Proceedings of

CHAT 2012

The 2nd Workshop on the Creation, Harmonization and Application of Terminology Resources

Co-located with TKE 2012

June 22, 2012
Madrid, Spain

Editor
Tatiana Gornostay

Linköping Electronic Conference Proceedings, No. 72

Linköping Electronic Conference Proceedings, 72
ISSN 1650-3740 (online)
ISSN 1650-3686 (print)

Linköping University Electronic Press
Linköping, Sweden, 2012
http://www.ep.liu.se/ecp_home/index.en.aspx?issue=072

Contents

Preface

The second workshop on the Creation, Harmonization and Application of Terminology resources (CHAT 2012) was held on 22 June, 2012 in Madrid, Spain. It was co-located with the conference on Terminology and Knowledge Engineering (TKE 2012). The workshop aimed at bringing together academic and industrial players in the terminology field and attracting holders of terminology resources. The workshop also focused on fostering the cooperation between EU projects and research and development activities in the area of terminology along with sharing experience and discussing recent advances of the consolidation, harmonization and distribution of terminology resources, as well as their practical application in various domains.

Every day, the volume of terminology is growing along with the increasing volume of information available on the web. Efficient terminology acquisition and management has become an essential component of intelligible translation, localization, technical writing and other professional language work. The current models for finding, sharing and using terminology data cannot keep up with a growing demand in multilingual Europe. The role of terminology however is today more important than ever to ensure that people communicate efficiently and precisely. Consistent, harmonized and easily accessible terminology is an extremely important prerequisite for ensuring unambiguous multilingual communication in the European Union and throughout the world.

The workshop was organized by the FP7 projects TaaS (Terminology as a Service)[1] and TTC (Terminology Extraction, Translation Tools and Comparable Corpora)[2], and the ICT-PSP project META-NORD (Baltic and Nordic Branch of the European Open Linguistic Infrastructure)[3] as a continuation a series of meetings that started as the first workshop CHAT 2011 on 11 May, 2011 in Riga, Latvia.[4]

We are delighted to hereby present the proceedings of CHAT 2012.

Altogether, 7 papers have been selected for presentation (4 regular papers and 3 short papers). The workshop papers cover various topics on automated approaches to terminology extraction and creation of terminology resources, compiling multilingual terminology, ensuring interoperability and harmonization of terminology resources, integrating these resources in language processing applications, distributing and sharing terminology data, and other.

We are also pleased to present four invited speakers at CHAT 2012.

Prof. Dr. Klaus-Dirk Schmitz is a full professor of terminology studies and language technology at the Institute for Translation and Multilingual Communication at the Cologne University of Applied Sciences, Managing Director of the Institute for Information Management at Cologne University of Applied Sciences, Vice-President of the German Terminology Association and Chairman of the German National Standards Committee "Systems for managing terminology, knowledge and content". At CHAT 2012, Klaus-Dirk Schmitz gave an invited speech on “Terminological Needs of Language Workers: a User Group Analysis for the TaaS Platform”.

[1] http://www.taas-project.eu
[2] http://www.ttc-project.eu
[3] http://www.meta-nord.eu
[4] http://www.tilde.eu/tilde-research/workshop-creation-harmonization-and-application-terminology-resources/2011

Prof. Dr. Alan K. Melby is a full professor at the Brigham Young University (BYU), Provo campus, the Department of Linguistics and English Language, and is a member of the Board of Directors and chair of the Translation and Computers committee of ATA (American Translators Association) and a member of the US delegation to ISO/TC37 (International Organization for Standardization, Technical Committee 37 for Terminology and Other Language Resources). At CHAT 2012, Alan K. Melby gave an invited speech on "Term Base eXchange: Status and Future".

Dr. Georg Rehm is a Senior Consultant at the Berlin site of DFKI GmbH, Germany, and is the Network Manager of the EC-funded network of excellence META-NET[5]. At CHAT 2012, Georg Rehm gave an invited speech on "META-NET and META-SHARE: Language Technology for Europe".

Dr. Andrejs Vasiļjevs is a co-founder and Chairman of the Board at Tilde, the project coordinator of the FP7 TaaS project and the ICT-PSP META-NORD project, and a member of the Intergovernmental Council and Bureau for the UNESCO Information for All Programme (IFAP), the Vice-Chairman of Latvia Information and Communications Technology Association, and a member of the Commission of the State Language of Latvia. At CHAT 2012, Andrejs Vasiļjevs gave an invited talk on "EuroTermBank – towards dedicated terminology services for European Linguistic Infrastructure".

Finally, we are glad to present the three presentations of the terminology tools made at CHAT 2012 by Béatrice Daille "TermSuite: an UIMA Type System for Bilingual Term Extraction from Comparable Corpora" (University of Nantes, LINA), Mārcis Pinnis "Toolkit for Multi-Level Alignment and Information Extraction from Comparable Corpora" (Tilde), and Rodolfo Maslias "Terminology management tools in the EP and cooperation and information sharing among the EU Institutions managing IATE" (Terminology Coordination Unit, European Parliament).

The organization of CHAT 2012 is a joint effort of several institutions, projects and their representatives. We would like to thank all of the Programme Committee members for fruitful collaboration during the preparation for the workshop and for their effort, time and attention during the review process. We would like to express our special gratitude to the workshop Organizing Committee – our colleagues from Tilde (Latvia)[6], the TaaS project, the TTC project, and the META-NORD project.

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

Tatiana Gornostay
Programme Committee Chair
CHAT 2012

[5] http://www.meta-net.eu
[6] http://www.tilde.com

Committees

PROGRAMME COMMITTEE & REVIEWERS

Tatiana Gornostay (Chair), Tilde, Latvia
Larisa Belyaeva, Herzen University, Russia
Gerhard Budin, University of Vienna, Austria
Béatrice Daille, University of Nantes, France
Patrick Drouin, University of Montreal, Canada
Judit Freixa, Universitat Pompeu Fabra, Spain
Ulrich Heid, University of Stuttgart, Germany
Sigrún Helgadóttir, The Árni Magnússon Institute for Icelandic Studies, Iceland
Marie-Paule Jacques, Stendhal University, France
Marita Kristiansen, Norwegian School of Economics, Norway
Klaus-Dirk Schmitz, Fachhochschule Köln and GTW, Germany
Inguna Skadiņa, Tilde, Latvia
Koichi Takeuchi, Okayama University, Japan
Rita Temmerman, Erasmushogeschool Brussel, Belgium
Andrejs Vasiljevs, Tilde, Latvia

ORGANIZING COMMITTEE

Andrejs Vasiļjevs, Tilde, Latvia
Tatiana Gornostay, Tilde, Latvia
Roberts Rozis, Tilde, Latvia
Klaus-Dirk Schmitz, Fachhochschule Köln and GTW, Germany
Béatrice Daille, University of Nantes, France

WORKSHOP ORGANIZERS

Tilde, Latvia
META-NORD project (CIP ICT-PSP)
TaaS project (FP7)
TTC project (FP7)

Workshop Programme

CHAT 2012: the second workshop on the Creation, Harmonization and Application of Terminology resources

June 22, 2012
Universidad Politécnica de Madrid
Madrid, Spain

MORNING SESSION

9:00-9:30 **Opening**
Welcome and workshop presentation
Workshop organizers: *Tilde, META-NORD, TaaS, TTC*
Presenters: *Andrejs Vasiljevs, Tatiana Gornostay, Klaus-Dirk Schmitz, Béatrice Daille*

9:30-10:30 **Invited presentations**
(Chair: Tatiana Gornostay)

9:30-10:00 *Klaus-Dirk Schmitz*
Terminological Needs of Language Workers: a User Group Analysis for the TaaS Platform

10:00-10:30 *Alan Melby*
Term Base eXchange: Status and Future

10:30-11:00 Coffee break

11:00-12:20 **Paper presentations**
(Chair: Béatrice Daille)

11:00-11:20 *Lara Sanz Vicente*
Searching for Patterns in the Transfer of Multiword Units: a Corpus-Based Contrastive Study on Secondary Term Formation

11:20-11:40 *Jorge Vivaldi and Horacio Rodriguez*
Using Wikipedia for Domain Terms Extraction

11:40-12:00 *Thierry Declerck and Dagmar Gromann*
Extraction of Multilingual Term Variants in the Business Reporting Domain

12:00-12:20 *Anne-Kathrin Schumann*
Towards the Automated Enrichment of Multilingual Terminology Databases with Knowledge-Rich Contexts – Experiments with Russian EuroTermBank Data

12:20-13:20 **Terminology tool demonstrations**
(Chair: Alan Melby)

12:20-12:40 *Béatrice Daille*
TermSuite: an UIMA Type System for Bilingual Term Extraction from Comparable Corpora

12:40-13:00 *Mārcis Pinnis*
Toolkit for Multi-Level Alignment and Information Extraction from Comparable Corpora

13:00-13:20 *Rodolfo Maslias*
Terminology Management Tools in the EP and Cooperation and Information Sharing Among the EU Institutions Managing IATE

13:20-14:30 Lunch

AFTERNOON SESSION

14:30-15:30 **Invited presentations**
(Chair: Roberts Rozis)

14:30-15:00 *Georg Rehm*
META-NET and META-SHARE: Language Technology for Europe

15:00-15:30 *Andrejs Vasiļjevs*
EuroTermBank – Towards Dedicated Terminology Services for European Linguistic Infrastructure

15:30-16:10 **Discussion panel: Terminology Resources as part of European Open Linguistic Infrastructure**
(Moderator: Tatiana Gornostay)

Participants: *Andrejs Vasiļjevs, Georg Rehm, Klaus-Dirk Schmitz, Rodolfo Maslias, Hanne Erdman Thomsen*

16:10-16:30 Coffee break

16:30-17:30 **Paper presentations**
(Chair: Klaus-Dirk Schmitz)

16:30-16:50 *Gruffudd Prys, Tegau Andrews, Dewi B. Jones and Delyth Prys*
Distributing Terminology Resources Online: Multiple Outlet and Centralized Outlet Distribution Models in Wales

16:50-17:10 *Dagmar Gromann and Thierry Declerck*
Terminology Harmonization in Industry Classification Standards

17:10-17:30 *Antti Kanner*
Bank of Finnish Terminology in Arts and Sciences

17:30 **Closing**

Regular Papers

Using Wikipedia for Domain Terms Extraction

Jorge Vivaldi[1] and Horacio Rodríguez[2]

[1] Universitat Pompeu Fabra, Barcelona, Spain
jorge.vivaldi@upf.edu
[2]Technical University of Catalonia, Barcelona, Spain
horacio@lsi.upc.edu

Abstract. Domain terms are a useful resource for tuning both resources and NLP processors to domain specific tasks. This paper proposes a method for obtaining terms from potentially any domain using Wikipedia.

Keywords: term extraction, domain terminology, Wikipedia

1 Introduction

Even though many NLP resources and tools claim to be domain independent, its application to specific NLP tasks uses to be restricted to specific domains. As the accuracy of NLP resources degrades heavily when applied in environments different from which they were built; a tuning to the new environment is needed.

The basic knowledge sources, KS, needed for performing this tuning are domain restricted corpora and terminological lexicons. The latter is specially challenging and this is the goal of the work described here. Manual acquisition is costly and time consuming due to an extremely low level of agreement among experts [14]. Terminology extraction is more serious in domains in which the distinction between real terms and general words is difficult to establish preventing us of using un-restricted out of domain documents.

In this paper we present an approach for extracting terminological information for a given domain using the Wikipedia (WP) as main KS. It is domain/ language independent, we have applied it to two languages (Spanish and English) and to some randomly chosen domains. In section 2 we introduce both term extractions and WP. Then, in section 3 and 4 we present both our approach for obtaining the terminologies and its evaluation. Finally, in section 5 we present some conclusions and future work.

2 State of the art

Terms are usually defined as lexical units that designate concepts of a thematically restricted domain. As shown in [2] and [10], many methods have been proposed to extract terms from a corpus. Some of them are based on linguistic knowledge, like in [6]. Others use statistical measures, such as ANA [4]. Some approaches combine both linguistic knowledge and Statistics, such as [3] or [5]. A common limitation of most

extractors is that they do not use semantic knowledge, therefore their accuracy is limited. Notable exceptions are Metamap [1] and YATE [11].

WP is the largest on-line encyclopaedia; its information unit is the *Page* that basically describes a concept. The set of pages and their links in WP form a directed graph. A page is assigned to one or more WP categories in a way that categories can be seen as classes linked to pages. At the same time, a category is linked to one or more categories structuring themselves too as a graph. WP has been largely used as KS for extracting valuable information ([8]).

3 Our approach

In previous works we developed two alternative methods for extracting terminology for a domain using WP categories and pages as KS. The aim is to collect these units from WP such that their titles could be considered terms of the domain.

The first approach ([13]) follows a top down strategy starting in a manually defined top category for the domain. The problem of this approach was its limited recall due to the absolute dependence of the extracted term candidates on such category.

