of statistical machine translation systems by integration with provided terminological data.

3 Service for Term Candidate Identification

Terms in user-provided monolingual documents are identified using the term tagging system TWSC (Pinnis et al., 2012b), which identifies term candidates in three steps:

1. At first, documents are pre-processed using part-of-speech or morpho-syntactic taggers (and optionally also lemmatisers if such exist for a language).
2. Then term candidates are extracted using linguistic filtering and statistical ranking methods. The filtering is performed with morpho-syntactic term phrase regular expressions and the ranking is performed with co-occurrence measures (e.g., log likelihood, modified mutual information etc.) for terms of two or more tokens and the TF*IDF (Spärck Jones, 1972) measure for unigram terms.
3. Finally, identified and extracted term candidates are marked in the user-provided documents using n-gram prioritisation and the term rankings.

The system has been extended to the languages supported by the platform by integrating existing part-of-speech taggers (e.g., the OpenNLP \(^1\) models for Dutch, English, French, German, Italian, and Spanish, the system by Pinnis and Goba (2011) for Estonian, Latvian, and Lithuanian, and HunPOS\(^2\) for Hungarian and Portuguese), building projected part-of-speech taggers for under-resourced languages using parallel corpora (Aker et al., 2014), and generating term phrase patterns from parallel corpora following a similar approach to the part-of-speech tagger projection.

We have evaluated the quality of the system for four languages in two subject fields (information technology and mechanical engineering). Two annotators (language specialists with a focus on terminology) identified terms in two documents. The documents across all languages were on similar topics and of similar difficulty levels. Each of the annotators has a subjective view on what comprises a term in a given context and what does not. This is because termhood and unithood of terms can be very ambiguous as well as subjective to the opinions of specialists who work with the terminology. Therefore, in our evaluation we use a union of their annotations and performed a precision analysis of the documents tagged by the system (see Table 1).

<table>
<thead>
<tr>
<th>Language</th>
<th>Information Technology</th>
<th>Mechanical Engineering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>Total</td>
<td>Precision</td>
</tr>
<tr>
<td>English</td>
<td>213</td>
<td>365</td>
</tr>
<tr>
<td>German</td>
<td>198</td>
<td>338</td>
</tr>
<tr>
<td>Hungarian</td>
<td>147</td>
<td>605</td>
</tr>
<tr>
<td>Latvian</td>
<td>316</td>
<td>540</td>
</tr>
</tbody>
</table>

The results show that on average around 50\% of the identified terms are true positives. Although seemingly average, the results are acceptable considering that termhood and unithood simultaneous identification is very challenging. This difficulty is supported also by comparing the annotator outputs. The average agreement rate of the two Latvian annotators was only at 63.3\%. Also the remaining term candidates are not necessarily wrong. Because of the linguistically motivated term phrase filtering, the system produces syntactically justified term candidates, which can still be useful in some application scenarios, e.g., machine translation (Pinnis et al., 2012c).

For users who work on morphologically rich languages, term identification may produce very redundant term candidate lists. This can be due to the inflective nature of many languages. For example, in Czech, Latvian, Estonian, etc., nouns, verbs, adjectives (and other parts of speech) may have numerous different inflected surface forms. Terms are also affected by this inflective nature and, therefore, the platform addresses this issue with term normalisation. Term normalisation is a process of trans-

---

\(^1\) http://opennlp.sourceforge.net/models-1.5/  
\(^2\) http://code.google.com/p/hunpos/
forming terms from their surface forms into their corresponding canonical forms as they are found in dictionaries and term banks. We use rule-based methods for term normalisation that for each of the term phrase regular expressions define a rule for term normalisation. For single-word terms the normalised forms often correspond to the term lemmas, however, for multi-word terms the normalised forms in many cases differ from the corresponding token lemma sequences. For example, the Latvian term “datoru tīklu” (transl. “computer network”) is normalised as “datoru tīklis”, however, the lemma sequence is different – “dators” “tīklis”. Using a rule-based approach we can remove redundancy in the monolingual term lists.

4 Service for Translation Equivalent Retrieval from the Web

One of the main sources for translations is the Web. It contains a vast amount of multilingual information that can be used to acquire up-to-date knowledge. Using novel workflows for the collection of Web corpora in automatic and on-demand scenarios, extraction of multilingual terminology from the corpora, and integration of the acquired terminology into the platform, we can provide users with up-to-date term candidate translation equivalents. All the multilingual terminology that is acquired from the Web is stored in a Statistical Data Base (SDB), which is accessible for translation candidate lookup when users create their bilingual terminology collections. There are four distinct workflows for bilingual terminology extraction from the Web:

1. On-demand bilingual terminology extraction from parallel data. For terms, for which existing resources do not yield any translation equivalents, users can manually trigger a bilingual terminology collection task that collects parallel corpora from the Web by identifying Web sites simultaneously containing the unknown terms as well as parallel content, and extracts bilingual terminology from the parallel corpora using bilingual phrase alignment techniques.

2. On-demand bilingual terminology extraction from focused comparable corpora. As parallel corpora may be scarce and not in all subject fields identifiable on the Web, an alternative path is to search for comparable corpora, which is much wider available (e.g., news, encyclopaedias, press releases etc.).

3. Automated bilingual terminology extraction from comparable RSS news corpora.

4. Automated bilingual terminology extraction from Wikipedia.

The three comparable corpora processing workflows are depicted in Fig. 1 and we further describe the comparable corpora processing workflows in more details.

4.1 Corpora Collection

Comparable corpora are collected using the following three different methods:

- On-demand comparable corpora in a specific subject field are collected using the Focussed Monolingual Crawler – FMC (Mastropavlos and Papavassiliou, 2011).
The corpus collection can be triggered by the user after the completion of a bilingual terminology extraction task and it can be performed using two scenarios:

— If the user requires translation equivalents for unknown terms only, a monolingual target language corpus is collected with FMC using seed URLs provided by the user. The bilingual terminology alignment tools then identify translation candidates of the unknown source terms in the collected monolingual corpus.

— If the user requires bilingual terminology of a specific subject field, two monolingual corpora are collected using seed URLs and optional seed terms for both languages. Then the corpora are cross-lingually aligned at the document level to acquire document-aligned comparable corpora. We use the DictMetric tool (Su and Babych, 2012) to estimate the comparability between two lists of monolingual documents. All documents are translated into English and a vector-based similarity metric is applied to calculate the document pair comparability.

• Iterative (automatic) comparable corpora are collected from RSS feeds using the tool proposed by Aker et al. (2012). The motivation behind this approach is that focused subject field oriented data is produced on a daily basis in the news articles from different sources (e.g., medical news, IT news etc.). Such news are often published in a full or adapted form in multiple languages, which makes them a valuable source of comparable corpora. The platform utilises such news sources in an automatic scenario in order to acquire up-to-date terminology once per week. In this scenario, DictMetric is also used to acquire bilingual comparable corpora.

• Comparable corpora are acquired from Wikipedia using the Wikipedia Retrieval tool (Paramita et al., 2012). Due to Wikipedia’s multilingual nature where articles
are linked between different languages using inter-language links, Wikipedia offers access to the largest comparable corpus found on the Web. The platform iteratively downloads Wikipedia dump files and extracts comparable corpora with the Wikipedia Retrieval tool for all the languages supported by the platform.

4.2 Bilingual Term Extraction Workflows

After corpora collection, we use the term tagging system in order to identify term candidates. Then, for languages with term normalisation support, terms are normalised. Finally, the bilingual terminology is extracted using cross-lingual term mapping:

- Context independent term mapping proposed by (Pinnis, 2013) using maximised character alignment maps. The mapper maps terms in two steps: 1) at first, possible translation and transliteration equivalents of monolingual terms (in the respective other language) are identified using probabilistic dictionaries, and rule-based and statistical transliteration techniques; 2) then the translation and transliteration equivalents are analysed in order to identify the maximum character level content overlap between source and target terms. The method has been evaluated automatically for 22 language pairs and has shown to achieve a precision from 72.3% up to 91.1% (with an average of 84.8%) with recalls ranging from 33.7% to 71.5% depending on the source language when paired with English. Manual evaluation of the mapper on an English-Latvian term-tagged comparable corpus collected with FMC in the field of medicine has shown to achieve a precision of up to 91.3%.

- Supervised term mapping using the method proposed by Aker et al. (2013). Similarly to the first method, this method operates in two steps, however both methods differ significantly. The supervised method requires language-specific models, which are trained on term pairs and is, therefore, limited to language pairs for which such models exist. At first, the mapper analyses whole term pairs and tries to identify dictionary based and cognate-based features. Then for all term pairs of two documents, a binary classifier is used to estimate whether two term pairs can be considered valid translation equivalents or not. Although the authors report automatic evaluation results of up to 100% (the automatic evaluation scenarios between the mappers are not comparable), manual evaluation shows that the precision for true translation equivalents for English-German ranges from 63% up to 82.

Both mappers produce output data where each term is described by its surface form, sequence of lemmas, sequence of part-of-speech (or morpho-syntactic) tags, normalised form (if normalisation is available), sequence of the normalised form’s part-of-speech (or morpho-syntactic) tags, and a concordance (up to 5 words around the term) that is extracted from the input document. An example is given in Fig. 2.

Fig. 2. Example of a mapped term pair ("→" denotes a tabulation character)
4.3 Delivery of Raw Terminological Data to Users

After cross-lingual term mapping, bilingual term pairs are integrated into the SDB by simultaneously performing term pair morphological consolidation. Depending on the linguistic tool support for languages, term consolidation is performed in three levels:

- For languages, for which lemmatisation of words is not integrated in the term tagging system, terms are grouped together only by their surface forms and part of speech sequences. This level ensures that SDB can support term translation candidate lookup even if linguistic support for certain languages is scarce.
- For languages with lemmatisation support, terms are grouped by their lemmatised forms and part of speech sequences. This consolidation level ensures that for morphologically rich languages redundancy, which is caused by having numerous surface forms of a single word, can be eliminated. However, this method can also group together surface forms belonging to different terms. For example, the term candidates “personāļais dators” and “personāls dators” from Fig. 2 both have identical lemma sequences. This issue is solved by the third level.
- For languages with term normalisation support, different surface form terms are grouped by their normalised forms and the normalised form part of speech sequences. This level provides the highest level of morphological consolidation.

The consolidation levels are used in order to provide the most appropriate term translation equivalents for a term lookup query. If no translation equivalents are identified in the higher consolidation levels, the data from the lower levels is used.

5 Service for Collaborative Refinement of Raw Terminology

The platform provides facilities for collaborative refinement of raw term pairs that are noisy and need validation. Term validation can be regarded as a three-step procedure: 1) monolingual validation (deletion of “unwanted” or unreliable term candidates,
definition of termhood, term variant identification, deduplication, etc.), bilingual validation (checking whether translation candidates are reciprocal translations, defining the right translations, etc.), and 3) validation in context.

The platform also provides a service for sharing terminology among users. Sharing that typically involves an interchange of non-confidential, non-competing, and non-differentiating terminology is highly rated by users. A recent survey (Gornostay et al., 2013) showed that up to 60% of users would share their data with the community.

6 Service for Terminology Sharing and Application

Approved and enriched terms can be exported and then used in other working environments with the help of the TBX3, and comma or tab-separated value formats. Approved terms can also be used in other terminology projects by the user(s) who owns the data as well as other users, provided that the term collection is shared.

Terminology resources are important not only for language workers but also for various language processing applications (“machine users”) such as computer-assisted translation tools and machine translation (MT) systems. The platform provides an API for external systems to access the terminology services and terminology data. This API-level integration is currently implemented by the memoQ CAT tool and the LetsMT statistical MT system (Vasiljevs et al., 2013). The objective of this work is to create project specific terminology resources for dynamic adaptation of MT systems.

7 Conclusion

We have presented a service-based model and its implementation in a novel platform for translators, terminologists, and language workers that streamlines the work on multilingual terminology generation, management, sharing, and use in various working environments. We described the services and workflows provided by the platform and presented evaluation results for the separate platform components. Although term identification is very challenging even to human annotators, we can achieve comparable precision with automatic methods using the extended term tagging system. For example, for Latvian an average precision of 53.8% was achieved in comparison to an average annotator agreement rate of 63.3%. The work within the TaaS project has received funding from the European Union under grant agreement n° 296312.

8 References


Theme Classification of Arabic Text: A Statistical Approach

Leila Fodil, Halim Sayoud, Siham Ouamour

USTHB University
fodilleila@gmail.com, halim.sayoud@uni.de, siham.ouamour@uni.de,

Abstract. The huge amount of textual documents that is stored in a lot of domains continues to increase at high speed; there is a need to organize it in the right manner so that a user can access it very easily. Text-Mining tools help to process this growing big data and to reveal the important information embedded in those documents. However, the field of information retrieval in the Arabic language is relatively new and limited compared to the quantity of research works that have been done in other languages (eg. English, Greek, German, Chinese …).

In this paper, we propose two statistical approaches of text classification by theme, which are dedicated to the Arabic language.

The tests of evaluation are conducted on an Arabic textual corpus containing 5 different themes: Economics, Politics, Sport, Medicine and Religion.

This investigation has validated several text mining tools for the Arabic language and has shown that the two proposed approaches are interesting in Arabic theme classification (classification performance reaching the score of 95%).

Keywords: Arabic Language, Text classification, Theme classification, Term weighting, TF-IDF.

1. Introduction

It is known that the volume of information available in Arabic on the World Wide Web and databases is continuously increasing. However, it has become almost impossible to take the full advantage of information embedded inside these documents without the help of various data mining techniques and tools.

Data mining is the process of extracting previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decision (Gharib 2009). Data mining can be applied on a variety of data types (structured data, multimedia data, free text, and hypertext).

Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output [1]. Text mining is well motivated, due to the fact that much of the world’s data can be found in text form (newspaper articles, emails, literature, web pages, etc.). Text mining tasks include text classification, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling.
Text Classification or Categorization (TC) is the task of automatically structuring a set of text documents into different categories [2], from a predefined set, according to a group structure that is known in advance. Text classification techniques are used in many applications, including e-mail filtering, mail routing, spam filtering, news monitoring, sorting through digitized paper archives, indexing of scientific articles, classifying news by subject or newsgroup, and searching for interesting information on the web.

There are two main approaches for text classification: the manual approach, and the automatic approach. In the manual approach, human experts classify the document manually or use classifiers [3]. Although it gives quite accurate results, it is still very difficult to continuously update the information. Furthermore, it is expensive to maintain the classifier.

In the automatic approach, some well-known classification methods exist such as decision trees Support Vector Machines (SVMs), K Nearest Neighbor (KNN), Neural Networks (NN), Naïve Bayes (NB), Decision Trees (DT) Maximum Entropy (ME) N-Grams and Association Rules [3].

Many of the TC researches have been implemented and tested with the English language. In addition to the English language, there are many studies in other European languages such as French, German, Spanish and Asian languages such as Chinese and Japanese. However, there is little current research in the automatic classification of text documents in Arabic, due to the specific morphological and structural changes in the language: polysemy, irregular and inflected derived forms, various spelling of certain words, various writing of certain combination character, short (diacritics) and long vowels, most of the Arabic words contain affixes [4].

In this paper, we propose a method of Arabic text classification in terms of theme using a statistical approach and want to compare the impact of text preprocessing, which has not been addressed before, on the classification performances. Our approach is based on the extraction of different keywords by two different methods and weighting schemes. In the first methods we use a different dataset for extracting automatically the keywords (pertinent terms), but in the second method, we use an in-house dictionary of keywords from each domain.

This paper is organized as follows: Section 2 presents an overview on related works, section 3 presents the proposed methodology, section 4 shows the experiments and results and finally a conclusion is given in section 5.

2. Related works

Different studies address the problem of text classification using different techniques to classify text documents, and different metrics to evaluate the accuracies of these techniques. We present in this section a number of studies and experiments of text classification conducted on Arabic language using different datasets, different features and different classification algorithms.

For instance, Sawaf et al. [5] used a statistical approach based on the maximum entropy technique to classify the Arabic NEWSWIRE corpus of the Linguistic Data Consortium (LDC) of the University of Pennsylvania which covers four classes: politics, economics, culture and sports. The main objective was to simplify Arabic classification difficulties by
avoiding morphological analysis and to use subword units (character n-grams). The best accuracy reported was 62.7%.

Also, in the same research field, El Kourdi et al. [6] used Naïve Bayes algorithm to classify 300 web documents into five classes (health, business, culture, science and sport). The average accuracy achieved was 68.78% and the best accuracy reported was about 92.8%.

On the other hand, Syiam et. al. [7], in their experimental results, show that the suggested hybrid method of statistics and light stemmers is the most suitable stemming algorithm for Arabic language and gives a general accuracy of about 98%.

Mesleh, A.A. [8] have applied a support vector machine (SVM) to classify Arabic articles with Chi-Square feature selection. The corpus contains 1445 documents that vary in length. They are collected from online Arabic newspaper archives and divided into nine categories: Computers, Economics, Education, Engineering, Law, Medicine, Politics, Religion, and Sport. They were focused on Feature Subset Selection (FSS) methods for TC tasks and, in particular, to the usage of an ant colony optimization algorithm to optimize the process of FSS. The author have compared Ant colony Optimization Based-Feature Subset Selection Method (ACO Based-FSS algorithm) with six state-of-the-art feature subset selection methods: Chi-square, Information Gain, Mutual Information, NGL coefficient, CSS score and Odds Ration in the classification of Arabic articles. Compared to the six classical FFS methods, the ACO Based-FSS algorithm achieved better text classification and was adapted to handle the large number of features in theme classification tasks.

In another attempt, Harrag and El-Qawasmah [3] applied neural networks (NN) on Arabic text. Their experimental results show that using NN with Singular Value Decomposition (SVD) as a feature selection technique gives better result (88.3%) than the basic NN (without SVD) (85.7%). They also experienced scalability problem with high dimensional text dataset using NN. Harrag collected his corpus from Hadith encyclopedia (موسوعة الحديثات) from the nine books. It contains 435 documents belonging to 14 categories. He applied light stemming and stopwords removal on his corpus.

In 2010, Hammouda et al. [9] have proposed a system of indexing and classification of Arabic texts enabling the text classification using a heuristic approach. The Arabic dataset is collected from Internet, newspapers and magazines, the collected documents were 25 documents, to form manually a corpus composed of around 12500 words, 5 documents for each of the following 5 categories: sports, medicine, politics, economics and agriculture. The system performance was evaluated on the different chosen domains, achieving 90% of precision and 85% of recall.

In this investigation, we propose two theme classification methods: a Semi-Automatic method (which we called: SACM) and an Automatic method (which we called: ACM) that are tested on an Arabic corpus containing five different themes.

3. Proposed Methods

This section presents our overall methodology for Arabic text classification. Hence, in our approach, we have proposed two methods: a Semi-Automatic Categorization Method (SACM) where the set of keywords (pertinent terms) is constructed manually for each
domain and an Automatic Categorization Method (ACM) where the set of keywords is extracted automatically.

In our TC methods, the text document pass through five main steps: data processing, stemming, feature selection, classification and evaluation. Figure 1 shows the different steps of this approach.

![Diagram of text classification approach]

Figure 1: The different steps of our text classification approach.

### 3.1 Text Preprocessing

Preprocessing is actually a way to improve text classification by removing worthless information.

After converting our text corpus into UTF-8 encoding, in the first step, it is necessary to clean the texts by removing punctuation marks, diacritics, numbers, non Arabic letters [6], and removing kashida except in the term Allah.

In addition, Arabic texts need the following processes:

- Normalizing some writing forms that include “ا” “ی”, “ی” to “ی”, “ء” and “ه”. The reason for this normalization is that all forms of hamza are represented in dictionaries as one form and people often misspell different forms of aleph.

- A Tokenizer is employed. It is responsible for scanning each document, word by word, and extracting words (token) in this document. In this step, the Arabic Tokenizer uses white space tokenization because the space is the only way to separate words in Arabic language. The documents were tokenized to produce bag of words (or vector of word) for each text to classify.

- Removing Arabic function words (stop words) (such as “فی”, “علی”, “لكن”, “فی”, “فیه” etc.). Stop words are common terms that provide only a little of meaning and serve only as syntactic function without indicating any important subject or matter. The remove of stop words changes the document length and reduce the memory of the process.
3.2 Stemming

The main goal of the stemming is to improve the efficiency of the classification by reducing the number of terms being introduced to the classifier [10]. This step aims to make all the terms, which have the same root, in a unique form (stem) to facilitate the treatment. We can take as an example the terms: 'I ...5DPW)'.

During the stemming operation, all those will be transformed into the term 'اقتصاد' to calculate the frequency with an efficient way.

![Example of stemming](image)

During this step, we compare every word (or token) of the text to classify with the list of predefined stems. If there is a correspondence, then the token will be replaced by the stem automatically.

3.3 Feature selection

3.3.1 Term weighting

After selecting the most significant terms in the super vector, each document is represented as a weighted vector of the terms found in this vector [11]. Term weighting is one of the important and basic steps in text classification based on the statistical analysis approach. In our experiments, we attempt to use two weighting schemes to reflect the relative importance of each term in a document (and a category) and to reduce the dimensionality of the feature space. In the first time, we have calculated the relative frequency TF of a term (denoted by t) in a text document (denoted by d) by the following arithmetical formula:

\[
TF(t, d) = \frac{\text{number of occurrence of } "t"}{\text{number of word in } "d"} \tag{1}
\]

In the second approach, we have used the TF-IDF: the Term Frequency (TF) of a word t is determined by the frequency of the word t in the document d. The Document frequency (DF) of a word t is the number of documents in the dataset where the word t occurs in at least once. The inverse document frequency (IDF) of the word t is generally calculated as follows:

\[
IDF = \log \left( \frac{N}{n} \right) \tag{2}
\]

where \( N \) is the total number of documents in the dataset and \( n \) is the number of documents that contain the concerned word.

The weight of word t in document d using TF-IDF is:

\[
\text{TF-IDF} = \text{TF}(t, d) \times \log \left( \frac{N}{n} \right) \tag{3}
\]
Thus, a term that has a high TF-IDF value must be simultaneously important in this document and must appear few times in the other documents. It is often the case where a term correlates to an important and unique characteristic of a document.

### 3.3.2 Feature Extraction

In this step, we propose two methods: in the first one called, which we called SACM (Semi-Automatic Classification Method), we use the local (in-house) dictionary; and in the second one, which we called ACM (Automatic Classification Method), we automatically extract this dictionary by a preliminary training from a specific dataset called ADTC1 dataset.

In the SACM method, we have constructed a dictionary for each theme and we have collected the specific keywords for each domain (see table 1).

<table>
<thead>
<tr>
<th>Theme of the dictionary</th>
<th>Number of words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Economics</td>
<td>35</td>
</tr>
<tr>
<td>Politics</td>
<td>41</td>
</tr>
<tr>
<td>Sport</td>
<td>31</td>
</tr>
<tr>
<td>Medicine</td>
<td>31</td>
</tr>
<tr>
<td>Religion</td>
<td>56</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>194</strong></td>
</tr>
</tbody>
</table>

In the other method (ACM), this dictionary has been extracted automatically. We have used a corpus composed of 25 texts (ADTC1), which undergoes the same steps described previously (processing, stemming and feature selection).

After calculating the frequency (or TF-IDF) of all the terms and in every domain, we extract the most pertinent terms. We have kept only the terms for which the value of the frequency exceeds a certain threshold (determined by experiment). Finally, the pertinent terms for every domain will be grouped in different subsets.

### 3.4 Classification

The main goal is the assignment of an Arabic text to one or more predefined categories (themes) based on their content.

For that purpose, the Apparition Probability AP of a keyword \( k \) in each text \( d \) will be computed by:

\[
AP(k) = \frac{\text{number of occurrence of keyword } "k"}{\text{number of words in } "d"} \quad (3)
\]

Once the probability of apparition is computed, we will be able to deduce the class of that text. Hence, the sum of apparition probabilities with regards to each domain "x" is determined.

This sum is called the Cumulative Thematic Probability or CTP.

\[
CTP(x) = \sum_k AP(k), \text{ where the keyword } "k" \in \text{ domain } x \quad (4)
\]
According to the CTP probability, the system will classify the text according to the appropriate domain: in other words, the theme having the highest CTP should correspond to the real theme of the unknown text.

4. Experiments and results

4.1 Dataset
The corpus used in this paper is collected from three sources: news (Al-Jazeera, BBC, CNN, France24, Al-Ahram...), newspapers (Al-Ahram, Al-watan, Elmoudjahid, Elkhabar Sport-Al-fagr...), and books (رياض الصالحين، الإنسان الوراثي، الطاعية...). It consists of five categories, which are sport, politics, economics, medicine and religion.

Two datasets have been built: ADTC1 (Arabic Dataset for Theme Classification, subset 1) and ADTC2 (Arabic Dataset for Theme Classification, subset 2). ADTC1 is used in order to get automatically the keywords. There are 25 texts (5 texts by theme) from Internet and newspapers, corresponding to 38605 words. ADTC2 is used for the testing. There are 150 texts (30 texts by theme) from three types of sources: internet, newspapers and books (50 texts by source), corresponding to 46060 words. Table 2 shows the number of documents present in the different categories.

<table>
<thead>
<tr>
<th>Categories</th>
<th>Number of texts</th>
<th>Number of words</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ADTC1</td>
<td>ADTC2</td>
</tr>
<tr>
<td>Economics</td>
<td>5</td>
<td>30</td>
</tr>
<tr>
<td>Sport</td>
<td>5</td>
<td>30</td>
</tr>
<tr>
<td>Medicine</td>
<td>5</td>
<td>30</td>
</tr>
<tr>
<td>Religion</td>
<td>5</td>
<td>30</td>
</tr>
<tr>
<td>Politics</td>
<td>5</td>
<td>30</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>150</td>
</tr>
</tbody>
</table>

4.2 Evaluation Criterion

In order to evaluate our methods, a recognition score R is calculated for each category and each method. This recognition score measures the percentage of documents that are correctly assigned in each category.

4.3 Results

We recall that the primary steps in our Arabic text classification system is preprocessing the document, tokenization and stemming. Thereafter, the selection of the keywords (features) is performed. Table 3 shows the number of selected features in the automatic method.
Table 3: Number of pertinent features that are selected

<table>
<thead>
<tr>
<th>Categories</th>
<th>Automatic method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TF.IDF frequency</td>
</tr>
<tr>
<td>Economics</td>
<td>34</td>
</tr>
<tr>
<td>Politics</td>
<td>45</td>
</tr>
<tr>
<td>Medicine</td>
<td>32</td>
</tr>
<tr>
<td>Sport</td>
<td>30</td>
</tr>
<tr>
<td>Religion</td>
<td>38</td>
</tr>
<tr>
<td>Total</td>
<td>179</td>
</tr>
</tbody>
</table>

Table 4 shows the different results obtained by the Automatic and Semi-Automatic methods and by using the two pertinence measures (TF-IDF and relative frequency). According to table 4, the Semi Automatic Method using TF-IDF presents the best performances with a score of about 95%, followed by the Automatic Method using the TF-IDF with a score of 92%. The other methods using the relative frequency come in last positions with a score not exceeding 88% of good classification.

Table 4: Scores of good classification by theme.

<table>
<thead>
<tr>
<th>Categories</th>
<th>Semi Automatic method</th>
<th>Automatic method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TF.IDF</td>
<td>Freq</td>
</tr>
<tr>
<td>Economics</td>
<td>93.33</td>
<td>83.33</td>
</tr>
<tr>
<td>Politics</td>
<td>96.66</td>
<td>93.33</td>
</tr>
<tr>
<td>Medicine</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Sport</td>
<td>93.33</td>
<td>80</td>
</tr>
<tr>
<td>Religion</td>
<td>90</td>
<td>83.33</td>
</tr>
<tr>
<td>Global score</td>
<td>94.67</td>
<td>88</td>
</tr>
</tbody>
</table>

Figure 3 is a graphical representation of the precedent table (table 4). We can see that, generally, the category medicine gets the highest score (100%); while the Religion category gets the lowest one (83.33%), in our case.
In a second experiment, we have computed the performances of each method (SACM and ACM) by varying the number of keywords (number of the most frequent consistent words that are kept for characterizing a theme). The numbers of these keywords are taken as follows: 10, 20, 30, 40, 50…, until 160 (see figure 4).

![Figure 4: Score of good classification vs number of keywords selected.](image)

Figure 4 shows that the semi automatic method using TF-IDF presents the best performances. We also notice that, when the number of features (keywords) increases, the classification accuracy increases too. An important point to note is that with 100 keywords (and more), the different methods begin to be powerful and accurate enough.

5. Conclusion

In this investigation, we have presented two approaches of Arabic text classification by theme. Five categories of themes were proposed and a special corpus has been built for that purpose.

During the text classification process, the document is coded into a vector of words (bag of words); this fact leads to a huge feature space and semantic loss. The proposed model in this paper adopts the keywords or “pertinent features” selected according to two approaches. These approaches extract these features statistically from the text and then the required theme is deduced from these selected features.

The comparison between the two proposed methods shows that the Semi-Automatic Method using TF-IDF achieves the best classification (score of about 95%), followed by the Automatic Methods using TF-IDF (score of 92%). On the other hand, the amount of time taken to build the dictionary of keywords is relatively greater for the semi automatic method, which makes the automatic method more interesting with regards to the time of execution.

These results show that there are some differences between these approaches according to two aspects, the classification score and execution time (to build the models).
Consequently, the user should decide which approach is suitable for him, according to his needs and constraints, before choosing the method to employ for the task of text categorization. In perspective, we plan to expand the number of themes and increase the size of our Arabic textual dataset.

Reference


Information Categories in a Learner’s Glossary Entry

Boyan Alexiev

University of Architecture, Civil Engineering and Geodesy - Sofia

Abstract. Three types of information categories are proposed in a learner’s glossary entry: conceptual, lexical and pragmatic. The glossary is envisaged as a knowledge-oriented terminological collection which provides quick access to the conceptual structure of a narrow domain, the head terms functioning in it, their definitions, collocations, translations and contextual use. The conceptual information is identified by analyzing the conceptual relations using a classification scheme of the general aspects of the basic concept (top term). The top-down procedure continues with establishing the system-structuring characteristics with possible graphic representation of types. The lexical information refers to terminological collocations identified by a lexico-semantic analysis. The pragmatic information is provided by corpus-extracted contexts and translation equivalents specified by a systematicity-based terminological contrastive analysis. A model is proposed for organizing the data obtained in a learner’s glossary entry. It is concluded that the proposed model allows maximum concentration of terminological knowledge applicable in technical translation.

Keywords: learner’s glossary; conceptual information categories; lexical information categories; pragmatic information categories; systematicity-based terminological contrastive analysis;

1 Introduction

Modern Terminology is an interdisciplinary field of study and practice closely related to the latest achievements in linguistics, information science and computing. Terminologists nowadays rely heavily on large machine-readable corpora and software tools which are capable of processing those corpora to extract terminological data used in terminology research and terminographic projects. The activity of gathering and ordering such data is getting more knowledge-oriented due to the increasing need of translators and other users of terminology services for knowledge-based mono- or multilingual information sources in a knowledge-based global economy. This tendency has resulted in developing large-scale terminology projects involving research teams of terminologists, ontologists and subject specialists for designing rich terminological knowledge bases. At the same time most universities in the world where languages for specific purposes are taught cannot afford to support the implementation of such projects. Moreover, when performing a specific translation assignment such as conference interpreting or translating a text in a narrow domain, technical translators actually need quick access to a small-scale bilingual terminographic source in printed or computerised form which can help them get some general knowledge of the subdomain in question and the special language used to express it in the respective target language.

The aim of this paper is to present a methodology for compiling a learner’s glossary based on 30-40 narrow-domain head terms, which can be defined as a knowledge-based terminological collection providing quick access to the conceptual structure of that domain, the relationships between the terminological units in it as well as their combinatorial capacities and communicative use. The focus is on the microstructure of the proposed glossary realized through the information categories represented in the glossary entry which can be grouped in three major types: conceptual, lexical and pragmatic. A methodology is proposed for providing and organizing terminological data into those categories by making use of conceptual, lexico-semantic and systematicity-based contrastive analyses. A knowledge-oriented model of a learner’s glossary entry is constructed based on the analytical data obtained.

2 Conceptual Analysis for Identifying Terminological Relations

One basic problem modern terminologists have to solve is how to search for information about terms for various terminographic purposes. Terminological data involve linguistic information about the terms, conceptual information referring to conceptual relations between terms and pragmatic information concerning the use of terms in contexts. Of particular interest are the terminological relations that have always been a major concern of modern terminologists and terminographers. Such relations also find
application in information retrieval and knowledge representation. Here I need to differentiate between the terms ‘terminological relation/ship/ships’, ‘conceptual relation/ship/ships’ and ‘semantic relation/ship/ships’. While agreeing with L’Homme and Marshman [1] that most authors use them as synonyms, in view of the terminographic project I present and the two types of analyses (conceptual and lexico-semantic) I deem necessary for capturing the broad variety of relations holding between terms, I assume that the term ‘terminological relations’ is superordinate to the terms ‘conceptual relations’ and ‘lexico-semantic relations’, the latter two being co-ordinate terms.

After extracting the terminological data from a textual corpus, the final list of terms to enter a learner’s glossary can be specified by analysing the conceptual relations between the candidate terms using available reference materials and term definitions. These sources can also be used to identify some additional terms worth including in the glossary which are related hyponymically or meronymically to the head terms but for some reason do not occur among the automatically extracted terms. In this way a reliable set of narrower terms (types) can be provided as well as other terms expressing concepts that enter into partitive, functional, causal etc. relations to the key concepts.