The second ([14]) follows a bottom up strategy. It starts with a list of TC, obtained from some domain specific text. In this approach both precision and recall are affected: i) the TC set is reduced to the list and ii) requires a top category that conditions the process as in the first approach.

In this paper we propose to combine both approaches to overcome these limitations. For accessing WP we have used Gurevych's JWPL [15]. Scaling up our methodology implies four additional not independent tasks over the work done previously, namely: i) choosing an appropriate domain taxonomy; ii) selection of category tops corresponding to the domains considered; iii) obtaining an initial set of TCs and iv) allowing a neutral automatic evaluation.

As domains taxonomy we use Magnini's Domain Codes, *MDC* [7]. Such codes enrich WordNet[1]. We can use WN for a cheap, though partial, evaluation of our method.

Our claim is that our method could be applied to any language owning a relatively rich WP. However, the results presented in this paper are reduced to English and Spanish and a randomly[2] selected subset of MDC consisting of 6 domains is presented and discussed. Figure 2 presents the overall process, it is organized into 8 steps (step 6 is iterated until convergence). The overall process is repeated for the two languages and domains involved (Agriculture, Architecture, Anthropology, Medicine, Music and Tourism). From now on let *lang* be the language considered and *dc* the Magnini's domain code, in *MDC*. The first step of our method consists of extracting from the WN corresponding to the language *lang* all the variants contained in all the synsets tagged with domain code *dc*. This results on our first set of TC, $terms_0$.

The second step consists of mapping *dc* to a set of WP categories. First we look whether dc occurs in the WP category graph (CG). If it is the case (it is true for 90% of *dc* for English), the set {dc} is selected. Otherwise we look if *dc* occurs in the WP

[1] http://wordnet.princeton.edu/

[2] Medicine has been included for allowing an objective evaluation, as reported in section 4.

page graph (PG). If this case we obtain the categories attached to the page. Otherwise a manual assignment, based on an inspection of WP is performed. The step results on an initial set of categories $categories_0$.

$categories_0$ contains mostly a unique category but when it has been built from a page it can contain noisy categories. In the third step $categories_0$ is cleaned by removing neutral categories and categories attached to domain codes placed above *dc* in MDC taxonomy.

The basis of our approach consists of locating two subgraphs, *CatSet* in CG, and *PageSet* in PPG having a high probability of referring to concepts in the domain, our guess is that the titles of both sets are terms of the domain.

Step 4 builds the initial set of categories, $CatSet_0$, expanding the tops. Starting in the top categories of dc, CG is traversed top down, avoiding cycles, performing cleaning as in step 3[3]. The categories in this initial set are scored, using only the links to parent categories, as shown in formula (1), then all categories with scores less than 0.5 are removed from the set resulting in our initial set, $CatSet_0$, as shown in Figure 2.

$$score_{cat} = \frac{\left|parents_{cat}^{ok}\right|}{\left|parents_{cat}^{ok}\right| + \left|parents_{cat}^{ko}\right|} \qquad (1)$$

$parents_{cat}^{ok}$, $parents_{cat}^{ko}$: set of parents categories under/outside domain tops

In step 5 the initial set of pages, $PageSet_0$, is built. From each category in $CatSet_0$ the set of pages, following category-page links, is collected in $PageSet_0$. Each category is scored according to the scores of the pages it contains and each page is scored according both to the set of categories it belongs to and to the sets of pages pointing to/from it. Three thresholding mechanisms are used: Microstrict (accept a category if the number of member pages with positive score is greater than the number of pages with negative score), Microloose (similarly with greater or equal test), and Macro (using the components of such scores, i.e. the scores of the categories of the pages). Formula (2) formalizes the scoring function.

$$score_{page} = comb(score_{pag}^{ocats}, score_{pag}^{input}, score_{pag}^{output}) \qquad (2)$$

where

$$score_{cpagt}^{ocats} = \frac{\sum_{\forall cat \in cats(page)} score_{cat}}{\left|cats(page)\right|}$$ with cats(page)= set of categories of page

$$score_{pag}^{input} = \frac{\sum_{\forall p \in input(page)} score_{p}}{\left|input(page)\right|}$$ with input(page)= set of pages of pointing to page

$$score_{pag}^{output} = \frac{\sum_{\forall p \in output(page)} score_{p}}{\left|output(page)\right|}$$ with output(page)= set of pages pointed from page

and *comb* is a combination function of their arguments

Then, in step 6, we iteratively explore each category. This way the set of well scored pages and categories reinforce each other. Less scored categories and pages are

[3] WCG was preprocessed for attaching to every category the depth in the categories taxonomy.

removed from the corresponding sets. As seen in (2) and (3), a combination function is used to compute a global score of each page and category from their constituent scores. Several voting schemata have been tested. We choose a decision tree classifier using the constituent scores as features. A pair of classifiers, *isTermcat* and *isTermpage*, independent of language and domain, were learned. The process is iterated, leading in iteration i to $CatSet_i$, $PageSet_i$, until convergence[4]. All the sets $CatSet_i$ and $PageSet_i$, are collected for all the iterations for performing the following step.

$$score_{cat} = comb(score_{cat}^{strict}, score_{cat}^{loose}, score_{cat}^{micro}) \quad (3)$$

where

$$score_{cat}^{strict} = \frac{count_{\forall page = pages(cat)}\left(score_{page} > 0.5\right)}{\left|pages(cat)\right|}$$

$$score_{cat}^{loose} = \frac{count_{\forall page = pages(cat)}\left(score_{page} \geq 0.5\right)}{\left|pages(cat)\right|}$$

$$score_{cat}^{micro} = \frac{\forall page \in pages(cat)}{\left|pages(cat)\right|}$$

with *pages(cat)* = set of pages of cat

and *comb* is a combination function of their arguments

In step 7 a final filtering is performed for selecting from all the $CatSet_i$ and $PageSet_i$ corresponding to all the iteration the one with best F1. According to the way of building these sets (in step 6) it is clear that precision increases from one iteration to the following at a cost of a fall in recall, as some TC are removed in each iteration Before computing F1 both category and pages sets are merged into a unique term candidate set for each iteration (there are more elements in $PageSet_i$ than in $CatSet_i$ and the intersection of both sets is usually not null. Finally, we evaluate the results as shown in section 4.

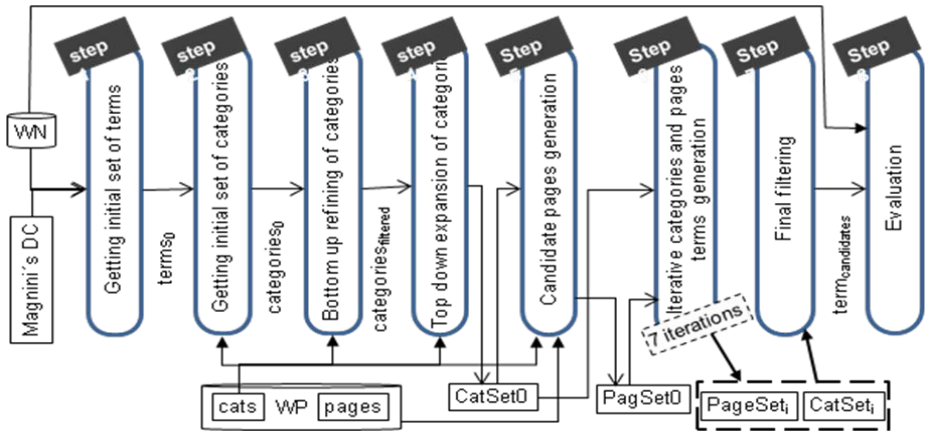


Fig. 1. Methodology

[4] In all the cases, convergence was reached in less than 7 iterations.

4 Evaluation

Evaluation of a terminology is a difficult task ([14]) due to a) the difficulty in doing it through human specialists, b) the lack/incompleteness of electronic reference resources and c) disagreement among them (specialists and/or reference resources).

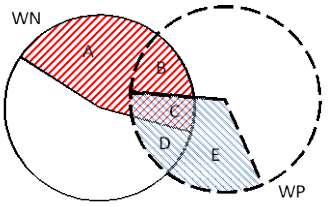
For this reason, we set two scenarios for evaluation. In the first one we analyze the results of Medicine for which we use SNOMED[5] as gold standard. In the second one, as we lack references our evaluation is only partial. Our thought is that the results in the Medicine domain related can be extrapolated to the others domains.

We use for comparison two baseline systems, one based on WN (Magnini) and the other based on the alignment of WN senses to WP pages in NG, [9].

Magnini baseline consists simply on, giving a domain code, *dc*, of Magnini's taxonomy, collecting all the synsets of WN assigned to *dc*, and considering as TCs all the variants related to these synsets. This approach has the obvious limitation of reducing coverage to the variants contained in WN; also it is rather crude because no score is attached to TCs, despite their degree of polisemy or domainhood.

NG map WP pages with WN synsets reaching a 0.78 F1 score. Our baseline is built collecting all the synsets corresponding to *dc* and from them all the WP pages aligned with the synset.

In the first scenario, the set of obtained TCs is compared with the two baselines for English and with the first one for Spanish and with the SNOMED repository. In the second scenario (covering the other domains) the comparisons are reduced to baselines. For both evaluations we need to consider the information shown in Figure 2[6].



$$\text{Precision} = \frac{C}{C+D} \qquad \text{Recall} = \frac{C}{B+C}$$

- A: WN domain variants not found in WP;
- B: WN domain variants found in WP but not considered in the domain by Magnini;
- C: WN domain variants found in WP;
- D: WN variants belonging to the domain according the WP but not according WN;
- E: WP pages/categories belonging to the domain but not found in WN;
- A+B+C: WN variants for a given domain;
- C+D+E: WP pages/categories discovered.

Fig. 2. Terms indirect evaluation

As shown in Figure 2, our system starts from the set of WN variants defined by [8], as belonging to the domain. Then it finds a number of WP pages and categories. Some of them are included in the set of variants already defined by Magnini but it also discovers new TC in WP. The evaluation can only be done using the terms already defined by Magnini and assuming their correctness. It is expected that terms discovered in WP will have similar precision values.[7]

[5] A comprehensive repository of Spanish/English terminology. See http://www.ihtsdo.org/

[6] The figure reflects Magnini's baseline, reflecting Niemann_Gurevych's is similar.

[7] Magnini assignment has been done in a semiautomatic way; therefore, they are not error free.

Using the sets of terms defined in Figure 2 we calculate the corresponding precision/recall values shown in Table 1. For each language and domain the initial number of WN variants and the precision/recall values are presented. As mentioned above such values are calculated against information obtained from the Magnini's domains. The table include also the results obtained using SNOMED.

Table 1. Results of the experiments (* at the best F1 values, ** evaluated using SNOMED-CT)

Domain		Tourism		Architecture		Music	
Language		EN	ES	EN	ES	EN	ES
Terms in WN	Total	744	441	303	143	1264	747
	In WP	554	286	244	112	1035	567
Precision [%]*	Cat.	33.33	100.00	0.00	85.71	50.57	50.00
	Page	15.65	85.71	36.59	59.52	11.11	27.42
Recall [%]*	Cat.	0.36	0.70	0.00	5.36	4.25	1.94
	Page	4.15	2.10	6.15	22.32	6.37	3.00
New Terms		1061	42	122	189	7046	614

Domain		Agriculture		Anthropology		Medicine			
Language		EN	ES	EN	ES	EN	EN**	ES	ES**
Terms in WN	Total	396	209	1106	651	2451		1595	
	In WP	238	137	909	443	1783		954	
Precision [%]*	Cat.	7.14	20.00	24.49	60.00	47.64	100.00	72.48	100.00
	Page	6.10	10.94	5.16	25.93	19.86	100.00	40.53	100.00
Recall [%]*	Cat.	0.42	0.73	1.32	0.68	10.21	6.56	11.32	16.25
	Page	10.50	5.11	5.06	1.58	16.32	9.76	15.93	54.51
New Terms		1491	193	6100	973	7855	3541	2225	2413

Table 2. Comparison of the results for Medicine/English among different approaches

Approaches	EWN	SNOMED	Precision	Recall
Ours	450	279	62.00	42.02
Magnini	1257	664	52.82	100.00
NG	190	150	78.95	22.59

A first consideration to be taken into account in analyzing the results shown in Table 1 is the own characteristics of WP as a source of domain terms. In particular:

- CG may change across languages. See for example Medicine and Veterinary. Although definitions are similar in both Spanish and English WPs, the former considers both as siblings whilst the latter considers it as a subcategory of former. This difference causes a large difference in the TC direct/indirect linked to them;
- English WP is a densely-linked resource; this causes unexpected relations among TC. Consider for example the domain "Agriculture" and the terms "abdomen" or "aorta". Both TCs are considered to be related to the domain due to a link among "Agriculture" → "veterinary medicine" which may be considered wrong;
- WP is an encyclopaedic resource; therefore, the termhood of some TC may be controversial. See for example: "list of architecture topics" in Architecture.

Low recall shown in Table 1 is due to the way of computing it, relating to terms in both WN/ WP. So, most of the extracted terms do not account for recall, eg, for tourism in English 1061 terms are extracted but only 25 of them occurs both in WN/WP. Due to the difficulties in the evaluation of the term lists, the characteristics of MDC and WP we perform additional evaluation for some domains. The results for Tourism were evaluated manually by the authors and the results for Medicine has been evaluated using SNOMED. Below we describe and analyse such additional evaluations.

1. Tourism (Spanish). We performed a manual evaluation of the TCs proposed. Partial evaluation takes as reference the list of EWN variants found in WP although, such variants not always are considered by WP to belong to the domain. Therefore it is possible to perform such evaluation taking into account this fact. It has been performed in two different ways for DC thresholds values ranging from 0 to 0.2:
   i) Precision/recall calculation: recall rises from 1.7 to 50%.
   ii) Error ration calculation: error rate decreases 70.96% to 0%.
2. Medicine. The use of SNOMED allows a better evaluation. The results show a considerably improvement in the precision/recall values (see Table 1, columns tagged with ** and Table 2). Magnini's offers the highest score in recall because the terms considered are all under its *dc* (ie. B in Fig. 2 is null). NG obtains the best score in precision with a low recall. Our results are in the middle.
3. Nevertheless there are some problems in using this repository such as:

   - Complex term: Some terms in this database are coordinated terms. See for example the Spanish TC: *enfermedades hereditarias y degenerativas del sistema nervioso central* (genetic and degenerative disorders of the central nervous system). It causes that none of the coordinated term are detected.
   - Some entries exist only as specialized. See for example the Spanish TC *glándula* (gland), it only exists as a more specialized terms like *glándula esofágica* (esophageal gland) or *glándula lagrimal* (lacrimal gland).
   - Number discrepancies among a WP category and the related SNOMED entry.
   - Missing terms like: *andrología* (andrology) or *arteria cerebelosa media* (medial cerebellar artery), present only in WP snapshot used for this experiment.
   - The results for Medicine and English are low. It is due to the number of entries, in our version, is much lower than those for Spanish (852K *vs* 138K).

5 Conclusions and future work

In this paper we present a new approach for obtaining the terminology of a domain using the category and page structures of WP in a language/domain independent way. This approach has been successfully applied to some domains and languages. As foreseen the results evaluation is a difficult task, mainly due to issues in the reference list. Also the encyclopaedic character of WP conditioned the list of new terms obtained. The performance may also change according the domain/language considered.

The current definition of domain (a set of WP categories) could be problematic when considering subdomains or interdisciplinary domains (like law, environment or information science). This will be a topic for future research/improvement.

In the future we plan to improve the final list of terms by: i) improve the exploration of the WP in order to reduce the false domain terms, ii) using the WP article text as a factor of pertinence of a page, iii) a better integration of both exploration procedures and iv) enlarge the number of proposed TC by using interwiki information.

Acknowledgement

This research has received support from the projects KNOW2 (TIN2009-14715-C04-04) and "RicoTerm 4" (FFI2010-21365-C03-01). We also thank the anonymous reviewers for their comments and suggestions.

References

1. Aronson A., Lang F.: An overview of MetaMap: historical perspective and recent advances. *JAMIA 2010* 17, p. 229-236 (2010).
2. Cabré M.T., Estopà R., Vivaldi J.: Automatic term detection. A review of current systems. *Recent Advances in Computational Terminology* 2, p. 53-87 (2001).
3. Drouin P.: Term extraction using non-technical corpora as a point of leverage. Terminology 9(1), p. 99-115 (2003).
4. Enguehard C., Pantera L.: Automatic Natural Acquisition of a Terminology. *Journal of Quantitative Linguistics* 2(1), p. 27-32 (1994).
5. Frantzi K. T., Ananiadou, S., Tsujii, J.: The C-value/NC-value Method of Automatic Recognition for Multi-word Terms. LNCS, Volume 1513, p. 585-604 (2009).
6. Heid, U., Jauß, S., Krüger K., Hofmann, A.: Term extraction with standard tools for corpus exploration. Experiece from German. In Proceedings of TKE'96. Berlin (1996).
7. Magnini B., Cavaglià G.: Integrating Subject Field Codes In WordNet. In 2nd LREC (2000).
8. Medelyan, O., Milne, D., Legg C., Witten, I. H.: Mining meaning from Wikipedia. International Journal of Human-Computer Studies 67 (9), p. 716-754 (2009).
9. E. Niemann, Gurevych I.: The People's Web meets Linguistic Knowledge: Automatic Sense Alignment of Wikipedia and WordNet. In: Proceedings of the 9th International Conference on Computational Semantics, p. 205-214 (2011).
10. Pazienza M.T., Pennacchiotti M., Zanzotto F.M.: Terminology Extraction: An Analysis of Linguistic and Statistical Approaches. *StudFuzz* 185, Springer-Verlag, p. 225-279 (2005).
11. Vivaldi J.: Extracción de candidatos a término mediante combinación de estrategias heterogéneas. PhD Thesis, Universitat Politècnica de Catalunya (2001).
12. Vivaldi J., Rodríguez H.: Evaluation of terms and term extraction systems: A practical approach. *Terminology* 13(2), p. 225-248 (2007).
13. Vivaldi J., Rodríguez H.: Finding Domain Terms using Wikipedia. In 7th LREC (2010).
14. Vivaldi J., Rodríguez H.: Using Wikipedia for term extraction in the biomedical domain: first experience. In *Procesamiento del Lenguaje Natural* 45, p. 251-254 (2010).
15. Zesch T., Müller C., Gurevych I.: Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In 6th LREC p. 1646-1652 (2008).

Searching for patterns in the transfer of multiword units: a corpus-based contrastive study on secondary term formation

Lara Sanz Vicente

Erasmushogeschool Brussel, Brussels, Belgium
marialara.sanz.vicente@ehb.be

Abstract. The dominance of English in specialised communication is currently emphasizing the importance of secondary term formation. In this respect, studying the way English multiword terms are transposed into other languages becomes of great interest. This paper reports on a corpus-based contrastive study that describes how multiword terms are formed in English and transferred into Spanish in the field of remote sensing of forest fires. The study particularly focuses on identifying patterns among these units and their language equivalents. The results reveal the existence of certain regularities which could be useful when transferring other multiword terms, but also report on the great structural diversity of the equivalents found for each source-language term.

Keywords: secondary term formation, multiword terms, corpus-based study, transferring procedures

1 Introduction

Secondary term formation, defined by Sager [17] as the process that 'occurs when a new term is created for a known concept [...] as a result of knowledge transfer to another linguistic community', is closely related to the transfer of multiword terms. These units derive from different formation procedures, the most frequently used being the addition of modifiers to an already existing term to reflect its specific properties [7, 17]: *infrared* > *near infrared*, *mid-infrared*, *short wavelength infrared*.

Most of the new terms created are multiword units. Preference for these forms in specialised languages has been noted by many authors, including Sager et al. [18], in the case of English, Kocourek [9] for French and Cabré [3] for Spanish. There are studies that quantitatively measure their importance, according to which they represent around 80% or more of the total vocabulary of certain domains [2]. Furthermore, it has been demonstrated that they are more frequent in highly specialised texts, i.e., texts written by and for experts [8, 15].

From a contrastive perspective, it has been noted that these units are a common cause of trouble in specialised translation, specially between Romance and Germanic languages. Some contrastive studies in multiword units between

English and Spanish are those of Salager-Meyer [19] in the medical field, Pugh [14] and Montero Fleta [11] in IT, Ahronian [1] on the Internet specifically, and Quiroz [15] in a body of texts on the genome. The translation of these multiword units from English into Spanish presents great difficulty due to their syntactic-semantic complexity, the differing syntactic natures of the two languages involved and their word formation rules, but also due to the lack of comparative studies and reference sources to understand and solve them.

The study reported in this paper describes and compares these type of terms in a different and recent field, remote sensing of forest fires, where English is the dominant language, i.e., the language of primary term formation. Using an English-Spanish comparable corpus of research articles, the study wants to go deeper into the knowledge of the secondary term formation process in highly specialised texts. The final aim is to assist translators in the identification, understanding and transfer of multiword units by providing strategies and offering a bilingual database which presents the results derived from the analysis.

The study is based on the belief that the description of the structures of multiword terms (MWTs) – of their morphosyntactic patterns and semantic contents – in a body of real texts can allow us to establish generalisations that increase understanding of these units and offer strategies for their translation. Specifically, the basic hypothesis is that there are certain patterns in the formation of MWTs in English and in their transfer and translation into Spanish. The results presented here derive from a contrastive analysis of MWTs carried out as part of a PhD dissertation. A detailed description of the corpus and methodology and a complete account of the results can be found in Sanz [20].

2 Methodology

2.1 Corpus design and term extraction

The research is based on a descriptive analysis of multiword units and their equivalents in a tailor-made English-Spanish comparable corpus composed of highly specialised texts on remote sensing of forest fires. The corpus compiled contains two subcorpora, English and Spanish, of 193,893 and 128,823 tokens respectively. It is composed of several research papers (35 in English and 38 in Spanish) published between 1992 and 2008 in peer-reviewed journals and conference proceedings, dealing with a specific subfield, burned area mapping. In both languages, it contains original texts – not translations – of similar characteristics and compositions as regards text type and origin, topic, size, setting, date of publication, etc., meaning that cross-linguistic comparisons can be drawn. Topic has been a relevant criterion to guarantee comparability between subcorpora. Research papers had to contain a set of keywords in the title, abstract and within the full text in order to be selected.

A collection of 12 glossaries and dictionaries of different types and sizes was also formed manually with the objective of contrasting and complementing the results obtained from the corpus: half of them specialised in the field of remote

sensing, three in forestry sciences and forest fires, and the rest concerning larger fields including or relating to remote sensing, such as geomatics and aerospace sciences and technology. Five of them are monolingual English dictionaries, three Spanish-English, one English-Spanish, two French-English, and one multilingual (French, Catalan, Spanish, Galician, Italian, Portuguese and English).

The selection of MWTs from the corpus was conducted with the aid of WordSmith Tools [21] and was done in a two-step process: first, drawing up a list of English MWTs and then, attempting to identify correspondences in the corpus of texts in Spanish.

The extraction from the English corpus was mainly based on drafting single-word and multiword wordlists (lists of clusters)[1] and concordance searches using both automatic and manual processes at all times.

The first step was the extraction of English word clusters. We computed 2–12 word clusters with a minimum frequency of occurrence of 3 in the corpus and with a relevant distribution through it, i.e., involved in at least three different texts. From the list obtained, only noun sequences were selected with the aid of a single-word frequency list filtered using a stoplist of high frequency words without specific meaning (articles, pronouns...) that were excluded. Generating single-word frequency lists helped in finding the most relevant units in the field and identifying keywords that could possibly work as nuclei or modifiers of multiword term candidates. The resulting candidates were then grouped into lemmas (*active fire*, *active fires*; *burned area*, *burnt area*; etc.), and inflectional and orthographic variants were also detected and grouped under the same heading.

A total of 460 English MWTs with different levels of lexicalisation were subsequently identified, precisely those complying with he characteristics linked to MWTs, which refer to: i) its morphological structure, formed by a noun (nucleus) accompanied by one or several modifiers, ii) its unity and semantic specificity within the conceptual system of the targeted specialised field, iii) its syntactic function as a minimum independent component of a sentence, and iv) its proximity to specialised phraseological units. This was done not only by producing concordances but also with the help of the dictionaries and glossaries collected and by consulting experts in the field.

The process of searching for equivalents within the Spanish subcorpus was based on compiling concordance lists from possible translations into that language of the 460 English MWTs identified and from the possible translations of their nuclei or modifiers or their most representative collocates. This method is closely related to those proposed for automatic extraction of bilingual terminology from comparable corpora. Most of them are based on the idea that across languages there is a semantic correlation between the co-occurrences of words that are translations of each other [4–6, 16]. Searching for equivalents was also supported by single-word and multiword wordlists in Spanish and with the help of the glossaries and dictionaries collected. It enabled us to find corresponding Spanish terms for 80% of the English MWTs.

[1] In WordSmith Tools multiwords are called *clusters* and defined as 'words which are found repeatedly together in each others company, in sequence' [22].

2.2 Data analysis

The analysis performed centered, first, and both for English and Spanish MWTs, on the manual description of the morphologic structure and substructure of each term (*burned area mapping* > Adj+N+N > Adj$_{-ed}$+N+N$_{-ing}$), and on the identification of the role played by each component element (nucleus or modifier) to be able to represent their morphosyntactic scheme and intraterm semantic relation too (*burned area mapping* > [(Adj+N)$_{Mod}$+N$_{Nuc}$] > PATIENT – ACTION). The intraterm semantic relations of the multiword terms were manually identified and classified using Oster's typology of semantic relational schemas [12, 13] – slightly modified to take account of all of the relationships observed in MWTs in the field under study[2]. The analysis and understanding of the internal syntactical-semantic structure of these units was thus considered an essential step which first required the identification and categorisation of the modifiers linked to the nucleus or nuclei. The information on syntactic and semantic relationships could only be recovered by returning to the context (the text) in which the term was produced and is used, taking all extralinguistic parameters involved into account.