The terms pertaining to a given subject field, subfield or even a topic within that subfield (cf. the concrete topic within the subfield of ‘building materials’ as part of the field of Civil Engineering) are characterized by both internal and external systematicity [2], the first type relating to the internal structuring of the terms in a terminological system and the second one to the communicative function of that system. A good terminographic project is necessarily based on a careful analysis of the internal (inherent) systematicity of the set of terms envisaged as entries in the respective terminological collection. Different terminologists propose different models for describing that systematicity. For example, the representatives of the traditional Vienna School of Terminology focus primarily on taxonomic and meronymic relationships [3] and are often criticized by proponents of alternative approaches for overlooking the multifaceted and multidimensional nature of terms whose relationships can also be described by using linguistic models, i.e. within lexico-semantic frameworks. However, I consider the conceptual analysis to be indispensable for structuring terminologies which can lay a solid foundation for identifying the proper entries for any type of terminological collection.

A conceptual analysis of terms for terminographic purposes should start with adopting a certain typology of conceptual relationships. This is not a very easy task since a large variety of typologies have already been proposed. For example, Felber [3] presents the following basic types of relationships (I do not present the subtypes):

1. Logical relationships
2. Ontological relationships
3. Relationships of effect

A similar typology is proposed by Cabré [4] but she distinguishes between only two main types of relationships, viz. logical and ontological. An interesting point in this typology is the further subdivision of the two types into subtypes according to logical criteria. The basis for a logical relationship between two concepts, for example, is the fact that they share one or more characteristics. When a concept has at least one more characteristic in addition to the characteristics of another concept, then the first concept is specific in relation to the second which, in turn, is generic in relation to the first one. In this case we have logical subordination. If two specific concepts are subordinated to the same generic concept, then we have logical coordination between two specific concepts. Coordination and subordination put together constitute the hierarchical structure of a subject field. On the other hand, ontological relations are not based on the similarity between concepts but on the proximity of objects to each other in the real world. These relations are further subdivided into coordination (whole-part) and chain (cause-effect) relationships.

Another interesting and consistent typology of systematic relations between terms is proposed by Popova [2]. She postulates two types of systematicity (scheme of relations) among terms: implicational and classificational. The former consists of two subtypes, viz. partitive/meronymic, i.e. whole-part relations and associative, i.e. relations of contiguity between entities participating in a real situation semantically represented as a predicative ‘scene’ (Fillmore’s frames) where referents perform semantic ‘roles’ (agent, object, result, purpose, etc.) assigned by the predicate. A similar actantial structure, but based on a different theory, will be used in the lexico-semantic analysis described in the next section. For the purposes of the glossary envisaged I adopt Sager’s classification of conceptual relations most frequently used in terminology involving generic, meronymic and complex relations [5]:

---

88/228
1. **Generic (hyperonymic and hyponymic) relationships** which establish a hierarchical order; a broader (generic) concept is superordinate to the narrower (specific) concept(s) and, conversely, the respective narrower concept is subordinate to the generic concept. It is important to note here that in certain cases it is necessary to indicate the criterion by which types have been declared. Such type indicators are known in information science as ‘facets’. For example, building materials can be classified by properties: ceramics, composites, plasticizers, etc.; or by function: abrasives, adhesives, coatings, insulating materials, etc.

2. **Meronymic/partitive relationships** also referred to as ‘whole-part’ relationships which indicate the connection between concepts consisting of more than one part and their constituent parts. For example, concrete is a fundamental ingredient in cement.

3. **Complex relationships** such as: cause-effect; material-product; material-property; material-state; process-product; process-instrument; process-method; process-patient; phenomenon-measurement; object-counteragent; object-container; object-material; object-quality; object-operation; object-characteristic; object-form; activity-place. For example, aggregate, cement and water are mixed (process) to produce concrete (product).

As a matter of fact every system of terminological units is structured around a top term designating a ‘seed concept’ from which all other terms in the system stem through complex branching of its characteristics in a certain hierarchical order. Hence, a conceptual analysis of the top term can be expected to yield the basic candidate terms to enter a glossary or any other terminological collection envisaged to cover that topic. Since concepts consist of characteristics, the analysis of the conceptual structure of a term should involve specification of these characteristics. The latter are extracted by applying a simplified procedure for identifying concept characteristics in terminological definitions consisting of three steps:

**Step 1:** Developing a classification scheme of the general aspects of the basic concept. For example, with the help of subject experts these aspects for the building material concrete were reduced to types, composition, properties, technology and use.

**Step 2:** Presenting general aspects as deep predications: Concrete is a type of X (genus predication); Concrete is characterized by TYPES, COMPOSITION, PROPERTIES, TECHNOLOGY and USE (species predications).

**Step 3:** Matching deep predications to the linguistic structure of definitions. In other words, the species characteristics are identified by the five aspects specified above and presented as generalized (from all available definitions) characteristics arranged in a hierarchical order.

In fact, the generalized characteristics represent what can be termed ‘system-structuring characteristics’, namely, genus and species characteristics. For example, the genus characteristic for concrete is composite building material. Examples of generalized species characteristics are:

1. **Types (hyponyms)**
   - aerated concrete
   - cast-in-place concrete
   - freshly mixed concrete
   - precast concrete
   - prestressed concrete
   - reinforced concrete

2. **Composition (meronyms):**
   - cement binder
   - aggregate
   - admixture
   - additive

The other types of generalized species characteristics are exemplified in the learner’s glossary entry model presented in section 5 below.
3  Lexico-Semantic Analysis for Term Collocation Identification

In section 2 above I tried to show how a conceptual analysis can be applied for identifying the candidate entry terms for a provisional English-Bulgarian Learner's Glossary of Concrete Terms and the narrower terms within these entries. It is a well-known fact that translators of technical texts very often encounter difficulties when translating not the terms themselves but the words they usually co-occur, i.e. their collocates. In fact, terminological collocations can justifiably be considered terminological knowledge items representing some kind of conceptual "scenes". In other words, terminological collocations could be interpreted as concept combinations, i.e. knowledge items which can be subjected to some categorization/classification (e.g. concrete: mixed, placed, compacted, finished, cured and protected – stages in concrete manufacturing). This is why I have decided to include term collocations as an information category in our knowledge-oriented glossary. An approach that is appropriate for capturing collocational information on the entries of the glossary in question is the lexico-semantic approach to terminology structuring whose theoretical and methodological premises I will present below.

The theoretical basis for the lexico-semantic approach to structuring terminological data for terminographic purposes is provided by the Explanatory and Combinatorial Lexicology (henceforth ECL) [6] which is the lexico-semantic component of the Meaning-Text Theory. This theory proposes a formalized model of natural language, a Meaning Text Model representing a system of rules which simulate the linguistic behaviour of humans. That model is designed to perform the transition from meanings in general (any information/content a speaker transmits by using natural language) to texts (physical manifestation of speech) and vice versa. The ECL, in turn, proposes an apparatus, namely, lexical functions (henceforth LFs) for capturing semantic relations between lexical units. LFs are a means for a systematic description of the so-called "institutionalized" lexical relations. Some simple examples of institutionalized lexical relations are those between attention and pay, wolves and pack, etc. From our concrete terminological microsystem we can provide the following examples: concrete and mix, concrete and set, concrete and harden, concrete and batch, etc. LFs are based on de Saussure's dichotomy of paradigmatic vs. syntagmatic relations. Paradigmatic relations can be defined as all contrast and substitution relations holding between lexical units in specific contexts. Syntagmatic relations are relations holding between lexical units that can co-occur, i.e. appear together in the same phrase or clause. Mel'čuk [6] explains that the term 'function' in the theory is used in its mathematical sense f(x) = y where f is the function, x is the argument and y is the value expressed by the function when applied to a given argument.

There is no doubt that this theoretical framework has had and will have important repercussions for a broad variety of lexicological endeavours. For the purposes of the particular project envisaged I am interested in the extent to which these theoretical assumptions can be used for analysing terminological data for terminographic purposes. A number of terminologists have already explored these possibilities and proposed various adaptations of the ECL to the specificity of terminological units. For example, L'Homme [7], comparing the two different approaches to terminology, viz. conceptual and the lexico-semantic, points out their advantages and shortcomings. She argues that truly conceptual approaches do not allow a flexible integration of terms and relationships between terms. In contrast, lexico-semantic approaches are more compatible with data gathered from corpora. For the lexico-semantic analysis of the computer term 'program' L'Homme applies lexical functions to formalize the following relationships 'program' enters in:

- synonym: Syn (program_i) = computer ~ ;
- agent of program: S_i (program_i) = programmer;
- create a program: CauseFunc_i (program_i) = create [DET ~], write [DET ~];
- cause a program to function: CauseFact_i (program_i) = execute [DET ~];
- the program stops functioning: FinFact_i (program_i) = [DET ~] ends, [DET ~] terminates;

In my opinion, this analytical procedure shows clearly two disadvantages of that approach. On the one hand, the LF notation is very complicated and will obviously have to be simplified in order to be conveniently applied to the analysis of terminological items. On the other hand, the specificity of the terminological system may require the postulation of new specific lexical functions that have not been considered in the ECL. For example, there is no LF and notation, respectively, for the so-called 'self-running natural processes' expressed by verbs such as 'set', 'harden', 'bleed' which collocate with our top term concrete (see below).
Therefore, for the lexico-semantic analysis of our corpus in view of extracting and consequently presenting useful collocations in the learner’s glossary entries, I will use a methodology which is generally based on Frame Semantics and uses semantic/actantial roles. L’Homme and Bae [8] propose a lexico-semantic analysis of the actantial structures of predicative terms (verbs). The procedure is exemplified by representing the term browse in a tabular form (the original examples are in French):

<table>
<thead>
<tr>
<th>AGENT</th>
<th>LOCATION</th>
<th>INSTRUMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>User</td>
<td>Internet</td>
<td>Browser</td>
</tr>
</tbody>
</table>

As can be seen in Table 1, the actantial structure gives the position of actants and explains them in terms of actantial roles.

I will follow a similar procedure to identify the verb collocations of the head terms cement and concrete, leaving aside adjectival (A+T) collocations that I have already identified by the conceptual analysis described in the previous section since most of these actually designate generic or partitive relations.

The special collocations with the head terms concrete and cement have been extracted from contexts provided by the term extractor TermoStat\(^1\). The specialised lexical combinations with these terms analysed below are selected because they have specialised meaning within the field of construction, e.g. the meaning of ‘cure’ (make a person or animal healthy again) is altered within the specific combination ‘concrete is cured’. Two types of activities can be captured by the methodology described above, namely, self-running natural processes during concrete manufacturing and actions performed on cement and concrete. The results of the analysis are presented in Tables 2 and 3.

### Table 2

<table>
<thead>
<tr>
<th>Natural self-running process</th>
<th>Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bleeds</td>
<td>Concrete</td>
</tr>
<tr>
<td>Cures</td>
<td></td>
</tr>
<tr>
<td>Sets</td>
<td></td>
</tr>
<tr>
<td>Hardens</td>
<td></td>
</tr>
</tbody>
</table>

### Table 3

<table>
<thead>
<tr>
<th>Agent</th>
<th>Action (on)</th>
<th>Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>Builder</td>
<td>places</td>
<td>Concrete</td>
</tr>
<tr>
<td></td>
<td>compacts</td>
<td></td>
</tr>
<tr>
<td></td>
<td>levels/</td>
<td></td>
</tr>
<tr>
<td></td>
<td>screeds</td>
<td></td>
</tr>
<tr>
<td></td>
<td>floats</td>
<td></td>
</tr>
<tr>
<td></td>
<td>trowels</td>
<td></td>
</tr>
<tr>
<td></td>
<td>cures</td>
<td></td>
</tr>
<tr>
<td></td>
<td>sprays</td>
<td></td>
</tr>
<tr>
<td>Builder</td>
<td>mixes</td>
<td>Cement, Water, Aggregate</td>
</tr>
</tbody>
</table>

The analytical results confirm L’Homme’s conclusion [9] that “semantic classes in a given syntactic position could be used to discover typical ‘frames’ thus implying the usefulness of resorting to Frame Semantics [10] when classifying specialized lexical units “in a way that enables us to make generalizations about them” [9]. In other words, a terminographer doing a research into a terminological system or subsystem with the view to identifying collocations is very likely to be forced by circumstances to ‘discover’ (definitely with the help of specialists) new actantial structures typical of the particular specialised discourse.

\(^1\) A tool for automatic acquisition of terms and their contextual use, designed by Patrick Drouin (OLST, UdeM) that exploits a method of opposition in specialized and non-specialized corpus for the identification of terms.
4 Systematicity-Based Terminological Contrastive Analysis

As already mentioned, the pragmatic data constitute a very important part of the overall learner’s glossary entry structure. They can be expressed by contexts for head terms whose selection does not require special analysis. What should be subjected to contrastive analysis are pairs of translation equivalents in cases of inappropriate ones according to both semantic and structural criteria. Since the solutions to such translation problems almost always require expert advice, I subsume the results of that kind of procedure under the more general notion ‘pragmatic data’. I propose a systematicity-based terminological contrastive analysis for term translation problems which is described below.

I assume that the terms and terminological collocations to be contrasted are translation pairs expressing the same concepts. Then the differences in the source language and target language should be sought in the particular language-specific choice of lexical items and structural patterns. In this sense I can propose the following two-step model for contrasting domain-specific terms and their target language equivalents, which consists of two levels of analysis, viz. the level of common conceptual structure and the level of interlingual asymmetry:

![Diagram](Fig. 1 Model for Terminological Contrastive Analysis)

The procedural steps I suggest for contrasting source language and target language terminological items in a special subdomain are as follows:

- **Step 1** Grouping the glossary items into conceptual groups by analysing their definitions and/or consulting an expert.
- **Step 2** Identifying the conceptual groups with their corresponding term sets containing translation-problem SL-TL term/term collocation pairs (in our particular case English-Bulgarian term/term collocation pairs).
- **Step 3** Determining the lexico-structural patterns of the identified SL and TL term sets.
- **Step 4** Comparing the linguistic systematicity of the term/term collocation sets in the source and target language.
- **Step 5** Proposing solutions to term/term collocation translation problems based on systematicity and pragmatic criteria.

The lexico-structural terminological contrastive analysis is performed individually for each translation-problem term/term collocation pair. The analysis is not purely structural but lexico-structural because I make use of the so-called semantic roles (frames) in the lexico-structural patterns in order to explicate the semantic relations between the lexemes in the terminological collocations. The analytical procedure is exemplified in the following case study.

**Case Study:** Float Concrete → Trowel Concrete

**Problem:** nonexistent Bulgarian equivalent

**Analysis:** An interesting example of what I would call cultural domain specificity (a combination of cultural specificity and domain specificity) is the case with two concrete terminological verb collocations which according to the existing specialised English-Bulgarian dictionaries are to be translated as absolute
synonyms. When putting together similar term collocations to create a conceptual group with the respective term collocation sets and translation equivalents (see Steps 1 and 2 above), with the help of the expert I arrived at the following sequence of stages of concrete manufacturing performed as actions by the builder:

1. Concrete is placed – бетонът се² полага/betonat se polaga
2. Concrete is compacted – бетонът се уплътнява/betonat se uplatnyava
3. Concrete is levelled/screeded – бетонът се подравнява/betonat se podravnyava
4. Concrete is floated – бетонът се заглажда (?)/betonat se zaglazhda?
5. Concrete is trowelled – бетонът се заглажда (?)/betonat se zaglazhda?
6. Concrete is cured – бетонът се съхранява/betonat se sahranyava

To check whether the English verbs in the concrete context have the same semantics, I applied a pragmatic approach to solving the problem by first searching the Internet for a context where both terms are encountered and found the following text:

Floating produces a relatively even, but slightly rough, texture that has good slip resistance and is frequently used as a final finish for exterior slabs. If a smooth, hard, dense surface is required, floating is followed by steel trowelling³.

From the context it becomes clear that if we subsume the two actions, floating and trowelling, under the generic action smoothing, then the distinction between the two should be sought in the manner of action which in floating could be defined as incomplete compared with the complete action in trowelling. Hence, the lexico-structural patterns of the English term collocations could be represented in the following way:

4. Noun (patient) + Verb (event: incomplete action)
5. Noun (patient) + Verb (event: complete action)

I reported the results back to the expert who advised me to add an adverb after each verb thus distinguishing between the two actions, placing them in a sequence rather than equating them (see solution below). This is a good example of how even in a very narrow domain the knowledge continuum can be segmented differently by different language cultures.

Proposed solution to problem:

4. Concrete is floated – бетонът се заглажда грубо (lit. concrete is smoothed roughly)
5. Concrete is trowelled – бетонът се заглажда фино (lit. concrete is smoothed finely)

The lexico-structural patterns of the proposed Bulgarian term collocations correspond semantically, if not structurally, to their English counterparts:

4. Noun (patient) + Verb (action) + Adverb (manner of action: incomplete)
5. Noun (patient) + Verb (action) + Adverb (manner of action: complete)

The case study discussed above seems to support Sager’s claim that terminological systematicity cannot be a fully reliable criterion for predicting term formation. However, the results I presented above prove that he is only partially right in stating the “limited usefulness” of “discovering regularities in term formation” [5]. As far as term translation strategies are concerned, that enterprise is definitely worth the effort.

---
² The particle ‘се’ here is used to denote a passive construction in Bulgarian.
³ My emphasis
5 A Learner’s Glossary Entry Model

It is generally assumed that the purpose of a dictionary or glossary will determine the entry layout designed to meet the needs of the specific type of potential users. Having that in mind, I propose the following model of a learner’s glossary entry:

CONCRETE  Target Language Equivalent/TLE

DEFINITION: A composite building material composed of coarse and fine aggregate (sand, gravel, crushed rock, etc.) held together by a hardened paste of hydraulic cement and water with added admixtures, which is characterised by durability, high compressive strength and compaction, low water/cement ratio and workability and is used in building foundations, structural walls, columns, slabs, etc.

E.g.: The composition of concrete is determined initially during mixing and finally during placing of fresh concrete. The type of structure being constructed as well as the method of construction determines how the concrete is placed and therefore also the composition of the concrete mix or mix design.

CONCRETE TYPES

By strength:
- Prestressed ~ TLE (pre-compressed using high-tensile wires)
- Post-tensioned ~ TLE (steel tendons tensioned after the concrete has been cast)

By presence/absence of reinforcement:
- Plain/ordinary ~ TLE
- Reinforced ~ TLE

By weight:
- Lightweight ~ TLE (density is less than normal concrete)
- Heavyweight ~ TLE

By location of casting:
- Precast ~ TLE (cast in a reusable form, cured and transported)
- Cast/poured-in-place/situ ~ TLE (placed in a plastic state)

Other types:
- Aerated ~ TLE (formed using gas-forming admixtures)
- Air-entrained ~ TLE (contains air bubbles to resist freezing)
- Cellular ~ TLE (low density, holds trapped air)

Note 1: The list of types is not exhaustive and at the discretion of the compiler and/or expert consultant, it can be expanded.

---

4 Underlined terms will appear as head terms in the glossary
CONCRETE TECHNOLOGY

Actions:
(1) ~ is placed/poured/cast/laid  TLE
   Variants: placing/placement(s)/pouring – TLE
   E.g. (a) Place concrete as near to its final position as possible; (b) Prestressed concrete requires the application of a load to the steel before concrete placement.

Note 2: The subsequent actions (2-compacted, 3-screeded, 4-floated, 5-trowelled and 6-cured) are represented in a similar way.

Processes:
(1) ~ bleeds TLE
   E.g.: After it is placed, concrete bleeds, i.e. the solids settle down and the mix water rises up to the surface.

Note 3: The subsequent processes (2-cures, 3-sets, 4-hardens) are represented in a similar way.

Note 4: The following cause-effect and other complex relations are represented as contexts or multiword terms which exemplify them:

Cause-effect relations:
~ curing aids hydration; ~ consolidation eliminates concrete voids; ~ compaction eliminates flaws; air-entraining admixture increases ~ durability; plasticizers increase ~ plasticity, etc.

Complex relations:
1. material – quantitative measure → concrete batch TLE
2. material – mold for pouring → concrete formwork TLE
3. material – preparation device concrete mixer TLE
4. material – pouring device concrete pump TLE

The Concrete Use terminological collocations do not need definitions but just TLE. They should be subdivided into concrete members (~ slab, ~beam, ~ column, etc) and concrete products (~ wall, ~ foundation; ~pavement, etc).
The single-word terms designating Constituents and Properties should be represented under the respective headings (aspects) and should be provided with definitions and contexts in their capacity as head terms. In case they have hyponyms (e.g. strength $\rightarrow$ compressive $\sim$, tensile $\sim$, etc.), they should form a nest within the head term entry, each provided with a short definition, e.g.:

**Strength** TLE (the capacity of an object or substance to withstand great force or pressure)
- compressive $\sim$ TLE (the resistance of a material to breaking under compression)

When implementing the small-scale terminographic project (see step 4 above) which I have termed a learner’s glossary, a compiler should bear in mind that the set of terminological knowledge items entering the glossary is to be considered an open system. In other words, the number of vocabulary items and terminological relations may vary according to the needs of the respective users but should not go beyond the boundaries of the conceptual structure of the special subdomain treated.

**6 Conclusion**

A methodology is proposed for a learner’s glossary which provides quick access to the conceptual relations in a subdomain, the head terms with their collocations as well as pragmatic information including contexts and translation equivalents. The terminological data are specified and organized by performing conceptual, lexico-semantic and systematicity-based contrastive analyses. The analytical data obtained are used to construct a knowledge-oriented model of a learner’s glossary entry. The model is created by representation and further subdivision of basic conceptual categories, addition of contexts exemplifying the usage of individual terms and explicating some terminological relations. It is emphasized that the learner’s glossary entry model allows maximum concentration of terminological knowledge. And last but not least, some major relationships are graphically represented by using the knowledge organization semantic tool ‘concept maps’.

Finally, the applicability of the methodology I propose for extracting and organizing terminological knowledge items in a subdomain should be emphasized. The methodology has been tested with translation and ESP students. The test results in the form of skilfully made course assignments in the form of bilingual mini term banks are quite encouraging and providing solid grounds for the introduction of that terminological practice in the LSP and technical translation classrooms.

**References**

TERMINOLOGY MANAGEMENT SYSTEMS AND 
ITS ROLE IN THE DEVELOPMENT OF THE 
LANGUAGE INDUSTRY IN COLOMBIA 

Juan Carlos Diaz Vasquez1, Carolina Franco Arroyave2

1. Professor at the Department of International Business, Universidad EAFIT, 
Medellín, Col. International Information Centre for Terminology, InfoTerm, 
jdiazva@eafit.edu.co

2. Student of Master in International Business at Universidad EAFIT, Medellín, 
cfranco3@eafit.edu.co

Abstract. This paper reflects the importance of Information and Communication Technologies (ICTs), especially Terminology Management Systems TMSs, as a key factor for the development of the language industry in Colombia and its economic global projection. Services offered within the scope of the Language Industry are mostly related to TMS, including tools for terminology extraction, knowledge management and communication, throughout the productive chain. The article concludes upon a survey conducted in the three biggest cities of Colombia about the appropriation levels of such tools in this industry and its impact on firm development.

Keywords: Terminology Management Systems TMS. Language Industry. Linguistic Software. International Standards. Internationalization.

1 Introduction

This research paper highlights the importance and dependence of the language industry on the newest ICT. Such technologies are the foundations for the effective providing of such services, which include, among others: translation, interpretation, subtitling, dubbing, software localization, web site globalization, development of language technologies, language education, consulting [LTC, 2009]). And which make use of tools such as Terminology Management Systems, Translation Memories,
Lexicons, Thesaurus, Taxonomies, Ontologies, Web-Dictionaries, E-mails, Video-conferencing Systems, which are intrinsic to linguistic activities. Additionally, these tools have propelled the internationalization of the Language Industry, through the reduction of communication costs and facilitating access to information. They are in many cases, even with unawareness of the producer, the entry modes to foreign markets and therefore essential part of the internationalization process of companies providing these services.

In general terms, Terminology Management Systems are probably the most common tool for technical and scientific translation activities, which represent a large proportion of the language industry activities. In this sense, the research pretends to answer the following question: Which are the most common technologies used in the language industry in these cities in Colombia and what is their role in the development of the companies?

1.1 Context

The digital revolution, since its early beginnings of Internet access, has been considered as a driver of economic growth, directly influencing GDP levels as pointed out by Katz (2010).

For the service sector, which has a high component of human capital, the proper management and use of information are of significant relevance for organizational operations, this explains why ICT have leveraged the growth of the sector worldwide [Javalgi & Martin, 2007]. ICT access and proper connectivity is considered imperative for the internationalization process, not only as a tool for communication with suppliers, customers and strategic partners but as an enabler of commercial processes to enter new markets through e-commerce.

In addition to the effects explained by Katz [2010] in the previous figure, the penetration of the broadband has some extra worthwhile mentioning effects, such as the improvement of education, increased possibilities for freedom of speech, increase of political conscience by accessing diverse sources of information, access to scarce products and services, etc.

In 2007, Latin America presented low levels of internet access compared to more developed countries (OCDE) and compared to the world average. The bright side is that such levels have presented patterns of growth since 1993 [López & Hilbert, 2010]. Nevertheless, it is important to keep in mind that the positive social and economic effects deriving from the access to broadband and ICT do not depend exclusively from the number of users in the region; it also depends on the quality of the service, and the education and training of the social actors for its appropriate use.

Colombia has been recognized for its production of primary goods. However, as the service sector and knowledge intensive industries gain importance in the economy also do the technologies for information and content management. Nonetheless, specific tools for Knowledge Management – KM – have developed at a slow pace and
have received poor attention from the private, public and academic sectors [Baquero & Schulte, 2007]. In knowledge intensive industries such as the language industry, is of high relevance the documentation of practices, techniques and employees implicit knowledge, in order for such practices to be preserved and improved by the organization to establish more effective processes. The study conducted by Baquero & Schulte [2007], presented that only 22% of the large- and medium-size enterprises in Colombia had a KM program implemented (written policies and strategies of KM) in 2007. One limitation for this study is the lack of segregation by sector since the average results may vary from the primary, secondary and tertiary sector. We could however induce, that if the largest companies in Colombia have not prioritized the information management, the situation for micro- and small-size companies is less encouraging.

In 2009, the Ministry of Communications of the Republic of Colombia was transformed to the Ministry of Information and Communication Technologies. Its purpose is to increase the access of Colombians to ICT through the implementation of new policies and programs intended to improve social, economic and politic development [MinTIC, 2014a].

The different strategies developed by the Ministry and implemented with the support of local administrations, are centered in a policy named “Vive Digital” (Live Digital). Such policy looks for the creation and strengthening of a digital ecosystem based on four strategic axes: Infrastructure, services offered by operators, applications, and users; while strengthening the demand and supply of the digital market [MinTIC, 2014b]. One of the projects that captured our attention is “Gobierno en Línea” (On-line Government), which promotes the modernization of the State and of its interactions with its citizens through ICTs. It integrates sites of information, procedures and services of the different government entities (Gobierno en Línea, ND).

Despite the slow growth in the access to Internet and ICT, the country still presents lags in this subject. For diverse industries, both national and international, this represents a great barrier for the access to the Colombian market, where only 16% of the surveyed Colombians participated on e-commerce in 2012 [Ipsos MediaCT, 2012].

The service sector in Colombia is the sector that has presented the biggest growth between 1970 and 2011, according to Fedesarrollo [2011]. However, such growth has not been reflected in the Colombian exports. This presents a broad spectrum of opportunities, specially for products and services with high knowledge and technological content since in 2000 only 12.7% of Colombian exports belonged to these types [Rojas & Vargas, 2007]. For 2010, the percentage of manufactures with high and medium technology content was approximately 34% (Torres & Gilles, 2013). This sector along with services, represent then high potential to be exploited, with regional referents as Mexico 75%, Costa Rica 50% and Brazil 44%, in the same period [Ibid].
1.2 Terminology Management Systems

The technic side of translation has become one of two necessary conditions that determine the professional capabilities of translation activities and other linguistic services; the level of technical capabilities is often a mandatory requirement to be competitive in a market that is more and more driven by ICTs. Covering the topic of software tools for every linguistic aspect requires a broad analysis of many of them. A common denomination that covers several tools for such services is Computer Assisted translation, which is the main focus of this article. To be more specific in tracking technologies for translation purposes in the scientific and technical domain we will focus in tools for terminology management, whose use is expected to be more widely spread.

The Terminology Management Systems are the software intended to administrate terminology data [Arntz, Picht, & Mayer, 2004], which are necessary for computer assisted terminographic work and for terminology management. Since such systems were conceived at first as special data banks for translations, they allow, mostly, interactions with text processing systems [Ibid].

Terminology Management Systems are computerized programs that facilitate collection, preservation and access to terminological data for reference of linguistic professionals, translators, terminologists and technical editors in specialized fields. [Díaz Vásquez, 2010]. By using such systems, it is possible to create terminology databases with entries containing term-related information.

Since Information Technologies have changed over time, so is terminology doing and moving forward to address more users’ needs. Some of those IT’s changes is the virtualization of the content so it is no longer storage in a given unit but accessible through Internet. These trends allow the use of different kind of mobile devices accessing the content in the Internet at the same time. There is then the cloud-based terminology (Varga, 2013). It has turned TMS’s into collaborative tools on the web.

These Terminology Management tools have developed solutions to different user including creation of terminology management networks, educational tools (e-learning) and commercial solutions (product data classifications).

TermCoord (www.termcoord.eu) is a very good example of a terminology network. This initiative fosters international collaboration within the framework of the EU in order to develop the mega terminology database IATE that is the biggest interactive terminology database from Europe. Termwiki (www.termwiki.com) is also a social network initiative offering Terminology Management tools for open communities and commercial purposes.

Other initiatives bringing Terminology Management solutions in an interactive way is the Euro Term Bank (www.eurotermbank.com). This platform enables users to participate in collaborative terminology projects editing and sharing the same information.

All these examples represent trends in the development of Terminology Management Systems that are necessary to carry on translation related tasks. Knowing these developments is a requirement for firms offering technical translation.
METHODOLOGY

This research has an exploratory purpose. Several studies have been developed in the language industry in Europe and the United States [Kelly & DePalma, 2012; TAUS, 2011; LTC, 2009], but little research has been conducted in this topic in Latin America. This research is being conducted with a quantitative analysis of primary nominal variables obtained through a web-based survey, from a positivist epistemological consideration that allows the acquisition of new knowledge through data collection [Bryman, 2008]. Although considered as opposed positions, the constructivist epistemological consideration also plays an important role since the social actors according to cultural, social and economic conditions might influence the results. Although Bryman [2008] also proposes that, the logic of a quantitative research is deductive, our research, in line with its exploratory purpose is developed inductively, since there is no initial hypothesis and the theory was proposed once the results were analyzed.

The choice for this data collection method was based on the advantages and disadvantages proposed by Bryman [2008] from a variety of methods. The self-administered surveys via e-mail, allow us to reach a high number of subjects in a short term and with low costs, additionally this allows to the respondent to answer conveniently accordingly to the availability of technical information required, which offers an advantage compared to methods such as interviews. However, the same characteristic presented us with the inconvenience of a low response rate, which we tried to avoid by contacting by phone the subjects of interest, looking for their previous acceptance to participate in the research.

The survey was conducted with firms in the language industry in the three main cities of Colombia [Bogota, Medellin and Cali]. The section of the survey that interests this article was focused on obtaining information about the use of Information and Communication Technologies (ICT), its relevance for the growth and expansion of the firms and the impact of the policies of the Ministry of Information and Communication Technologies in the development of a National “Digital Ecosystem” [MinTIC, 2013].

Finally, secondary data was obtained to complement the understanding and analysis of primary information. The main source for such data was the Colombian Ministry of Information and Communication Technologies. An analysis of existing literature was also conducted to provide context to the results of this research in comparison to previous research.

2 Results

We identified a total of 251 firms in the language industry in the three main cities of Colombia: Bogota, Medellin and Cali. These cities were selected under two crite-
ria: the offer of academic programs related with the industry (bachelor, specialization, master and doctorate degrees) and the population. In such criteria, these three cities ranked highest in Colombia.

From this initial population, we discarded 29 firms in the first filter since they were mainly free-lancers; these were legally constituted firms where the owner is the only translator. From the remaining 222 firms we were only able to contact 106, from which only 30 answered our survey and only 29 answered it completely.

Such low response rate hinders our ability to make inferences about the whole industry; therefore, results need to be presented merely in a descriptive way rather than in a generalized one.

Table I. Results.

<table>
<thead>
<tr>
<th>Value</th>
<th>Count</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Email</td>
<td>30</td>
<td>100%</td>
</tr>
<tr>
<td>Web site</td>
<td>25</td>
<td>83.3%</td>
</tr>
<tr>
<td>Shared web site</td>
<td>2</td>
<td>6.7%</td>
</tr>
<tr>
<td>Social Networks</td>
<td>20</td>
<td>66.7%</td>
</tr>
<tr>
<td>Translation software</td>
<td>15</td>
<td>50%</td>
</tr>
<tr>
<td>Translation memories</td>
<td>14</td>
<td>46.7%</td>
</tr>
<tr>
<td>Terminology Management Sys-</td>
<td>8</td>
<td>26.7%</td>
</tr>
<tr>
<td>tems</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mobile Apps</td>
<td>7</td>
<td>23.3%</td>
</tr>
<tr>
<td>Other</td>
<td>5</td>
<td>16.7%</td>
</tr>
</tbody>
</table>

Statistics

| Total    | 30 |
| Responses|     |

3 Discussion

Despite the great efforts of the Colombian government to increase the appropriation of ICT and the inclusion of new technologies in the productive processes of companies, a wide use of them is not perceived among the industry. This might be explained by the fact that the main target for the ICT Ministry is the education sector rather than the productive sectors. Therefore, their impact in the productive sector is expected in the long run.