A comparative analysis was carried out afterwards between the English MWTs and their equivalents, which were interpreted as translation equivalents. This analysis was performed in the English-Spanish direction by describing the equivalents of the English MWTs as regards their morphosyntactic and semantic structure and the influence of English in them.

We compared the morphological and morphosyntactic structure of the English MWTs and their equivalents, and how the English MWTs' intraterm semantic relationships materialised in Spanish. That involved studying the correlation between the English MWTs' semantic relationship and the form of the equivalent terms in Spanish.

Finally, the Spanish equivalents were classified according to the strategies applied when importing them into Spanish. This classification, specifically defined for this analysis, included ten basic procedures: borrowing, calquing, paraphrasing, adaptation, transposition, modulation, synonymy, clarification, shortening and endogenous formation, and paid special attention to calquing, as the most important procedure regarding the transfer of MWTs. The classification, therefore, differentiates between calques of expression and structural calques, which, in turn, have been subdivided into two groups: full translation (literal or free) and half translations (literal or free). Attention was also drawn to the procedures most frequently used to import each of the elements of the MWTs separately.

[2] Oster [12, 13] defines semantic intraterm relations as the semantic relation between two concepts *a* and *b* expressed through the combination of the functions carried out by *a* and *b* with respect to each other. For example, *burned area mapping* will be understood as a PATIENT – ACTION relation, where *mapping* performs an action on the patient, *burned area.*

3 Results of the English-Spanish contrastive study

The comparison of the English MWTs' structures with those of their Spanish equivalents demonstrated that there are certain regularities in the translation of these units. As shown in Table 1, of the 30 solutions observed in Spanish for the English morphosyntactic structure [$N2_{Mod}$+$N1_{Nuc}$], the most frequent one, four are highly productive, accounting for more than 70% of the equivalents produced using this English construction: [$N1_{Nuc}$+(prep+N2)$_{Mod}$] (EN. *brightness temperature* → ES. *temperatura de brillo*), [$N1_{Nuc}$+(prep+art+N2)$_{Mod}$] (EN. *infrared band* → ES. *banda del infrarrojo*), [N_{Nuc}+Adj_{Mod}] (EN. *cloud pixel* → ES. *píxel nuboso*) and [$N1_{Nuc}$+$N2_{Mod}$] (EN. *difference image* → ES. *imagen diferencia*). Equally, the second most frequently-used structure in English, [Adj_{Mod}+N_{Nuc}], is matched with the reverse structure [N_{Nuc}+Adj/Pp_{Mod}] in 55% of cases in Spanish (EN. *ancillary data* → ES. *datos auxiliares*, EN. *contaminated pixel* → ES. *píxel contaminado*) and in 10% as [$N1_{Nuc}$+(prep+N2)$_{Mod}$] (EN. *contextual algorithm* → ES. *algoritmo de contexto*).

Table 1. English-Spanish structure correspondences of N+N and Adj+N English multiword terms

English multiword terms		**Spanish equivalents**			
Morphological structure	Morphosyntactic structure	Morphological structures	Morphosyntactic structure	N.	%
N+N	[$N2_{Mod}$+$N1_{Nuc}$]	N+prep+N	[$N1_{Nuc}$+(prep+N2)$_{Mod}$]	85	26.23
		N+prep+art+N	[$N1_{Nuc}$+(prep+art+N2)$_{Mod}$]	73	22.53
		N+Adj	[N_{Nuc}+Adj_{Mod}]	43	13.27
		N+N	[$N1_{Nuc}$+$N2_{Mod}$]	40	12.35
		N	[N_{Nuc}]	13	4.01
		other (25)		70	21.61
Adj+N	[Adj_{Mod}+N_{Nuc}]	N+Adj	[N_{Nuc}+Adj_{Mod}]	80	47.06
		N+prep+N	[$N1_{Nuc}$+(prep+N2)$_{Mod}$]	17	10.00
		N+Adv+Pp	[N_{Nuc}+(Adv+Pp)$_{Mod}$]	15	8.82
		N+Pp	[N_{Nuc}+Pp_{Mod}]	14	8.24
		N+prep+art+N	[$N1_{Nuc}$+(prep+art+N2)$_{Mod}$]	6	3.53
		other (17)		38	22.35

N: Noun; Adj: Adjective; prep: preposition; art: article; Pp: Past participle; Adv: Adverb; Mod: Modifier; Nuc: Nucleus

As for the MWT equivalents with three or more elements, it has been observed that their structures vary based on the syntactical dependency shown by the English MWTs. For example, the Adj+N+N MWTs with dependency [(C+B)$_{Mod}$+A_{Nuc}] are generally translated as N+prep+(art)+N+Pp/Adj (EN. *burned area mapping* → ES. *cartografía de (las) áreas quemadas*, EN. *spectral mixture analysis* → ES. *análisis de mezclas espectrales*), while the most frequent solution for compounds Adj+N+N with dependency [C_{Mod}+(B+A)$_{Nuc}$]

is N+Adj+prep+(art)+N (EN. *viewing zenith angle* → ES. *ángulo cenital de observación*).

Furthermore, the analysis by substructures has shown that in those cases where Spanish uses prepositional phrases to add the modifying element to the nucleus, the connecting preposition most often used is *de*, which is used as the wild card preposition sometimes replacing prepositions with a more specific meaning (EN. *omission error* → ES. *error de omisión/error por omisión*).

The analysis of the intraterm semantic relationships showed that the most frequently-used schema in English MWTs, PROPERTY – DETERMINED ENTITY, which is almost always expressed using the structure [Adj_{Mod}+N_{Nuc}], is essentially formulated with the reverse structure in Spanish, [N_{Nuc}+Adj_{Mod}], (EN. *spectral signature* → ES. *firma espectral*). The second most frequent in English, ORIGIN – DETERMINED ENTITY, mainly expressed in that language using the form [N2_{Mod}+N1_{Nuc}] to denominate remote sensing images according to the sensor or satellite they come from, is translated in Spanish as [N1_{Nuc}+N2_{Mod}], using the sensor or satellite's name as a direct modifier (EN. *AVHRR image* → ES. *imagen AVHRR*) or, sometimes, by connecting it with the preposition *de* plus an article (EN. *Landsat imagery* → ES. *imágenes del Landsat*). The third most often used schema, PATIENT – ACTION, expressed in English with N+N compounds (*change detection*) and Adj+N+N (*burned area mapping*), mainly gives rise to prepositional constructions with *de* in Spanish (*detección de cambios, cartografía de áreas quemadas*). In general, it has been observed that prepositional constructions with *de* serve to express all sorts of semantic relationships.

The results of the classification of MWTs by transferring procedures confirmed that the majority of Spanish equivalents (66%) are translated and imported as calques of expression with full translation of the English MWT, literal in most cases (EN. *active fire* → ES. *incendio activo*) and, to a lesser extent, free (EN. *active fire* → ES. *foco activo*). The second most used resource is explicative paraphrasing (13%), which reformulates the meaning of the English term (EN. *burn signal* → ES. *señal procedente de las áreas quemadas*). In third place, with 5%, are calques of expression containing unadapted loans (mainly initialisms and acronyms) which consist of a literal translation (EN. *AVHRR image* → ES. *imagen AVHRR*). These are followed by unadapted loans, which are not very numerous (4%) and which mainly correspond to the proper names of sensors and satellites expressed as initialism compounds (*NOAA-AVHRR, NOAA-11, Landsat ETM+*) and to some image analysis and interpretation techniques (*Maximum Value Composite, Normalized Burn Ratio*).

As regards the procedures most often used to import each of the elements of the MWTs separately, three are noteworthy: i) transpositions, among which changes from singular to plural prevail (EN. *cloud shadow* → ES. *sombra de nubes*) and noun to adjective (EN. *azimuth angle* → ES. *ángulo acimutal*); ii) clarifications, which involve the inclusion of some elements that were implicit in the English forms, such as prepositions (EN. *colour composite* → ES. *composición en color*); and ii) modulation, based, above all, on the use of partial synonyms (EN. *statistic* → ES. *índice*).

4 Conclusions

The results reveal the existence of certain regularities which guide the transposition of these MWTs into Spanish and which could be therefore useful in translation. Generalising greatly, it could be concluded that the prepositional construction N+*de*+N is mainly used to translate N+N English MWTs and N+Adj to translate Adj+N MWTs. This data, set out in this manner, could lead some to believe that a linear translation rule (right to left) exists, as suggested in some English-Spanish translation manuals [23, 10].

However, comparing the structures of the English MWTs with those of their Spanish equivalents clearly shows, that for each English structure there are many divergent structures in Spanish. The English structure N+N alone has up to 30 different corresponding structures in Spanish. Furthermore, where the MWT features two or more premodifiers in English, its Spanish equivalents' structures vary more widely mainly due to an increase in the variety of possible translations for each source-language term (EN. *burned area mapping algorithm* → ES. *algoritmo para la cartografía de áreas quemadas*, *algoritmo para cartografiar áreas quemadas*, *algoritmo para la producción de mapas de área quemada*), and sometimes because of difficulties in understanding English units (EN. *maximum value composite* → ES. *composición del máximo valor*, **máximo valor compuesto*).

Besides, translation arises as the most important procedure in transferring English MWTs to Spanish. The results have shown that the preferred mechanism in importing these units into Spanish is calques of expression, i.e., a mechanism that respects the syntactic structures of the target language and, more specifically, that consists of a literal translation of the English MWT. This demonstrates the influence English has on Spanish formation of these units within the area being studied. This preference for calques (loan translations) means we should consider to what extent they act as a terminologically innovative and enriching element in the language of secondary word formation.

References

1. Ahronian, C.: Les noms composés anglais français et espagnols du domaine d'Internet. PhD thesis, Université Lumière-Lyon 2 (2005)
2. Boulanger, J.C., Nakos-Aupetit, D.: Le syntagme terminologique: bibliographie séléctive et analytique 1960-1988. Reference materials - bibliographies – multilingual/bilingual materials, Centre international de recherche sur le bilinguisme (CIRB), Université Laval, Québec (1988)
3. Cabré, M.: Terminology. Theory, Methods and Applications. John Benjamins, Amsterdam/Philadelphia (1999)
4. Déjean, H., Gaussier, E., Sadat, F.: An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In: COLING'02 Proceedings of the 19th International Conference on Computational Linguistics., Taipei, Taiwan (2002) 1–7
5. Fung, P.: Parallel text processing. In Véronis, J., ed.: A statistical view on bilingual lexicon extraction - from parallel corpora to nonparallel corpora, Dordrecht, Kluwer Academic Publishers (2000) 1–17

6. Gaussier, É., Renders, J.M., Matveeva, I., Goutte, C., Déjean, H.: A geometric view on bilingual lexicon extraction from comparable corpora. In: Proceedings of the 42nd annual meeting of the Association for Computational Linguistics (ACL). (2004) 526–533
7. Guilbert, L.: La dérivation syntagmatique dans les vocabulaires scientifiques et techniques. In Dabène, M., Gaultier, M.T., eds.: Les langues de spécialité. Analyse linguistique et recherche pédagogique. Actes du Stage du Saint-Cloud, 23-30 Nov. 1967, Strasbourg, AIDELA (1970) 116–125
8. Horsella, M., Pérez, F.: Nominal compounds in chemical English literature: Towards an approach to text typology. English for Specific Purposes **10**(2) (1991) 125–138
9. Kocourek, R.: La langue française de la technique et de la science. 2nd (1991) edn. Brandstetter Verlag, Wiesbaden (1982)
10. López-Guix, J.G., Minett, J.: Manual de traducción: inglés/castellano. Gedisa, Barcelona (1999)
11. Montero-Fleta, M.B.: Technical communication: complex nominals used to express new concepts in scientific English. The ESP **17**(1) (1996) 57–72
12. Oster, U.: Las relaciones semánticas de términos polilexemáticos. Peter Lang, Frankfurt am Main (2005)
13. Oster, U.: Classifying domain-specific intraterm relations: a schema-based approach. Terminology **12**(1) (2006) 1–17
14. Pugh, J.: Contrastive conceptual analysis of noun compound terms in English, French and Spanish within a restricted, specialized domain. In Hartmann, R.R.K., ed.: Lexeter'83 proceedings. Papers from the International Conference on Lexicography at Exeter, 9-12 Sep. 1983, Tübingen, Max Niemeyer Verlag (1984) 395–400
15. Quiroz-Herrera, G.Á.: Los sintagmas nominales extensos especializados en inglés y en español: descripción y clasificación en un corpus de genoma. PhD thesis, Universitat Pompeu Fabra (2008)
16. Rapp, R.: Automatic identification of word translations from unrelated English and German corpora. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics (ACL), Maryland, Association for Computational Linguistics (1999) 519–526
17. Sager, J.C.: A Practical Course in Terminology Processing. John Benjamins, Amsterdam/Philadelphia (1990)
18. Sager, J.C., Dungworth, D., McDonald, P.F.: English Special Languages: Principles and practice in science and technology. Brandstetter Verlag, Wiesbaden (1980)
19. Salager-Meyer, F.: Syntax and semantics of compound nominal phrases in medical English literature: a comparative study with Spanish. English for Specific Purposes Newsletter **95** (1985) 6–11
20. Sanz-Vicente, M.L.: Análisis contrastivo de la terminología de la teledetección. La traducción de compuestos sintagmáticos nominales del inglés al español. PhD thesis, Universidad de Salamanca (2011)
21. Scott, M.: WordSmith Tools (version 5). Lexical Analysis Software, Liverpool (2008)
22. Scott, M.: WordSmith Tools Help. Lexical Analysis Software, Liverpool (2011)
23. Váquez-Ayora, G.: Introducción a la traductología; Curso básico de traducción. Georgetown U.P. (1977)