The lack of widespread use of technologies encouraging the systematization of processes is also reflected in the lack of accession to international protocols and standards. Only 8 out of 30 firms, comply with at least 1 national/international regulation.

Equally important, is the poor use of technologies for Knowledge Management, since, in line with the lack of systematization of process, it makes the firm’s opera-
tions highly dependent of its employees. Consequently, if there is a change of person-

It is very interesting to note how, despite the developments of software specific
for terminology management, some of these firms are still attached to tools such as
Web Dictionaries. These tools, although of easy access and availability, are not reli-

Such findings allow us to present these firms of the language industry in Colombia,
as firms with very informal production. There is little standardization of processes,
not enough application of international standards over terminology, knowledge and
content management such as ISO 26162, EN 15038, quality standards such as
ISO9001 and even national standards over service supply as the NTC5808, which
reflect a lack of professionalization.

This last characteristic is also reflected in the low recognition translation profes-
sionals have in the country. The study conducted by Clavijo, Duque, Franco, Mendo-
za & Rodriguez in 2008 reflected that in Colombia, the translation activities are, to a
great extent, performed by professionals of fields different of linguistics and have
empirical knowledge of other languages, but few knowledge of terminology manag-
ment and its impact in the translation process. This lack of knowledge of terminology
specificities, might then explain the lack of appropriation of facilitating technologies.

4 Conclusions

The Language Industry in Colombia, as represented by this few firms, is still an in-
dustry in development, which shows low levels of professionalization and standardi-
ization regarding specifically the use of ICTs. These are issues of high importance for
a higher impact of this industry in the economy. For such process is necessary the
accession and compliance with international standards and regulations establishing
quality parameters for the service and the processes of this particular industry.

In Colombia, there is low participation of technologies in this industry what led us
to conclude that firms in the language industry are highly dependent on the know-
ledge and expertise of their employees. The lack of appropriation of these technolo-
gies can be explained under the light of the above-mentioned conditions: informality
and lack of knowledge on international standards.

Although Colombia has made significant progress in terms of technology appropri-
ation, the Ministry of Communications was even changed for a more updated and
comprehensive entity: the Ministry of Information and Communication Technologies,
such efforts have been mainly focused on the education sector but its impact in the
productive sector is yet to be seen.

Consequently, in order to increase the appropriation of technology tools with the
purpose of increasing the productivity of firms, it is required a bigger joint action.
Such action should be conducted by the industry, the Ministry of ICT, software pro-
producers, international standard and regulation organizations (such as ISO) to inform, promote and support the inclusion of such tools in the operation and management of firms in the language industry.

The appropriation of Terminology Management Systems not only increases firm’s efficiency and productivity but allows companies within this industry to provide higher quality services assuring coherence and standardization of the language.

5 Bibliography


Fedesarrollo, ¿Se está des-industrializando Colombia? (R. Steiner, Ed.) Tendencia Económica (109), 6-11 (2011)


H. Galperin, & W. Peres (Edits.), Acelerando la revolución digital: Banda ancha para América Latina y el Caribe (págs. 51-78). Santiago de Chile, Chile: CEPAL (2010)


López, P., & Hilbert, M. De la brecha en banda ancha al blanco móvil del ancho de banda. En V. Jordán, H. Galperin, & W. Péres (Edits.), Acelerando la revolución digi-
tal: Banda ancha para América Latina y el Caribe (págs. 121-142). Santiago de Chile, Chile: CEPAL (2010)


MinTIC. Plan de acción 2013. Bogotá
http://www.mintic.gov.co/portal/604/w3-article-1454.html


MinTIC, Ecosistema Digital. 29 de Enero de 2014, Ministerio de Tecnologías de la Información y la Comunicación:

Rojas, A., & Vargas, O. Globalización, nueva economía y TICS: Colombia, un caso de rezago tecnológico. CENES, 47-88 (2007)


TAUS, Lack of Interoperability Costs the Translation Industry a Fortune (2011)

Torres, D., & Gilles, E. Estructura tecnológica de las exportaciones industriales en América Latina 1990-2010. Administrador (22), 96-111 (2013)


Abstract. The ThirdPlaceLearning (TPL) frame and its sub-frames, constructed by Cognitive Linguistics methods, is a visual representation of the interconnected concepts. A list of defined terms and their underlying concepts were developed to create this frame – a representation of connections and associations within this frame. The concepts were then clustered within mental constructs, named categories, and used to clarify the structure of the TPL theory. The TPL frame and its sub-frames represent an associative network of concepts and terms belonging to different categories. It is a visualization of the TPL theory that comprises a set of dialogically negotiated mental/conceptual constructs.

Keywords: Category, cognitive linguistics, conceptsystem, frame, mental representation, proposition, terminology, ThirdPlaceLearning, TPL relational criteria, cognitive-onomasiological modeling

1 Research Overview

The ThirdPlaceLearning (TPL) theory [18] of InterCulturalCommunication (ICC) describes the liminal phase between the interlocutors’ perspectives. In contrast with other cross-cultural or trans-cultural theories of ICC, TPL views ICC interaction from a “third” vantage point. In other words, TPL uses a vantage point different to either interlocutor. This is achieved through developing an understanding of each other’s perspective and thus a better understanding of each interlocutor’s own perspective, through Perspective Sharing and Perspective Taking (PSPT). This latter process is facilitated by a set of processes and conditions, we call the TPL relational criteria, [1], [2], [3], [18] as depicted in Figure 1. In this paper we describe how Cognitive Linguistics (CL) methods [9], [14], [7] were applied to analyze the TPL concepts and terminology.

Epistemological negotiation of “ways of knowing” among the authors was required to connect education, intercultural communication, philosophy, psychology, terminology, linguistics and cognitive science with various theories of knowledge, because of the interdisciplinary character of the TPL theory. To achieve deeper understanding of TPL theory, the authors collaboratively explored a glossary of terms and conceptual models (frames) that represent associations among the TPL concepts.

The focus of our terminology research of the TPL theory has moved from the prescriptive and descriptive approaches [22],[19] to the cognitive aspects of terminol-
ogy. These aspects include the study of mental mechanisms of an individual interacting with the world that are reflected in his/her language [5],[15],[17],[21].

The application of CL to the terminology research illustrates the value of modeling mental processes. This cognitive modeling primarily assigns a structure to language units in a corresponding terminology and explains the behavior of terminological units in language. The holistic system of knowledge, or worldview, is specific to an individual’s consciousness and facilitates his/her orientation in a scientific or professional sphere; it comprises cognitive structures and terminology. Construction of conceptual structures (i.e. frames) of a terminology by means of cognitive research of their linguistic units (i.e. terms) gave us a holistic picture of a certain sphere of knowledge. Further, this type of construction enabled us to learn about complex theories of an area of knowledge that is related to TPL. Through an ongoing dialog, we developed a CL perspective of the TPL theory and its terminology that comprised a thesaurus, categories and frames construction.

TPL Thesaurus. To investigate the TPL theoretical framework [18] based on the CL methods, we first constructed an English language thesaurus of TPL terms and definitions. Where we used a word specific to the TPL framework, we defined it in the thesaurus using the ISO 704 standards [11]. TPL theory and its relational criteria were recorded and stored by means of terms based on an awareness and consciousness of terminology construction that involved its cognitive-communicative functioning. The TPL system was characterized with following terminological processes: identification of the terms assigned to each TPL concept and adding new terms when necessary; analysis of each TPL concept, concept structure and associated connections; and finally, compilation and management of the resultant terminology.

TPL Categories. The TPL concepts and their corresponding terms were clustered into mental constructs named categories based on their properties and functions. According to Kövecses, “…the conceptual categories… are the backbone of language and thought…” [13, p. 17]. Therefore, categories are always constructed with regard to a specific sphere of knowledge, where categories reflect the essential properties and connections within reality, as represented by human thought processes.

TPL Frames Construction. The associative network of concepts and terms belonging to different TPL categories was used to construct a series of frames and sub-frames. They are a form of visualization of the TPL theory comprising a set of dialogically negotiated mental/conceptual constructs.

2 Frame Construction of the TPL Knowledge Base

Some research in the field of terminology science [8],[15],[17],[21] uses frame analysis to discover the mechanisms of human knowledge representation and to study language as a mental construction. Terminology research, in the context of CL, presupposes the construction of a knowledge sphere conceptual system. The latter comprises connections between knowledge structures and their language representations. Hence, the frame analysis allows us to assume that language is a mental formation that is generated by human thought; a system of human knowledge representation. Therefore, the cognitive-communicative approach helps us to conceptualize the connections between firstly, phenomena that exist in the symbolic representation of lan-
language and secondly phenomena that are hidden from our observation, deep within the human consciousness. Cognitive linguists use this approach to better understand processes and mechanisms of human cognition.

The aim of our research was to construct a frame[16] representation of the TPL terminology system and to understand existing types of relationships between different concepts. The term frame was first used by Minsky in 1975 as a paradigm to understand visual reasoning and natural language processing. According to Fillmore [9], a frame is an abstraction (an “idealization”) of an individually construable form of mental representation encompassing one’s experience, knowledge, perception, etc. Barsalou [4] defined frames as complex conceptual structures that are used to, “…represent all types of categories, including categories for animates, objects, locations, physical events, mental events and so forth…” (p. 29). Incorporation of knowledge about a fragment of real life into a frame involves encoding of information about it (e.g., participants, their behavior, etc.). In other words, a frame is a unified knowledge structure or specific schematization of experience that provides a conceptual basis for a significant lexical corpus. “Frames are continually updated and modified due to ongoing human experience, and are used in reasoning in order to generate new inferences” [6, p. 223].

Research of knowledge organization of any sphere of human activity by means of the frame analysis is an effective way of terminology systems’ study where a frame is a structural unit that helps us to visualize integration of concepts into a real communicative situation. The frame of a knowledge sphere can perform three functions: (i) reflect the structure of a knowledge sphere, (ii) result in knowledge sphere cognition, and (iii) record knowledge in human mind (mental representation).

According to Minsky [16], the structural organization of frames is a network of nodes and their relations where nodes contain specific instances of data. A conceptual system (frame) captures an overarching, holistic representation of a knowledge sphere. Different frames and sub-frames of a system share the same terminals where activation of higher or lower levels of slots/nodes is initiated, i.e. one and the same node of a lower level may be included in different frames. As pointed out in some studies [20], sub-frames may contain information of different complexity, from a simple marker of special background knowledge to encyclopedic data. Every sub-frame reflects conceptualized knowledge about a certain entity in the world but is not a simple enumeration of entity’s markers. Some slots are filled with general information and other slots with specific information. Activating an existing frame by means of data at the lower level, we can reconstruct the structure of a situation as a whole.

Terminology categorization is reflected in the TPL categories, and it allowed us to understand yet another representation of the way concepts are organized within these categories. The main relations among TPL concepts were identified by means of cognitive-onomasiological modeling. The propositional analysis of this modeling is an information compression process that allows us to demonstrate that reality is structured in a certain way by language.
2.1 Visual Conceptualization of the TPL Theory

TPL frame construction described here was based on the visual representation of the TPL theory components, which encompass relational criteria, contexts, learning domains, etc. [1],[2] and [3]. The visual conceptualization of the TPL theory is an initial generalized version of terminology representation and relationships among the main terms used to represent the TPL conceptual network via frames. In Fig. 1 the following TPL relational criteria are presented: active listening, holistic mindfulness, dialectic thinking, critical co-reflection, intercultural sensitivity and conscientization.

Fig. 1. Concept map of the TPL theory

The visual conceptualization of the TPL theory is an initial generalized version of terminology representation and relationships among the main terms used to represent the TPL conceptual network via frames. In Figure 1 the following TPL relational criteria are presented: active listening, holistic mindfulness, dialectic thinking, critical co-reflection, intercultural sensitivity and conscientization [2].

2.2 TPL Frame

For each frame we compiled a list of terms for representation of frames or sub-frames and organized them according to their categories. To visualize the process of cognitive modeling, we include, below, the results of terminology categorization and propositional analysis for the general TPL frame (see Table 1):
### Table 1. TPL frame: Categorized terms and related propositions

<table>
<thead>
<tr>
<th>Term</th>
<th>Category</th>
<th>Proposition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Third Place Learning</td>
<td>Abstract</td>
<td>[CONCEPT - BE ABOUT – WHOLE/PART]</td>
</tr>
<tr>
<td>TPL Relational Criterion</td>
<td>Concept</td>
<td></td>
</tr>
<tr>
<td>Disorienting Dilemma</td>
<td>Abstract</td>
<td>[CONCEPT – BE OF – TYPE/KIND]</td>
</tr>
<tr>
<td>Misconception</td>
<td>Abstract</td>
<td>[CONCEPT – BE – WRONG]</td>
</tr>
<tr>
<td>Miscommunication</td>
<td>Concept</td>
<td></td>
</tr>
<tr>
<td>Preconception</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iceberg effect</td>
<td>Abstract</td>
<td>[CONCEPT1 – CHANGE – CONCEPT2]</td>
</tr>
<tr>
<td>Self-Identity</td>
<td>Abstract</td>
<td>[CONCEPT – BE ABOUT – SELF]</td>
</tr>
<tr>
<td>Active Listening</td>
<td>Process</td>
<td>[PROCESS – BE OF – TYPE/KIND]</td>
</tr>
<tr>
<td>Critical [Co]Reflection</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perspective Sharing/Taking/Shift</td>
<td>Process</td>
<td>[PROCESS – DEAL WITH – STATE/ABSTRACT CONCEPT]</td>
</tr>
<tr>
<td>Intercultural Sensitivity</td>
<td>State</td>
<td>[STATE (relation) – BE OF – TYPE/KIND]</td>
</tr>
<tr>
<td>Conscientization</td>
<td>State</td>
<td>[STATE1 – (NOT) BE AWARE OF – STATE2]</td>
</tr>
<tr>
<td>Holistic Mindfulness</td>
<td>Hybrid</td>
<td>[PROCESS/STATE – BE OF – TYPE/KIND]</td>
</tr>
<tr>
<td>Dialectic Thinking</td>
<td>Hybrid</td>
<td>[PROCESS/CONCEPT – BE OF – TYPE/KIND]</td>
</tr>
<tr>
<td>ICC Context</td>
<td>Hybrid</td>
<td>[TIME/SPACE/RELATION – BE OF – TYPE/KIND]</td>
</tr>
<tr>
<td>Multiple Perspectives</td>
<td>Hybrid</td>
<td>[STATE/CONCEPT – BE OF – TYPE/KIND]</td>
</tr>
<tr>
<td>Point of View</td>
<td>Hybrid</td>
<td>[STATE/CONCEPT – BE OF – TYPE/KIND]</td>
</tr>
<tr>
<td>Perspective</td>
<td>Hybrid</td>
<td>STATE/CONCEPT</td>
</tr>
<tr>
<td>Liminal Phase</td>
<td>Hybrid</td>
<td>[TIME/SPACE – BE PART OF – WHOLE]</td>
</tr>
<tr>
<td>Discourse (of learning)</td>
<td>Space</td>
<td>[SPACE – BE LIMITED BY – BOUNDARY]</td>
</tr>
<tr>
<td>Learning Domains: Cognitive; Psychomotor; Interpersonal and Affective</td>
<td>Space</td>
<td>[SPACE – BE OF – TYPE/KIND]</td>
</tr>
</tbody>
</table>
The TPL frame (see Figure 2) is a mental representation of interconnected concepts, which also includes six subframes/slots of each TPL relational criterion. ThirdPlaceLearning (ABSTRACT CONCEPT₁) occurs in a particular environment (SPACE₁) which comprises the discourse of learning within the four learning domains (cognitive, psychomotor, interpersonal, affective) (SPACE₂). Discourse of learning is grounded in critical social and cognitive theories and includes certain aspects of TPL, such as environment (formal, informal, class, home, etc.), power relations (employer, employee), culture (Khazakstani, Lithuanian), communication (introverted, extroverted). At the same time, the six TPL criteria occur within four learning domains. For example, critical co-reflection involves the cognitive learning domain, whereas bodymindfulness involves the affective and cognitive domains.

TPL is ICC context-dependent, where ICC context encompasses spatial, temporal, relational, and historical subcontexts (TIME/SPACE/RELATION). A disorienting dilemma, misconception, miscommunication, preconception (ABSTRACT CONCEPT₁,₂,₃) together with an iceberg effect (ABSTRACT CONCEPT₁) might initiate the transformational process of ThirdPlaceLearning [18]. The iceberg effect based on the iceberg’s deep, (subconscious) levels—meaning structure—is a cultural
lens through which we can see other cultures. Our worldview of other cultures is subjectively formed due to the influence of our individual meaning structures.

TPL relational criteria (ABSTRACT CONCEPT) is a combination of:

- particular cognitive and communicative PROCESSES, (dialectic thinking, active listening; critical co-reflection);
- awareness and somatic-emotional STATES, (intercultural sensitivity, conscientization); and
- as well as a conceptual hybrid STATE & PROCESS with somatic-emotional process and awareness state (holistic mindfulness).

When TPL relational criteria are in play, a disorientation may trigger a transformation of the initial self-identity of an individual (ABSTRACT CONCEPT) including an individual’s perspective, point of view, or multiple perspectives (STATE & ABSTRACT CONCEPT) to a new self-identity (ABSTRACT CONCEPT) with a new augmented perspective, point of view, or multiple perspectives (STATE & ABSTRACT CONCEPT). This transformation happens by means of the communicative and cognitive PROCESSES (perspective taking, perspective sharing, perspective shift), during or within a liminal phase (TIME & SPACE).

As an example of a sub-frame analysis, we would like to focus on conscientization, or critical consciousness, a critical social theory concept developed by Freire [10]. We define conscientization as, first, an awareness of privilege, oppression, deprivation and power difference between individuals and, second, taking action to minimize the power difference so as to improve communication. Further, Freire explains critical consciousness as a sociopolitical educative tool that engages learners in questioning the nature and consequences of their historical and social situation. A simple example is the acceptability of students asking questions or sharing opinions in a school classroom. In some classrooms this is encouraged and in others it is seen as an affront to the teacher. In the latter case, a conscientized student is aware of the freedoms of the former case, and realizes the lack of freedom imposed by the teacher. The conscientized student may want to change this situation.

2.2.1. Conscientization Frame

The frame of conscientization, a sub-frame of the TPL frame, constructed by CL methods, is a visual representation of the interconnected TPL concepts. A list of defined terms and their underlying concepts was developed to create this frame—a representation of connections and associations within the conscientization criterion. The concepts were then clustered within mental constructs, named categories, and used to clarify the structure of the conscientization criterion within the TPL theory. The conscientization sub-frame represents an associative network of concepts and terms belonging to different categories. It is a form of visualization of the conscientization criterion comprising a set of dialogically negotiated mental/conceptual constructs.

The sub-frame Conscientization is represented by the list of terms (see Table 2), which includes their categories and propositions.
To illustrate our results (Fig. 3), we described the associative connections among the terms and CONCEPTS standing behind them. *Conscientization* (STATE₁) is an awareness of *oppression* (STATE₄) and AGENT B taking action to move towards a position that minimizes the effect of *power difference* (STATE₂) on communication with AGENT A who has *privilege* (PROPERTY). Within the context of a *power relationship* (STATE₂), an oppressed AGENT B needs to develop the ABILITY₁ (having a voice) and ABILITY₂ (courage) to lower/overcome power difference.

**Table 2.** Conscientization subframe: Categorized terms and related propositions

<table>
<thead>
<tr>
<th>Term</th>
<th>Category</th>
<th>Proposition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conscientization</td>
<td>State</td>
<td>[STATE₁ – (NOT) BE AWARE OF – STATE₂]</td>
</tr>
<tr>
<td>Power relationship</td>
<td>State</td>
<td>[STATE (relation) – BE OF – TYPE/KIND]</td>
</tr>
<tr>
<td>Oppression</td>
<td>State</td>
<td>[STATE – BE USED FOR – OPERATION]</td>
</tr>
<tr>
<td>Privilege</td>
<td>Property</td>
<td></td>
</tr>
<tr>
<td>Conformity</td>
<td>Abstract concept</td>
<td>[CONCEPT – BE USED FOR – OPERATION]</td>
</tr>
<tr>
<td>Having a voice</td>
<td>Ability</td>
<td>[ABILITY – BE IN – COMMUNICATION]</td>
</tr>
<tr>
<td>Courage</td>
<td>Ability</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 3. Conscientization sub-frame
Having a voice is the ability to participate in a dialog and freely express ideas, opinions and perspectives while being respected and appreciated. Courage is the capability to ask about the other person's perspective or express your opinion, irrespective of the prevailing power relations and social conformity.

3 Conclusions

In conclusion, the concepts and methods of CL are applicable to many disciplines, such as terminology, applied linguistics, descriptive linguistics, lexicography, cultural linguistics, cultural studies, language acquisition, discourse studies and ideology, sociolinguistics, visual communication, stylistics, poetics, pedagogical linguistics, computational linguistics, signed language, philosophy, pedagogy, psychology and within interdisciplinary frameworks [12].

In our research, we presented a multidisciplinary study based on the intersection of terminology, education, intercultural communication, cultural studies, psychology and linguistics, where the methods of CL were widely applied. We described and presented the main TPL categories that may be used to assist learners to connect appropriately the abstract concepts of the TPL theory with certain linguistic categories and acquire the terminology belonging to this theory.

Also, we used the methods of cognitive-onomasiological modeling to construct the propositional models for all the terms representing the TPL theory. It gave us an opportunity to distill the main mental concepts that underpin the TPL lexicon and bring them together in an associative network in the form of conceptual models or frames described in the previous passages. In other words, the application of these CL methods allowed us to visualize the associative relation of terms as well as mental concepts standing behind them in the form of frames. The frames will assist our students to understand more deeply and precisely the conception of the TPL theory and help them allocate this network of concepts in certain categories.

Thus, these CL methods provided an effective explanation and visualization of TPL, which is a complex and abstract scientific concept that is described by a relevant terminology system. Moreover, by means of these methods, the TPL concepts at the mental level and the terms representing them at the language level were constructed into an associative network, which is an indicator of a well-developed and mature terminology where all the terms and concepts are interconnected.

References

TBX goes TEI

Implementing a TBX based extension for the Text Encoding Initiative guidelines

Laurent Romary

Abstract
This paper presents an attempt to customise the TEI (Text Encoding Initiative) guidelines in order to offer the possibility to incorporate TBX (TermBase eXchange) based terminological entries within any kind of TEI documents. After presenting the general historical, conceptual and technical contexts, we describe the various design choices we had to take while creating this customisation, which in turn have led to make various changes to the actual TBX serialisation. Keeping in mind the objective to provide the TEI guidelines with, again, an onomasiological model, we try to identify the best compromise in maintaining both isomorphism with the existing TBX standard and the characteristics of the TEI framework.

Keywords
Terminology, XML, TBX, TEI

Why adapt TBX for the TEI guidelines?

A brief history of the TBX lineage
TBX (TermBase eXchange) is the name of a standards for the representation of terminological data that has been published, in its full form, as ISO standard 30042 in 2008. It is the result of more than twenty years of standardisation activities in the domain of computerized terminologies that have taken place in ISO, the TEI\(^1\) consortium and the LISA\(^2\) association. It is currently being revised and updated in the LTAC/Terminorgs environment, from whence it will be published as an open standard under the ETSI framework. It is difficult to understand both the importance of this standard and the necessity to provide a subset of TBX integrated in the TEI guidelines without keeping in mind the overall history that lead to its publication. The major steps in such a history can be outlined as follows:

- ISO 6156:1987 (Mater), which described a format for representing terminological information on magnetic tapes, is probably the first international standard ever that dealt with the interchange of lexical data. Designed as a flat field-based representation in a pre-SGML world, the format is further developed to be applicable on microcomputers (MicroMater; see Melby, 1991);
- The experience gained with the work on Mater and MicroMater lead Alan Melby and several other colleagues\(^3\) to take part to the Text Encoding Initiative and put together a specific chapter\(^4\) of the TEI guidelines dedicated to the representation of terminological data. This first attempt to have an

---

\(^1\) www.tei-c.org

\(^2\) The Localization Industry Standards Association (1990-2011)

\(^3\) Those interested in seeing how the group actually worked can have a look at the minutes from the corresponding meetings (see TEI A17a, TEI A17b, TEI A17c)

\(^4\) See the P3 edition under http://www.tei-c.org/Vault/GL/P3/TE.htm
SGML-based representation integrated in the TEI framework remained there until the P4 edition when the TEI guidelines switched to an XML conformant representation. The chapter was considered as obsolete in comparison to ongoing developments within ISO (see below). Still, as can be seen from Figure 1, the TEI representation, despite its missing language section, contains most of the elements that one can find in the current TBX framework;

- ISO 12200 (Martif), published in 1999, resulted from a wish to take up and improve the TEI proposal, but also to make it an independent initiative under the auspices of ISO. Beyond the introduction of a language section, Martif also embeds terminological entries within a document structure strongly inspired from the TEI (e.g. the header-text organisation; entries embedded within a <text> and <body> hierarchy);
- In 1999, ISO also published an important standard, ISO 12620:1999, which compiled a set of reference descriptors (or data categories) to be used in any kind of terminological database. This standard represented the first abstraction over specific terminological formats as well as a tool for specification and customisation in the domain of termbank design. All descriptors we will use in this paper and in the TEI customisation we describe originate in this standard;
- ISO 16642 (TMF), published in 2003, brought another level of abstraction by describing a meta-model and modelling mechanisms independently of any specific XML format. Interoperability across terminological formats is basically ensured through compliance with ISO 16642 and ISO 12620:1999;
- In the period that followed, further work was carried out within LISA to define an optimal follower to Martif, which lead to the publication of the first version of TBX (TermBase eXchange) in 2003, which in turn was published as ISO standard 30042 in 2008.

As published in 2008, the TBX standard appears to be a comprehensive integration of the progress made over the years in understanding the variety of needs in the domain of digital terminology management. Still, the cumulative design that we have tried to outline above made it a complex object with two many options (e.g. two elements <tig> and <ntig> for implementing the term section component) and above all a proprietary technological implementation (reflected in the so-called XCS parameter files) for customising the necessary subsets needed specific implementations. This situation has slowed down the actual up-take of the format dramatically, with the emergence of other initiatives such as SKOS within the W3C that could completely ignore the benefits of a proper terminological representation of the corresponding data. In parallel, the quick uptake of digital methods in the humanities, where several research communities such as field linguists, epigraphists or classicists (among others) have a need to record onomasiological data associated with their primary sources, made it necessary to consider providing them back with an accessible chapter in the TEI guidelines.

<termEntry>
  <admin type='domain'>appearance of materials</admin>
  <tig lang=en>
    <term>opacity</term>
    <gram type=pos>n</gram>
    <descrip type='definition'>degree of obstruction to the

5 Martif aimed at reaching out to the translation and localisation industry, which were not part of the target audience of the TEI guidelines.
6 ISO 12620:1999 was further revised to describe the general setting of a data category registry (see Ide & Romary, 2004) for ISO/TC 37 activities, hence the dated reference.
The TEI perspective

The Text Encoding Initiative was initiated in 1987, when a group of experts in charge of several major literary text archives met to establish the foundation of what was about to become the most ambitious standardisation activity in the humanities. Under the auspices of the two major learned societies of that time, the ACH\(^7\) and the ALLC\(^8\), it produced reference guidelines that, since their first stable edition in 1992 (see Burnard & Sperberg-McQueen, 1995), evolved to incorporate both the most recent technological evolutions and the continuous feedback from its constantly growing user community.

There are several reasons why the paths of both terminology standards and the TEI crossed at some point:

- The Text Encoding Initiative had created a dynamics combining both a strong political (and financial) support as well as a high level of technological awareness. It brought together experts from many different fields ranging from poetics to computational linguistics, as well as with informed contribution from the library community;
- The first important decision related to the design of the TEI guidelines was to adopt the recently published ISO SGML standard (ISO 8879, published in 1986!) which offered, despite its complexity, a then unique platform for specifying and customising complex document structures;
- From a content point of view, the TEI was organised into working groups dedicated to various domains of text technology and representation. For lexical data, there was an ambition to cover both semasiological and, after some discussion, onomasiological representation models with on the one hand a group dedicated to “print dictionaries” and, on the other hand, one working on a future chapter “terminologies”.

---

\(^7\) Association for Computers and the Humanities
\(^8\) Association for Literary and Linguistic Computing
In this somehow favourable double context, the present paper describes what we think is an optimal strategy to both enrich the TEI guidelines with a proper means of representing onomasiological (terminological) data and contribute to the wider dissemination of best practices in this domain as reflected by the existing portfolio of international standards.

**Filius prodigus**

The main idea is to take up the simplified representation of a terminological entry in a way close to the TBX Basic proposal (See Melby, 2008) and incorporate it at any place in a TEI document where a semasiological representation (mainly embodied in the TEI guidelines by using the <entry> element) could also take place. The choice of TBX Basic, a small subset of TBX originally developed by LISA, as an inspirational source is justified by the need to achieve the best compromise between simplicity and conformance to the ISO standard. It is indeed essential, when considering this proposal, to keep in mind that the target user group is that of scholars with enough digital awareness to use the TEI guidelines, but with little or no background in terminology management. The resulting format should thus be both easy to use and compatible with the TEI cultural background.

The choice thus made can be seen as a start for further iteration between the TEI and TBX since it is based on two approximate assumptions. First, to take up TBX content at <termEntry> level and use the TEI document structure as the encompassing environment. In a way, we use TBX as the basis to bring back terminological entries within the TEI architecture in more or less the same conditions. Second, we initially limit our work to a very small number of data categories an will iterate with the TEI community to see which type of semasiological use cases are actually needed.

A basic overview of the available building blocks

The TEI document architecture

The TEI guidelines contain more than 500 elements to mark-up textual documents at various level of structural granularity and to take into account many different textual genres such as manuscripts, dramas, speech transcriptions or dictionaries. Its main document structure, depicted in Figure 2, combines a mandatory header (<teiHeader> element), grouping together all the meta-data attached to the document, with the actual content (<text>). This content can be further decomposed as a <front>, a <body> and a <back> element, in reference to the traditional book structure.

![Figure 2: TEI document architecture](image-url)

The TEI vocabulary provides a variety of means to encode a document, which we can sketch according to the following main categories:

- Description of the structure of a text by means of the generic <div> element;
- Organisation of the content along paragraph level objects such as lists, figures, tables, etc.
• Inline annotation elements to mark-up specific linguistic segments (highlighted object, foreign expressions) or reference to entities (names, dates, numbers);
• Domain specific constructs for dealing with turns in speech transcription, dictionary entries, etc.
• General-purpose representation objects such as bibliographical descriptions.

All these elements are part of a reference framework from which a given project should select which elements are needed for its purpose. Indeed, the TEI guidelines provide specific mechanisms to express such customisations as described in the next section.

The TEI specification framework

The TEI guidelines can be seen from two different angles. First, as the basis of an XML representation format, they provide the technical constraints to control the validity of TEI conformant document instances. Second, they are delivered with an extensive prose description that informs users about the logic of the guidelines as well as the most appropriate way(s) to use them to represent specific textual phenomena. 

Still, these two views are not split into two separated objects, but are indeed integrated within one single specification, from which one view and the other can be automatically generated. This mechanism, in line with the concept of literate programming (Knuth, 1992), relies in the existence of an underlying specification language named ODD (One Document Does it all), which is itself expressed in TEI.

In the TEI infrastructure, each element is thus defined as an ODD specification providing all the necessary information both to control its (XML) syntactic behaviour and to generate the corresponding documentation. Such information comprises a gloss, a definition, the technical description of its content model, the various attributes it can bear and one or several example of its usage.

In addition, the TEI framework offers two additional mechanisms that are central in providing the global coherence of the TEI guidelines: Classes and Modules.

There are two types of classes: attribute classes, which groups together attributes used in the same way across various elements and model classes, which group together elements that have a related semantics and occur at the same places in a document. The latter are means to simplify the expression of content models, but also to facilitate the customisation process by simply adding or removing an element from a class.

Modules are more global objects intended to group together coherent sets of elements designed for a similar purpose. Typically all the specific elements for dictionary encoding are grouped together in one single class.

These various mechanisms form the basis of the customisation mechanisms in the TEI guidelines based upon the ODD language, namely by selecting a group of modules for a given representation objective and within each module, keeping or discarding elements in the content model by editing the appropriate model classes.

---

9 The TEI guidelines contain in particular a wealth of examples for each element and the major constructs they allow.
10 For instance, the class att.global, which contains general purpose attributes such as the W3C @xml:id, @xml:lang or the TEI generic @n (local numbering), @rend (rendering information).
The TBX entry model

The TBX entry model is organised in strict compliance with ISO 16642 (TMF) such that the three core components of TMF, namely Terminological Entry, Language Section and Term Section are implemented respectively by means of the three elements: <termEntry>, <langSet> and <tig>.