Terminology Harmonization in Industry Classification Standards

Dagmar Gromann[1] and Thierry Declerck[2]

[1] Vienna University of Economics and Business,
Nordbergstrasse 15, 1090 Vienna, Austria
dgromann@wu.ac.at
[2] DFKI GmbH, Language Technology Department,
Stuhlsatzenhausweg 3, D-66123 Saarbruecken, Germany
declerck@dfki.de

Abstract. Terminology as a research area has shifted from portraying terms as lexical units to a concept-oriented approach. Accordingly, the process of terminology harmonization has to cope with the concept orientation of term entries. One approach to harmonization is the integration of several terminologies into one centralized terminology repository, which is either formalized as a conceptual system or points to such systems. In contrast, we propose an approach adopting the linked data strategy by linking resources that preserve the initial terminologies with the corresponding lexical items and the related ontology concepts. As ontologies traditionally link concepts but not the natural language designation of concepts, we propose a model that utilizes terminologies for terminological and ontology lexicons for morpho-syntactic information. We illustrate our suggested approach, applying it to closely related but competing industry classification standards.

Keywords: Terminology, lexicon, ontology, harmonization, industry classification

1 Introduction

Industry classification standards allow for a thorough analysis of the industrial landscape. Investors and asset managers rely on the transparency these standards offer by means of global comparisons by industry. But despite of very similar categories, (competing) systems of industry classification often employ different terminology. Harmonization of these systems experiences issues not only on the terminological level, also on the hierarchical level various degrees of granularity can be observed. For instance, the Industry Classification Benchmark (ICB)[3] defines and refers to *Banks*, whereas the Global Industry Classification Standard (GICS)[4] differentiates between *Diversified Banks*, *Regional Banks*, and *Thrifts & Mortgage Finance*. A strategy for harmonization could consist in subsuming

[3] http://www.icbenchmark.com/
[4] http://www.standardandpoors.com/indices/gics/en/us

these categories under one concept or modifying the existing classifications in order to make them interoperable.

Alternatively, our approach suggests a strategy based on the linked data [10] framework in that harmonization is achieved by interlinking terminologies, including their associated lexicons and related ontology concepts. Connecting these resources by means of formal languages, such as the Resource Description Framework (RDF)[5] and the Simple Knowledge Organization System (SKOS)[6], enables the preservation of the original classification ID for all terms and their variants, as well as the concepts they are associated with.

At the end of the day, nothing can be said against still opting for a new, centralized and unique terminology in case the linking mechanisms reveal consistent overall similarities and/or suggest the possibility of an integrative re-organization of the various knowledge sources.

2 Research Background

Term banks initially portrayed terms as lexical units [8], overloading the term with different meanings. Gradually, a concept-oriented approach developed, emphasizing the relationship of one concept per term entry [3]. Recent developments view terminological resources as expert systems, focusing on a knowledge-oriented approach [8]. For instance, César et al. [12] harmonize a wide variety of standards regarding the improvement of software processes with a focus on terminology. Ontologies are applied to the task of eliminating inconsistencies on a semantic and conceptual level, implicitly harmonizing the terminology [12].

The TermSciences initiative [17] establishes semantic relations among medical terminologies, by means of TMF-compliant metadata. Ontologies or high-level terminologies serve the unification process of different resources. Nevertheless, the project centers around merging, grouping, restructuring resources, converting term-centered representations to concept-oriented ones. Our proposal focuses on the benefit of different conceptualizations, i.e. ontological, terminological, lexical, to the process of harmonization with a very clear emphasis on terminology rather than controlled vocabularies and a preservation of its integrity and origination.

Several models exist to account for the terminological dimension of ontologies such as ontoterminology [16], termontography [14], or the TERMINAE method [15]. Whereas the latter two focus on the establishment of one terminology for or in combination with an ontology, the former emphasizes the differences. Roche et al. [16] highlight the importance of separating the linguistic and the conceptual dimension of terminology and ontology, as terms cannot simply be reduced to the textual content of `rdfs:label` or `rdfs:comment` annotation properties without any linguistic layer.

The model for the integration of conceptual, terminological and linguistic objects in ontologies (CTL) [1] uses the TERMINAE method [15] and the *LexInfo* metamodel [4] to obtain a modular and multi-layered linguistic annotation of

[5] `http://www.w3.org/RDF/`
[6] `http://www.w3.org/2004/02/skos/`

ontology labels, further detailed in [2]. Expanding on the CTL model [1] and formalizing the approach, we focus on separating the lexical, syntactic, terminological and (domain) semantic levels into adequate resources, linking them with RDF and SKOS. Lexical and syntactic descriptions will be provided using *lemon*, a Lexicon Model for Ontologies [11]. The *lemon* model offers a formal representation of linguistic information to be associated with the word forms contained in the `rdfs:label` annotation property of ontology classes, and with a clear referential mechanisms to ontology classes, thus defining the semantic of such linguistic expressions by their references to concepts.

3 Industry Classification Systems

Industry classification systems aim at providing a comparison of companies across nations. Due to numerous and often competing classification systems, the resulting overlapping and inconsistent terminologies require harmonization on a conceptual and term level, including the harmonization of the linguistic properties of the tokens building the term. In the following, we suggest a linking approach for harmonizing two major industry classification systems.

The Global Industry Classification Standard (GICS) represents a taxonomy of industry sectors developed by MSCI and Standard & Poor's[7]. The GICS structure consists of 10 sectors, 24 industry groups, 68 industries and 154 subindustries into which all major companies have been categorized. The ten main industries are: Energy, Materials, Industrials, Consumer Discretionary, Consumer Staples, Healthcare, Financials, Information Technology, Telecommunication Services, Utilities.

Similar to the GICS, the Industry Classification Benchmark (ICB) developed by Dow Jones and FTSE[8] consists of four major levels. The system is organized into 10 industries, 20 supersectors, 41 sectors and 114 subsectors. The ten main industries are: Oil & Gas, Basic Materials, Industrials, Consumer Goods, Healthcare, Consumer Services, Telecommunications, Utilities, Financials and Technology.

In comparison, both systems classify a company according to its principal business, apply four major levels to their structure and have a comparable number of subcategories. In both cases the categories are organised in a hierarchical tree. Intermediate nodes are labelled with short natural language strings and the leaf nodes are equipped with (partly lengthy) definitions. Both systems are delivered in several languages

One major difference is to be found in the consumers section. GICS differentiates between staples and discretionary containing both goods and services, whereas ICB distinguishes consumer goods from consumer services. As this regards the top-level classification, it is an important aspect to be considered in

[7] See respectively `http://www.msci.com/products/indices/sector/gics/` and `http://www.standardandpoors.com/indices/gics/en/us`

[8] See `http://www.ftse.com/Indices/Industry\_Classification\_Benchmark/index.jsp`

the harmonization strategy. Naturally, the terms used to designate equivalent categories differ substantially.

4 Three-layered Model for Harmonization

Conceptual structures in an ontology differ from those in terminologies. The ontology links on the basis of domain knowledge, whereas the terminology links on a linguistic and language-related background. The combination of both types of information seems to be beneficial to the process of harmonization. Our model illustrated illustrated in Fig. 1 utilizes terminologies – complying with the Terminological Markup Framework (TMF) [7] – in combination with ontologies to create a net of labels interlinked with SKOS and RDF(S). In order to clearly distinguish between terminological and morpho-syntactic information, we additionally include a lexicon level to be represented using *lemon*.

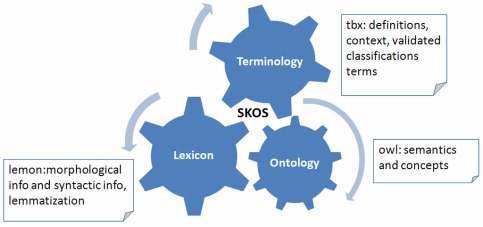


Fig. 1. Three-layered model for harmonization

Each component of the system represents different aspects of the net of labels. Firstly, the lexicon mainly provides information on basic lexical, morphological and syntactic information. Secondly, the terminology (in TMF) as such represents the validated terms and "soft" variants [9] such as synonyms, acronyms and orthographic variants. Finally, ontologies provide the (domain) semantic layer. The suggested layered model allows thus to state whether a term varies morphologically or semantically. The resulting net of labels contains the original classification ID of each term, whether it is a preferred term or normalized form, etc., rich linguistic information and a thorough conceptual basis provided by the ontology.

In detail, within the process of creating the terminologies we apply general principles such as concept orientation [3], term consistency, etc. to validate the classifications' terminology. The harmonization strategy is a two-fold contrastive approach considering the conceptual level of terminology and its designations. Term harmonization either refers to the designation of one concept by terms or the establishment of equivalences across languages or term variations in one language [5].

5 Harmonizing Industry Classification Systems

Subsequent to obtaining the multilingual taxonomies from the respective web presences of the industry classifications, we utilized the source data to create terminologies and ontologies, lexicalizing the latter. This entire process abides to the current ISO standards for terminology (ISO1087, ISO704) and harmonization [5], proposing an extension of the latter.

5.1 From Source Data to Terminology

Based on the resources provided by ICB and GICS we created one TermBase eXchange (TBX) [6] format term base for each classification, which allows for a semi-formal representation of the multilingual terminology and for a validation of the classifications' terminology. The initial analysis of the input data necessitated the harmonization of terms on several levels. At times designations provided pleonastic information as illustrated in the following example:

```
<termEntry id="ICB1779">
  <descrip type="subjectField">mining</descrip>
  <descrip type="definition">ICB sector</descrip>
  <langSet xml:lang="en">
       <descrip type="definition">Companies producing and exploring platinum,
         silver and other precious metals not defined elsewhere.</descrip>
       <tig>
              <term>Precious Metals</term>
              <termNote type="partOfSpeech">noun</termNote>
       </tig>
  </langSet>
</termEntry>
```

[Simplified TermBase eXchange (TBX) example of the ICB terminology.]

As the definition clearly classifies platinum as precious metal, it represents a case of pleonasm. Thus, the entry was adapted to "Precious Metals" in the term base. Similarly, the use of homonymous designations for different categories on the same hierarchical level has to be avoided in the terminologies, such as the ICB classification containing two sibling sectors both defining mining.

Concept orientation refers to the fact that each term entry contains the full terminological data for the respective concept [3]. GICS designates a mining category "Steel," but the definition clearly states that it classifies "Producers of iron and steel and related products" - only referring to steel could infringe the integrity of this terminological entry. Additionally, term consistency is often an issue in combination with concept orientation. In contrast to its sibling, the ICB subsector "Exploration & Production" does not refer to its supersector *Oil & Gas Producers* in its designation. On the basis of the definition provided it can be adapted to "Oil & Gas Exploration & Production" in order to improve both consistent terminology and concept orientation.

The presented methodology clearly employs a bottom-up approach, analyzing the leaf nodes first. This initial analysis represents a prerequisite step for the actual harmonization on a terminological and conceptual level.

5.2 Harmonization Steps

The process of concept harmonization usually precedes the process of term harmonization [5]. In case the concepts are equivalent, a correspondence between them can be established. For instance, the definition "Residential Retail Estate Investment Trusts (REITs)" can be aligned directly in both classifications by `skos:exactMatch` as they are orthographically and semantically identical. However, most cases are more complicated.

Lexical information are represented by the *lemon* model. Although the representation of term variation is not the primary objective of the lexicon-ontology model, it is generally possible [13]. *lemon* creates sense objects that refer to one ontology concept (semantic by reference). The whole *lemon* entry is used to refer to a concept, not the canonical or the alternative form of the term. But one would like to be able to state that a term used in a category of a classification system is an alternate form of a term that is used in a category of another classification, while the two categories can be related by an equivalence relation. In *lemon* two lexical entries have to be created for this purpose.

TMF neither provides a solution to this problem of including two terms in one term entry while preserving the original source by means of the reference ID of both terms as they are used in their respective classification system. TBX allows for the inclusion of synonyms in an entry and also variants, but each entry has one ID. As with lemon, two entries are needed to establish the equivalence or relation between the terms by means of a cross-reference.