The descriptive objects associated with these three levels are made of a) specific elements such as <term>, <note>, <ref> and <xref> and b) so called meta data-elements that may express a wide range of possible data categories, namely <admin>, <descrip> and <termNote>. For instance, <descrip type="definition"> is the TBX construct to represent a definition at the level of <termEntry> or <langSet>.

Figure 3 shows a typical (reduced) example of a TBX-basic entry\(^{11}\), where we can see how specific descriptions can also be embedded within grouping elements (<descripGrp>), when they need to be refined by additional information.

```xml
<termEntry xmlns="http://www.tbx.org">
  <descrip type="subjectField" xml:lang="fr">Industrie mécanique</descrip>
  <langSet xml:lang="de">
    <descripGrp>
      <descrip type="definition">endloser Riemen mit trapezförmigem Querschnitt, der auf zwei Riemscheiben mit Eindrehungen läuft
      </descrip>
      <admin type="source">De Coster, Wörterbuch, Kraftfahrzeugtechnik, SAUR, München, 1982</admin>
    </descripGrp>
    <note>Wird zum Antrieb der Lichtmaschine, des Ventilators und der Wasserpumpe benutzt</note>
  </langSet>
  <tig>
    <term>Keilriemen</term>
    <admin type="source">De Coster, ...</admin>
  </tig>
</langSet>
</termEntry>
```

Figure 3: Example of a TBX Basic entry (source: http://iate.europa.eu)

Technical integration

TBX in ODD

In order to integrate the TBX entry construct into the TEI framework, we have actually created a full ODD specification for it, which facilitates and documents the merging process. This ODD specification of TBX, which is delivered with the preprint of this publication\(^{12}\), is based on a simple architecture depicted in Figure 4 and strongly inspired from the TBX-basic organisation. Each level of the TBX terminological entry model is specified as a combination of members from the class model.auxInfo (whose member are, by default, <admin>, <descrip>, <descripGrp>, <transacGrp>, <note>, <ref>) and element(s) required at the child level.

\(^{11}\) All TBX elements are presented here as attached to a fake TBX namespace: http://www.tbx.org. The lack of a well-defined namespace for TBX is one of the difficulties associated with the current version of the standard.

\(^{12}\) http://hal.inria.fr/hal-00950862
For instance, the content model of `<tig>` is simply described as ¹³:

```xml
<rng:group>
  <rng:ref name="term"/>
  <rng:zeroOrMore>
    <rng:ref name="termNote"/>
  </rng:zeroOrMore>
  <rng:zeroOrMore>
    <rng:ref name="model.auxInfo"/>
  </rng:zeroOrMore>
</rng:group>
```

Figure 4: Overall architecture of the ODDD model of a TBX entry

**Modernizing and complementing the TBX technical setting**

This first level of specification of the TBX entry in ODD has some immediate impact on its actual XML properties. First, all elements in this specification, like any element from the TEI guidelines, are made members of the att.global attribute class in order to supply them with so-called global attributes, in particular, the generic W3C attributes `@xml:id`, `@xml:lang`, `@xml:base`, `@xml:space`. Second, all elements originally bearing a `@target` attribute in TBX are made members of att.pointing, the corresponding attribute class where this attribute is defined. The consequence here is that the old-fashioned ID/IDREF pointing mechanism in TBX is refreshed into a general URI based reference system. This change also makes the difference between internal (<ref>) and external (<xref>) reference obsolete leading to the disappearance of `<xref>`.

**Conflicting elements and attributes**

Several elements ¹⁴ and attributes ¹⁵ have the same name as other elements and attributes in the TEI guidelines. This could be seen as a non-issue when dealing with them in their appropriate namespaces, thus avoiding any syntactic conflict. Still, we have considered it important to provide a vision for making such objects in TBX and in the TEI converge, and this for the following reasons:

- For historical reasons, the TEI has been a strong inspiration for TBX, and the various TBX instances of the duplicate elements are indeed clones of the corresponding TEI ones, with a very similar underlying semantics;

¹³ The content models are expressed in ODD as RELAX NG snippets in the corresponding namespace: [http://relaxng.org/ns/structure/1.0](http://relaxng.org/ns/structure/1.0), and marked here with the prefix “rng” (see [http://relaxng.org](http://relaxng.org))

¹⁴ term, ref, hi, foreign

¹⁵ type, target
• From a user point of view, it would be difficult to systematically keep both of these elements and attributes in the same XML document model without spending a lot of energy trying to elicit the actual differences;
• The TBX instances are often simplified versions of their TEI counterparts, both from the point of view of their content models and the attributes they may bear;
• Furthermore, some of the TBX elements or attributes have departed from their original semantics (<hi>) or lost base with the XML technological evolutions (<ref> or @target, which still rely on the deprecated ID/IDREF mechanism);
• Finally, the perspective of merging part of the TBX vocabulary into the TEI framework in the long run should lead us to anticipate a necessary convergence.

These various arguments have lead us to replace all TBX instances of the duplicates by their TEI counterparts, even if we are fully aware of the drawback resulting from the frequent alternation of namespaces within a terminological entry.¹⁶

**Enriching element contents**

One important aspect of the TEI infrastructure is that it provides a large range of elements for the inline annotation of textual content. Such elements are indeed essential for the precise mark-up of all sorts of entities mentioned in a text (persons, places, etc.), linguistic phenomena (cited references, foreign expressions) but also technical means like inline cross-referencing. In this respect, the TBX framework is relatively poor since it only allows the following elements in the content model of plain text components¹⁷:

• A group of so-called meta-markup elements (<bpt>, <ept>, <ph>)¹⁸ designed to encapsulate non XML mark-up from an primary text source. These elements have also become quite standard practice in localisation or translation memory documents;
• Two specific inline elements: <hi> for highlighting segments of texts, optionally cross-referencing an external object (with a @target attribute) and <foreign> to annotate foreign words or expressions inline.

The process of creating a customisation incorporating TBX entries in TEI documents offers the perspective of enlarging the number and types of such components while preserving backward compatibility with the legacy model of such textual elements. The proposed customisation implements this extension in two steps:

• Create a class `model.metaMarkup` grouping together all above-mentioned meta-markup elements;
• Express the content model of textual elements as a combination of text, members of the `model.metaMarkup` class and the existing `model.limitedPhrase` from the TEI infrastructure.

†¹⁶ Alternation that should disappear if the TBX vocabulary could be taken up within the TEI framework.
†¹⁷ `<admin>, <descrip>, <note>, <ref> and <termNote>; <term> has a specific status as it only allows <hi> in its content.
†¹⁸ These elements are part of the TMX standard, which is the most widely accepted XML application in the localization industry (see http://en.wikipedia.org/wiki/Translation_Memory_eXchange)
Doing so changes the content model of text elements in several ways:

- The meta-markup elements are available in the TBX namespace;
- Textual elements are provided with a wide variety of annotation objects;
- The TBX <hi> and <foreign> are replaced by their TEI counterparts. This removes the @target attribute from <hi>, which is somehow fortunate since it guides users towards adopting more appropriate elements (such as <ref>) for cross-referencing.

**How can TBX benefit from going TEI**

Beyond the obvious gain in visibility that the TBX standard can benefit from being linked to the widely adopted TEI framework, we would like to focus on the various ways the work described in this paper could positively impact on the future technical definition of TBX.

The main proposal for TBX in managing its evolution within a more sustainable framework would be to re-align its document macro-structure to that of the TEI. Indeed, the whole document model down to <termEntry> in TBX is strongly inspired from the TEI guidelines, but suffers from both its oversimplification and from the fact that it cannot automatically benefit from existing components, and further improvements, in the TEI guidelines that would immediately be valuable to TBX (e.g. software description, bibliographical sources, etc.). By re-using the TEI model in its reference namespace, with the appropriate customisation in order to avoid cluttering the header with constructs that have nothing to do with terminology management, TBX would benefit from a rich and maintained meta-data environment while keeping the freedom of defining the standard as a specialized subset.

One real challenge for the experiment we are carrying out here is to find a strategy for the various domains where the TEI provides alternative, and essentially more precise, constructs than TBX for the same types of objects.

A first example are the grammatical features in the TEI guidelines, which offer a wide variety of elements for providing part of speech, gender, number, etc. information, possibly grouped together within a <gramGrp> container. For users accustomed to the TEI background, such descriptors would be obvious alternatives to the use of the <termNote> element from TBX. Should we provide both possibilities over time? Should we suggest the use of <gramGrp> in TBX? Or should we just consider that the use of one or the other corresponds to two dialects (TEI-TBX vs. mainstream TBX) with explicit mappings between the two.

In the same way, the bibliographical descriptions in TBX are very shallow and only allow plain text content such as:

```
<admin type="source">De Coster, Wörterbuch, Kraftfahrzeugtechnik, SAUR, München, 1982</admin>
```
whereas the TEI could possibly offer any level of representation for such object by means of unstructured (<bibl>) or structured (<biblStruct>) representation, for instance:

\[
\text{<bibl><author>De Coster</author>, <title>Wörterbuch, Kraftfahrzeugtechnik</title>,<publisher>SAUR</publisher>,<pubPlace>München</pubPlace>,<date>1982</date></bibl>\]

There are several similar cases that we could point out, in particular with regards the descriptions of the forms associated with a term (pronunciation, transliterations, inflexions). The precise analysis of all the potential conflicts goes beyond the scope of this version of the paper, but we consider that the answer should not be made issue per issue, but be based upon global scenarios as to what is the optimal implementation in the TEI framework, what level of interoperability we want to keep with mainstream TBX and what the TBX standard itself could actually incorporate in the future.

Where should we go from here?

We have described in this paper an attempt to create a missing component in the TEI guidelines that would provide an onomasiological representation for lexical data as a complement to the existing “dictionaries” chapter. This component is strongly based upon the TBX standard, but we have made the choice of defining a specific blend of TBX rather than just plugging-in TBX entries within a TEI document. This is, to our view, an essential step if we want acceptance of such representations within the largest possible TEI user community covering all fields in the humanities. Still, in conformance to the principles of ISO 16642 (TMF) we have tried to ensure maximal isomorphism between the blend we have defined and the original TBX standard, while identifying the situations where the TEI could bring even more encoding precision.

The vision that needs to be developed in the long term for the TEI guidelines should provide clear guidance as to which model can be used for which purpose in a Digital Humanities scenario. In this context, it is important to remember that TBX-Basic was conceived for a specific class of applications in terminology management and not to offer a simple framework for onomasiological representations. Further work should also contribute to have a better understanding of the optimal articulation between semasio-lexical representations and onomasio-terminological ones in the humanities. Such work could be inspired by seminal works such as (Melby & Wright, 1999) or the ongoing Ontolex initiative within W3C.

References


ISO 6156:1987 Magnetic tape exchange format for terminological/lexicographical records (MATER)

\[19\] We have not gone into the details of the TBX-Vasic data categories, but we can observe how such categories as projectSubset, with its industrial background, probably are out of scope of a humanities application.

ISO 12200:1999 Computer applications in terminology — Machine-readable terminology interchange format (MARTIF) — Negotiated interchange

ISO 16642:2003 Computer applications in terminology — Terminological markup framework

ISO 30042:2008 Systems to manage terminology, knowledge and content — TermBase eXchange (TBX)


TEI A17a, TEI A17M5, Minutes of Meeting of June 11, 1991 [http://www.teic.org/Vault/AI/ai7m05.txt](http://www.teic.org/Vault/AI/ai7m05.txt)


TEI A17c, A17-Minutes of the 3rd TEI A17 meeting of 1991-11-15/16 [http://www.teic.org/Vault/AI/ai7m17.txt](http://www.teic.org/Vault/AI/ai7m17.txt)

**Acknowledgements**

This work has benefited from the support of the French ANR TermITH project ([http://www.atilf.fr/ressources/termith/](http://www.atilf.fr/ressources/termith/)).

Many thanks also for the feedback provided by the various TKE reviewers of this paper.
A Model and Method to Terminologize Existing Domain Ontologies

Dagmar Gromann
Vienna University of Economics and Business
Welthandelsplatz 1, 1020 Vienna, Austria
dgromann@wu.ac.at

Abstract. In today’s shrinking world, the need to automatically process multilingual knowledge becomes increasingly pressing, particularly in specialized communication. Domain ontologies enable automated computing of structured knowledge, but feature little and mostly English natural language content. Terminological resources, on the other hand, provide rich multilingual data, but differ in their distribution format, data semantics, representation language, and approach to terminology science. This diversity makes it difficult to interchange their data and link them to ontologies. To address this issue, we propose a terminology interchange model that supports sharing terminologies and linking them to ontologies. The proposed model is a formalized ontology on terminology that represents the main elements of terminology science derived from ISO TC 37 standards and best practices. A methodology for applying the model to domain ontologies and merging the resulting terminological resource with available multilingual terminological data is proposed.

Keywords: Terminology, Domain Ontologies, Standards, Metamodeling, Data Interchange and Integration

1 Introduction

Ontologies are crucial to writing consistent and formalized definitions within a specialized domain, relying on formal semantics [7, 13]. Human users, natural language processing applications, semantic indexing, and information retrieval based on ontologies additionally require natural language content in the ontology [13, 5]. However, few resources provide rich multilingual natural language information and formal semantics. If available at all, natural language content is usually restricted to annotation properties, such as rdfs:label, and to the English language [5]. To address this problem, the proposed modular terminology interchange model (T-Mint) represents the main elements of concept-oriented terminologies as a formal ontology, a so-called ‘terminology of terminology’ [4]. It is designed to facilitate the data integration and interchange of terminological data and link them to domain ontologies.

In line with ISO TC 37 standards, the proposed terminology interchange model consists of a conceptual, language, and term level. To those three main
levels, administrative and descriptive information is added by re-using data categories from ISOcat\(^1\). To make the proposed model itself highly re-usable, it is represented as a modular ontology. Its three main modules are the core-structure module, representing the three main levels, a data category selection module, and a sub-term module. Thereby, connecting the core structure with any customized data category subset is facilitated. Instances of the proposed model provide dereferenceable terminological data - such as terms, definitions, contexts, etc. - that can be related, queried, and further annotated.

Established solutions to terminological modeling in ontologies frequently restrict natural language data to annotation properties, whereby they cannot be related, annotated, or used for reasoning. Section 2 specifies current practices and similar models at the terminology-ontology interface. To be truly useful, any such terminological metamodel needs to be based on requirements of the community. This is why Section 3 briefly summarizes the most important requirements taken from standards, best practices, and current modeling practices. The proposed model is detailed in Section 4. A methodology for applying it to a domain ontology and interchanging its data with other terminologies is suggested in Section 5 prior to some concluding remarks.

2 Background

RDFS's label properties link synonymous strings to an ontological entity. Because such annotation properties cannot be annotated or related, no information other than XML language tags can be added. The Simple Knowledge Organization System (SKOS) model \[8\] allows for a differentiation between 'preferred', 'alternative', and 'hidden' labels. With its SKOS-XL extension, terms are instances of OWL Full class and can thus be related to each other. However, modern terminological resources require a higher complexity of terminological data and modeling decisions than provided by these vocabularies. Furthermore, all references to 'terminology' were removed in the current version of the W3C SKOS recommendation for a reason \[15\]. In fact, SKOS is a representation format for controlled vocabularies targeted towards human users \[8\] and not machine-readability \[14\] or representing terms in use.

Descriptive and administrative properties are frequently added as metadata. Metadata vocabularies tend to define all properties in their vocabulary, even if this leads to duplication. For instance, depending on the vocabulary notes can be represented as rdfs:comments, skos:note \[8\], omv:description \[10\] to name a few. Without explicitly mapping one to the other their content is not detected as identical by automated programs. Tao et al. \[14\] tackle this problem by proposing terminology guidelines on vocabularies and not just their metadata. ISO TC 37 handled this problem by introducing data categories for metadata and data, which have been collected in the ISOcat repository.

The objective of the proposed terminological interchange model is to represent a large variety of terminological data in general and for an arbitrary ontolo-

\(^1\) http://www.isocat.org/
ogy. To meet this goal, the proposed model is based on standardization efforts within the terminology community, in particular the Terminology Markup Framework (TMF) (ISO 16642:2003) and TermBase Exchange Format (TBX) (ISO 30042:2008). Together TMF and TBX define the basis for a family of terminological markup languages, but have no mechanism for relating terminological data to ontologies.

TERMINAE [2] presents the most comprehensive and well-established approach for modeling terminological information in ontologies, but focuses mainly on learning informal ontologies from text and represent terms as annotation properties. Thus, a separate manipulation of terms and concepts in TERMINAE is not granted. Other terminological models for ontologies reduce natural language content to natural language definitions [12, 11].

3 Terminology Metamodel Requirements

A metamodel on the elements of terminology science needs to address the requirements of its intended users and community. Requirements to the terminology interchange model derive from a systematic analysis of the state of the art in terminological modeling. First, standardization efforts of the terminology community were analyzed, in particular TMF, TBX, and ISO 704:2009 on terminology work. Second, an extensive literature review focused on theoretical methods (e.g. [1, 4, 7, 15]), guidelines (e.g. [14, 9]), and existing analysis of terminological resources (e.g. [3]). Lessons learned and limitations of existing models (e.g. [11]) and available resources (e.g. SNOMED CT) helped gather principles and requirements for the proposed model. The major requirements identified from this systematic analysis are:

1. Multi-Purpose and Multi-Domain Applicability: While the representation format itself needs to be independent of any specific purpose, the model it represents should be applicable to various purposes, such as indexing, natural language processing, capturing findings, etc., and domains, such as finance, biomedicine, law, etc. This entails that the model needs to be highly re-usable for various settings.

2. Concept Orientation: The meaning of a concept is unique in its concept system. Terms related to the concept have at least one and not more than one meaning. Although subordinate concepts inherit characteristics from superordinate ones, differentiation between parent and child need to exist.
   (a) Concept Permanence: Concepts and terms can be manipulated separately so that the concept system is maintained even if terms evolve.
   (b) Unique Non-Semantic Identifier: Reclassification of concepts or polyhierarchies make hierarchical numbering (e.g. 1010 is subclass of 10) difficult. Thus, the unique identifier needs to be void of any (implicit) meaning.
   (c) Hierarchical Ordering: Each concept has at least one parent. Top-concepts are children of a unique empty top concept that facilitates extending the resource. Polyhierarchies may only be added when qualifying the subsumption relation (IS A) with a subdivision criterion, e.g. IS_Agent.
3. **Term Autonomy**: All terms (abbreviations, symbols, variants, etc.) can be documented with all, i.e., unlimited number of, necessary term-related details and data, including term use and context. This means that the representation format needs to provide for extensions at any time.

4. **Accessibility**: Terminological data need to be accessible to humans and machines. This means that definitions and descriptions of meanings need to be formalized as well as represented in natural language.

5. **Interoperability**: Any terminological metamodel needs to be available in a format that facilitates data interchange across applications and resources.

## 4 Terminology Interchange Model

The proposed terminology interchange model (T-Mint) illustrated in Fig. 1 represents a metamodel for terminology. One terminological data collection aggregates several terminological entries and can reference one domain ontology. The proposed model is represented as a modular ontology, each module being a logically consistent separate ontology with alignments across the modules. The proposed model consists of three major modules: the core-structure module (CSM), the data category selection module (DCSM), and the sub-term module (STM). Components of the core-structure module correspond to the Terminological Markup Framework (TMF) metamodel and are further described by data categories, which are derived from ISOCat and represented in the data category selection module. The core-structure module currently represents data categories as specified in the default subset of TBX with some adaptations based on technological change and best practices. Terms that consist of several words, so-called multi-component terms, can be separated into term components and sub-terms. Elements in in Fig. 1 marked as <<auxiliary>> are abstract and only included for demonstrative purposes, but have no corresponding class or relation in the actual model implementation.

**Terminological Data Collection** The terminological data collection is realized as a resource, i.e., an ontology file containing terminological data. Terminological entries in the collection are subclasses of a unique empty top concept `owl:thing`. Administrative details are added to the file directly by means of re-using pre-existing meta-data from repositories, such as the Dublin Core\(^2\) elements `dc:creator` or `dc:title`.

**Terminological Entry** It represents the conceptual level of a terminology and is equivalent to `Meaning` of the semiotics.owl\(^3\). Each entity of the input ontology might be referenced by one terminological entry and represents the meaning of this entry. Thus, the object property `reference` - re-used from the ontolex lexicon model for ontologies\(^4\) with a relation of the same name - is modeled as functional with the class `TerminologicalEntry` in its domain.

---

\(^2\) [http://dublincore.org/](http://dublincore.org/)

\(^3\) [http://www.ontologydesignpatterns.org/cp/owl/semiotics.owl](http://www.ontologydesignpatterns.org/cp/owl/semiotics.owl)

\(^4\) [http://www.w3.org/community/ontolex/wiki/Main_Page](http://www.w3.org/community/ontolex/wiki/Main_Page)
Each terminological entry can be related to other terminological entries by means of concept relations, such as antonymConceptOf.

**Language Section** Synonymous terms in one language are all grouped into one language section, which is a part of a terminological entry. The relation oboRO:part_of is re-used from the OBO Relation Ontology[^5], which is transitive, reflexive, and anti-symmetric. Terminological entries can be monolingual with one or multilingual with multiple language sections.

**Term Section** Each term section consists of one term in one language and is a oboRO:part_of a language section. The class term section is instantiated by an actual term, e.g. 'securities' is the label of an owl:Individual of the class term section. Term sections can be related to other term sections by means of term relations, some of which are predefined. For instance, if two terms are spelling variants of each other, they are either related by variant or its subproperty spellingVariant.

**Term Component Section** This section represents individual components of multi-component terms. Thereby, descriptive and administrative details can also be added to a component of a term. Each component has to be linked to at least one term.

**Descriptive Data Categories** Simple data categories (DCs) are atomic and can neither be assigned a value nor contain other DCs, so they are modeled

[^5]: http://obofoundry.org/ro/
as individuals. Complex closed DCs provide picklists of predefined values in the form of simple DCs, such as `contextType` with a value 'associated context'. They are modeled as classes with an enumeration of their individuals by means of `ObjectOneOf`. Complex open DCs have a predefined datatype but open data value, so they are modeled as data type properties specifying the data type, which mostly are plain literals or integers. For instance, `geographicalUsage` is a complex open DC and has the datatype restriction plain literals.

**Administrative Data Categories** Administrative data are represented as annotation properties of the core-structure components the data category belongs to. For instance, `sourceType` can be added to the conceptual, language, and term level of the core-structure module. For this purpose, the proposed ontology re-uses existing annotation properties of various repositories, such as the Dublin Core\(^6\) or the Ontology Metadata Vocabulary (OMV)\(^7\). Some administrative data can alternatively be modeled as relation, for instance, `conceptOrigin` can relate the concept in the terminological data collection to the concept in the original resource if the latter has a dereferenceable URI.

**Sub-Terms** Sub-terms are components of multi-component term which themselves are terms in the same domain as the term they are part of. For example, 'liquidity risk symptom' in the financial domain might be dissected to 'liquidity', 'risk', 'symptom' and combinations thereof. While 'liquidity' could still pertain to the financial domain, 'symptom' on its own is less likely to be included in a terminology on finance. So the former might be considered a sub-term because it is domain-specific. Sub-terms can be identical to term components, but term components do not need to be domain-specific. This idea of creating a repository of sub-terms has the objective of providing a full account of terms within a domain ontology.

The proposed terminology interchange model relates terminological data to domain ontologies and allows for a description of their semantics in more detail. Furthermore, it constitutes a terminological data collection which can be re-used for other purposes, such as natural language processing. Terminologies using T-Mint allow for terms and definitions to be addressed and annotated directly, related to each other, and be manipulated separate from terminological entries. It has to be noted here that T-Mint is not merely a syntactic conversion of the TMF meta-model and TBX-Default. Its representation in the Web Ontology Language (OWL) required the addition of formal semantics and several modeling decisions, which were based on best practices, existing models, and existing resources.

An increasing tendency to model terminologies and linguistic models based on description logic could be observed \[^3\]. One reason for this might be that high-quality available DL reasoners can be used for automated consistency checking, e.g. to avoid terminological cycles. Furthermore, OWL makes the open world

\(^6\) [http://dublincore.org/](http://dublincore.org/)

\(^7\) [http://omv2.sourceforge.net/](http://omv2.sourceforge.net/)
assumption, which means that knowledge not represented or inferred is simply unknown and not wrong or false. Thereby, OWL facilitates a modification or extension of existing resources. This is why the proposed model is represented in the DL-based EL subprofile of OWL 2, which has been chosen for its scalability, ample tool support, and successful record with biomedical terminological ontologies. It has to be noted that OWL 2 EL is specifically orientated towards resources with high classification needs, which is the case for terminological resources with very large subsumption hierarchies.

The proposed model itself is domain-independent and adequate for various purposes and domains. Each of its entries is identified by a unique numeric identifier to ensure concept permanence. Concepts and terms are related but modeled separately so that evolving terms can be changed without having to alter the concept. Its modular structure allows for the creation of customizable subsets of data categories as well as unlimited addition of new categories to ensure term autonomy and multi-purpose adequacy. While the human-driven aspects are ensured by the usage of natural language names, the machine-readability requirement is met by the chosen representation language. While OWL is fit to facilitate syntactic interoperability, semantic interoperability relies on formal and explicit representations of meaning provided in T-Mint and by reference to a formal ontology. Moreover, the strong emphasis on re-using existing vocabularies contributes towards the model's interoperability with existing models and resources, e.g. resources using Dublin Core, the OBO Relation Ontology, GOLD, or semiotics.owl.

5 Methodology

This section presents methodologies for two possible use cases of the proposed model, namely generating a terminological resource and merging it with other, informal terminologies. Both cases serve the purpose of showing potential application scenarios of T-Mint, but still have to be validated and evaluated in a separate paper.

5.1 Terminologizing a Domain Ontology

One of the main goals of T-Mint is to be used to generate terminological resources to describe domain ontologies. It represents the meta-structure of a terminological resource and needs to be instantiated for the specific domain. A top-down methodology for creating an instance of the T-Mint model is described below.

Methodology This methodology presupposes a formal domain ontology represented in OWL as input. The methodology requires an ontology editing method, such as the OWL API, for manipulating the input ontology and the created T-Mint instance, and one off-the-shelf NLP tool for POS tagging and tokenization.

http://owlapi.sourceforge.net
Random generation of integer as building block for unique non-semantic URI to create terminological entry and use the reference relation to establish connection to a top hierarchy class of the input ontology.

Use language tag from RDF or SKOS label to create a language section - the URI for the language section uses the number created above and adds the xml:lang tag to it, e.g. 1423en for English.

Extraction of label from RDF, SKOS, or URI fragment if no label is available and creation of a term section identified by a number added to the language section identifier, e.g. 1423en1.

Tokenization produces subcomponents of the label, which are linked to the term section.

Part of speech tags are represented as complex closed data categories, extracted from the created tags, and related to the term components.

Identified noun phrases are suggested as sub-terms, which have to be evaluated manually for their reference to the domain of the input ontology.

Definitions are, if available, instantiated and related to the terminological entry.

The described methodology produces a collection of terminological entries - of which one entry is exemplified below - that depend on the input ontology for formal semantics and concept relation. Frequently, ontologies do not feature any natural language definitions. In such cases, a generation of the natural language definition based on the formal definition is one alternative. One way to achieve this goal is a combined method of ontology verbalization and ontology design patterns as we describe elsewhere [6].

5.2 Merging existing resources with T-Mint instances

Existing multilingual and concept-oriented terminological resources, such as EuroTermBank⁹, are frequently available as alphabetical listings in formats restricted to human readability. Nevertheless, if they contain proper terminological definitions, the superordinate concept of a terminological entry can be extracted from the natural language definition. The superordinate concept of the terminological entry in the T-Mint instance is obtained by way of the subsumption hierarchy of the input ontology. Merging the entry of the existing resource with the entry of the T-Mint instance creates a machine-readable multilingual resource.

⁹ http://www.eurotermbank.com
Methodology The following components can be used to evaluate whether two concepts can be considered equivalent:

- All terms - entry as well as semantically related terms
- Term components
- POS tags assigned to terms and term components
- Nouns, adjectives, and verbs extracted from natural language definitions
- Superordinate concepts extracted from definition and subsumption hierarchy

Instead of string-matching each of these components from an entry of the T-Mint instance to the potential matching entry in another resources separately, we suggest representing them as vectors. Vector Space Models are used to measure the similarity of keywords in documents in information retrieval. As the merging process can be compared to a keyword search, the Vector Space Model is used as a means of similarity measure of both entry vectors. Each dimension of the vector represents one term or expression from the list above. Its value depends on the frequency of occurrence of the term. The dot product of both vectors is divided by the product of their norms. The threshold for merging entries should be lower for exact matches of entry terms than for partial matches. Merging means adding new term sections based on the extracted content of the resource grouped by language to the terminological entry of the T-Mint instance. Thereby, more languages are added to the ontology by reference.

6 Conclusion

In this paper we propose a terminological metamodel build on standards and best practices from the terminology community. Established vocabularies frequently provide terminological data as informal RDF resources and/or annotation properties (e.g. SKOS). Both cases do not support the relation of terminological data, their annotation, and machine-readability. This is why, the proposed model is a formalized ontology that allows modeling highly complex terminological data in relation to a formalized domain ontology. We provide a methodology for terminologizing domain ontologies as well as for merging available multilingual terminologies building on the proposed terminology interchange model. The major motivation for terminologizing ontologies is the ability to relate, annotate, and directly address natural language elements while maintaining the reference to formal semantics. First experiments still need to be extended to provide a full evaluation and validation of the proposed method and model.

References


Towards modeling lexicons compliant LMF in OWL-DL

Abstract. Elaborating reusable lexical databases and especially making interoperability operational are crucial tasks effecting both Natural Language Processing (NLP) and Semantic Web. With this respect, we consider that modeling Lexical Markup Framework (LMF) in Web Ontology Language Description Logics (OWL-DL) can be a beneficial attempt to reach these aims. This proposal will have large repute since it concerns the reference standard LMF for modeling lexical structures. In this paper, we study the requirement for this suggestion. We first make a quick presentation of the LMF framework. Next, we define the three ontology definition sublanguages that may be easily used by specific users: OWL Lite, OWL-DL and OWL Full. After comparing of the three, we have chosen to work with OWL-DL. We then define the ontology language OWL and describe the steps needed to model LMF in OWL. Finally, we apply this model to develop an instance for an Arabic lexicon.

Keywords: Lexical Markup Framework LMF, Web Ontology Language Description Logics OWL-DL, Interoperability.

1 Introduction

Consistent lexical resources represent a crucial requirement for several NLP tasks. This necessity arises by the rising need of automatic tools to deal with Information retrieval, Information Filtering, Information Extraction, Question-Answering, etc. However, these tasks suffer from the lack of reusable linguistic, and in particular lexical, resources. These deficiencies vary quite a great deal from one language to another. Arabic is one of the languages which suffers most from this shortcoming. The common problem of the majority of the languages is that the lack of resources limits any progress in the computational linguistic sciences for these languages [1].

From another angle, the need of standardized lexicons is even harder to achieve because standardization requires significant time resources. First, human resources are needed to ensure compatibility with the chosen standards making the task of putting together a conformant lexical structure more complex. LMF is one of the standards in language technology that intends to cover all languages in the world. Providing compliance to such a standard thus makes our work comparable with similar endeavours worldwide.

In this paper, we thus propose an initiative enabling us to model the LMF standard in the OWL-DL ontology language, with the aim to facilitate the elaboration of reusable lexical data bases and make interoperability operational in future works. As a matter of fact, there are few standards dedicated to digital lexica in comparison to available standards for language resources at large. International Standardization Organization (ISO) lexicons have critical effect in NLP. Indeed, this standardization identifies an informative common coverage for all lexicons. The developed coverage is fundamental for introducing tools allowing the exchange and the share of lexical
resources. So that interoperability can be easily introduced. This notion means that information and communication systems will be able to exchange data and enable sharing knowledge [2, 3]. Nowadays, having interoperable framework is so required then before. It will be a mixture of standards and guidelines such as Text Encoding Initiative (TEI) [4]. So, standards will be consistently correlated and guidelines will explain application of standards specification. However, between these two axis (standards and guidelines), a transformation prototype should be present. This prototype should have a lot of characteristics which will be explained later. However now, we can prove that OWL-DL can be an important factor in this prototype [5].

In this paper, we will present first of all a scope for LMF in order to make sure that this standard will be able to be mapped to OWL-DL [6]. Secondly, we will present OWL with its three sublanguages: OWL Lite, OWL-DL and OWL Full and we will prove our choice for the OWL-DL. The next section will be the most important one while it interests our transformation prototype from LMF to OWL-DL. Finally, we will instantiate this model to develop an instance for an Arabic lexicon.

2 LMF overview

After successful scientific activities and teamworks in developing lexicons, NLP and Machine Readable Dictionaries (MRDs) communities decided to start ISO tasks in 2003. Several theoretical divergences in structures languages make these activities hardly achieved in 2008. In fact, a group of 60 researchers was behind LMF standard creation [3]. LMF is an ISO standard covering monolingual and multilingual lexica. LMF specification follows UML modeling principles defined by Object Management Group (OMG). It is composed of a core model and extensions packages. The modeling principles of LMF take up the general principles developed in ISO committee TC 37 and allow a lexical database designer to combine any component of the LMF meta-model with data-categories [7] in order to create an appropriate model. These data categories function as UML attribute-value pairs in the diagrams. The core model covers the backbone of a lexical entry. It specifies the basic concepts of vocabulary, word, form and sense. LMF core model is a hierarchical structure consisting on several components. Lexical Entry is one of the components that represents the basic resource in the lexicon. In fact, this unit represents the lexeme and contains associated form and sense.
Fig. 1. Inflected forms of the verb "نام" "nAm" (to sleep)

The example in Fig. 1 illustrates one prototype among the entire inflected forms of the verb "nAm" (to sleep). This example is an instance from the Arabic LMF core model. Extensions are used according to the requirements of the users. Thus, lexicons developers have to choose packages that are useful for their needs. However, an extension package can not be drawn regardless of the core package [3].

Fig. 2. Syntactic extension of the verb "نام" "nAm" (to sleep)

\[1\] http://www.qamus.org/transliteration.htm
The example illustrated in Fig. 2 shows the syntactic extension concerning the verb "nAm". This example is a part of the Arabic LMF extensions packages.

There have been some works dealing with LMF lexicons for Arabic language. In [8], authors have studied the importance of the reuse of syntactic Arabic lexicons. They have consequently encoded a lexicon compliant LMF after the examination of HPSG and LTAG lexicons specificity [9]. In the same context, we can mention the work concerning Arabic lexicons projection from HPSG to LMF [10].

As we have already explained previously, LMF and lexical standard in general can make the notion of interoperability more operational if we use a transformation prototype of LMF in OWL-DL. In the next section, we will provide a presentation of this language in order to describe the main lines for the prototype.

3/owl overview

In general, ontology is a philosophy concept that allows studying the existing [11]. Yet, in the computer sciences, it must be defined as a structured and formal set of concepts offering meaning to informations. Particularly, OWL, recommended by W3C and strongly inspired from DAML+OIL, is a language which represents ontologies. These ontologies are quite useful on the Semantic Web [12]. In fact, data occurring there could be easily published and shared. From a technical point of view, OWL includes comparison tools of properties and classes that match properly with LMF such as identity, cardinality, inheritance. In fact, OWL presents more capacity for content web interpretation than RDF and RDFS due to the largest vocabulary and the right formal semantic [6], [12]. While modeling in OWL, we will notes that OWL supplies three increasingly expressive sublanguages that may be easily used by specific users: OWL Lite, OWL-DL and OWL Full [5]:

OWL Lite is the sublanguage which is used by those who need simple classification hierarchy and constraint features. For instance, if we study cardinality constraints supported by OWL Lite, we note that only the values of 0 or 1 are allowed. Indeed, OWL Lite looks simpler than OWL-DL and OWL FULL which seem to be more expressive.

OWL-DL is the sublanguage whose users look for highest number of expressiveness. In spite of this expressiveness, completeness and decidability are assured as well [13]. It means that all inferences are computed and in limited time. Technically, OWL-DL involves absolutely all paradigms of OWL. These concepts have some restrictions. For example, a class can not be considered as an individual or property. This phenomenon is called type separation. As well, a property can not be considered as an individual or a class. This restriction is allowed in OWL FULL and consequently makes it non decidable [14]. OWL-DL is named from its underlying logical formalism, description logics, which offers adequate inference capabilities while preserving expressive power. And for this purpose, we want to model LMF with this expressive sublanguage.
Finally, OWL Full is destined for those who want the greatest expressiveness without caring about how much completeness and decidability are guaranteed. This liberty makes one use the syntactic of RDF with large freedom. From a technical corner, a class can be discussed at once as a collection of individuals and as an individual in its own right. In a word, we can say that the maximum of expressiveness limits decidability and makes reasonable mechanisms more and more complex.

Works dealing with the generation of OWL resources in Arabic language are so few. We can mention the approach proposed for the generation of domain ontology from LMF standardized dictionaries. An additional alternative of automatic domain ontology enrichment based on the semantic component of LMF has been suggested [15]. We can mention the lexicon model for ontologies “lemon”[2]. This RDF model, dedicated for representing lexical information relatives to ontologies, is totally the opposite of our model. In fact, the developed prototype consists on developing a consistent ontology for lexicons. In the following section, we will explain details for the developed model for the transformation of LMF to OWL-DL model since it will facilitate having interoperable framework in the future. In fact, this prototype will play the role of the pivot between standards and guidelines. Currently, projects require such a construction; otherwise they will be out of business. A recent report by TAUS declares that: “The lack of interoperability costs the translation industry a fortune” [16]. Fortune is compensated for adjusting data formats. Thus, setting up an interoperability framework will gain us much more time and fortune. The step before building this framework is to seek for a pivot language. This language will be described as a dynamic environment where standards will be consistently related and guidelines evidently explain the specifications application to several types of resource.

4 Modeling LMF in OWL-DL

The built of interoperable framework is our future target. However, the construction of a similar framework requires an environment making possible interoperability between applications exchanging non formal and non structured informations through the web. So, it helps exchanging data and simplifies documents description. These characteristics are available in OWL [17]. Then, we have chosen OWL-DL because of its expressiveness and decidability as we have described in the previous section. Thus, modeling LMF in OWL-DL is a crucial task in the process of making an interoperable managing lexicons framework. In fact, we have to check the possibility of mapping the whole LMF concepts to OWL-DL ones. This task is hard to carry out since the components in LMF model are so nested and complex. In this part, we are going to describe the prototype of transformation LMF into OWL-DL. This prototype is divided in seven parts:

---

4.1 Building OWL-DL Entities

In order to simplify some entries in OWL modeling, we have to use first of all required entities. Here, we have defined the following entities:

```xml
<!DOCTYPE rdf:RDF [  
<!ENTITY xsd "http://www.w3.org/2001/XMLSchema">  
<!ENTITY owl "http://www.w3.org/2002/07/owl">  
<!ENTITY rdf "http://www.w3.org/1999/02/22-rdf-syntax-ns">  
<!ENTITY rdfs "http://www.w3.org/2000/01/rdf-schema">  
<!ENTITY lmf "http://www.lexicalmarkupframework.org">  
]
```

The xsd, owl, rdf and rdfs entities are related to the OWL-DL language. Yet, "lmf" entity is related to the LMF model.

4.2 Used Namespaces

In order to make ontologies more comprehensible and non-ambiguous, OWL offers the possibility of a new component definition: namespaces. This component is an indication for specified vocabularies used in the ontology.

```xml
<rdf:RDF  
   xmlns:owl="#owl;"  
   xmlns:rdf="#rdf;"  
   xmlns:rdfs="#rdfs;"  
   xmlns:xsd="#xsd;"  
   xmlns:lmf="#lmf;" >
</rdf:RDF>
```

The above namespaces can be useful for the terms related to the LMF standard. Thus, the first component of the ontology is the definition of a declaration set of XML namespace contained in an opening tag `<rdf: RDF>`. These statements are used to interpret identifiers and make the following presentation of the ontology much more readable.

4.3 LMF Header and Classes

A set of assertion should be described after the definition of namespaces. These assertions adorn the modeling file by comments, labels, version control and inclusion of other ontologies.

```xml
<owl:ontology  
   rdf:about="http://www.lexicalmarkupframework.org">  
   <rdfs:comment>Ontology LMF in OWL</rdfs:comment>  
   <rdfs:label>LMF Ontology</rdfs:label>  
</owl:ontology>
```
Classes in OWL are considered as basic components. All these classes are in fact members of “Thing class”. Concerning our LMF core model, we have defined eight classes as follows:

```xml
<owl:Class rdf:ID="Lexical Resource"/>
<owl:Class rdf:ID="Global Information"/>
<owl:Class rdf:ID="Lexicon"/>
<owl:Class rdf:ID="Lexical Entry"/>
<owl:Class rdf:ID="Form"/>
<owl:Class rdf:ID="Sense"/>
<owl:Class rdf:ID="Definition"/>
<owl:Class rdf:ID="Statement"/>
```

The classes described above concern only the LMF core model. Classes in other extensions packages have to be mapped also to OWL-DL.

### 4.4 LMF SubClasses

Generally, all ontologies contain a list of restrictions. Subclasses are one of those restrictions. In LMF core model, Form Representation are restrictions of the class Representation:

```xml
<owl:Class rdf:ID="Representation"/>
<owl:Class rdf:ID="Form Representation">
  <rdfs:subClassOf rdf:resource="#Representation"/>
</owl:Class>
```

Only two subclasses are defined above. The LMF model includes several subclasses that should be converted to OWL-DL.

### 4.5 LMF Properties

Some general and specific information are interpreted as attributes in LMF core model. For example, Global Information is an administrative class involving general attributes, such as /language coding/ or /script coding/ which are suitable for the whole lexical resource:

```xml
<owl:DatatypeProperty rdf:ID="language coding ">
  <rdfs:domain rdf:resource="#Global Information"/>
  <rdfs:range rdf:resource="#&xsd; String"/>
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="script coding ">
  <rdfs:domain rdf:resource="#Global Information "/>
  <rdfs:range rdf:resource="#&xsd; String"/>
</owl:DatatypeProperty>
```

All attributes figured in the LMF model have to be converted to OWL-DL.
4.6 LMF Relations

LMF relations define a list of domain and co-domain restrictions. For instance, “has lexicon” is an “ObjectProperty” restriction:

![Diagram of LMF Relations](image)

Fig. 3. LMF Relations

The Fig. 3 shows the name relation added for the LMF model. This name will allows us making the transformation to OWL-DL. Modeling the restriction “has lexicon” in OWL, we obtain:

```xml
<owl:ObjectProperty rdf:ID="hasLexicon">
  <rdfs:domain rdf:resource="#LexicalResource"/>
  <rdfs:range rdf:resource="#Lexicon"/>
</owl:ObjectProperty>
```

We have to add relations names in the LMF core model with the aim of modeling relations restrictions in OWL-DL.

4.7 LMF Cardinalities

Cardinalities are transformed to restrictions in OWL-DL. Thus, they are defined as follows:

```xml
<owl:Class rdf:ID="LexicalResource">
  <owl:Restriction>
    <owl:onProperty rdf:resource="#hasLexicon"/>
    <owl:minCardinality rdf:datatype="xsd:nonNegativeInteger">1</owl:minCardinality>
  </owl:Restriction>
</owl:Class>
```

These cardinalities make the designed model richer in term of restrictions. Thus, individuals created later will have number constraints.

5 Instantiation for Arabic lexicon

The instantiation part is done according to the OWL-DL built scheme. So, in this section, we are going to choose morphological extension in LMF extensions
packages. This choice is based upon the importance of this part for the most lexicons in NLP. Morphological extension is treated by two different ways in LMF. The first represents explicitly inflected forms. The second uses the paradigms of flexions to generate different forms derived from the Lexical Entry. We represent a part from the entire inflectional description of the verb "جلس" “jls” (to sit). This description is in fact an instantiation of the built prototype which is explained in the previous section:

```xml
<LexicalEntry rdf:ID="جلس فعل"/>
 <PartOfSpeech>verb</PartOfSpeech>
 <Root>جلس</Root>
 <Scheme>جلس</Scheme>
 <WordForm rdf:resource="">
   <writtenForm>جلس</writtenForm>
   <grammaticalTense>inaccomplished</grammaticalTense>
   <grammaticalMood>apocated</grammaticalMood>
   <person>2</person>
   <grammaticalGender>feminine</grammaticalGender>
   <grammaticalNumber>dual</grammaticalNumber>
 </WordForm>
</LexicalEntry>
```

The example shows a prototype of one possible inflected form from a set of 56 Word Form that an Arabic verb could take.

6 Discussions

The proposal of modeling Arabic lexicons compliant LMF in OWL-DL is based upon several paradigms: (i.e. header, classes, subclasses, properties). Applying these concepts, we have built a new ontology designed on OWL-DL. However, constructing such a comprehensive ontology is hindered by the complexity of the LMF model. Once the entire LMF ontology (core model and extension packages) is already built, we have to populate it with individuals. With this prototype, making lexicons, in any language, is so easier to build since we have just to instantiate the OWL-DL prototype with the appropriate individuals. However, such modeling includes shortcomings: mapping prototype can lead to the loss of certain informations such as aggregation.

7 Conclusion

We have studied the structure and representation of the LMF mode in order to design an OWL-DL ontology that would be able to match its components maximally. The next step will be to use this model as a tool to check the actual coverage of existing LMF serialisations such as the one anticipated in [18] on the basis of the TEI framework. The underlying vision is to create an interoperable framework describing a dynamic environment among standards and guidelines. Such environments should be both internally coherent and facilitate the continuous update of modeling standards.
and their serializations when use cases and associate tool development provide new representational needs.

References

18. Romary, L. TEI and LMF crosswalks, to appear in Gradmann S. (Ed.) Digital Humanities: Wissenschaft vom Verstehen, Humboldt Universität zu Berlin - http://hal.inria.fr/hal-00762664 (pre-print, 2013)
TBX-Min: A Simplified TBX-Based Approach to Representing Bilingual Glossaries

Arle R. Lommel¹, Alan K. Melby², Nathan Glenn², James Hayes², Tyler Snow²

¹German Research Center for Artificial Intelligence (DFKI), Berlin, Germany
arle.lommel@gmail.com
²Brigham Young University, Provo, Utah
{alan.melby,garfieldnate,james.s.hayes,tylerasnow}@gmail.com

Abstract. TermBase eXchange (TBX) has provided a successful mechanism for exchanging complex terminological data. Because TBX defines a family of related formats for representing terminological data rather than a single format, it can be adapted to many needs. However, use within commercial production environments has remained limited due to the perception that TBX is too complex for particular use cases. This paper describes the development of a new derivative of TBX, called TBX-Min, that is designed to represent the sorts of columnar tables of terms in two languages widely used by practicing translators, in a TBX-compatible fashion. Through TBX-Min, translators will be able to send and receive simple, machine-processable bilingual terminology while still gaining access to the wider ecosystem of TBX-compliant tools.

Keywords. terminology management • TermBase eXchange • TBX • XML • translation • localization

1 Introduction

The proper use of terminology is considered one of the most important aspects of translation quality. A recent examination of translation quality assessment metrics and tools in the QTLaunchPad project conducted by one of the present authors found that the only error category included in all metrics examined was adherence to terminology guidelines. A translation process that does not include access to domain-, company-, or discourse type-specific terminology will produce incorrect results.

To address the requirement for correct terminology, many organizations maintain mono- or multi-lingual terminology databases (“termbases”). Termbases often have very complex internal metadata structures that are used to facilitate knowledge management processes and linguistic processes such as information on the legal status of terms, guidance on what translations must (or must not) be used for specific terms, links to additional information, etc. Such termbases may easily define twenty or more such “data categories” for a given concept and each term tied to that concept. The principles of concept-oriented terminology behind such systems have been established for many years and are defined by ISO standards 704:2000 (“Terminology work —

Translators also require terminology information, but their requirements are considerably more modest. In general they need to know how specific terms should be translated (or not translated) and do not (usually) need the detailed metadata found in complex organizational termbases. (One exception involves those cases in which translators are asked to find or create translations for terms that have not previously been translated, but as this is a separate research task, this case is not addressed in this paper.) Translators for many years have used lists of terms and translations, initially hand-written, and later stored in word processors or spreadsheet applications, to document their terminological preferences and requirements. Such lists generally consist of rows, each containing a term and its translation(s). They may, additionally, contain a part of speech, general notes, an indication of what customer uses the term, and the term’s status (such as whether it should be used or not); these additional items, however, with the exception of notes, are generally not found in such spreadsheets.

Although statistics are not available on this subject, from anecdotal discussion, we believe that spreadsheet files containing bilingual terminology lists account for the substantial majority of all terminology resources in the language industry. Even those translators who use the terminology-management capabilities of computer-assisted translation (CAT) tools, also known as Translation Environment Tools (TEnTs), often still create, send, and receive spreadsheet-based terminology lists since they are easy to use and manipulate.

These spreadsheets are not without their drawbacks, however. When exchanged between different spreadsheet applications they are frequently exported as comma-separated value (CSV) files. CSV is not a single format, but rather a loose descriptor of a set of heterogeneous formats that use a comma to define column boundaries. One particularly common problem is that CSV files can appear in a variety of character encodings and the encoding is not indicated in the file, leaving the interpretation ambiguous. Microsoft Excel, perhaps the most popular spreadsheet program, for example, assumes ISO Latin-1 encoding and requires workarounds to load CSV files in other encodings, rendering CSV files problematic in Excel for many languages.

The development of standards for representing and exchanging terminology data has largely focused on the needs of organizational users, leaving a gap between the needs that standards address and the requirements of translators and project managers. The remainder of this article will describe the development of TBX-Min, a new format for representing bilingual glossaries that helps bridge the gap between spreadsheet glossaries and complex concept-oriented terminology resources.

(Note that the description of TBX-Min in this article represents the current working draft as of May 2014 and is subject to change. Please visit http://www.tbxinfo.net for the latest version.)
2 Abbreviated History of Terminology-Interchange Formats

Before going into the technical details of TBX-Min (short for TBX-Minimal), it is important to situate TBX-Min in the history of terminology-interchange formats. This section will not provide a full description of all formats, but instead provides the reader with an overview of the relevant formats. Readers interested in more detail on the history of terminology interchange standards are encouraged to consult [2], which describes this topic in more detail.

2.1 Pre-TBX

Although the most relevant starting point for this discussion is the development of TermBase eXchange (TBX) in the first decade of the twenty-first century, there were a number of earlier formats used for terminology interchange. These formats include MATER (ISO 6156:1986) and MicroMATER, and the SGML-based MARTIF (ISO 12200:1999) and GENETER formats. Of these, MARTIF is the most relevant as it served as the basis for the development of TBX. However, all of these have since been superseded by TBX.

Although not directly used for interchange, the Terminological Markup Framework (TMF, ISO 16642:2003) standard [3], provides “guidance on the basic principles for representing data recorded in terminological data collections” (ISO 2003). It defines an abstract “metamodel” for the structures to be used in specific terminological markup languages, and serves as the basis for the model in the TermBase eXchange (TBX) standard. While using TMF does not guarantee that all metadata will be exchangeable between systems, using it does guarantee a degree of compatibility between systems.

2.2 TermBase eXchange (TBX)

TBX is an XML-based family of formats for representing the structure and content of termbases. Initially developed in the European Union-funded SALT project and later by the OSCAR standards group of the now-defunct Localization Industry Standardization Association (LISA) and published in 2002, TBX replaced MARTIF with a similar, but updated, XML format. TBX was subsequently adopted by ISO Technical Committee (TC) 37 as ISO 30042:2008 and co-published with LISA, and is thus now the primary international standard for the exchange of structured, concept-oriented terminology data. (More information on the need for TBX can be found at [4].)

Although TBX has been implemented by a number of large organizations and translation tool developers, overall uptake among language service providers and individual translators has been lower than desired. In the authors’ discussions with implementers and users of terminology management tools, one of the primary reasons cited for not using TBX is that it is “too complex”. TBX’s descriptive vocabulary contains many more data categories (types of metadata) than required for any individual termbase and uses a mechanism—the eXtensible Constraint Specification (XCS)
file—to declare the specific data categories allowed in a given “dialect” (also called “variant”) of TBX. It is thus flexible, but it is impossible to know without consulting the XCS file which data categories an arbitrary TBX file will use. This complexity means that TBX import routines need to be able to support and interpret arbitrary TBX dialects and inform users of problems when data categories present in a TBX file cannot be represented in the destination termbase. In addition, a typical TBX file will contain far more information than a translator is likely to use.

It is important to note that TBX, in order to support its flexibility requirements without the need to create new document type definitions (DTDs) for each dialect, has a structure that declares data categories as attributes rather than elements. For example, rather than declaring “part of speech” as an XML element (e.g., `<partOfSpeech>noun</partOfSpeech>`), the 2008 version of TBX declares data categories as attributes (e.g., `<termNote type="partOfSpeech">noun</termNote>`). This decision leaves the core structure of TBX very compact and allows easy subsetting via the XCS file, which constrains the allowable types via attributes, allowing all TBX dialects to share the same basic XML structure. This style (called “DCA” for “data categories in attributes”), although used in TEI, is somewhat unusual when compared to other XML formats.

1.1.1 TBX Basic

In response to requests from tools developers for a format that would be easier to implement than arbitrary TBX dialects and for guidance about what features of TBX would be needed in typical localization scenarios, the LISA Terminology Special Interest Group introduced the TBX-Basic specification in 2009 [5]. TBX-Basic is a fully compliant subset of the default set of TBX data categories that reduces the available data categories from 117 to 24. The only mandatory data categories in TBX are the term itself and its language. However, implementation guidance in the specification strongly recommends that TBX-Basic files include the part of speech for each term as well. Like the full TBX, it maintains the DCA style of XML. While TBX-Basic is considerably easier to implement than the default TBX with its set of data categories, it is still more complex than spreadsheet-type resources and has met with limited adoption.

1.1.2 Multiple Rows per Concept (MRC)

In an effort to provide a bridge to the spreadsheet world, a spreadsheet-style representation of TBX was developed called MRC (“Multiple Rows per Concept”). The MRC format allowed data to be stored in a spreadsheet, but proved difficult to use because TBX is fundamentally a relational format that cannot be easily stored in a 2-D table. Although MRC can be stored and manipulated in a spreadsheet, it is not a typical spreadsheet format and does not meet the requirement for a simple equivalent to a multicolumn spreadsheet-based list of terms. Therefore, while MRC can represent TBX-Basic in full in a spreadsheet, it does not fulfill working translators’ requirements from a spreadsheet-type format.
2.3 Universal Terminology eXchange (UTX)

At around the same time that TBX was moving to the ISO framework, an independent effort within the Asia-Pacific Machine Translation Association resulted in the Universal Terminology eXchange (UTX) specification [6], originally called UTX-Simple, in 2009. UTX was focused specifically on Machine Translation (MT) systems (although it has since found broader application). UTX was seen as an alternative to “heavier” formats for MT lexicons like Olif (http://www.olif.net). It was intended to be a very lightweight format with a tab-delimited structure that could be easily viewed in a spreadsheet. Accordingly, it does not use an XML structure. Because it came from an MT perspective and developed independently from TBX, UTX has very little similarity to TBX. Although it does fill the requirements of a simple spreadsheet-style format, UTX’s structure does not allow for it to be easily integrated with structured concept-oriented terminology formats.

3 TBX-Min: The TBX Format for Glossaries

Since TBX was adopted by ISO and the creation of TBX-Basic, it has become clear that these formats were too complex for use as a spreadsheet replacement for working translators. Even the spreadsheet-oriented MRC format did not meet their requirements since it contains far more information than is typical in a spreadsheet glossary. While UTX meets these requirements, it does not provide the linkage to terminology standards that would be needed to provide a “migration path” for moving spreadsheet glossaries into structured terminology environments or for exporting subsets of organizational termbases as glossaries for translators.

To address these needs for translators while still using standards-based approaches, the informal TBX steering committee, which continues the work previously done within LISA, has now created TBX-Min for representing bilingual glossaries in a TBX-compatible format, based on previous work in this area [7]. The purpose of TBX-Min is to represent extremely simple termbases, such as spreadsheets, and to be as human-readable as possible. TBX-Min is, as the name implies, a very minimal dialect of TBX. Its feature set is minimal, providing just enough to convert most UTX and simple spreadsheet documents losslessly while still conforming to the TMF metamodel. Additionally, a valid TBX-Min document contains information only for the source and target languages. It provides a simple method by which a project manager may send a bilingual glossary to an assigned translator rather than unnecessarily sending a (potentially) large multilingual glossary. If the multilingual glossary is a TBX-Basic file, it can be converted using the appropriate method described below.

A cursory examination of a TBX-Min file shows that it does not look like a traditional TBX file. Figure 1 shows a side-by-side comparison of a single term entry in both TBX-Basic and TBX-Min (with spacing added to keep content parallel).

The most obvious difference is that TBX-Min does not use the “DCA” style of tagging. Instead it uses a “DCT” (for “Data Categories as Tag names”) style (see [2]) that uses elements for the data category names. DCT style is easier to validate in some cases since the most common version of the XML schema language XSD does not
allow the content of elements to be restricted based on attribute values (an important requirement for DCA-style TBX since many data categories may share one element name in DCA), but does require a custom DTD or schema for each dialect. Because TBX-Min is intended for widespread use, the use of a custom schema for this dialect does not pose a problem. (It is possible to automatically convert a DCT-style file to a DCA-style file and the two structures are semantically equivalent. Although TBX-Min files do not look like traditional TBX, they can be easily converted to validate against the core TBX structure.)

<table>
<thead>
<tr>
<th>TBX-Basic</th>
<th>TBX-Min</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;termEntry id=&quot;C003&quot;&gt;</td>
<td>&lt;entry xml:id=&quot;C003&quot;&gt;</td>
</tr>
<tr>
<td>&lt;descripGrp&gt;</td>
<td></td>
</tr>
<tr>
<td>&lt;descrip type=&quot;subjectField&quot;&gt;</td>
<td></td>
</tr>
<tr>
<td>Restaurant Menus</td>
<td></td>
</tr>
<tr>
<td>&lt;/descrip&gt;</td>
<td></td>
</tr>
<tr>
<td>&lt;/descripGrp&gt;</td>
<td></td>
</tr>
<tr>
<td>&lt;langSet xml:lang=&quot;fr&quot;&gt;</td>
<td></td>
</tr>
<tr>
<td>&lt;tig id=&quot;C003fr1&quot;&gt;</td>
<td></td>
</tr>
<tr>
<td>&lt;term&gt;</td>
<td></td>
</tr>
<tr>
<td>poulet</td>
<td></td>
</tr>
<tr>
<td>&lt;/term&gt;</td>
<td></td>
</tr>
<tr>
<td>&lt;termNote type=&quot;partOfSpeech&quot;&gt;</td>
<td></td>
</tr>
<tr>
<td>noun</td>
<td></td>
</tr>
<tr>
<td>&lt;/termNote&gt;</td>
<td></td>
</tr>
<tr>
<td>&lt;termNote type=&quot;grammaticalGender&quot;&gt;</td>
<td></td>
</tr>
<tr>
<td>masculine</td>
<td></td>
</tr>
<tr>
<td>&lt;/termNote&gt;</td>
<td></td>
</tr>
<tr>
<td>&lt;/langSet&gt;</td>
<td></td>
</tr>
<tr>
<td>&lt;langSet xml:lang=&quot;en&quot;&gt;</td>
<td></td>
</tr>
<tr>
<td>&lt;tig id=&quot;C003en1&quot;&gt;</td>
<td></td>
</tr>
<tr>
<td>&lt;term&gt;</td>
<td></td>
</tr>
<tr>
<td>chicken</td>
<td></td>
</tr>
<tr>
<td>&lt;/term&gt;</td>
<td></td>
</tr>
<tr>
<td>&lt;termNote type=&quot;partOfSpeech&quot;&gt;</td>
<td></td>
</tr>
<tr>
<td>noun</td>
<td></td>
</tr>
<tr>
<td>&lt;/termNote&gt;</td>
<td></td>
</tr>
<tr>
<td>&lt;/langSet&gt;</td>
<td></td>
</tr>
</tbody>
</table>

**Fig. 1.** Comparison of a TBX-Basic termEntry and its corresponding TBX-Min entry.

Note as well that TBX-Min does not have elements for all of the data categories seen in TBX-Basic. As a result any information about data categories not available in TBX-Min has been rendered using the <note> element in the TBX-Min example, as is the case with the information about grammatical gender. The content of this element should be displayed to the translator, who can use it for guidance.

Figure 2 shows a small but complete TBX-Min file with two term entries (one of which corresponds to that shown in Figure 1). In it example, the simplicity of the format is clear.
Lossless conversion of the data in TBX-Min to a tabular representation for viewing is straightforward, accomplished by mapping the individual elements within each termGroup element contained in a langGroup element to specific columns. Because a langGroup element can contain more than one termGroup element, as can be seen with the termGroups for chick peas and garbanzo beans, 1-to-n TBX-Min entries converted to tabular formats require that information be duplicated across rows (e.g., repeating the information about pois chiches in rows for chick peas and garbanzo beans to indicate that both have the same French translation). N-to-n cases (e.g., where a source langGroup contains three synonyms stored in termGroup
elements and the target langGroup has two synonyms) are more complex and require special attention.

While TBX-Min is not a tabular format, its logical structure corresponds quite closely to the spreadsheet glossaries used by translators. Because of its simple and predictable structure it is much easier to implement than previous TBX variants. As an XML format its semantics are clear and it can more readily be integrated into modern translation workflows and tools than can spreadsheets. It avoids the problems caused by the lack of standardization of CSV and because it uses the default XML encoding of UTF-8, problems with variant encodings can be avoided. Terminology disseminated in TBX-Min can be easily converted to UTX or viewed in a tabular format as needed by translators. In addition, terminology stored in TBX-Min (or converted to it) can be easily integrated into TBX-based structured terminology resources, providing a growth path for individuals interested in migrating from simple spreadsheet formats to more robust and complex terminology management solutions. If a particular TEnT already supports TBX-Basic, TBX-Min can be automatically converted to TBX-Basic using a free and open-source utility and imported as TBX-Basic. If a TEnT does not currently support TBX-Basic, a TBX-Min import/export feature can be implemented with a modest expenditure of software developer resources. Implementing TBX-Min in a TEnT does not preclude subsequent support for TBX-Basic or other DCA-style TBX dialects.

3.1 TBX-Min Structure

Because TBX-Min did not evolve directly from any other TBX or XML dialect, it does not have certain historical artifacts such as the <martif> element found in TBX-Basic (not shown here, but see [2] for details), and it was possible to design it so that it is immediately apparent what information a given element contains. The combination of DCT, a minimal feature set, and succinct and intuitive element naming makes TBX-Min documents very readable. The hope is that the minimal and intuitive structure of the TBX-Min dialect will encourage its proliferation among both end-users and implementers. Note that TBX-Min files are strictly bilingual.

The structure of TBX-Min is as follows (required elements/attributes in bold):

- The root element, **TBX**, contains a **header** element and one or more **entry** elements. It has a **dialect** attribute that distinguishes the dialect (and allows TBX-Min files to be distinguished from other DCT-style TBX-compliant files).
- The **header** can contain all of the information a UTX file contains in its header as optional elements: author, ID, date, description, directionality, license, and languages.
- The **entry** elements contain a **subjectField** and one or more **langGroup** elements.
- A **langGroup** element contains **termGroup** elements. It also has a mandatory **xml:lang** attribute that defines the language for the entry.
- The **termGroup** element contains a required **term** element and the following optional elements: note, status, customer, and partOfSpeech.
4 Interfacing with Other Formats

To facilitate adoption of TBX-Min, the TBX development team has provided a number of resources at http://tbxinfo.net. All utilities described in this section are available from links available at this URL. Included at this site are converters to and from other formats (UTX and TBX-Basic), validators, and documentation of the format. As it is anticipated that implementers of TBX-Min will need to interact with other formats, this section provides an overview of how to deal with other formats, using the resources at the TBX-Min converter page where appropriate.

4.1 Converting UTX to TBX-Min

The mapping between UTX and TBX-Min is straightforward (see the TBX-Min resource page for details). Note that a subset of UTX has been implemented in the XLIFF:doc format (see http://interopability-now.org), and thus it should also be straightforward to convert between glossaries stored in XLIFF:doc files and TBX-Min.

4.2 Converting Spreadsheets to TBX-Min

The authors will provide a Perl tool to convert spreadsheet glossaries into TBX-Min, provided they follow certain format requirements (a totally generic converter is not possible since the column semantics of arbitrary tabular formats cannot be known in advance). The converter reads in a tab delimited UTF-8 glossary pre-configured with TBX-Min-specific column headings. The conversion process is similar to the UTX conversion process.

4.3 TBX-Basic to TBX-Min

The TBX development team has also created a Perl tool to extract TBX-Min glossaries from TBX-Basic files. After specifying the source language and target language and a TBX-Basic file, it extracts the corresponding TBX-Min file. Considerable information is lost in the process since TBX-Min cannot represent all aspects of a TBX-Basic file. A log file informs the user of what information is converted into note elements (such as all of the unsupported data categories that appear at the termGroup level) and what is simply ignored (such as notes at levels other than termGroup).

4.4 TBX-Min to TBX-Basic

Conversion from TBX-Min to TBX-Basic is quite straightforward, although there is some data loss because the TBX-Min header was designed with UTX in mind. Those elements that are unsupported in TBX-Basic are placed in the TBX
sourceDesc element. The ID is turned into the title, since the title is the closest thing to a uniquely identifying string.

Although not currently supported, development is planned to allow multiple TBX-Min files to be combined into one TBX-Basic file, essentially reversing the process by which the TBX-Basic to TBX-Min converter separates bilingual files from multilingual TBX-Basic files. Another planned development is a viewing utility that will permit a translator or project manager who receives a TBX-Min file but does not have access to a TEnT that already supports TBX-Min to view the information without looking directly at XML.

5 Conclusion

TBX-Min provides a simple and straightforward XML representation for basic terminological data of the sort commonly exchanged in spreadsheets. It provides an easy entry point for freelance translators to access and utilize TBX data without the need to invest in tools that support the full range of TBX functionality. A variety of tools will assist implementers to use this simple format.