The objective is to obtain two equivalent and equal terms referring to their original ID and to establish the harmonization by means of relations. Thus, it is up to the user to decide to which term entry the information extracted by the ontology-based system should be mapped. The harmonization is accomplished by means of relations utilizing SKOS and RDF(S), as illustrated below `mfo` meaning "Multilingual Financial Ontology".

```
tbx:ICB rdf:type skos:ConceptScheme.
mfo:ICB rdf:type skos:ConceptScheme.

lemon:full_line_insurance rdf:type skos:Concept;
   lemon:canonicalForm [lemon:writtenRep "Full line insurance"@en  ] ;
   lemon:reference <http://icb.org/ICB8532>  ;
   skos:inScheme mfo:ICB ;
   skos:inScheme tbx:ICB.

tbx:GICS rdf:type skos:ConceptScheme.
mfo:GICS rdf:type skos:ConceptScheme.

lemon:multi_line_insurance rdf:type skos:Concept;
   lemon:canonicalForm [lemon:writtenRep "Multi-line insurance"@en  ] ;
   lemon:reference <http://gics.org/GICS40301030> ;
   skos:inScheme mfo:GICS ;
   skos:inScheme tbx:GICS.

<http://icb.org/ICB8532> skos:closeMatch <http://gics.org/GICS40301030>.
```

[Linking the labels of a GICS and an ICB concept, by means of SKOS.]

The example shows how the ontology concept points to the terminology, which in turn is linked with the lexicon. The `closeMatch` indicates that the two concepts are sufficiently aligned to be used interchangeably. And so the associated labels (lemon entries that refer to the concepts) can be interlinked. We can not apply the skos matching mechanisms directly to the lemon entries, since we want to establish a semantic interoperability, and not a string-based one. The aspect of multilingualism represents an additional challenge, as terms in different languages might not be truly harmonized within one entry, even if it is less an issue with such a standardized representation of terms.

Finally, we created frequency lists for each classification and found that several phrases or words are only mentioned in the definition, but not in the designations of the classifications. Whereas the ICB definitions contain the term "company" 62 times, it is not to be found once in the designations of the classification. Similar statistics apply to manufacturer, producer, distributor to name but a few. Due to the predominance of company, we decided to add the term to the ontology and apply it to labels where no other business activity is predominant. In case of several types of business activity, consistency calls for the use of company again. However, the major basis for this decision is provided by the definition. One example of GICS is the subsector "Aluminum," which as such can clearly not be identified as ontologically valid or conceptually sound, as it does not provide any information on company. Thus, we decided to introduce a superordinate node for the concept *company*.

6 Conclusion

Confronted with a variety of competing schemes in the field of industry classification, we investigated the possibility to harmonize their respective terminology, also for the benefit of a multilingual information extraction task, which has to map textual data in the financial domain to concepts described in such classification systems. We opted for an approach that proposes a three-fold model, clearly separating lexical, (morpho-)syntactic, terminological, and (domain) semantic levels. Using SKOS and RDF(S), we designed intra-model relations by interlinking the lexicon entries, the terms, and concepts in and betweeen each resource. These links preserve the original source information and thus document the role of terminology within the process of harmonization. As an addtional result we see the emergence of a net of conceptual labels that can be organized independently from the ontological sources in which they were introduced.

Acknowledgements. The DFKI part of this work has been supported by the Monnet project (Multilingual ONtologies for NETworked knowledge), co-funded by the European Commission with Grant No. 248458.

References

1. Declerck, T., Lendvai P.: Towards a standardized linguistic annotation of the textual content of labels in knowledge representation systems. In: The seventh international conference on Language Resources and Evaluation. LREC-10, Malta (2010)
2. Declerck, T., Lendvai, P., Wunner, T.: Linguistic and Semantic Features of Textual Labels in Knowledge Representation Systems. In: Harry Bunt (ed.): Proccedings of the Sixth Joint ISO - ACL/SIGSEM Workshop on Interoperable Semantic Annotation, Oxford, United Kingdom, ACL-SIGSEM (2011)
3. Bassey, A., Budin, G., Picht, H. Rogers, M., Schmitz, K.D., Wright, S.E.: Shaping Translation: A View from Terminology Research. Translators' Journal 50:4 (2005)
4. Buitelaar, P., Cimiano, P. Haase, P., Sintek, M.: Towards linguistically grounded ontologies. In: Proceedings of the 6th European Semantic Web Conference, pp. 111-125, Springer Berlin/Heidelberg (2009)
5. ISO 860: Terminology work - Harmonization of concepts and designations (2005)
6. ISO 30042: Systems to manage terminology, knowledge, and content - TermBase eXchange (TBX) (2008)
7. ISO 16642: Computer applications in terminology - Terminological markup framework (2003)
8. Vasiljevs, A., Gornostay, T., Skadina, I.: From Terminology Database to Platform for Terminology Service. In: Proceedings of the CHAT 2011, Vol. 12, pp. 16-21, NEALT Proceedings Series (2011)
9. Delpech, E., Daille, B.: Dealing with lexicon acquired from comparable corpora: validation and exchange. In: Proceedings of the TKE 2010, pp. 211223, Dublin, Ireland (2010)
10. Bizer, C., Heath, T., Berners-Lee T.: Linked data - the story so far. International Journal on Semantic Web and Information Systems (IJSWIS). 5:3, 1-22 (2009)
11. McCrae, J., Spohr, D., Cimiano, P.: Linking Lexical Resources and Ontologies on the Semantic Web with Lemon. The Semantic Web: Research and Applications. Volume 6643 of LNCS, pp. 245-259. Springer, Berlin (2011)
12. César, P., Pino, F.J., García, F., Piattini, M., Baldassarre, T.: An ontology for the harmonization of multiple standards and models. Computer Standards & Interfaces, 34: 1, pp. 48-59 (2012)
13. Montiel-Ponsoda, E., Aguado-de-Cea, G., McCrae, J.: Representing term variation in *lemon*. WS 2 Workshop Extended Abstracts, TIA 2011, pp. 47-50, Paris (2011)
14. Temmerman, R., Kerremans, K.: Termontography: Ontology Building and the Sociocognitive Approach to Terminology Description. CIL17, Prague (2003)
15. Aussenac-Gilles, N., Szulman, S., Despres, S.: The Terminae Method and Platform for Ontology Engineering from Texts. In: Proceedings of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge. IOS Press, pp. 199–223, IOS, Amsterdam (2008)
16. Roche, C., Calberg-Challot, M., Damas, L., Rouard, P.: Ontoterminology: A new paradigm for terminology. In: International Conference on Knowledge Engineering and Ontology Development, pp. 321-326, Portugal (2009)
17. Khayari M., Schneider, S., Kramer, I., Romary, L.: Unification of Multi-Lingual Scientific Terminological Resources Using the ISO 16642 Standard, The TermSciences Initiative. In: Acquiring and Representing Multilingual, Specialized Lexicons: the Case of Biomedicine (LREC 2006), Genoa (2006)

Towards the Automated Enrichment of Multilingual Terminology Databases with Knowledge-Rich Contexts – Experiments with Russian EuroTermBank Data

Anne-Kathrin Schumann

University of Vienna / Tilde
anne.schumann@tilde.lv

Abstract. Although knowledge-rich context (KRC) extraction has received a lot of attention, to our knowledge few attempts at directly feeding KRCs into a terminological resource have been undertaken. The aim of this study, therefore, is to investigate to which extent pattern-based KRC extraction can be useful for the enrichment of terminological resources. The paper describes experiments aiming at the enrichment of a multilingual term bank, namely EuroTermBank, with KRCs extracted from Russian language web corpora. The contexts are extracted using a simple pattern-based method and then ranked by means of a supervised machine learning algorithm. The internet is used as a source of information since it is a primary means for finding information about terms and concepts for many language professionals, and a KRC extraction approach must therefore be able to deal with the quality of data found online in order to be applicable to real tasks.

Keywords: computer-aided terminography, knowledge-rich contexts, web as corpus, Russian language, multilingual terminology databases

1 Introduction and Related Work

In recent years, knowledge-rich context (KRC) extraction has been put forward as a means for enriching existing multilingual terminology resources with concept definitions and explanations while keeping the acquisition effort on a justifiable level. KRCs can be defined as follows (see [10], [13]):

Definition 1. Knowledge-rich contexts are naturally occurring utterances that explicitly describe attributes of domain-specific concepts or semantic relations holding between them at a certain point in time, in a manner that is likely to help the reader of the context understand the concept in question.

KRC extraction aims at identifying contexts that provide semantic information about *concepts* (as opposed to linguistic information about *terms*) in text corpora and to feed the results of this process into a terminological resource. It therefore touches upon

aspects of terminology research that remain yet unresolved: although the different types of contexts have been described in ISO 12620 ([6]), many terminological resources do not distinguish between various context types and often restrict themselves to linguistic contexts and more or less informative usage examples. In other cases, contexts are completely omitted.

The extraction of KRCs has been actively researched for several languages. Seminal work for English was carried out by [11] and [10], and recent studies providing a contrastive perspective on English and French are [8] and [9]. Recent work on other languages are [4] for Catalan, [16] for Spanish, and [7] for French. [17] studies the topic of definition extraction from German court decisions, whereas [13] gives a first evaluation of KRC extraction patterns for Russian and German.

KRC extraction generally requires high precision, while specialized corpora from which KRCs can be extracted are typically small or must be crawled from online sources, a process that often outputs messy data. What is common to many studies in the field, therefore, is the fact that they employ a pattern-based method. A systematic overview over pattern-based work is given by [1]. Often, extraction patterns are acquired manually, but some groups ([2], see [5]) also devise a bootstrapping procedure for automated pattern acquisition similar to methods developed in information extraction ([18]).

As for the ranking of extraction output, [17] gives a detailed account of his experiments in the ranking of definition candidates using supervised machine learning techniques. The features used in his experiments can be divided into five groups:

- *Lexical*, such as boost words or stop words and features that are specific for legal language, such as subsumption signals
- *Referential*, such as anaphoric reference or definiteness of the definiendum
- *Structural*, such as the position of the definiendum relative to the definiens
- *Document-related*, such as the position of the definition candidate in the document and whether there are other candidates in its immediate context
- Others, such as sentence length or TF-IDF

2 Towards the Enrichment of EuroTermBank

2.1 EuroTermBank

EuroTermBank[1] ([12]) is a multilingual term bank that was released in 2007. More specifically, it is a terminology repository binding together specialized terminology collections in 27 European languages. The terminology collections represented in EuroTermBank (ETB) consist of electronic collections contributed from various partners as well as digitalized versions of print dictionaries. Special attention was paid to providing resources for small and under-resourced languages especially from the new EU member-states, such as the Baltic languages. In terms of entries, the 5 best-resourced languages in EuroTermbank are English, Russian, German, Latvian, and Polish (in this order).

[1] http://eurotermbank.com/.

2.2 Knowledge-Rich Context Extraction in Russian

Previous studies of KRC extraction from Russian web corpora ([13]) were based on a pattern-based extraction approach using 47 mainly predicative Russian patterns. These patterns had been combined either with target terms or morpho-syntactic term formation patterns to form regular expressions. In our present experiments, we used a similar approach, but extraction was applied to lemmatized text in order to facilitate the process and extraction patterns were used without any kind of term representation. Example 1 illustrates a lexical extraction trigger and a valid KRC extracted in the course of our experiments. The underlined term is an ETB target term, whereas the lexical extraction trigger is marked in bold.

Example 1. Эстафетная палочка **представляет собой** цельную, гладкую, полую трубку, круглую в сечении, сделанную из дерева, металла или другого твердого материала.
*(The relay baton **is** a one-piece, smooth, hollow, and round tube made from wood, metal or another hard material.)*

Semantic relations are elementary building blocks of KRCs. We therefore devised a typology of semantic target relations that make up a valid KRC. Table 1 gives an overview over these relations along with examples of lexical extraction triggers:

Table 1. Semantic relations and Russian extraction triggers

Relation	Explanation	Patterns	Translation
Hyperonymy	Generic-Specific	Относить к, включать в себя	Belong to, include
Meronymy	Part-Whole	Состоять из	Consist of
Process	Temporal neighbourhood	Воздействовать	Act upon
Position	Spatial neighbourhood	Располагать	Locate
Causality	Cause-Effect	Обусловить	Determine
Origin	Material or ideal origin	Состоять из	Is made of
Reference	General predication or definition	Представлять себя, называть	Is, call
Function	Purpose or aim	Служить, позволять	Serve, allow

2.3 Ranking

KRC candidates are extracted using the patterns described in the previous section. They are then ranked directly according to the values outputted by a Naïve Bayes classification algorithm. The Perl Algorithm::NaiveBayes module[2] is used to carry out this procedure based on the following 13 features:

[2] http://search.cpan.org/~kwilliams/Algorithm-NaiveBayes-0.04/lib/Algorithm/NaiveBayes.pm.