6 References

A standard TMF modeling for Arabic patents

Chihebeddine Ammar, Kais Haddar, and Laurent Romary

1 Department of Computer Science, University of Sfax, Laboratory MIRACL, Route de Soukra, B.P. 1171, Sfax, 3000, Tunisia
ammarchihebeddine@hotmail.com kais.haddar@fss.rnu.tn

2 INRIA & Humboldt Universität zu Berlin, Institut für Deutsche Sprache und Linguistik
laurent.romary@inria.fr

Abstract. Patent applications are similarly structured worldwide. They consist of a cover page, a specification, claims, drawings (if necessary) and an abstract. In addition to their content (text, numbers and citations), all patent publications contain a relatively rich set of well-defined metadata. In the Arabic world, there is no North African or Arabian Intellectual Property Office and therefore no uniform collections of Arabic patents. In Tunisia, for example, there is no digital collection of patent documents and therefore no XML collections. In this context, we aim to create a TMF standardized model for scientific patents and develop a generator of XML patent collections having a uniform and easy to use structure. To test our approach, we will use a collection of XML scientific patent documents in three languages (Arabic, French, and English).

Keywords: TMF, Operability, Patents, Terminological databases

1 Introduction

Works on how to define a database’s standard models are abundant in literature, especially in fields such as data warehouse. A standardized modeling terminological databases consists in integrating, homogenizing and giving terminology data a unique sense understandable by all users. It provides us a tool to integrate and merge terminology data from multiple source systems while improving terminology data quality and maintaining maximum interoperability between different applications. There are several standardized modeling terminological databases: TMF [4] and [5], TEI [11], etc.

One of the very rich in terminology work streams are the scientific patents. They are similar, for example, to a scale repository. They also cover several scientific and technical fields, while offering rich interdisciplinary relations. That is why we will need several terminological databases, one for each field.

In fact, scientists inventors are the best to present the technical words of a field. Since, when drafting their patent applications, they will carefully choose words and named entities of a specific domain. In addition, patents may contain extreme examples of noise, deliberately vague and misleading wording for the
title, abstract and claims while maintaining relatively standard technical terminologies in the body description of the patent. But on the other side, in the same field, there is a risk that terms will be represented in different ways from one patent to another.

Indeed, standardized modeling patent allows us to maintain a standard for the representation of texts in digital form, so that we protect patents data by bringing them in digital databases. It will provide a single common data model for all terminological data regardless of the data’s language, source, field, etc. Also, we will be able to build collections of uniform patents which facilitate the extraction and the exploitation of patents data and the extraction of links between valid terms. Standardized modeling patent ensure also interoperability between applications. Finally, it will allow us to easily enrich other terminological databases.

Another motivation was to decide which standard will we choose to model our terminological databases, which standard will best represent patents terms and which approach to use, onomasiological or semasiological approach?

Patents are available in different formats: Full text, PDF document, set of images, XML, etc. They have heterogeneous components that require different modelings. Also, patents have linguistic structures like text and titles, and non-linguistic structures like figures, citations, tables and formulas. In fields such as mechanics, automatic extraction based only on the text will fail.

In addition to the text, figures and citations information, all patent publications contain a relatively rich set of well-defined metadata. These metadata are often found in the cover page of patents and titles of figures and tables. To cope with the large volume of data and metadata, we will develop a patents terminological editor to generate terminological databases. This allows us to develop heuristics, based on metadata such as the applicant(s) name(s), the inventor(s) name(s) or priority documents, etc, for finding interesting documents.

The structure of the XML documents may be used for the processing performed to differentiate various elements according to their semantic. Thus, a section title, a summary, bibliographic data, or examples can be used to identify different aspects of the text. Indeed, scientific patents can be easily processed as XML documents. So we can treat their structures\(^1\) as a source of information.

We propose in this paper to treat the problem of extraction and operating information from a collection of Arabic scientific patents. In order to achieve this, we will propose a standardized model for multilingual patents and generate terminological databases from patent collections. It is an original idea because nobody treated terminology in Arabic patents in previous works.

Our learning collection includes a small number of multilingual patent documents. Each patent is associated with one of the three languages: Arabic, English, and French. In this paper, we will focus on the Arabic patents. Some of them have their translation in one (or two) of the other languages. Others have a translation of technical words or keywords of the invention and even a literal

\(^{1}\) Remind that an XML document is structured as a tree consisting of hierarchical elements which may have one or more attributes, the leaf nodes have information.
A standard TMF modeling for Arabic patents

This article is organized as follows. Section 2 is devoted to the presentation of the previous works. In section 3, we present our TMF standardized model for patents. Section 4 is devoted to the evaluation and discussion and we conclude and enunciate some perspectives in section 5.

2 Previous works

Information retrieval techniques in multilingual patents are not lacking in previous works, the question is whether the results of this works remain valid if one expands the collection by documents into other languages (Arabic, for example), and if they will be affected by changing the type of the documents collection, calculating noise, redundancy, cost, precision, recall, silence, etc.

2.1 CLEF initiative

The most recent previous works have been performed under the CLEF initiative on non-Arabic patents. The CLEF [12] initiative 2000-2014 (Conference and Labs of the Evaluation Forum, formally known as Cross-Language Evaluation Forum) promotes research and stimulates development of multilingual and multimodal IR systems for European languages. It provides tracks to evaluate the performance of systems for: for example, from 2009, the intellectual property: The aim is to encourage and facilitate research in the field of data mining in patents by providing a large database of experimental data. This database is formed of patent documents from the European Patent Office and it is called MAREC (MAtrixware REsearch Collection) which is a standard corpus of patent data available for research purposes. It consists of 19 million of patent documents in different languages (English, French, German) in a standardized XML schema highly specialized.

Previous works [2], [3] and [8] were mainly based on purely statistical approaches. They used standard techniques of information retrieval and data extraction. But there are others who have worked on purely linguistic or hybrid approaches. [6] and [7] used claims section as a bag of words and information source. In [9], the authors developed multilingual terminological database called GRISP covering multiple technical and scientific fields from various open resources.

2.2 Comparison between various Intellectual property offices

Patent applications are similarly structured worldwide. They consist of a cover page, a description, claims [1], drawings (if necessary) and an abstract.

The cover page of a published patent document usually contains bibliographic data such as the title of the invention, the filing date, the priority date, the names and addresses of the applicant(s) and the inventor(s). It also has an
abstract, which briefly summarizes the invention, and a representative drawing. Bibliographic data are extremely useful for identifying, locating and retrieving patent documents. The patent description must describe the claimed invention and give technical informations. The claims determine the patentability and define the scope of the claimed invention.

The European Patent Office (EPO) [14] offers inventors a uniform procedure of application, and a register of multilingual patents (English, French, German). In the Arabic world, there is no North African or Arabian Intellectual Property Office and therefore no uniform collections of Arabic patents. In Tunisia, for example, the INNORPI [13] (National Institute for Standardization and Industrial Property) does not propose a digital collection of patent documents and therefore no XML collections.

As a result, Arabic patents have no unique structure. For the Tunisian patents, the cover page doesn’t have abstracts and patent documents could be in one of the three languages (Arabic, English or French). In the regional office\textsuperscript{2} for the Gulf Cooperation Council (GCC Patent Office) [15], there is only Arabic patents and there is an Arabic abstract in the cover page. The layout of the description part varies also from place to place. For example, the summary and the background of the invention could not exist in some patent descriptions. The Tunisian patents themselves have no unique structure in that some of them have no abstract, have missing bibliographic data and even no cover page. For these reasons, a normalization phase for Arabic patents (e.g. Tunisian patents) is necessary.

3 TMF standardized model

The onomasiological terminological resources, usually go from one sense to the various embodiments of the term in different languages. Classic examples of onomasiological dictionaries are thesaurus, synonym dictionaries, etc. The terminology is interested in what the term means: notions, concepts, and words or phrases that nominate. This is the notional approach. Ideally, a concept corresponds to one term and a term corresponds to a concept. Motivated from the industrial practice terminology, the Terminology Markup Framework (TMF, ISO 16642) was developed as a standard for these resources. This Standard also allows the modeling of lexical resources, but also contains a serialization, in this case, an XML format. So if we have onomasiological resources such as glossaries, thesaurus or words networks, we refer to TMF (ISO 16642) modelling.

In this paper, we treat patents as networks of terms and citations. So we consider that TMF is the most appropriate for patent modeling.

The meta-model (Fig. 1) of TMF is defined by a logical hierarchical levels. It thus represents a structural hierarchy of the relevant nodes in linguistic description. Each structural node or level can be described using basic or complex unit of information. The meta-model describes the main structural elements and

\textsuperscript{2} Certificates of Patents granted by the GCC Patent Office secure legal protection of the inventor’s rights in all Member States.
A standard TMF modeling for Arabic patents

their internal connections. It is combined with data categories [10] from a data category selection (DCS). And finally, this model is matched with a model defined by the user. So we can appropriate the model according to our needs. The aim of the meta-model is to act as a reference regarding possible interoperability requirements. So it defines a principle of interoperability between two Terminological Mark-up Languages which guarantees equivalence since they are based on the same set of data categories. This principle is guaranteed thanks GMT (Generic Mapping Tool) (Fig. 2) advocated by TMF to allow passage between two TMLs.

In the following, we will present our TMF standardized model for bibliographic and application terminology. The structure of the patent can be divided into two parts: bibliographic data taken from the cover page and application data from the rest of the patent document. Fig. 3 shows the class diagram of
Fig. 3. The class diagram of patent bibliographic data.

Fig. 4. The class diagram of patent application data.
the patent bibliographic data in which all associations are a strong composition associations. It contains, Bibliographic Data class which includes the Filing Number and Date, the Publication Date and the Language and Type of the patent. Bibliographic Data object is associated with one or more Title of Invention in different languages, zero or more Priority patent applications, one or more Inventor(s) and Applicant(s), zero or one Representative and zero or more Internation Publications (PCT).

The Fig. 4 above shows the class diagram of the patent application data in which all associations are also a strong composition associations, because, if a composite is removed, all of its component parts will be removed with it. It presents, the association of the Invention class with one or more Abstract in different languages, one Claims and Description parts and zero or one Drawings part. The two above presented diagrams allow us to introduce a DTD for scientific Tunisian patents. The DTD given in the following is much more simple and basic than European patents DTD. In fact, European patents contain more bibliographic data.

```xml
<!ELEMENT Invention (Bibliographic_Data, Description, Claims, Drawings?, Abstract+) >
<!ATTLIST Invention
  State CDATA #IMPLIED
  Lang (AR|FR|EN) "AR"
  File CDATA #REQUIRED >
<!ELEMENT Bibliographic_Data (Title_Invention+, Priority*, PCT?, Applicant+, Inventor+, Representative?) >
<!ATTLIST Bibliographic_Data
  Num_Filing CDATA #REQUIRED
  Date_Filing CDATA #REQUIRED
  Date_Publication CDATA #REQUIRED
  Lang (AR|FR|EN) "AR"
  Type CDATA #REQUIRED >
<!ELEMENT Title_Invention ( #PCDATA ) >
<!ATTLIST Title_Invention Lang (AR|FR|EN) "AR" >
<!ELEMENT Priority EMPTY >
<!ATTLIST Priority
  Priority_Num CDATA #REQUIRED
  Priority_Date CDATA #REQUIRED
  Priority_State CDATA #REQUIRED >
<!ELEMENT PCT EMPTY> 
<!ATTLIST PCT
  Application_Num CDATA #REQUIRED
  Application_Date CDATA #REQUIRED
  Publication_Num CDATA #REQUIRED
  Publication_Date CDATA #REQUIRED >
<!ELEMENT Applicant EMPTY>
<!ATTLIST Applicant
  Applicant_Num CDATA #REQUIRED
  Applicant_Date CDATA #REQUIRED
  Applicant_State CDATA #REQUIRED >
```

163/228
A terminological record is a structured presentation that allows us to provide all information relating to a term in a clear and orderly way. Our terminological record includes: the entry identifier, subject field, definition, term, sub-term, synonym, example and abbreviation. Fig. 5 shows an example of bibliographic terminological entry (Title of invention) in the form of an XML document conforming GMT in the three languages (French, Arabic and English).

4 Evaluation and Discussion

Our main obstacle is that the structure of patents differs from an intellectual property office or institute to another in the Arabic world. The cover page of a Tunisian patent differs from the Egyptian or Moroccan patent cover page. We conducted a TMF modeling for multilingual patents based on the forms of patents published in the Arabic world in general and precisely in Tunisia.

We did not have a collection of document in digital form because it is not the official in Tunisia for example. So we created our small collection of multilingual patents from various fields to generate our terminology database.

To cope with the large volume of patents data and metadata, we developed a patents terminological editor to automatically generate terminological databases.
A standard TMF modeling for Arabic patents

Fig. 5. Terminological entry in the form of an XML document conforming GMT.

This will enable us to facilitate the extraction and information retrieval tasks from the cover pages (metadata), and the other parts (data) of patents. The results of our terminological database are presented in Table 1. It concerns Tunisian and Gulf Arabic patents and it can be easily merged with other terminological databases. We hope that our terminology database will improve patent search.

<table>
<thead>
<tr>
<th>Collection</th>
<th>Number of Patents</th>
<th>Number of terms</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Full text</td>
</tr>
<tr>
<td>INNORPI</td>
<td>28</td>
<td>6924</td>
</tr>
<tr>
<td>GCCPO</td>
<td>30</td>
<td>7632</td>
</tr>
</tbody>
</table>

Our terminological database contains terms of different technical and scientific fields and various patents with different structures. We can distinguish two categories of terms; the scientific and technical terms and the other terms. Scientific and technical terms in their turn were divided according to their technical and scientific fields.
5 Conclusion

In this paper, we have proposed a TMF modeling for Arabic patents which provide us a single common data model for all terminological data and we developed a patents terminological editor to automatically generate terminological databases. In future, we plan to enlarge our patents collection and then our terminological database. Patent documents are often difficult to understand and have a variety of structures. So, we aim to develop a patent editor which automatically generates a collection of XML patent documents having a similar structure. It will facilitate the task of terms and keywords extraction. We will merge several terminology databases of patents. We aim to better extract information from a collection of multilingual scientific patents and to combine onomasiological and semasiological models. We are also developing a new annotation procedure, to annotate our learning and test collections.

References

1. Hong, S.: Claiming what counts in business: drafting patent claims with a clear business purpose, SMEs Division, WIPO
3. Lopez, P., Romary, L.: Experiments with Citation Mining and Key-Term Extraction for Prior Art Search, In CLEF 2010 Workshop. Padua Italy (2010)
TermWise: Leveraging Big Data for Terminological Support in Legal Translation

Kris Heylen, Stephen Bond, Dirk De Hertog, Hendrik Kockaert, Frieda Steurs, and Ivan Vulić

QLVL, Linguistics Department (KU Leuven), Belgium
{kris.heylen, stephen.bond, dirk.dehertog, hendrik.kockaert, frieda.steurs}@kuleuven.be

LIIR -Department of Computer Science (KU Leuven), Belgium
ivan.vulic@cs.kuleuven.be

University of the Free State, Bloemfontein, South Africa

Abstract. Increasingly, large bilingual document collections are being made available online, especially in the legal domain. This type of Big Data is a valuable resource that specialized translators exploit to search for informative examples of how domain-specific expressions should be translated. However, general purpose search engines are not optimized to retrieve previous translations that are maximally relevant to a translator. In this paper, we report on the TermWise project, a cooperation of terminologists, corpus linguists and computer scientists, that aims to leverage big online translation data for terminological support to legal translators at the Belgian Federal Ministry of Justice. The project developed dedicated knowledge extraction algorithms and a server-based tool to provide translators with the most relevant previous translations of domain-specific expressions relative to the current translation assignment. In the paper, we give an overview of the system, give a demo of the user interface and then discuss, more in general, the possibilities of mining big data to support specialized translation.

Keywords: Legal Terminology, Automatic Knowledge acquisition, Big Data, Context Sensitive Suggestion

1 Introduction

Translators in specialized domains are confronted with source texts that are teeming with highly specific terminology and other domain-specific expressions. Even the most experienced of translators regularly needs to check the translation of such expressions against a reliable resource. Although (online) specialized dictionaries and state-of-the-art Computer Assisted Translation (CAT) tool offer some terminological support, the coverage of Translation Memories (TM), Term banks and Term Bases is often insufficient. Typically, translators turn to online collections of bilingual documents and search these with a general-purpose search engine (see [1] for a discussion of typical search behavior). However, finding relevant examples is often hard and time-consuming and the

* TermWise was funded by KU Leuven IOF grant KP/09/001. Special thanks to Martine Perpet of the FOD Justitie.
reliability of online sources is not always guaranteed. In this paper we present the outcome of the TermWise project, which tries to leverage big online collections of bilingual documents to offer additional terminological support for domain-specific translation in a user-friendly way. The TermWise project adds an extension to existing CAT-tools in the form an extra cloud-based database, which we call a Term&Phrase Memory. It provides one-click access to translations for individual terms and domain-specific expressions that stem from known, trustworthy online sources and that are sorted for relevance to the translator’s current translation assignment. The Term&Phrase Memory has been compiled by applying newly developed statistical knowledge acquisition algorithms to large parallel corpora harvested from official, public websites. Although these algorithms are language- and domain-independent, the tool was developed in a project with translators from the Belgian Federal Justice Department (FOD Justitie/SPF Justice) as end-user partners. Therefore the tool is demonstrated in a case study of bidirectional Dutch-French translation in the Belgian legal domain. In this paper, we first describe the specific needs that our end-user group expressed and how we translated them into the new functionality of the Term&Phrase Memory. Next, we summarize the term extraction and term alignment algorithms that were developed to compile the Term&Phrase Memory from large parallel corpora. Section 4 describes how the Term&Phrase Memory functions as server database that is now, in this proof-of-concept phase, accessed via a lightweight stand-alone tool, but that is designed to be fully integrated with a CAT user-interface so as to provide context-sensitive terminological support in the normal translation work-flow. In Section 5, we present the user-based evaluation of the tool that was carried out by students of Translation Studies and professional translators at the Belgian Federal Justice Department. Section 6 concludes with a discussion of how TermWise is an example of a dedicated linguistic search tool that allows translators to exploit Big Data that takes the form of large online bilingual document collections.

2 User needs of Legal Translators

Like other domain-specific translators, the translators at the Belgian Ministry of Justice are confronted with source texts full of domain-specific terminology which requires exact (as opposed to interpretative) translation and which even skilled translators need to check against a reference source once in a while. However, existing (online) Belgian legal dictionaries have limited coverage and are outdated. Also in the commercial CAT-tool used by the Ministry, the support for terminological look-up is quite limited. As with most CAT-tools, it does come with a Term Base functionality, but this type of terminological dictionary is initially empty and entries have to be added manually. Even a large organization like the Ministry cannot afford to invest much time in Term Base compilation. They acquired an externally compiled Term Base, but its coverage is limited and it contains no informative examples of the idiomatic usage of terms in contexts. Such proper phraseological usage of terms is especially important in legal language, where validity of a text depends on the usage of the appropriate formulae. Although the commercial tool’s Translation Memory (TM) automatically gives translation suggestions, its retrieval on the level of entire sentences or even paragraphs is too coarse-grained for finding examples of individual words and phrases. A concor-
dancer does allow for a manual look-up of a specific expression, but occurrences are not sorted for relevance, nor do they come with meta-data about the source document that could allow translators to assess its relevance and reliability. Additionally, the TM only keeps track of the Ministry’s in-house translations, and does not include the vast body of relevant bilingual legal documents translated at other departments. The translators therefore often resort to doing Google searches for terms and phrases in open on-line legal document repositories to check previous translations in specific contexts. However, also here, the relevance of the search hits must be assessed manually. Based on this situation, we identified the following user needs:

- Access to previous translations of domain-specific single and multi-word expressions
- Examples of usage in context to infer correct phraseology
- Information about the source documents of the translation examples
- Examples from all relevant documents that are available online
- Sorting the examples by relevance to the current translation assignment
- Easy access to the examples from within the CAT-tool

To our knowledge, this combination of functionalities is not implemented in any existing CAT-tool [12]. In TermWise they are grouped in a separate module, which we will call a Term&Phrase Memory, so that in principle this module can be integrated in existing CAT-tools. However, commercial CAT-tool developers do not readily allow plug-ins by third parties. Also, the focus of the TermWise project was to deliver a proof-of-concept for the Term&Phrase Memory’s functionality, not to develop a fully functional CAT-tool. Therefore, we opted to implement a stand-alone, lightweight tool to showcase the new functionality of the Term&Phrase Memory, but in such a way that it can easily interact with the current commercial CAT software of the Belgian Ministry of Justice. In the next section, we discuss which type of information is included in the Term&Phrase Memory and how it was compiled. Section 4 describes the user interface.

3 Corpus and Knowledge Acquisition

A number of official bilingual legal document collections are put online by the Belgian Federal Government (e.g. juridat4, De Kamer/La Chambre5) but for our case study, we focused on the largest collection, viz. the online version of the Belgian Official Journal (Belgisch Staatsblad/Moniteur Belge6), which publishes laws, decrees, and official communications of both the national state and the federal entities, in both French and Dutch. We implemented a web crawler in python to systematically download all the issues to our server. For the case-study, we only use the issues from 1997 to 2006 because they have been published as a curated, open-source corpus (100M words)[18]7. However, in a next stage, the aim is to continually update the corpus with new issues.

4http://www.cass.be/
5http://www.dekamer.be/
6http://www.ejustice.just.fgov.be/cgi/welcome.pl
7http://opus.lingfil.uu.se/MBS.php
All issues were language-checked\(^8\), tokenized and POS-tagged\(^9\), and sentence-aligned\(^{10}\) with publicly available tools. The webcrawler also retrieved the source department (e.g. ministry, agency) for all documents. Both the Dutch and French corpus were POS-tagged with TreeTagger. To extract domain-specific expressions and their translations, we followed the *extract-then-align* paradigm that is predominant in the literature on bilingual terminology extraction (e.g., see [3]; [6]; [5]; [8]; [10]). In this paradigm, terms are first extracted for the two languages separately and then in a second step aligned cross-lingually. Although both tasks are well known in NLP and have many existing implementations, most current tools are geared towards delivering intermediate results for a Machine Translation system or further manual lexicon compilation. In the Term&Phrase Memory, however, the output has to be usable directly by end-users. We therefore developed our own statistical algorithms for term extraction and term alignment to accommodate the specific user needs above. The knowledge acquisition proceeded in two steps.

**STEP 1: Domain-Specific N-gram Extraction**

Following [9], we consider expressions of variable length as relevant for the legal domain. These do not only include single and multi-word terms that refer to legal concepts (typically NPs), but also phraseologies (e.g. typical verb-NP combinations), and formulaic expressions that can comprise entire clauses. The term extraction algorithm therefore considers n-grams of variable length without imposing predefined language-specific POS patterns as is the case in most term extraction algorithms. Instead, the relevance of an n-gram is assessed based on its external *independence* and its internal *coherence*. Independence is the extent to which an n-gram can occur in different contexts. Following [16], this is operationalized as a maximazation of frequency differences relative to the n-1 and n+1 grams in an n-gram expansion progression. Coherence is the extent to which the lexemes within an n-gram tend to co-occur in an informational unit. This is measured as the Mutual Information of the n-gram’s POS-sequence. The algorithm is described in more detail in [4]. Note that the expressions extracted do not necessarily correspond to theoretically motivated, concept-based terminological units, but rather to domain-specific expressions in general that are of practical use to a translator. The extraction step resulted in a list of 649,602 n-grams for French and 639,865 n-grams for Dutch.

**STEP 2: Bilingual N-gram Alignment**

The goal of the alignment step was to provide for each Dutch n-gram a ranked subset of likely translations from the French n-grams list and vice versa. To build these ranked subsets, we developed a statistical algorithm for bilingual lexicon extraction (BLE) from parallel corpora, called SampLEX, and adapted it to handle n-grams of variable length. In a pre-processing step, the aligned sentences in the corpus are represented as a bag-of-terms taken from the French and Dutch input lists. SampLEX uses a strategy of data reduction and sub-corpora sampling for alignment. For more details about the algorithm

---

\(^8\) TextCat: http://odur.let.rug.nl/~vannoord/TextCat/
\(^9\) TreeTagger [15]
\(^{10}\) Geometric Mapping and Alignment system [11]
and its properties, and benchmarking against other BLE models, we refer the reader to [20]. Running SampLEX results for each Dutch n-gram in the list of French n-grams sorted by translation probability and vice versa. Also, the document and sentence ID of each occurrence of a candidate translation-pair in the corpus is returned. As a post-processing step, a hard cut-off of the output ranked lists of translation candidates is performed. Some example output is displayed in Table 1.

<table>
<thead>
<tr>
<th>sur la proposition du conseil d’administration</th>
</tr>
</thead>
<tbody>
<tr>
<td>op voorstel van de raad van bestuur</td>
</tr>
<tr>
<td>op voordracht van de raad van bestuur</td>
</tr>
<tr>
<td>16 mai 1989 et 11 juillet 1991</td>
</tr>
<tr>
<td>16 mei 1989 en 11 juli 1991</td>
</tr>
<tr>
<td>sur la proposition du ministre</td>
</tr>
<tr>
<td>de voordracht van de minister</td>
</tr>
<tr>
<td>op voorstel van de minister</td>
</tr>
<tr>
<td>op voordracht van de minister</td>
</tr>
<tr>
<td>...</td>
</tr>
</tbody>
</table>

Table 1. Example output of the SampLEX algorithm for n-grams. Translation direction is French to Dutch.

4 Context-sensitive Database Querying

The Term&Phrase Memory is conceived to function as an additional database accessible from within a CAT-tool’s user-interface, next to the Translation Memory and Term Base. As with terms contained in a manually crafted Term Base, the terminological expressions included in the Term&Phrase Memory are highlighted in the source text of the translator’s new assignment. By clicking on them, their previous translations-in-context are shown in a separate pane. Figure 1 illustrates this for the expression méthodes particulières de recherche in segment 5 of a Belgian-French legal document. The examples are ranked by relevance, defined as the similarity of their respective source documents to the current source text. The meta-data of the examples’ source documents (e.g. issuing ministry or agency, state or federal level) and a link to the online version is also provided, both in html and pdf. This way, the user can assess the relevance and reliability of the translation’s source. If the user agrees with a suggested translation, a button click copies it to the active segment in the target text pane.

Although the Term&Phrase Memory is meant to be integrated into a CAT tool, in the current test phase, it is implemented as a stand-alone tool. However, to make the tool easily usable next to a CAT tool, it is possible to upload the xli file that CAT tools use to store translation projects in a segmented format. This makes sure that the segmentation of the source text in the TermWise tool is compatible with the one in CAT
tool. A translator can easily navigate from segment to segment and then copy-paste translation examples from the TermWise tool to the CAT Tool.

Figure 2 shows the architecture behind the TermWise tool. The system consists of a server, which handles translation requests, and a client, which issues the requests and displays the returned results in a GUI. When handling a translation request, the server takes as input a xlfiff-file or plain txt file and returns an XML file containing the segmentized document, translation suggestions for each segment, the n-grams found in the document, and translation suggestions for each n-gram together with context-sensitive annotated usage examples. The translation suggestions for segments correspond to the fuzzy matches from Translation Memories in traditional CAT-tools, but in this case the entire online document collection of the Belgisch Staatsblad/Moniteur Belge functions as a TM. The fuzzy matching algorithm is similar to that in existing software will not be further discussed here. Instead we will focus on handling of n-grams for the new Term&Phrase functionality.

The Term&Phrase Memory consists of (a) a list of paired, sentence-aligned documents from the Belgian Official Journal annotated with their source department, and (b) a dictionary of the n-grams found in those documents. In the latter, each n-gram is associated with a list of translation candidates of a given translation probability, and each n-gram translation pair is associated with the list of documents and line numbers in which that translation is found.

When the server receives an input document in xlfiff format the segmentation is checked. If it is in plain txt, it is first segmentized using the Alpino tokeniser [17]. N-grams are extracted from the segmentized input document by consulting the n-gram dictionary of the same language. A ranked list of similar corpus documents and their

Fig. 1. Screen cap of TermWise GUI with n-grams highlighted in the source text and translation examples displayed in the Term&Phrase Memory pane
respective source departments is retrieved by calculating the number of n-grams in common with the input document.

N-gram translations to be suggested are chosen on the basis of the given translation probabilities and on document similarity. The list of documents that are similar to the input document is compared with the list of documents for each n-gram translation pair. The relevance value for an n-gram translation pair is determined by a weighted interpolation of its given translation probability and the cosine similarity of the highest-ranking document on its list (based on a “set of n-grams” vector space model). If the relevance value exceeds a configurable threshold, that n-gram translation pair is displayed and suggested to the user. Example sentences are extracted from the highest-ranking document and from other high-ranking documents from the same source department.

5 Evaluation

The TermWise tool is evaluated by two end-user groups. In December 2014, 19 students of legal translation at the KU Leuven, campus Antwerp were made acquainted with the tool and then asked to translate a legal document from French into their native Dutch with the help of the TermWise tool alongside SDL Trados Studio 2011 that had the legal Translation Memory and Term Base of Belgian Federal Justice Department loaded. More specifically, the students were asked to record all the expressions in the source text that they normally would look up outside of the CAT tool and report whether they were present in the TermWise tool. The result are shown in Figure 3. Although not all desired expression were covered, students reported significant gains in look-up time.

Currently, seven professional translators at the Belgian Ministry of Justice are assessing the usability of the tool in their daily translation practice. First, legal translators are invited to make use of the tool to translate an unseen legal text and give comments and feed-back on the Term&Phrase Memory functionality and coverage as they are
translating. Afterwards, they are also asked to fill in a survey on the general usability of the tool and the new functionality it offers. Results are expected by September 2014. The results of this qualitative evaluation will be used to improve the tool’s user-friendliness and to fine-tune the parameters of the knowledge acquisition algorithms and the context-sensitive search function.

Fig. 3. Evaluation results with students

6 Big Data for Translation

Big Data is a buzz word in ICT in general and also in the translation industry. Discussions on the opportunities that big parallel corpora offer for translation, usually focus on three aspects:

1. Sharing translation memories as open data (e.g. [13])
2. More data to improve (statistical) Machine Translation (e.g. [2], p. 60)
3. More data to improve term extraction for the compilation of multilingual term bases or ontologies (e.g. [7])

However, these approaches deal with derived products (TM’s, MT systems or Term ontologies) and do not acknowledge that the translators themselves might want to exploit the data directly to help them in their translation process. Actually, professional translators are often very good at assessing applicability of a translation by comparison to previous examples and only resort to dictionaries when real conceptual confusion is at stake (for which good terminological work is still crucial). However, translators do need support to find translation examples that are informative and relevant to their current

---

11 For an overview of the exploitation of comparable corpora, see [14].
translation assignment in the deluge of available parallel data. Additionally, meta-data about the source of a previous translation is crucial to assess the reliability and appropriateness of the example. Clearly, general search engines like Google are not optimized for this type of linguistic search, but also parallel corpus search tools like Linguee only allow for the context-insensitive look-up of expressions that do not take into account the specific assignment that a translator is working on. The TermWise project aims precisely to combine access to large and constantly expanding online bilingual document collections with support for highly context-specific translation needs. More specifically, the Term&Phrase Memory functionality presented in this paper improves over current practices in the following ways:

- Highly domain-specific expressions are identified for the translator, whereas in concordance searches in current CAT and corpus search tools, translators have to select expressions for look-up themselves. Thanks to the dedicated term extraction algorithm, these expressions go beyond traditional noun phrases and include phrasemes and typical formulae.
- Moreover, the domain-specific expressions have already been looked up for the translator beforehand as the source text is submitted to a pre-search when it is uploaded to the tool. The translator just has to click the expression in the source text to get to the examples.
- Like in state-of-the-art parallel corpus search tools, the domain-specific expressions are aligned to their translation and the translator does not have to locate the relevant passage in a bilingual document.
- Unlike with corpus search tools, the examples are sorted for relevance to the current translation assignment: Searches for expressions are not executed in isolation, but the context of the source text is taken into account.
- Unlike in a general search engine, the translator only gets translation examples from selected reliable sources and the meta-data of the source is readily provided.

We believe this type of functionality complements other resources that translators have available. Machine Translation can reduce translation time, but post-editing will remain necessary for the foreseeable future, and post-editors need easy access to online repositories to check translations. Also, high quality term banks and specialized (online) dictionaries remain a crucial resource for translators, but these are time-consuming and expensive to compile and maybe not necessary for all terminological needs of translators. Informative translation examples from qualitative and reliable sources can go a long way. In short, we argue that Term&Phrase Memory offers a novel functionality that is highly useful for specialized translation.

References

2. Rahzeb Choudhury and Brian McConnell 2013. TAUS Translation Technology Landscape Report, Published by TAUS BV, De Rijp.

see [19] for an overview and comparison of current tools


Segueing from a Data Category Registry to a Data Concept Registry

Sue Ellen Wright, Menzo Windhouwer, Ineke Schuurman, Daan Broeder

Kent State University
The Language Archive – DANS
CLARIN-NL – KU Leuven
The Language Archive – MPI

sellenwright@gmail.com
menzo.windhouwer@dans.knaw.nl
ineke@ccl.kuleuven.be
daan.broeder@mpi.nl

Abstract.
The terminology Community of Practice has long standardized data categories in the framework of ISO TC 37. ISO 12620:2009 specifies the data model and procedures for a Data Category Registry (DCR), which has been implemented by the Max Planck Institute for Psycholinguistics as the ISOcat DCR. The DCR has been used by not only ISO TC 37, but also by the CLARIN research infrastructure. This paper describes how the needs of these communities have started to diverge and the process of segueing from a DCR to a Data Concept Registry in order to meet the needs of both communities.

Keywords: data categories, concepts, semantic registries, communities of practice

1 Introduction

For more than a decade now ISO TC 37 has been transitioning from a static paper-based list of data categories for terminology management (ISO 12620:1999) to a more dynamic Data Category Registry (DCR, i.e., http://www.ISOcat.org) designed to serve a broad range of language resource communities [1]. This paper describes these communities and their needs and how they are leading to a new vision.

Terminology management and concept registries have been developed by a variety of Communities of Practice (CoP). Efforts have been made to characterize these CoPs over the last decade and to create a taxonomy of knowledge organization resources [2, 3, 4, 5, 6, 7]. In the context of the ISOcat repository, we distinguish the following:

Discourse-purposed terminology (and concept) management: Lexicographers document the many definitions associated with words and special language terms, using the head word as their core element, to produce monolingual dictionaries and multilingual glossaries. In contrast, terminologists write careful definitions to document
concepts in special fields and link them to the many designations (terms, synonyms, multilingual equivalents, formulae, symbols, etc.) associated with each concept. They often produce multilingual resources where lexical approaches pose problems for semantic mapping of concept systems and interlingual equivalencies. Despite differences, lexicographers and terminologists provide linguistic and semantic information for humans using language to speak and write. These terms are elements of discourse.

Subject-purposed terminology (index languages): Librarians and archivists use terms and definitions to create controlled vocabularies, classification systems, subject catalogues, and thesauri in order to document knowledge and objects in collections and archives (hence the term: documentary languages). Terms are used to retrieve known objects, such as books or art works in a collection, but they are now also used for information retrieval from open heterogeneous archives. Discourse-purposed term/concept pairs function differently from subject-purposed terms. Svenonius writes: “In an index language the naming function of terms works somewhat differently from the same function in ordinary language. In ordinary language, the word ‘butterflies’ has as its denotational … meaning, that is, its referent, the set of all butterflies, past, present, and future, real or imaginary. In an index ‘butterflies’ names a subject and its denotational meaning is the set of all documents about butterflies” [8]. Here, terms are identifiers used to retrieve objects or information.

Data dictionaries: Here terms are called data element names designating elements in data or metadata models, e.g., database schemas or tagging systems for tagged corpora. Together with their underlying data element concepts, they are defined in conjunction with conceptual domains, i.e., their permissible values. The combination of data model and data category (DC) specifications is used to create, retrieve, and map elements for developing and sharing compatible resources. Data dictionaries (DDs) vary in scope and purpose, from very specific DDs that describe shared application models and elements to Metadata Registries (MDRs). At their most rigorous, DDs prescribe data types, data elements, and enumerated values, in order to facilitate precise data interchange and interoperability. At their most flexible, DDs focus on semantic content in order to retrieve and integrate data from heterogeneous sources.

Semantic Web and Linked (Open) Data (LOD): the LOD approach connects distributed data sets over the web by sharing URIs. Data are represented using RDF/ RDFS-based languages. Semantic Web technologies, such as OWL and SKOS, are used to represent knowledge and/or thesauri and other controlled vocabularies, potentially enabling automated reasoning on top of LOD. Here terms act as classes and properties in knowledge and/or data representation systems.

All these approaches used by the various CoPs share the need to describe the semantics of terms so users can determine whether a term applies to a given use case. The more data-oriented approaches also provide representation information, e.g.: does one term have values (a conceptual domain) or is it a value in such a domain? For instance, does /grammatical gender/ have masculine and feminine as values, or neuter as well? This paper describes the data-oriented ISOcat DCR and how its use by various CoPs steers it towards becoming a Data Concept Registry.
2 The ISOcat Data Category Registry

DCs have a long history especially in the ISO TC 37 community [9]. This section describes this and more recent history revolving around the development of the ISOcat DCR [10] at the Max Planck Institute for Psycholinguistics (MPI-PL) and its use in the wider community, especially CLARIN.1

2.1 ISO TC 37, Terminology and other language and content resources

The evolution of the DCR reflects the convergence of multiple purposes and subsets of experts within the framework of the broader community of practice represented by resource and application developers in linguistics and the social sciences.

ISO TC 37 specifies DCs for use in terminology databases and (as expanded in ISOcat) as tags for marking up language resources. The documentation of standardized DC names (and originally, their abbreviations) began when terminologists were still recording information on paper fiches. Computerized DDs led to initial efforts to collect and document data element concepts associated with terminology management as a part of the development of a terminology interchange format (originally called MARTIF, then XLT, and now known as the TermBase eXchange format or TBX) during the SALT project (Standards-based Access service to multilingual Lexicons and Terminologies; see [11, 12]). After evolving through the SYNTAX pilot project [13, 14], this effort emerged as a Metadata Registry in the sense of the ISO 11179 family of metadata standards [15], called in TC 37 parlance a Data Category Registry.

The primary focus of this effort was originally the definition of DCs representing data element concepts used as semantic units in terminological databases, such as term, part of speech, definition. These elements are used in modeling and creating databases, and in manipulating data in exchange environments requiring interoperability, not only in terminology management, but also in a variety of text and corpus annotation frameworks, such as syntactic or semantic information. As a consequence, they are rigorously defined and generally conform to a variety of metamodels ([11, 16, 17, 18], etc.). Given the metamodels used in the various environments, definitions created for use with these resources are ideally rigorously linked to their respective metamodels and reflect relationships, particularly between parent and child DCs, expressed in the DCR as open, closed, simple, and constrained DCs (see Section 3.1).

2.2 MPI-PL and CLARIN

The Max Planck Institute for Psycholinguistics has a long history in cooperation with ISO TC 37. During the LIRICS project they developed a web-based lexicon tool to support the Lexical Markup Framework (LMF), while INRIA created the SYNTAX DCR [19, 20]. When LIRICS ended, the MPI-PL started developing ISOcat as the successor to SYNTAX. Around the same time, the preparatory phase of the CLARIN

1 A research infrastructure for scholars in the human and social sciences, cf. http://www.clarin.eu
infrastructure started and the ISOcat DCR was introduced as a foundation for semantic interoperability. One of the aims of the European CLARIN infrastructure is to allow scholars to easily find and integrate language resources (LR) and language technology (LT) from a wide range of sources. For this purpose CLARIN set out to develop (1) a joint domain of LR & LT metadata and (2) a federated content search domain allowing users to perform queries on corpora of annotated texts or media housed at different sites in parallel. Differences between sub-community descriptive terminologies dictate that CLARIN address semantic interoperability.

In the description of terminology and corpus management models cited above, interoperability involved adherence to shared metamodels, but in the CLARIN context, interoperability is not so much a function of compatible data design, but rather of data retrieval from potentially heterogeneous resources. In this environment, differences in data description require reinterpretation when retrieved data from different sources are to be processed as one set or when they have to be semantically ‘normalized’ for a specific tool, although the community is encouraged to use one of the various available description standards, cf. above. The ISOcat DCR has been used in this infrastructure as a recommended resource for the purpose of providing linkage between the heterogeneous (meta)data models in order to enable integrated data retrieval.

3 Converging and Diverging Communities of Practice

3.1 The ISO TC37 CoP

As noted, in ISOcat DCR practice, DCs specify a data element name assigned to the definition of a data element concept. As such, they play an important role in terminology management and the creation of annotation schemes used to mark up text and speech corpora. The evolution of ISOcat coincided with an expansion of ISO TC 37 to include a range of ‘other language resources’, many of which share DCs across sub-communities of practice. In its original configuration, the data and organizational model of the DCR was designed to comply with then-current ISO directives pertaining to the standardization of concept-related items cited in ISO standards. This approach dictated the strict identification of so-called Thematic Domain Groups (TDGs). Only a few of these established expert groups have become active:

- Terminology – ISO 12620 and 30042
- Morphosyntax – LMF [20]
- Metadata – CMDI (see Section 3.2; [21])

The ISO requirements cited above imposed a complex standardization process on both the theoretical framework of the DCR and (perhaps more importantly or even unfortunately) on the actual data model and instantiation of the resource. In practice, these structures have proven not only unworkable in terms of human computing conventions, but also unwanted because no actual DC standardization has taken place inside the DCR. Instead, Data Category Selections (DCSs) specified for any given sub-CoP are being simply listed in the related standards [17]. Within the DCR itself,
consensus-based recommendations have proven more effective than formal balloting procedures as prescribed in the now-rejected cumbersome ISO approach.

The original ISOcat design was wedded to the terminological view of linguistic data and categorizes DCs based on their function(s) in various metamodels:

- Open DCs that can take values that conform to the abstract definition of the DC (example: /writtenForm/ (isocat.org/datcat/DC-1836));
- Complex DCs, subdivided into:
  - Closed DCs whose values are constrained to an enumerated set of values (examples: /part of speech/ (isocat.org/datcat/DC-1345));
  - Simple DCs, which serve as those enumerated values (example: /adposition/ (isocat.org/datcat/DC-1231));
- Constrained DCs whose values are defined by automatically parsable rules (example: /breath alcohol concentration/ (isocat.org/datcat/DC-4359) specifies a regular expression to limit the value domain);
- Container DCs, which can be used as high-level container components in compliance with various metamodels (example: /description/ (isocat.org/datcat/DC-3868)), whereby descrip can contain multiple other DCs, such as /definition/, /context/, /source/, /note/, etc.

For the terminology community in particular, the relationships between closed and declared simple DCs is critical to ensure rigorous interoperability in industrial environments. The DCR was originally designed to allow for multiple sets of enumerated values depending on the requirements of different sub-communities, but the need to declare data types and data element categories imposes unwanted constraints for users who may want to use specific data concepts in a variety of ways. For instance, noun can function as a simple DC dependent on the parent part of speech in one environment, but in another it might have its own sub-categories, e.g., proper, common, count, mass, etc. These concerns suggest that the relations currently expressed in the DCR be moved outside the system to external Relation Registries (RRs [22]), so that DCs within the DCR would be unconstrained by these relations (see Section 4.2).

3.2 The CLARIN(-NL) CoP

With respect to the current recommended use of the ISOcat DCR for LRs and LT in the CLARIN domain, we must distinguish between instances of LR & LT metadata and DC use within LR content such as annotations. For the CLARIN joint metadata domain, the Component Metadata Infrastructure (CMDI [21]) actually references ISOcat using links to ISOcat Persistent Identifiers (PIDs) [23]. CMDI allows CLARIN to deal with the wide variety of metadata needs within the LR and LT domain. DC references have been used there to indicate semantic overlap between metadata components, elements and values. Tools like the Virtual Language Observatory (VLO) [24, 25] metadata catalogue use such references to do semantic mapping for kindred metadata attributes. However, similar use of such references for LR content schemas has not progressed far, partially because of the problems of exhaustively describing all LR content schemas used, and partly because providing accurate DC specifications is a hard task, as explained below.
The core of CMDI consists of reusable components. These components group metadata elements and possibly other components, which are managed by a Component Registry (CR). To describe a resource type, a metadata modeler combines components from the CR into a metadata profile. Due to the flexibility of this model, the metadata structures can be very specific to an organization, project or resource type. Although structures can thus vary considerably, they are still within the domain of metadata for linguistic resources and thus share many key semantics. To deal with this variety, general CMDI tools, e.g., the VLO, operate on this shared semantics layer. To establish these shared semantics, CMDI components, elements and values can be linked to concept registries. The major concept registries currently used by CMDI are the Dublin Core metadata elements and terms [26] and the ISOcat DCR. While Dublin Core is closed, ISOcat is an open registry, which means that anyone can register new concepts as needed. Recent visualization experiments have shown an increasing amount of semantic overlap between various sub-communities in the CLARIN joint metadata domain [27, 28].

In principal mapping capability is good between the building blocks of CMDI and Data Category types:

- Components can be linked to container DCs;
- Elements can be linked to complex DCs;
- Values can be linked to simple DCs.

The CR edit utility attempts to adhere to this mapping if one uses the CR’s ISOcat search interface, but it has always been possible to override this feature and include any concept or DC reference, which has resulted in a growing type mismatch between the content of the CR (components, elements and values) and their referenced DCs:

- 165 elements and 72 components are linked to simple DCs;
- 778 components are linked to complex DCs;
- 4 elements are linked to container DCs.

These data indicate that the metadata modelers assessed the applicability of a DC based on the semantic specification only and did not take the associated representation information, i.e., the data category type, into account. To map totally compatible DCs to the content of the CR, in some cases it becomes necessary to create DCs that are semantically redundant, but that are assigned to different DC types (e.g., noun as an open DC, or noun as a simple DC that is a value of part of speech. This practice can lead to significant proliferation in the DCR, which would also make it harder for users to select the proper DC. Current practice makes it impossible to rely on the typing info in the DCR; instead generic tools (e.g., VLO) rely on inspecting the CMDI profile metadata schema to determine the actual status in a given use case. Hence, the insight has been growing that this typing is, for CLARIN’s purposes, counterproductive in the registry, as it can always be, and can better be, gleaned from the actual application involved. This circumstance supports the notion of removing DC typing

---

2 Statistics from December 2013. Thanks to Matej Durco.
from the DCR proper and moving this information to specialized RRs for those communities that rely on type categorizations in compliance with ISO 11179.

In addition to metadata DCs, CLARIN-NL has also created DCs for resource content, which is even more diverse, meaning that DC typing in this area can lead to even more proliferation. In many cases non-technical domain experts are asked to create or select relevant DCs, and for them the more technical details of a good specification, e.g., DC type and data type, are very hard. Within CLARIN-NL, and increasingly throughout the broader CLARIN community, these users are supported by an ISOcat content coordinator. She informs them about good patterns, reviews specifications and selections, and recommends DCs for reuse. Nevertheless, the complexity of the current DCR data model and its management processes has become a burden [29].

4  A new focus for ISOcat

The previous sections have shown that there are many problems with the current ISOcat setup. This section describes a leaner focus for ISOcat, while still providing modalities for expressing the additional information that some users need.

4.1 Towards a Data Concept Registry

In general, for communities that are not able to provide expert terminologists, using a complicated model such as ISO 12620 for DCs has proven unusable. Currently CLARIN is investigating alternatives based on a simpler data-model, which is more focused on specifying Data Concepts and leaves the representation of these concepts to the data models. To summarize developments up to this point, we have seen a divergence between the needs and applications of sub-communities using the DCR, specifically between the ISO TC 37 and the CLARIN communities. Where TC 37 experts may need specifically constrained data categories with strictly specified data types and DC types, CLARIN now realizes the need for a repository of data concepts that are unencumbered by the constraints of any specific data modeling environment. Instead of using the DCR as a prescriptive tool for data modelers who need rigorous data definitions, CLARIN users are better served by more semantically suggestive information units. This means that for each data element concept, we need to create a concept specification with a reusable definition, but without the constraints of declaring data type and DC type. This transition in needs also dictates an evolution in the criteria required for writing adequate definitions, which means not only that definitions must be well-conceived (which is not always the case in the current DCR), but they should also be less dependent on any one view of individual data concepts, thus making them “reusable” across applications. Adding to this, the CLARIN community has only fully realized its requirements and also the limitations of the CLARIN community involvement in the last few years. While the current configuration has seemed clear for the terminology community, it has not been truly integrated into the work of other sub-groups within TC 37. So the divergence between the various groups has only come to light with the coming of age of CLARIN and the evolution of TC 37.
The CLARIN CoP (see Section 3.2) clearly sees the need to focus its efforts on describing the concepts underlying the (meta)data of LRs and LT, and hence the need to relieve the registry of the complexity in the data model associated with the assignment of DC types. The future plan is to create an optional open or free area in the data concept specification where it is still possible to retain and add this kind of information. The core registry will not interpret this optional information, which is left to the communities that need to use it.

Furthermore, as the ISO standardization process has stalled, a community-based recommendation system has already been put into place, which is seen as an easier way to help users select or create data concepts appropriate for their resources or tools. The system provides the ability for multiple sub-communities, including ISO TC 37, to designate individual DCs as “recommended”.

4.2 Relation Registry

Ontological relationships between DCs had already been banned from the DCR data model in its early design stages. This was due to the fact that these relationships are heavily context dependent, i.e., they change with regard to application context or domain. Despite this rule, relationships between closed Data Categories and simple Data Categories have always been a core part of the data model. However, the CLARIN experience has shown that even these relations are also very context-dependent, i.e., different applications need different value domains, and in the current system it is hard to extend these domains due to DC ownership or the fear of further proliferation within the registry.

The Relation Registry (RR) was originally envisioned as a way to align multiple DCRs (once those would start to appear), but due to the increasing amount of proliferation in ISOcat itself, such a Relation Registry is even necessary when there is only one DCR, be it a Data Category Registry or a Data Concept Registry. In addition to (loose) equivalence ((quasi-) same-as) relationships, it has been clear that other ontological relationships could be stored as well, e.g., generic and partitive relationships [22, 30]. But the RR can also be a place to store the value domain relationships, which are currently stored in ISOcat. The combination of information from both the Data Concept Registry and the RR can thus result in a complete DC specification. They can even be broader, covering full taxonomies [31, 32], and may be even be configured as ontologies [33]. Placing these resources outside the DCR proper also accommodates the reality that different users may wish to produce different ontological systems using the DCs.

5 Conclusion and future work

The first stage in developing a Data Concept Registry appropriate for our needs involves (hopefully) finding the ideal off-the-shelf, semantically oriented software platform that can be used to meet our requirements or can be modified with minimal investment. In parallel, we must complete the development of a rich, user-friendly RR
utility to support the supplemental definition requirements of those who want to use
the DCs for more rigorous data modeling. Conceptually, RR software could run in
multiple specialized environments in the periphery of the Data Concept Registry un-
der the control of the individual sub-CoPs who need this capability. Also in parallel,
CoPs should be encouraged to expand their recommendations with the Data Concept
Registry and to improve DC definitions in order to enhance the perceived value of the
resource. Finally, when the new configuration is clearly defined and in place, the old
ISO 12620:2009 must be revised accordingly.

References
1. ISO 12620: Terminology and other language and content resources – Specification of data
categories and management of a Data Category Registry for language resources (2009)
2. Souza, R. R., Tudhope, D., and Barcellos Almeida, M.: Towards a taxonomy of KOS: Di-
pages/research/hypermedia/nkos/nkos2007/presentations/NKOS%202007-HLava.ppt
(2007)
5. Tudhope, D. A.: tentative typology of KOS: towards a KOS of KOS? In: NKOS Work-
shop, ECDL. (2006)
and Aural Languages. In: Journal of the American Society for Information Science, vol
45(8), pp. 600-606 (1994)
9. Wright, S.E.: Data Categories for Terminology Management. In: The Handbook of Termin-
ology Management. Amsterdam and Philadelphia: John Benjamins Publishing Company,
pp. 552-571 (2001)
10. ISOCat: Data Category Registry, Defining widely accepted linguistic concepts.
http://www.isocat.org
11. ISO 30042: Systems to manage terminology, knowledge and content – TermBase eX-
change (TBX) (2008)
12. Melby, A.K.: SALT: Standards-based Access service to multilingual Lexicons and Termi-
13. Ide, N. and Romary, L.: A Registry of Standard Data Categories for Linguistic Annota-
14. Wright, S.E.: A Global Data Category Registry for Interoperable Language Resources. In:
15. ISO/IEC 11179: Information Technology – Metadata registries (MDR). http://metadata-
standards.org/11179/)
23. ISO 24619: Language resource management – Persistent identification and sustainable access (2011)
Abbreviations

CLARIN  Common Language Resources and Technology Infrastructure
CMDI   Component Metadata Infrastructure
CoP    Community of Practice
CR     Component Registry
DC (DCs) data category, -ies
DCR   Data Category Registry
DCS   Data Category Selection
DD (DDS) data dictionary, -ies
INRIA Institut national de recherche en informatique et en automatique
ISO    International Organization for Standardization
LIRICS Linguistic Infrastructure for Interoperable Resources and Systems
LMF    Lexical Markup Framework
LOD    Linked Open Data
LR     language resource
LT     language technology
MARTIF Machine-Readable Terminology Interchange Format
MPI-PL Max Plank Institute for Psycholinguistics, Nijmegen
OWL    Web Ontology Language
PID    Persistent Identifier
RDF/RDFS Resource Description Framework/RDF Schema
SALT   Standards-based Access to multilingual Lexicons and Terminologies
SKOS   Simple Knowledge Organization System
SYNTAX (DCR pilot project)
TBX    Termbase eXchange Format
TC     Technical Committee (ISO)
TDG    Thematic Domain Group
URI    Uniform Resource Identifier
VLO    Virtual Language Observatory
XLT    eXchange format for Lex/Term data
Exploring the Use and Usefulness of KRCs in Translation: Towards a Protocol

Emmanuel Planas¹, Aurélie Picton², Amélie Josselin-Leray³

¹LINA, Université de Nantes & IPLV, Université Catholique de l'Ouest
²TIM, Faculty of Translation and Interpreting, University of Geneva
³CLLE-ERSS, Université Toulouse 2 & CNRS (UMR589)

emmanuel.planas@univ-nantes.fr, aurelie.picton@unige.ch, josselin@univ-tlse2.fr

Abstract: This paper outlines the design of and some methodological conclusions drawn from a pilot study conducted among trainee translators to measure the use and usefulness of Knowledge-Rich Contexts (KRCs) in the translation process. After discussing the issue of context and KRCs in translation, it reviews the literature on previous observation protocols and tools designed for the study of the translation process. It then presents the customized software designed for the experiment to record the translator’s activity. It describes the details of the pilot study, and, finally, some preliminary results and methodological changes planned for the subsequent final experiment(s).

Keywords. Knowledge-Rich Contexts; Terminology; Empirical studies; Translation Process; CAT tools; Logging

1 Introduction

Knowledge-Rich Contexts (KRCs), defined by Meyer [1] as “context[s] indicating at least one item of domain knowledge that could be useful for conceptual analysis”, are a well-known notion in terminology and knowledge extraction. Although the existing studies about KRCs originally focused mainly on text-based terminology or ontology-building [2][3], more recently, several papers (e.g. [4]) have shed light on the importance of such contexts for translators: having access to usage information for a given term or to semantic and conceptual relationships between terms –be it in the source language or in the target language– is essential for translators. The (semi-) automatic extraction of Knowledge-Rich Contexts thus seems very relevant.
This is what the CRISTAL project\textsuperscript{1} aims at doing, by retrieving KRCs from bilingual comparable corpora and integrating them into CAT tools. However, tailoring this kind of tools to suit the translator’s needs implies refining what underlies the notion of KRC. In this paper, we are thus testing a protocol that precisely aims at providing insights into various elements to better understand what a good KRC for translators is. In the medium term, the experiments to come should allow us to develop a typology of the most useful KRCs for a translator, to gather details on their required extension and structure, to get some information about the stages of the translation process in which KRCs are most needed and about the way they are used in relation to other resources (such as dictionaries or term banks). For the time being, we focus on the protocol itself. To meet this ultimate objective, the protocol relies on a combination of different technologies to record the translator’s use of KRCs in an environment thought to be as “ecological” as possible \cite{5}. The first part of this paper (section 2) focuses on the issue of context in translation; section 3 then provides a short review of existing methods to observe the translation process. This review provides the basis of a new interface we designed to better identify the use of resources by translators: Argos (section 4). In section 5, we present the pilot study led at the University of Geneva to validate our protocol. The preliminary results are provided in section 6.

\section{KRCs and Translation}

Even though it is generally agreed that context is an essential component of translation, the definition of that fuzzy notion remains somewhat unclear, maybe due to the fact that it is used in many fields, e.g. philosophy, psychology, and linguistics. Following Melby & Foster \cite{6}, we define the context of a lexical unit as the text that surrounds it, i.e. the units that precede/follow it, at sentence level or on a larger scale.

A number of shortcomings regarding context can be identified in the tools that translators generally have at their disposal: dictionaries, term banks, and CAT tools. As underlined by Varantola \cite{7} and Bowker \cite{4}, since dictionaries try to provide general information that can be applied to a wide array of situations, they usually provide “context-free descriptions of word-use”, i.e. prototypical information, which is of limited use to translators who need context-specific information. Moreover, when provided, the context-related data is usually presented in a very condensed version, while translators “also need information relating to longer stretches of text than a single lexical item” \cite{7}. Paradoxically enough, despite the advances in terminology research about context, and in particular KRCs, Bowker \cite{4} notes that what translators usually find in term banks are “terms presented out of context, or in

\textsuperscript{1}CRISTAL (“Contextes RIches en connaissanceS pour la trAduction terminoLogique”) is an original French project involving linguists, computer researchers and a firm specializing in multilingual text management. The CRISTAL project is a three-year project funded by the French National Agency for Research (ANR; ANR-12-CORD-0020).
only one single context” (which is usually provided only for the “best” term), while what they need is actually “information that would allow them to see all possible terms in a range of contexts and thus find the solution that works best in the target text at hand”. Barrière [8] shows that very simple IR techniques on the biggest corpora available provide better terminological support than the biggest term banks available. Finally, terms automatically provided by term databases in CAT tools are not shown in context, but in a small window providing a translation proposal, a comment and some non-linguistic data like the date or author. However, in tools such as Transit2 and Multitrans3, translators can intentionally search in translation memory databases (parallel texts) for some concordance-like contexts for a given term.

What makes a Knowledge-Rich Context in the field of translation? According to Bowker [9], the notion of KRC can be widely understood as “any context that contains useful information” for the translation process. In Bowker [9][4], she draws a list of those items of information that can prove useful for the translator which can be summed up as follows: (i) information about usage; this of course includes collocations, in particular which general-language words collocate with terms, (ii) information about the frequency of use of a particular word or term, (iii) information about lexical and conceptual relations (such as synonymy, meronymy, hyperonymy etc.), (iv) pragmatic information about style, register and genre –something which was already underlined by Varantola [7] back in 1998, (v) information about usages to avoid. This list can seem really extensive, and Bowker [4] even adds that “translators might not even know what they need: they are seeking inspiration, associations, similar examples, parallel situations that can be adapted.” She concludes by saying that “it is often a case of I don’t know what I’m looking for, but I’ll recognize it when I see it”. According to her, the information needed by translators could/should be provided through corpus-based “word-clouds” (with frequency data), “collocate clouds” and a large number of corpus-based contexts that could be presented as KWIC concordances [9][4]. While Barrière [10] proposes a tool to help terminologists collect corpora and build KRCs semi-automatically, a hands-on experiment to test the use and usefulness of pre-selected KRCs thus seems very welcome. That is what we propose, through the observation of translators and the recording (log) of their actions while they translate.

3 Observation of Translators in Action: a Brief State of the Art

In order to gather information about the KRCs translators resort to when translating, it seemed necessary to first examine previous observation protocols and tools designed for the study of the translation process.

---

2 http://star-group.net/ENU/group-transit-nxt/transit.html (last consulted 02.28.14)
3 http://multicorpora.com/products-services/other-available-products/ (last consulted 02.28.14)
Before technology was available for recording translators' activity on the fly, researchers favored methods where the translator would express his processes either orally or on paper. Göpferich & Jääskeläinen [11] and Ehrensberger-Dow & Massay [12] identify: 

(i) Think Aloud Protocols (TAPs) where the translator, while translating, comments aloud his choices, which are recorded on a tape recorder; 

(ii) dialogue protocols where the decisions about translation are taken through a dialogue between peer translators, which is also tape-recorded; 

(iii) retrospective interviews, where the translation is explained just after being done (for short memory matters); 

(iv) integrated problem and decision reporting (IPDR), where translators write down and explain points they think critical; 

(v) questionnaires, interviews, and diaries. It was common—and still is—to ask about the translators’ background and translation habits through pre-questionnaires or interviews [13][14]. These methods have often been used to evaluate the differences in the use of translation resources between experimented and trainee translators [15][13][5]. Varantola [7], Künzli [13], Desilets et al. [16] noted down which resource was used during translation. Bowker [17] and Delpech [18] used the separation into two or more groups to study the influence of the use of specific resources on translation quality.

Yet, some of these methods were proved [19][12] to be invasive enough to disturb the translator's natural translation process. That is why, following the idea of “ecological validity” introduced by Ehrensberger & Massey [5], we prefer to use a method fostering the respect of the translator's natural environment, such as keylogging software. Since the late 2000s, key logging software (such as Inputlog [20] or Translog [21]) have allowed recording the translator's textual production without intrusion. The QRedit interface of the MNH-TT platform [22] gives an alternative for logging a collaborative translation on the web. For studying the use of translation resources, key logging software are fully useful only if there is a means to set a link between the text that is typed and the resources used by the translator. This is the case when the screen activity as a whole is recorded at the same time. Pieces of software like Camtasia or BB Flash Back can help in that respect. In addition to these, eye-trackers (ET) became precise enough around 2010 to map which word was looked at, and at what time, by the translator's eyes (e.g. [23]). Even though ET could give us more precise information about a translator’s decision, at this stage of our research, we are more interested in textual information that we can post-process automatically.

4 Argos: a New Interface for Translation Process Observation

In order to observe and record the translators’ use of KRCs, we looked at the existing software. Logged interfaces like Translog or InputLog show the participants

---

4 http://www.techsmith.com/camtasia (last consulted: 28 February 2014)
5 http://www.bbsoftware.co.uk (last consulted: 28 February 2014)
the source text in a source window and records all the edition changes the translator inputs in a target window. But such tools do not provide any logged interface for the translation resources used by the translator. CAT tools like OmegaT ⁶, or TradosStudio⁷, do provide a complete interface for the use of translation resources, but the editing activity is not recorded in a log.

Using a combination of such a CAT tool with existing logging software like Camtasia Studio or BB Flashback would provide us with a recorded video of the screen, a log of the typed text, and possibly the changes of software window. But it would not record which resource was used and which proved useful. This led us to the conclusion that we had to design a new logged interface that would meet our precise needs. For it to be as close as possible to the translator’s usual environment, we studied existing CAT tools because they are organized in an ergonomic design translators are familiar with. With all this in mind, we created Argos⁸.

![Fig. 1. The Argos interface](image)

Argos is composed of (i) a source window in which the source text (ST) is displayed; (ii) a target window in which the translator can type in his translation (TT); (iii) a window where a list of source contexts are displayed when the translator selects

---

⁶ http://www.omegat.org (last consulted: 28 February 2014)
⁷ http://www.translationzone.com/trados.html (last consulted: 28 February 2014)
⁸ Argos is coded in Java 1.7 and has been tested on Linux, Mac OS and Windows.
a term in the ST (like cinder cones in the screen shot); (iv) a window where a list of target contexts is displayed when the translator types a target text in a specific input field; (v) a dedicated window for a bilingual specialized dictionary, where target terms are displayed when source terms are entered in a specific field; (vi) 4 logged tabbed windows connected to specific translation resource URLs (e.g. Termium) (see §5.3).

All keyboard activity is recorded, whichever window is being used: characters, deletion, etc. When the translator selects a (simple or complex) term in the ST, a list of KRCs is displayed in the KRC window. This blocks the TT window until the translator chooses at least one KRC with a simple click, forcing him to explicit which KRC was useful. The same mechanism is set for the target KRCs. The translator’s queries about terms and his KRC choices are recorded into the log.

5 Pilot Study

The experiment we present here is a pilot study that aimed at testing our protocol and the translation interface with a small number of participants. Two larger scale studies are also planned, involving 20 participants each.

5.1 Participants

7 students from the Faculty of Translation and Interpreting at the University of Geneva –4 Master’s students, and 3 PhD students– participated in the experiment. We felt that these students would be good candidates as they all had French as a mother tongue, had followed translation courses from English to French and were all familiar with CAT tools. The PhD students had some professional experience in translation.

5.2 Text

The text to be translated was chosen based on previous experiments with translators (e.g. [13][17][7]). Its main features are: (i) written in English, to be translated into French; (ii) 150-word long, to be translated in less than 2 hours; (iii) dealing with a subject that was (a) technical enough for fostering terminology search, but (b) not highly technical for the students, and (c) familiar enough to us to ensure we would be able to assess the quality of the translations at the end of the experiment, (iv) containing a number of collocational and syntactic difficulties, (v) structured in a very logical way. We picked an extract from a popular-science book on volcanology9 that describes the 2 phases in which cinder cones are built.

9 What’s so hot about volcanoes?, Wendell A. Duffield (2011), Mountain Press.
5.3 Resources

Lexicographic resources. Participants had access to: the Robert & Collins (English-French, French-English), Termium, the Grand Dictionnaire Terminologique, (all three online), and some entries of a specialized bilingual dictionary of volcanology.\(^{10}\)

KRCs. For some terms in the text, we selected different types of supposed KRCs\(^ {11}\). First, some contexts were selected in the source language (English). Second, we tried to anticipate possible equivalents in the target language (French) for each term and provided contexts for each. A dozen one-sentence long KRCs were provided to the participants for each term. We tried to put together different types of KRCs for each term, such as definitions, hyperonymy, synonyms, collocations, etc. We added some “Knowledge Poor Contexts”, supposed to be of no use to the translator (Table 1).

<table>
<thead>
<tr>
<th>Definition</th>
<th>Scoria is very vesicular, low density basalt.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hyperonymy</td>
<td>Volcano type: scoria cone, shield volcano, stratovolcano.</td>
</tr>
<tr>
<td>Property + collocation</td>
<td>Many scoria cones are monogenetic in that they only erupt once, in contrast to shield volcanoes and stratovolcanoes.</td>
</tr>
<tr>
<td>&quot;Knowledge poor context&quot;</td>
<td>The second moai has a Pukao which is made of red scoria.</td>
</tr>
</tbody>
</table>

Table 1. Examples of KRCs for scoria (source language, English)

Internet Access. Unlike Master’s students, PhD students had access to Google through the interface, in order to search for resources we would have not thought of.