Table 2. Shallow features used for ranking

Feature name	Explanation
Word tokens	The number of word tokens in the sentence.
Subscore	The normalized sum of the term relevance scores of terms constituting the subject.
Subpos	1 if the sentence starts with the subject, else 0.
Term score	The normalized sum of the term relevance scores of all other terms.
Nr. of terms	The number of terms in the sentence.
Position	1 if the subject is located before the extraction pattern, else 0.
Adjacent term	1 if there is a term directly adjacent to the extraction pattern, else 0.
Distance	The token distance between subject and pattern.
Negation	1 if the extraction pattern is preceded by a negation particle, else 0.
Boost words	1 if the pattern is preceded by a generalization signal, else 0.
Pattern score	A pattern reliability estimate.
Stop words	Number of negative markers normalized by word tokens.
Definite Subject	1 if the subject is preceded by markers of definiteness or anaphora.

In order to identify the subject of a sentence, a heuristic using the rich annotation provided by the Russian TreeTagger tagset ([15]) and syntactic noun phrase formation patterns as observed in our corpus was devised. As for the term scoring method, we achieved the best results not by using a classical TF-IDF score, but a slightly modified score that takes into account relative term frequency as well as the occurrence of the target term in the extraction corpus and a reference corpus[3]. This score outputs values higher than zero for all terms that occur in at least one of the corpora and always ranks frequent terms higher than less frequent terms, which corresponds to the hypothesis that the existence of a valid KRC is more likely, if the target term is highly frequent. The development and adaptation of the best term scoring method will be further studied in future experiments. The positional features in our ranking scheme are based on the hypothesis that even in a language with relatively free word order such as Russian sentences that contain definitional information favour a regular word order. Boost words are generalization signals such as *часто* (often) or *обычно* (usually), whereas stop words include outdated language such as *СССР* (USSR) and *советский* (soviet).

[3] The Russian Internet Corpus ([14]) was used as a reference corpus. A search interface to this corpus is available here: http://corpus.leeds.ac.uk/ruscorpora.html.

3 Experiments on Enriching EuroTermBank with Knowledge-Rich Contexts

3.1 Resource Selection and Corpus creation

We selected a rather small ETB resource, namely the athletics domain. For Russian, this domain comprises 665 entries from which the target terms were harvested. The final term list has 667 target terms. Some of these terms are verb phrases, others are rather generic terms such as *скорая помощь* (first aid) and „*ветер*“ (wind), or polysemic such as *построение* (which often means “formation” or “construction”, but in ETB’s athletics domain is translated to English as “line-up”) and “*Нет!*” (No!), which is given as a synonym for *прыжок не засчитан* (the jump was not counted).
We used some of the target terms harvested from ETB as seeds in a corpus crawling process. The corpus crawler was Babouk ([3]). However, the term list obtained from ETB had to be cleaned in order to remove the following shortcomings:

- Some entries contain synomyms or near synonyms separated by commas. In such cases, the synonyms were treated as two separate target terms.
- If very general terms are fed into Babouk, the obtained corpus is likely to contain a high percentage of out-of-domain texts, since the seed terms are polysemic. Therefore, most unigrams were removed from the seed list.

Moreover, for each seed term, more than one word form was supplied in order to improve the performance of Babouk. The crawling process had to be repeated several times. The resulting corpus has 517.266 running words and 28.448 sentences after cleaning. Table 3 gives an overview over the 10 most frequent ETB term concordances in the corpus.

Table 3. Overview over 10 most frequent ETB terms in corpus

Term	Translation	Count	Term	Translation	Count
бег	running	4254	подготовка	preparation	1353
техника	technique	1648	дистанция	distance	1336
соревоноание	competition	1467	прыжок	jump	1106
скорость	speed	1396	шаг	step	998
спортсмен	athlete	1229	выносливость	endurance	843

Out of our initial 667 terms, 420 were found in the corpus, and out of those, 209 had at least 10 and 102 at least 50 concordances.

3.2 Experimental Setup and Results

KRC extraction for term bank enrichment besides filtering KRC candidates from unseen data includes two more tasks, namely the attribution of the explanation provided by the KRC candidate to a specific target term and the filtering of KRCs that are not related to any of the relevant target terms.

To test the performance of our current method on these tasks, we extracted KRC candidates from the sports corpus. This process outputted 3068 KRC candidates. Unlike the experiments described in [13] no morpho-syntactic target term representation was used in this step, resulting in a very simple extraction method and a large amount of data. On this data, we conducted two experiments. In the first setting, ranking was performed only on those KRC candidates, for which the feature extraction step revealed a target term in subject position. In a second experiment, we applied the ranking algorithm to all KRC candidates that matched at least one term. Since the ranking algorithm is based on supervised learning, each data set had to be split into a training and a test set. Table 4 gives an overview over the data sets.

Table 4. Datasets used in experiments

Setting	Overall size of data set	Size of training set	Size of test set
Subject setting	521 KRC candidates	100	421
Term setting	1813 KRC candidates	300	1513

The ranking algorithm was applied to select valid KRCs from the datasets and simple heuristics were devised in order to find the target term of each KRC candidate: in the subject setting, the subject of each sentence was set to be the target term, whereas in the term setting a cascaded procedure for target term selection was applied:

- If there was a term in subject position, this term was set to be the target term.
- Otherwise, a term directly adjacent to the extraction pattern – if applicable – was set to be the target term.
- If none of these conditions was met, the first matching term in the sentence was set to be the target term.

Results were manually evaluated by picking and evaluating the highest ranked sentence for each term. Sentences with very low ranks were not evaluated. For target terms that are verb phrases, a relaxed setting was applied by accepting sentences that contain valid collocations, since it is yet unclear how the concept of KRCs can be applied to verbs. Table 5 presents the results.

Table 5. Results obtained in two experimental settings

Setting	Number of evaluated sentences	Unique KRC candidates for ETB target terms	Correct unique KRCs	Precision of attribution of KRC candidate to target term
Subject setting	407	82	57	0.96
Term setting	1504	197	112	0.91

4 Discussion and Future Work

The results of our experiments suggest that even in a very relaxed extraction setting, the current KRC extraction method achieves only limited coverage. More specifically, only for roughly 18% of our initial 667 target terms and 28% of all ETB terms in the corpus unique valid KRCs could be found including KRCs found during the manual annotation of the training sets. For higher recall, the pattern-based method may need to be supplemented by other methods that might be applied to the data in an iterative fashion. The systematic use of term variants may also help to retrieve more relevant contexts from the corpus.

The fact that the more relaxed term setting outperforms the subject setting in terms of coverage suggests that future research efforts should concentrate on the use of more linguistic information for higher precision and better ranking results to support the selection of valid candidates: In our view, the improvement of the current method by applying deeper linguistic knowledge such as syntactic information and making wider use of morphology will help establish a link between an ETB term and a lexical extraction trigger, thus eliminating noise and resulting in better ranking and target term selection. Other aspects that deserve to be mentioned are term-inherent polysemy affecting the process starting already upon corpus crawling. Moreover, more sophisticated processing such as the filtering of proper names and ambiguity resolution for polysemic terms may improve results.

Last but not least, the results outlined in this paper show that KRC extraction can be just one means of term bank enrichment: the current method deals but weakly with terms that are verbs and verb phrases and other kinds of information, e.g. collocations, may indeed be the better choice for this particular kind of terms.

Acknowledgement. The research described in this paper was funded under the CLARA project (FP7/2007-2013), grant agreement n° 238405.

References

[1] Auger, A., Barrière, C.: Pattern-based approaches to semantic relation extraction. Terminology. 14 (1), 1-19 (2008)

[2] Condamines, A., Rebeyrolle, J.: Searching for and identifying conceptual relationships via a corpus-based approach to a Terminological Knowledge Base (CTKB). In: Bourigault, D., Jacquemin, C., L'Homme, M.-C. (eds.) Recent Advances in Computational Terminology, pp. 127-148. John Benjamins, Amsterdam/Philadelphia (2001)

[3] De Groc, C.: Babouk: Focused web crawling for corpus compilation and automatic terminology extraction. In: IEEE/WIC/ACM International Conference on Web Intelligence (2011)

[4] Feliu, J., Cabré, M.: Conceptual relations in specialized texts: new typology and an extraction system proposal. In: Proceedings of TKE 2002, pp. 45-49. INRIA, Nancy (2002)
[5] Halskov, J., Barrière, C.: Web-based extraction of semantic relation instances for terminology work. Terminology. 14 (1), 20-44 (2008)
[6] International Organization for Standardization. International Standard ISO 12620: 2009 – Terminology and Other Language and Content Resources – Specification of Data Categories and Management of a Data Category Registry for Language Resources. ISO, Geneva (2009)
[7] Malaisé, V., Zweigenbaum, P., Bachimont, B.: Mining defining contexts to help structuring differential ontologies. Terminology. 11 (1), 21-53 (2005)
[8] Marshman, E.: Towards strategies for processing relationsips between multiple relation participants in knowledge patterns. An analysis in English and French. Terminology. 13 (1), 1-34 (2007)
[9] Marshman, E.: Expressions of uncertainty in candidate knowledge-rich contexts. A comparison in English and French specialized texts. Terminology. 14 (1), 124-151 (2008)
[10] Meyer, I.: Extracting Knowledge-Rich Contexts for Terminography: A conceptual and methodological framework. In: Bourigault, Jacquemin, L'Homme (eds.), pp. 279-302 (2001)
[11] Pearson, J.: Terms in Context. (Studies in Corpus Linguistics 1). John Benjamins, Amsterdam/Philadelphia (1998)
[12] Rirdance, S., Vasiļjevs, A. (eds.): Towards Consolidation of European Terminology Resources. Experience and Recommendations from EuroTermBank Project. Tilde, Riga (2006)
[13] Schumann, A.-K.: A Bilingual Study of Knowledge-Rich Context Extraction in Russian and German. In: Proceedings of the Fifth Language & Technology Conference, pp. 516-520. Fundacja Uniwersytetu im. A. Mickiewicza, Poznan (2011)
[14] Sharoff, S.: Creating general-purpose corpora using automated search engine queries. In: Baroni, M., Bernardini, S. (eds.), *WaCky! Working papers on the Web as Corpus*. Gedit, Bologna (2006)
[15] Sharoff, S., Kopotev, M., Erjavec, T., Feldmann, A., Divjak, S.: Designing and evaluating Russian tagsets. In: *Proceedings of LREC* (2008)
[16] Sierra, G., Alarcón, R., Aguilar, C., Bach, C.: Definitional verbal patterns for semantic relation extraction. Terminology. 14 (1), 74-98 (2008)
[17] Walter, S.: Definitionsextraktion aus Urteilstexten. PhD thesis in Computational Linguistics. Saarland University Saarbrücken (2010)
[18] Xu, F.-Y.: Bootstrapping Relation Extraction from Semantic Seeds. PhD thesis in Computational Linguistics. Saarland University Saarbrücken (2007)

Short Papers

Distributing Terminology Resources Online: Multiple Outlet and Centralized Outlet Distribution Models in Wales

Gruffudd Prys[1], Tegau Andrews[1], Dewi B. Jones[1], Delyth Prys[1]

[1]Language Technology Unit, Bangor University, Bangor, Wales, UK
{g.prys,t.andrews,d.b.jones,d.prys}@bangor.ac.uk

Abstract. This paper describes a system which enables the targeted distribution of terms from a single centralized terminology development environment to multiple online outlets. These dissemination outlets include distinct subject-specific websites, a master terminology portal, and aids such as a rollover terminology lookup function that can be added to existing websites.

Keywords: terminology, distribution, dissemination, online, Welsh, dictionary

1 Introduction

A key player in the standardization of terminology in Wales is the Language Technologies Unit (LTU) [1]. Based in the Canolfan Bedwyr centre for Welsh language support at Bangor University, the LTU has been active, in one form or another, since the early 1990s. It marries terminological expertise with technical expertise to enable the creation of online dictionaries and a range of other products, including a Welsh grammar checker, translation memory system, and machine translation system, into which currently over 20 terminology dictionaries can be integrated.

Much of the terminology standardization work currently being undertaken in Wales is accomplished within a centralized online terminology development environment created by the LTU, known as *Maes T*. The most recent online terminology resources in Wales have been generated using this system, including national terminology projects in both Higher Education and in Secondary and Vocational Education [2]. The Maes T system is concept-based and language-neutral, and conforms to the ISO standards listed in the guidelines for Welsh terminology work [3]. It was developed in order to facilitate the collaboration of geographically dispersed teams of subject specialists and terminologists, that they might develop, maintain and standardize Welsh-language terms for any and all required domains [4]. Maes T is not itself, however, a vehicle for the public dissemination of terms. This paper will discuss an implemented solution to distributing terms from a centralized standardization system to separate, non-homogenous outlets such as websites and web services.