5.4 Questionnaire and Interviews

The translation task was completed by an online questionnaire about the main translation difficulties, the use of resources and KRCs, the relevance of KRCs, the stages of the translation process when KRCs were needed most, the interface, and general information (age, experience, degrees, etc.). Then, an approx. 20-minute semi-structured interview was conducted with all the participants.

5.5 Experiment

After a 15-minute test of the environment, the students were allocated 2 hours to translate the text, and to indicate which KRCs were the most useful. Their activity

\(^{10}\) Dictionnaire bilingue des Sciences de la Terre (anglais / français) (2013), Michel J.-P. et al., Dunod, 5th edition. Relevant entries were converted into electronic form.

\(^{11}\) These were taken partly from a comparable, French-English, popular-science corpus compiled by Josselin-Leray [24], partly from reliable documents found on the Internet.
was recorded and saved. Immediately after the translation task, we asked them to fill in the questionnaire. We then conducted the recorded interviews.

6 Preliminary Results

The nature of the preliminary results of the pilot study is twofold: (i) they provide feedback regarding the validity of our protocol; (ii) they allow us to identify some preliminary tendencies about the use of KRCs during the translation process.

Data Analysis. The expected analysis of the results obtained through this protocol relies on the complementarity of different types of data: questionnaires, video recording, logs, and final translations. Alves [24] showed how the combination of these different techniques—which he calls “triangulation”—leads to more explicit results. To help us read the logs, we created automatic post-logging compilation processes. These gather all the translators’ individual logs in one file containing: the terms that were searched, the resources they were searched in, the KRCs that were selected, the (anonymized) translators that selected them.

Validation of the protocol. Our protocol is operational and everything went smoothly during the experiment, without any interfering on the translation process. All the data was saved, and the log compiled all the results to be observed. All participants warmly welcomed the protocol they considered user-friendly and respectful of most of their environment, especially regarding (i) the resources provided, (ii) the usability of the interface, (iii) the appearance of the interface which was close to existing CAT tools and (iv) the level of difficulty of the text. The difficulties we had anticipated in the text we chose were identified as such and treated by all the participants with all the resources provided.

First results on KRCs. The most important finding is that the participants overwhelmingly chose knowledge-rich contexts and discarded “knowledge-poor” contexts: out of 92 contexts that were selected by the participants, only 5 were “knowledge-poor” contexts. In addition, 6 participants out of 7 clearly stated that the KRCs selected in the interface were very useful and used them to translate, especially KRCs that contain information about collocations. However, even if KRCs are indeed valuable, their usefulness decreases when the information they present is either irrelevant or not easily accessible. It is then of prime importance to work on the diversity of KRCs, on their quality, but also on their layout.

Future adjustments. Feedback from the participants leads us to operate several adjustments for the full-scale experiments to come. Among them, participants suggested to better select the “target” KRCs, and to complete the list provided. We are working on better anticipation of the types of target KRCs to display. Second, even if the size of the text was suitable, the participants voiced concern about the fact
that it was only an excerpt of a chapter, which hampered their translation. We will then provide the full text, but ask the participants to translate only the chosen excerpt.

Last but not least, some features suggested by the participants will be added to the interface, such as an electronic notepad and keyboard shortcuts.

7 Conclusion Remarks and Perspectives

The pilot study presented in this paper is a stand-alone experiment designed to test logistics and gather information prior to a larger study, in order to improve its quality and reliability. The results from the pilot study show that our protocol and the Interface Argos sound promising to assess the use and usefulness of KRCs in translation on a large scale (over 40 participants), in an environment which is as “ecological” as possible. What makes the future results worthy of interest is the quantity, the quality and the diversity of the data (cf. §6 “triangulation” [24]). At this point of our research, the analysis of all the data collected still needs to be refined. The subsequent experiments will enable us to quantify and generalize some tendencies and complement this with fine-grained observations especially through the analysis of the quality of the translations obtained, the viewing of the video recordings and the semi-structured interviews. Finally, the replicability of the method, which also allows one to compare several groups of translators, several types of KRCs, and several types of texts, seems to guarantee a refined comprehension of the linguistic phenomena that are at stake in specialized translation.

8 Bibliography


TBX between termbases and ontologies

Detlef Reineke
Universidad de Las Palmas de Gran Canaria
TiT – Terminology & Translation Institute

Introduction

For the purpose of reuse, terminology is exchanged between language applications, mostly translation-orientated, and TBX. Data exchange with description formalisms of the Semantic Web like RDF, RDFS, OWL or SKOS constitute an increasingly important scenario for TBX. However, when migrating terminology between TBX and the serializations of the mentioned formalisms, the question arises as to which kind of information is relevant to data exchange. It would surely be presumptuous to try to determine an approximate number of exchange constellations, since data exchange between terminological databases and knowledge databases can range from relatively flat-structured and simple to very comprehensive and complex data models.

The present article describes a nearly automated, well-defined conversion routine based on a maximal TBX data model. Smaller data models may be derived from the model depicted here. Potential use cases for the maximal data model are knowledge databases targeted also for terminologists, translators or content authors who wish to enrich their knowledge database with meaningful terminological information. In the reverse case, such terminology-enriched knowledge databases provide an added value for translation-orientated or monolingual language production environments so that data exchange from RDF, RDFS, OWL or SKOS serializations to TBX is also highly desirable.

Data model

The conversion routine described in this article uses TBX and the RDF/XML serialization. On the TBX-side, the sample data model instantiates the following data categories:

- term entry level: id, subject field, definition, source (of definition), figure, source (of figure), concept position, superordinate concept generic and subordinate concept generic
- language section: xml:lang
- term section: id, term, part of speech, grammatical gender, grammatical number, source (of term)
- term component section: id, termComp, part of speech, grammatical gender and grammatical number...

The aim was to completely map the TBX data model to RDF/XML and to reconvert the RDF/XML output file document back to TBX without data loss. Additionally, the RDF/XML document included the ISOcat persistent identifiers of the TBX data categories, the data category names (subject field), the data category identifiers (subjectField) as well as data type of the data category values for the purpose of further data processing. The complete TBX document instance and the corresponding RDF/XML output file can be found in annex A and annex B.

TBX

The core structure of TBX document instances is based on the metamodel of ISO 16642, i.e. every terminological entry (= termEntry element) must contain at least one language section (= langSet element) and the latter also ought to have at least one term section (= tig or = ntig element). In general, terminological data categories are instantiated as a value of a type attribute associated with a metadata category (descrip, admin, xref, etc.). A small number of data categories is instantiated in form of element names (term, date, note) or as attributes (id, xml:lang).
The structure and the content of terminological entries usually differ from termbase to termbase. Therefore, TBX has been designed as a framework in order to meet user-defined requirements with a maximum possible number of data models. ISO 30042 recommends a concrete instance of TBX called TBX-Default. This TBX dialect comprises a total of 117 data categories. TBX-Basic is a leaner version (29 data categories, plus the ntig element is not allowed) and, at the moment, an expert group is developing an even smaller TBX dialect called TBX-Min which will include only about a dozen data categories. To achieve a maximum of interoperability among the three mentioned standard TBX dialects, it is intended to define the smaller dialects as exact subsets of the bigger dialects. Another objective of the expert group is to establish syntactic variants of these dialects where data categories are instantiated as elements and not as value of an attribute (<definition>...</definition> instead of <descrip type="definition">...</description>). In the following, "TBX" always refers to the three standard dialects in attribute style.

RDF

RDF (Resource Description Framework) allows making logical statements about any kind of things (resources). In principle, every statement consists of a triple (subject, predicate, object). The RDF data model is based on formal semantics where the relations between the resources are represented by means of a graph. Subjects and objects (ellipses) constitute the nodes of the graph whereas the predicate is represented by a labeled arrow. Fig. 1 shows a simple graph that describes the relation between a circuit-breaker and the media used for arc extinction.

As illustrated, the resources are allocated with URIs (Uniform Resource Identifiers, ASCII-based) or IRIs (Internationalized Resource Identifiers, Unicode-based). These identifiers need not necessarily be linked to existing resources (on the internet, on a computer, etc.). URIs or IRIs are merely used to unambiguously designate the resources and, therefore, be able to address them efficiently. As these identifiers can hardly be processed, resources can be issued as literals (string in boxes) (s. Fig. 2).

Fig. 1.: Simple RDF graph describing a relation between a circuit-breaker and its extinction media

Fig. 2.: RDF graph with two literals
As can be seen by the example of the predicate, resources are not only physical objects; concepts, designations, data categories and the like can also be expressed in a RDF data model. Fig. 3 illustrates a graph which represents the relations between a termbase, the concept oil circuit-breaker, the language section and the term “oil circuit-breaker”.

For the purpose of this article, the URIs of the RDF/XML serialization have been designed in such a way that they indicate the structure of the terminological entry or the path within TBX document instance to a certain extent. In order to keep URIs relatively short (and therefore, human reader-friendler), concepts, terms and term components are represented by the values of their corresponding id attributes within the TBX document instance. As in TBX, the language sections are represented by ISO language codes. The predicates are labeled using the node names of the TBX core structure and the corresponding data category identifiers (tit:hasTermEntry, tit:hasLangset, tit:hasTerm, tit:hasPartOfSpeech, etc.). The graph in Fig. 3 corresponds to the following RDF/XML coding:

Example 2

```xml
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"

<rdf:Description rdf:about="http://www.tit-institute.de/termbase">
  <tit:hasTermEntry>
    <rdf:Description rdf:about="teid_1114">
      <tit:id>teid_1114</tit:id>
      <tit:hasLangSet>
        <rdf:Description rdf:about="teid_1114/en">
          <tit:langSet>en</tit:langSet>
          <tit:hasTerm>
            <rdf:Description rdf:about="teid_1114/en/tid_2114_1">
              <tit:id>tid_2114_1</tit:id>
              <tit:term>oil circuit-breaker</tit:term>
            </rdf:Description>
          </tit:hasTerm>
        </rdf:Description>
      </tit:hasLangSet>
    </rdf:Description>
  </tit:hasTermEntry>
</rdf:RDF>
```
In RDF/XML, the resource identifiers of subjects and objects are instantiated as value of a rdf:about attribute. The URIs of predicates appear as element names and literals have no further child element. To shorten URIs and improve the readability of the code, an xml:base attribute may be inserted to declare a base URI (in the present case “http://www.tit-institute.de/”). When generating the graph, the relative paths in the RDF/XML document (for example, the value “teid_1114” of the rdf:about attribute) are then expanded to complete URIs (“http://www.tit-institute.de/teid_1114”).

**TBX > RDF/XML**

The starting point for the development of the present conversion routine is the migration of terminological data to RDF/XML. The main focus during XSLT transformation was on how to handle the persistent identifiers of the data categories and the data type information for the values of the data categories. The transformation process is exemplarily presented here. For further details on XSLT, see the website of the World Wide Web Consortium.

**Persistent identifiers**

As stated above, the data model should include information about the persistent identifiers of TBX data categories defined in ISOcat in order to improve the stringency and the reliability of the RDF statements. Fig. 4 shows a graph which describes the relations between a data category instantiated in a TBX document (http://(...).de/.../subjectField), the value of the data category (switchgear, controlgear and fuses), the concept of the data category registered in ISOcat (http://www.isocat.org/datcat/DC-489), the data category name (subject field) and the data category identifier (subjectField).

![Fig. 4: RDF graph describing data categories](image)

At the moment, persistent identifiers are included in ISOcat, but are not specified in the standard TBX data category constraints (XCS file and RNG Schema). The XCS file still contains the identifiers from the withdrawn ISO 12620:1999 (<descripSpec name="definition" datcatId="ISO12620A-0501">). The RNG Schema does not even have identifiers for data categories. The following mechanisms could be used to prepare the XSLT mapping of persistent identifiers to RDF/XML:

- Implement persistent identifiers in terminology tools and export them to TBX
- Include persistent identifiers and extract them from the XCS file / RNG Schema during transformation
- Allocate persistent identifiers in the XSLT style sheet

Implementing persistent identifiers in terminology tools could lead to the following results:
Example 3

<descrip type="subjectField" pid="http://www.isocat.org/datcat/DC-489">switchgear, controlgear and fuses</descrip>

If the pid instances are to be validated against TBX content models, the core-structure DTD, for example, would have to be adapted to include a pid attribute where necessary, and the name of the corresponding entities should be extended accordingly by adding “Pid”. Example 4 shows a matching DTD entity for the instance in example 3:

Example 4

<!ENTITY % impIDLangTypTgtDtypPid
    id ID #IMPLIED
    xml:lang CDATA #IMPLIED
    type CDATA #REQUIRED
    target IDREF #IMPLIED
    datatype CDATA #IMPLIED
    pid CDATA #IMPLIED>

The value of the pid attribute could be selected from the TBX document instance during transformation:

Example 5

<xsl:if test="descrip/@type='subjectField'">
...
<tit:dataCategory>
  <rdf:Description>
    <xsl:attribute name="rdf:about">
      <xsl:value-of select="descrip/@pid"/>
    </xsl:attribute>
    ...
  </rdf:Description>
</tit:dataCategory>
...
</xsl:if>

The transformation would generate the following RDF/XML code:

Example 6

<tit:hasSubjectField>
  <rdf:Description rdf:about="teid_1112/subjectField">
    <tit:subjectField>switchgear, controlgear and fuses</tit:subjectField>
    <tit:dataCategory>
      <rdf:Description rdf:about="http://www.isocat.org/datcat/DC-489">
        ...
      </rdf:Description>
    </tit:dataCategory>
  </rdf:Description>
</tit:hasSubjectField>

In the RNG Schema, entities and their names should also be adapted accordingly adding a ref element and a name attribute with the corresponding attribute value (pid.attributes):

Example 7

<define xmlns="http://relaxng.org/ns/structure/1.0" name="IDLangTgtDtypPid.attributes">
  <ref name="lang.attributes"/>
  <ref name="id.attributes"/>
  <ref name="target-IDREF.attributes"/>
  <ref name="IDLangTgtDtyp.attribute.datatype"/>
  <ref name="pid.attributes"/>
</define>

As an alternative, the persistent identifiers might be selected from the data category constraints (XCS file or RNG Schema). The coding of the XCS file and the corresponding DTD, for example, should be changed as follows:
In the style sheet, the selection of the pid attribute value would be coded as shown in example 9:

Example 9

```xml
<xsl:if test="descrip/@type='subjectField'">
  ...
  <xsl:value-of select="document('TBXXCSV02.xcs')/TBXXCS/datCatSet/
                  descripSpec[@name='subjectField']/@pid"/>
  ...
</xsl:if>
```

A third and leaner variant for mapping persistent identifiers to RDF/XML consists of incorporating them directly in the XSLT style sheet:

Example 10

```xml
<xsl:if test="descrip/@type='subjectField'">
  ...
  <tit:dataCategory>
    <rdf:Description>
      <xsl:attribute name="rdf:about">
        <xsl:text>http://www.isocat.org/datcat/DC-489</xsl:text>
      </xsl:attribute>
      <tit:dcName>subject field</tit:dcName>
      <tit:dcIdentifier>subjectField</tit:dcIdentifier>
    </rdf:Description>
  </tit:dataCategory>
  ...
</xsl:if>
```

**Data types**

In the previous examples untyped literals like `<tit:dcCategoryName>subject field</tit:dcCategoryName>` were used. However, RDF literals are rarely processed as mere character strings, but instead are interpreted for further processing as specific data types (for example, as a date or an identifier for referencing purposes). In RDF/XML, the data type can be instantiated by means of the `rdf:datatype` attribute. The data type instantiation requires the corresponding namespace declaration in the RDF/XML document instance (`xmlns:xs="http://www.w3.org/2001/XMLSchema#"`):

Example 11

```xml
<tit:id rdf:datatype="http://www.w3.org/2001/XMLSchema#ID">teid_1114</tit:id>
```

Data type information cannot be derived without further manipulation from TBX document instances or TBX content models, since TBX does not provide data type description mechanisms in the classical sense. Of course, the TBX content models contain data type attribute declarations, but the attribute values are basically used to determine the type of text handling (plainText, basicText, noteText) or to point to the existence of closed data categories (picklist).

Similar to the mapping of persistent identifiers, the following preparatory measures would be applicable when transforming data type information to RDF/XML:
• Data typing in the terminology tools and export the data type information to TBX
• Include data type information and extract it from the in XCS file / RNG Schema during transformation
• Allocate data type information in the XSLT style sheet

If the values of the data categories were already typed within the terminology tools and then exported to TBX, the instantiation of data category value for “subjectField” in example 3 would result in the following coding:

Example 12

<descrip type="subjectField" pid="http://www.isocat.org/datcat/DC-489" datatype="string">switchgear, controlgear and fuses</descrip>

Such an instance where the value domain is interpreted in a broader sense would correctly validate against the TBX core structure DTD. Since the data type attribute is processed as CDATA, character data like „string“, „ID“, „date“, etc. would be valid. Nevertheless, this instantiation variant does not avoid the necessity for declaring new attributes to enable the instantiation of certain type of data type information. This is due to the fact that an attribute can only be associated with one element at a time. But frequently different data types appear in a single element – for example, when a descrip element contains a text value (data type “string”) as well as a target attribute, for example, whose value would be of ID type. In such cases, additional attributes would be necessary (for example target-datatype, id-datatype, etc.). And even if the originally intended value domain of the datatyp attribute is preserved and an alternative attribute is declared (for example, xs-datatype or similar), additional attributes have to be created to enable the description of the target or id attribute values.

The TBX core structure entity “% impIDLangTypTgtDtyp” is the only entity that provides a datatype attribute. For data elements instantiated at the basic nodes of the TBX core-structure (termEntry-, langSet-, term- and termComp-Elemente), this attribute is not allowed in the current DTD. In order to be able to generate valid TBX document instances, the corresponding components of TBX DTD should be extended to include the data type attribute.

The manipulation of the XCS file and the RNG Schema is no ideal solution either because similar data typing mechanisms have to be implemented. As with the persistent identifiers, the leanest variant would consist of linking the instantiation of the data type information to a conditional. The mapping of the data type information of data category values to RDF/XML could be generated using an xsl:if element:

Example 13

<xsl:if test="descrip/@type='subjectField'">
  ...
  <tit:subjectField>
    <xsl:attribute name="rdf:datatype">
      <xsl:text>http://www.w3.org/2001/XMLSchema#string</xsl:text>
    </xsl:attribute>
    <xsl:value-of select="descrip[@type='subjectField']/text()"/>
  </tit:subjectField>
  ...
</xsl:if>

Further issues

When transforming from TBX to RDF/XML, the tig and ntig elements can be easily distinguished by means of xsl:if. In relation to the term components the question arises if it would be desirable to allocate persistent identifiers as with terms.
The TBX sample document instance (see annex A) includes sources for figures. The sources point to external resources and are described using the data category “xSource”. This data category is neither listed in ISOcat’s TBX-Default data category selection nor in the Terminology profile. Therefore, a provisional persistent identifier was created (http://www.isocat/datcat/DC-xSource).

There might be need for discussion with regards to the ISO 30042 data category names/identifiers “elementIdentifer” und “lang”. The question is whether both data categories names/identifiers should be explicetly mapped to RDF/XML. In ISOcat, the data category name/identifier “lang” corresponds to „language identifier“/“languagelIdentifer” and has a persistent identifier (http://www.isocat.org/datcat/DC-279). “elementIdentifer”is not listed in ISOcat. If needed, a provisional persistent identifier should be created (for example “http://www.isocat.org/datcat/DC-id”).

Another issue that might need further clarification is the handling of data category names when data is not directly allocated in the style sheet, but extracted form external resources. For the moment, TBX standard resources do not contain data category names that could be used for mapping.

RDF/XML > TBX

In the XSLT style sheet for the transformation of the generated RDF/XML document instance back to TBX all those elements have been omitted that do not lead to valid TBX document instances, such as the persistent identifiers, the data type information as well as the data category names. The transformation had no errors and generated a valid TBX document instance.

Conclusion and outlook

A user-defined transformation style sheet can be relatively easily designed. But it would be interesting to create a style sheet as a kind of content model to be able to map a maximum or all possible TBX document instances to RDF/XML automatically without the necessity of further user-specific adaption. Such a style sheet should completely reflect the recommended TBX dialects.

Another issue concerns the conversion between TBX and more powerful semantic description formalisms such as RDFS or OWL. Constructs like rdfs:Class, rdfs:subClassOf, owl:Class, owl:equivalentClass or owl:intersectionOf are surely of interest when dealing with subject-specific queries and reasoning:

Example 14

```xml
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:tit="http://www.tit-institute.de/examples#">
  <rdfs:Class rdf:about="http://www.tit-institute.de/examples/mechanicalSwitchingGear">
    <rdfs:label xml:lang="de">mechanisches Schaltgerät</rdfs:label>
    <rdfs:label xml:lang="fr">appareil mécanique de connexion</rdfs:label>
  </rdfs:Class>
  <rdfs:Class rdf:about="http://www.tit-institute.de/examples/circuit-breaker">
    <rdfs:label xml:lang="de">Leistungsschalter</rdfs:label>
    <rdfs:label xml:lang="fr">disjoncteur</rdfs:label>
    <rdfs:subClassOf rdfs:resource="http://www.tit-institute.de/examples/mechanicalSwitchingGear"/>
  </rdfs:Class>
  <rdfs:Class rdf:about="http://www.tit-institute.de/examples/vacuumCircuit-breaker">
    <rdfs:label xml:lang="de">Vakuumleistungsschalter</rdfs:label>
    <rdfs:label xml:lang="fr">disjoncteur à vide</rdfs:label>
    <rdfs:subClassOf rdfs:resource="http://www.tit-institute.de/examples/circuit-breaker"/>
  </rdfs:Class>
</rdf:RDF>
```
However, it seems that TBX itself offers few ontological constellations (relation between closed data categories and their values, relation between data categories and the metamodel levels of ISO 16642, and so on). The mapping of TBX-specific ontologies (key word: relational register) and conversions between TBX and RFDS/OWL/SKOS serializations will be part of future research work at the Terminology & Translation Institute.

References


ISO 30042:2008 Systems to manage terminology, knowledge and content -- TermBase eXchange (TBX). Geneva: ISO.


mechanical switching device

switchgear, controlgear and fuses

a mechanical switching device, capable of making, carrying and breaking currents under normal circuit conditions and also making, carrying for a specified time and breaking currents under specified abnormal circuit conditions such as those of short circuit

switchgear, controlgear and fuses

1.2.5

mechanical switching device

air circuit-breaker

gas-blast circuit-breaker

sulphur hexafluoride circuit-breaker

air-blast circuit-breaker
Leistungsschalter

disjoncteur

air circuit-breaker

oil circuit-breaker

switchgear, controlgear and fuses

a circuit-breaker in which the contacts open and close within a highly evacuated envelope

vacuum circuit-breaker
Vakuumleistungsschalter

Disjoncteur à vide

gas-blast circuit-breaker
sulphur hexafluoride circuit-breaker

air-blast circuit-breaker
Annex B (RDF/XML document instance)

<?xml version="1.0" encoding="UTF-8"?>
  <rdf:Description rdf:about="http://www.tit-institute.de/termbase">
    <tit:hasTermEntry>
      <rdf:Description rdf:about="teid_1111">
        <tit:id rdf:datatype="http://www.w3.org/2001/XMLSchema#ID">teid_1111</tit:id>
        <tit:hasLangSet>
          <rdf:Description rdf:about="teid_1111/en">
            <tit:hasTerm>
              <rdf:Description rdf:about="teid_1111/en/tid_2111_1">
                <tit:id rdf:datatype="http://www.w3.org/2001/XMLSchema#ID">tid_2111_1</tit:id>
                <tit:term rdf:datatype="http://www.w3.org/2001/XMLSchema#string">mechanical switching device</tit:term>
                <tit:dataCategory>
                  <rdf:Description rdf:about="http://www.isocat.org/datcat/DC-508">
                    <tit:dcName rdf:datatype="http://www.w3.org/2001/XMLSchema#string">term</tit:dcName>
                    <tit:dcIdentifier rdf:datatype="http://www.w3.org/2001/XMLSchema#string">term</tit:dcIdentifier>
                  </rdf:Description>
                </tit:dataCategory>
              </rdf:Description>
            </tit:hasTerm>
          </rdf:Description>
        </tit:hasLangSet>
      </rdf:Description>
    </tit:hasTermEntry>
    <tit:hasTermEntry>
      <rdf:Description rdf:about="teid_1112">
        <tit:id rdf:datatype="http://www.w3.org/2001/XMLSchema#ID">teid_1112</tit:id>
        <tit:hasSubjectField>
          <rdf:Description rdf:about="teid_1112/subjectField">
            <tit:subjectField rdf:datatype="http://www.w3.org/2001/XMLSchema#string">switchgear, controlgear and fuses</tit:subjectField>
            <tit:dataCategory>
              <rdf:Description rdf:about="http://www.isocat.org/datcat/DC-489">
                <tit:dcName rdf:datatype="http://www.w3.org/2001/XMLSchema#string">subject field</tit:dcName>
                <tit:dcIdentifier rdf:datatype="http://www.w3.org/2001/XMLSchema#string">subjectField</tit:dcIdentifier>
              </rdf:Description>
            </tit:dataCategory>
          </rdf:Description>
        </tit:hasSubjectField>
        <tit:hasDefinition>
          <rdf:Description rdf:about="teid_1112/definition">
            <tit:definition rdf:datatype="http://www.w3.org/2001/XMLSchema#string">a mechanical switching device, capable of making, carrying and breaking currents under normal circuit conditions and also making, carrying for a specified time and breaking currents under specified abnormal circuit conditions such as those of short circuit</tit:definition>
            <tit:dataCategory>
          </rdf:Description>
        </tit:hasDefinition>
      </rdf:Description>
    </tit:hasTermEntry>
  </rdf:Description>
</rdf:RDF>
gas-blast circuit-breaker

sulphur hexafluoride circuit-breaker

air-blast circuit-breaker

circuit-breaker
disjoncteur
<rdf:Description rdf:about="teid_1112/fr/tid_4112_1/source">
  <tit:dataCategory>
    <rdf:Description rdf:about="http://www.isocat.org/datcat/DC-471">
      <tit:dcName rdf:datatype="http://www.w3.org/2001/XMLSchema#string">source</tit:dcName>
      <tit:dcIdentifier rdf:datatype="http://www.w3.org/2001/XMLSchema#string">source</tit:dcIdentifier>
    </rdf:Description>
  </tit:dataCategory>
</rdf:Description>
</tit:hasSource>
</rdf:Description>
</tit:hasTerm>
</rdf:Description>
</tit:hasLangSet>
</rdf:Description>
</tit:hasTermEntry>
</rdf:Description>
</tit:hasTermEntry>
</rdf:Description rdf:about="teid_1113">
  <tit:id rdf:datatype="http://www.w3.org/2001/XMLSchema#ID">teid_1113</tit:id>
  <tit:hasLangSet>
    <rdf:Description rdf:about="teid_1113/en">
      <tit:hasTerm>
        <rdf:Description rdf:about="teid_1113/en/tid_2113_1">
          <tit:id rdf:datatype="http://www.w3.org/2001/XMLSchema#ID">tid_2113_1</tit:id>
          <tit:dataCategory>
            <rdf:Description rdf:about="http://www.isocat.org/datcat/DC-508">
              <tit:dcName rdf:datatype="http://www.w3.org/2001/XMLSchema#string">term</tit:dcName>
              <tit:dcIdentifier rdf:datatype="http://www.w3.org/2001/XMLSchema#string">term</tit:dcIdentifier>
            </rdf:Description>
          </tit:dataCategory>
        </rdf:Description>
      </tit:hasTerm>
    </rdf:Description>
  </tit:hasLangSet>
</rdf:Description>
</tit:hasTermEntry>
</rdf:Description rdf:about="teid_1114">
  <tit:id rdf:datatype="http://www.w3.org/2001/XMLSchema#ID">teid_1114</tit:id>
  <tit:hasLangSet>
    <rdf:Description rdf:about="teid_1114/en">
      <tit:hasTerm>
        <rdf:Description rdf:about="teid_1114/en/tid_2114_1">
          <tit:id rdf:datatype="http://www.w3.org/2001/XMLSchema#ID">tid_2114_1</tit:id>
          <tit:dataCategory>
            <rdf:Description rdf:about="http://www.isocat.org/datcat/DC-508">
              <tit:dcName rdf:datatype="http://www.w3.org/2001/XMLSchema#string">term</tit:dcName>
              <tit:dcIdentifier rdf:datatype="http://www.w3.org/2001/XMLSchema#string">term</tit:dcIdentifier>
            </rdf:Description>
          </tit:dataCategory>
        </rdf:Description>
      </tit:hasTerm>
    </rdf:Description>
  </tit:hasLangSet>
</rdf:Description>
<tit:hasTermEntry rdf:about="teid_1115">
  <tit:id rdf:datatype="http://www.w3.org/2001/XMLSchema#ID">teid_1115</tit:id>
  <tit:hasSubjectField>
    <rdf:Description rdf:about="teid_1115/subjectField">
      <tit:subjectField rdf:datatype="http://www.w3.org/2001/XMLSchema#string">switchgear, controlgear and fuses</tit:subjectField>
    </rdf:Description>
  </tit:hasSubjectField>
  <tit:hasDefinition>
    <rdf:Description rdf:about="teid_1115/definition">
      <tit:definition rdf:datatype="http://www.w3.org/2001/XMLSchema#string">a circuit-breaker in which the contacts open and close within a highly evacuated envelope</tit:definition>
    </rdf:Description>
  </tit:hasDefinition>
  <tit:hasSource>
    <rdf:Description rdf:about="teid_1115/definition/source">
    </rdf:Description>
  </tit:hasSource>
</tit:hasTermEntry>
Vakuumleistungsschalter
disjoncteur à vide
<rdf:Description rdf:about="teid_1118">
  <tit:id rdf:datatype="http://www.w3.org/2001/XMLSchema#ID">teid_1118</tit:id>
  <tit:hasLangSet>
    <rdf:Description rdf:about="teid_1118/en">
      <tit:hasTerm>
        <rdf:Description rdf:about="teid_1118/en/tid_2118_1">
          <tit:id rdf:datatype="http://www.w3.org/2001/XMLSchema#ID">tid_2118_1</tit:id>
          <tit:dataCategory>
            <rdf:Description rdf:about="http://www.isocat.org/datcat/DC-508">
              <tit:dcName rdf:datatype="http://www.w3.org/2001/XMLSchema#string">term</tit:dcName>
              <tit:dcIdentifier rdf:datatype="http://www.w3.org/2001/XMLSchema#string">term</tit:dcIdentifier>
            </rdf:Description>
          </tit:dataCategory>
        </rdf:Description>
      </tit:hasTerm>
    </rdf:Description>
  </tit:hasLangSet>
</rdf:Description>
Annex C (circuit-breaker)
Authors Index

Alagic, Mara.................................................................................................................................106
Alexiev, Boyan............................................................................................................................87
Ammar, Chihebeddine..................................................................................................................157
Bjork, Kaj-Mikael.........................................................................................................................11, 58
Bond, Stephen..............................................................................................................................167
Broeder, Daan...............................................................................................................................177
De Hertog, Dirk.............................................................................................................................167
Diaz Vasquez, Juan Carlos...........................................................................................................97
Fodil, Leila......................................................................................................................................77
Forss, Thomas.................................................................................................................................11, 58
Franco Arroyave, Carolina............................................................................................................97
Glenn, Nathan...............................................................................................................................147
Gornostay, Tatiana........................................................................................................................67
Gromann, Dagmar.........................................................................................................................127
Haddar, Kais....................................................................................................................................137, 157
Hayes, James.................................................................................................................................147
Heyer, Gerhard..............................................................................................................................39
Heylen, Kris.....................................................................................................................................167
Hitcheva, Dobromira....................................................................................................................20
Josselin-Leray, Amélie...................................................................................................................188
Kockaert, Hendrik..........................................................................................................................167
Lhioui, Malek...................................................................................................................................137
Liu, Shuhua.....................................................................................................................................11, 58
Lommen, Arle..................................................................................................................................147
Maslias, Rodolfo.............................................................................................................................49
Melby, Alan.......................................................................................................................................147
Niekler, Andreas.............................................................................................................................29, 39
Nuopponen, Anita...........................................................................................................................1
Orel, Tatiana.....................................................................................................................................106
Ouamour, Siham.............................................................................................................................77
Picton, Aurélie..................................................................................................................................188
Pinnis, Marcis....................................................................................................................................67
Planas, Emmanuel............................................................................................................................188
Reineke, Detlef..................................................................................................................................198
Rimmington, Glyn...........................................................................................................................106
Romary, Laurent............................................................................................................................116, 137, 157
Sayoud, Halim..................................................................................................................................77
Schuurman, Ineke...........................................................................................................................177
Snow, Tyler............................................................................................................................................ 147
Steurs, Frieda...................................................................................................................................... 167
Vasiljevs, Andrejs.......................................................................................................................... 67
Vuli?, Ivan.......................................................................................................................................... 167
Wiedemann, Gregor....................................................................................................................... 29, 39
Windhouwer, Menzo..................................................................................................................... 177
Wright, Sue Ellen........................................................................................................................... 177