2 Distributing terms

The main commissioners of Welsh-language terminology are public sector organizations which are legally obligated to provide services to the public in both English and Welsh [5]. Such commissioners include the Justice Wales Network, the Local Health Boards and the Welsh Language Board. It is often the case that such clients require

the commissioned terminology resources to be available from their own (possibly pre-existing) websites rather than from a centralized terminology-specific portal such as the Welsh National Terminology Portal [6][1].

There are a number of reasons for this. Firstly, commissioners may wish to communicate ownership of a terminology resource by having the resource situated on their own web domain, using their own corporate branding. Secondly, the intended user groups for the terminology must be considered. Terminology resources may cater to a number of different audiences, including the professional linguist, the subject specialist, professionals working in a specific domain and the ordinary user of the commissioner's services. Centralized terminology portals that aggregate many different subject fields in a manner that is useful to professional linguists may overload with potentially irrelevant terms those users whose work focuses on a particular domain. As a result, may different user groups may require differing methods of delivering terms.

Thirdly, certain financing bodies require their contribution to specific terminology resources to be acknowledged alongside the resources. On centralized terminology portals which consolidate all terminology resources, this can be difficult to accomplish with sufficient granularity to satisfy the terms of the original grant. Finally, it is important that bodies offering services in a minority language, alongside a dominant language, do so in a prominent manner, giving each language equal visibility on their website. Providing vital terminology on an institution's own websites makes Welsh content easily accessible to the institution's users and can help underline the institution's commitment to operating through the medium of the language.

It is evident therefore that a centralized terminology development system requires the creation of an additional component that can distribute terminology resources to both distinct terminology dissemination outlets and a centralized dissemination outlet such as the Welsh National Terminology Portal. Note that it is not a case of choosing to either disseminate terms through a single central outlet or through many distinct commissioner outlets: most commissioners welcome the chance to make terms available both on their own website and on the Welsh National Terminology Portal.

3 The LTU's distribution solution

The many disparate term distribution outlets in Wales call for an efficient method of distributing both the term data and the language-specific functionalities (such as lemmatization of search queries) which are required to aid users in finding relevant terms.

The solution devised by the LTU was developed using Google Web Toolkit, an open source development toolkit for building and optimizing complex browser-based applications [7]. Google Web Toolkit enables client-side applications to be written in Java and then be deployed as JavaScript. This ensures that the application can be distributed to the external websites of commissioners of terminology work, irrespective of the server technology used, as the it runs within the website visitor's

[1] This portal, named Porth Termau in Welsh, brings together the content of all publicly available terminology dictionaries developed since 1993 by the Centre for the Standardization of Welsh Terminology (now merged with the LTU) and its approved partners.

browser (the vast majority of which are JavaScript compatible) rather than on the website's server.

3.1 Installation on existing websites

Where the commissioner of the terminology work wishes to distribute terms from Maes T on their own pre-existing website, a Javascript "include" can be placed within the header of the web page. This replaces empty elements within the page of the commissioner's website which have been set aside to be the home of the terminology dictionary. However, this requires the co-operation of the website developer, often a subcontractor, who may not always be readily available to the institution.

3.2 Installation on Content Management Systems such as WordPress

When the external website is based on a Content Management System (CMS) such as WordPress, the involvement of the web developer may be avoided by creating a dedicated plugin for the CMS containing the required component code.The LTU uses WordPress, an open source CMS, to create its own term distribution websites in addition to websites that it hosts on behalf of commissioners. WordPress provides a platform that is easy to use and adapt, and it can be localized and adapted for multilingual use using the WPML (WordPress Multilingual) plugin. WordPress' open-source licence and support for plugins enables the platform to be customized as required.

To facilitate the dissemination of terms to WordPress websites, the LTU has created a WordPress plugin, *Porthydd*. The plugin includes code that adds searchable access to the terms found in Maes T to any self-hosted WordPress page that possesses the appropriate API key. The API key, provided by the LTU, establishes whether the website has permission to access the data and controls the settings associated with the search facility and the displaying of search results. These settings include determining which subject-specific terms are displayed, which data fields are displayed, and whether or not to display features such as A-Z lists of a dictionary's contents.

4 An example implementation: *Y Termiadur Addysg*

The example implementation of the LTU's distribution solution chosen to be demonstrated at CHAT 2012 is the *Y Termiadur Addysg* (Education Terminology Dictionary) website. The website is part of a project funded for three years by the Welsh Government with the aim of standardizing and distributing Welsh-language terms to be used in resources such as examinations and course textbooks as well as in general classroom use in primary, secondary and further education in Wales. This is the third project in the *Termiadur* series of terminology dictionaries, and first that will not be published as a print edition.

The Y Termiadur Addysg is a fully bilingual Welsh/English online dictionary built on the WordPress platform, featuring guidelines, amended terms, a contact form, a random term feature and game. The search functionality is provided by an instance of the Porthydd plugin which can be installed via the WordPress administrator screen by an administrator with very little technical knowledge. Porthydd was used to add two pages for accessing terminology to the Y Termiadur Addysg website.

The first is a simplified interface that auto-detects the key word language, providing a dual result display for word forms that exist in both languages. The search facility uses language-specific lemmatization rules to return relevant entries even when the search terms feature mutated or conjugated froms. Terms that are alphabetically adjacent to the search term are also displayed in a side panel. The second page features an A-Z letter-by-letter list of all of the dictionary's terms intended to aid those who wish to browse the content of the dictionary. Elements such as abbreviations or parts of speech possess rollover tooltips that provide additional information to the user. Porthydd provides a dynamic link to the terms that are stored in Maes T, so that when a terminologist pushes a button to publish a term in Maes T, it will immediately be available to websites such as Y Termiadur Addysg.

Maes T and Porthydd therefore combine to produce an efficient method of standardizing and distributing terms from a central standardization hub to any number of distribution outlets, including a central multi-domain portal and domain-specific sites.

References

1. Welsh Government: A Living Language: A Language For Living, Welsh Language Strategy 2012-17 (2012), p. 49, http://wales.gov.uk/docs/dcells/publications/122902wls201217en.pdf .
2. The Welsh-Medium Higher Education Terminology Project, http://www.porth.ac.uk/termau. Y Termiadur Addysg, http://www.termiaduraddysg.org.
3. Prys, D., Jones, D. B.: Guidelines for the Standardization of Terminology for the Welsh Assembly Government Translation Service and the Welsh Language Board (2007), http://www.byig-wlb.org.uk/English/publications/Publications/5338.pdf.
4. Andrews, T., Prys, G.: The Maes T System and its use in the Welsh-Medium Higher Education Sector in Wales. In: Gornostay, T., Vasiļjevs, A. (eds), CHAT 2011: Creation, Harmonization and Application of Terminology Resources, NEALT Proceedings Series, Vol. 12, pp. 49-50. Riga: Northern European Association for Language Technology (2011), http://dspace.utlib.ee/dspace/bitstream/handle/10062/17370/proceedings.pdf?sequence=1.
5. Great Britain: Welsh Language Act 1993, Chapter 38. London: Her Majesty's Stationery Office (1993), http://www.opsi.gov.uk/ACTS/acts1993/ukpga_19930038_en_1.
6. Cymdeithas Edward Llwyd, Llên Natur Project, http://llennatur.com. Coleg Cymraeg Cenedlaethol (National Welsh College), http://www.porth.ac.uk. National Terminology Portal, http://termau.org.
7. Google, Google Web Toolkit Overview, https://developers.google.com/web-toolkit/overview

Extraction of Multilingual Term Variants in the Business Reporting Domain

Thierry Declerck[1] and Dagmar Gromann[2]

[1] DFKI GmbH, Language Technology Department,
Stuhlsatzenhausweg 3, D-66123 Saarbruecken, Germany
declerck@dfki.de
[2] Vienna University of Economics and Business,
Nordbergstrasse 15, 1090 Vienna, Austria
dgromann@wu.ac.at

Abstract. Within the context of the European research project "Monnet", which implements among other activities ontology-based multilingual information extraction, we tackle the the issue of recognizing variants of concept labels in business reports that guide the information extraction process. In this short paper, we describe two related experiments in finding variants of multilingual taxonomy labels used in business reporting – across distinct reporting legislations and languages. A core taxonomy developed by the XBRL-Europe Association provides a starting point, as we map multilingual term variant candidates we extract from the web presence of relevant players in the field of business reporting to its labels.

Keywords: Terminology extraction, variants, ontology, multilingualism, business reporting

1 Introduction

Within the context of the European research project "Monnet"[3], which implements among other activities the ontology-based extraction of multilingual information to be used in the field of business reporting, we face the challenge of detecting relevant terms and their variants in a variety of document types. Afterwards, these terms and variants as well as their associated data have to be transformed into domain facts that can be stored as instances of classes of an integrated financial and reporting ontology.

In the European context the fact that each country is marked by different legislations as regards the description of information companies have to provide represents a particular challenge as well as the fact that the corresponding financial statements to be reported are mainly based on so-called national General Accepted Accounting Principles (GAAP). Fortunately, most of these GAAPs

[3] See http://www.monnet-project.eu for more details.

are nowadays encoded using a standard representation language, called XBRL[4], which provides relatively harmonized taxonomies listing the main concepts and associated natural language labels (using the xml:lang attribute) containing the official reporting terminology. A simplified example from the taxonomy of the Belgian National Bank is provided below.

```
<loc xlink:label="Assets_loc" xlink:type="locator"
   xlink:href="pfs-2011-04 01.xsd#pfs_Assets"/>
<labelArc xlink:from="Assets_loc" xlink:to="Assets_lab"
   xlink:type="arc"xlink:arcrole="http://www.xbrl.org/2003/arcrole/concept-label"/>
<label xlink:label="Assets_lab" xlink:type="resource"
   xlink:role="http://www.xbrl.org/2003/role/label" xml:lang="fr">Total de l'actif</label>
<label xlink:label="Assets_lab" xlink:type="resource"
   xlink:role="http://www.xbrl.org/2003/role/label" xml:lang="nl">Totaal van de activa</label>
<label xlink:label="Assets_lab" xlink:type="resource"
   xlink:role="http://www.xbrl.org/2003/role/label" xml:lang="de">Summe der Aktiva</label>
<label xlink:label="Assets_lab" xlink:type="resource"
  xlink:role="http://www.xbrl.org/2003/role/label" xml:lang="en">Total assets</label>
```

[Simplified excerpt from the Belgian taxonomy for reporting: The concept `pfs_Assets` with labels in four languages.]

Although the representation of such information in XBRL, also allowing machine readability of the data, already marks a substantial progress towards more transparency in financial reporting, the cross-country and cross-lingual comparison still continues to be an issue. A working group of the XBRL-Europe Association started investigating this problem, and developed a core taxonomy, called xEBR (eXtended European Business Registers)[5], which connects concepts used in different legislations with common concepts, using also SKOS descriptors to indicate whether the mappings are exact, broad, or narrow, as can be seen in the code example below.

```
pfs_GainLossPeriod              exactMatch      xebr_ProfitLossForThePeriodTotal
pfs_FormationExpenses           narrowMatch     xebr_FixedAssetsTotal
pfs_AccumulatedProfitsLosses    broadMatch      xebr_ProfitLossForThePeriod
```

[Example of xEBR mappings from concepts of the Belgian National Bank, indicated by using the namespace "pfs", to the core concepts of xEBR.]

This work on the core taxonomy constitutes a very valuable step towards conceptual interoperability across reporting legislations. The xEBR core taxonomy has been semantically "upgraded" in our project to become an ontological module in a set of ontologies describing a class hierarchy and related properties in the broader financial domain. The labels of the core taxonomy (only available in English) are encoded in our ontology by means of the `rdfs:label` annotation property. Results of the information extraction procedure applied to local XBRL

[4] XBRL stands for "eXtensible Business Reporting Language, see `http://www.xbrl.org/` for more details.

[5] This taxonomy, which has not been published yet, is briefly described at: `http://www.monnet-project.eu/Monnet/Monnet/English/Navigation/XBRLEuropeanBusinessRegisterxEBR`.

instance documents are transformed into xEBR and stored as instances of the classes of this ontology.

Nevertheless, the aspect of multilingual terminology has not been resolved. It would be nice to offer a financial analyst not only the concept IDs (and the associated English labels) of the core xEBR taxonomy we can identify in business reports, but also the terms as they are used both in the source taxonomies and in the corresponding documents.

2 Linking Labels of National Taxonomies on the Basis of xEBR

On the basis of conceptual mappings, as displayed in Table 2, we implemented a procedure that extracts all the labels associated to national concepts from national taxonomies as a first step. Thereby, we achieve a mapping between the terms in these labels that is similar to the mapping between national taxonomies and xEBR. So if we, for example, detect the (Belgian) concept *pfs_Intangible FixedAssets* in an XBRL instance document of the Belgian National Bank, this concept is mapped to the xEBR concept *xebr_IntangibleFixedAssetsTotal*. However, in addition to the xEBR English label *Intangible fixed assets*, our procedure delivers all Belgian labels (*Immobilisations incorporelles@fr*, *Immaterielle Anlagewerte@de*, etc.)[6], and interlinks these labels using the SKOS descriptors applied to the corresponding concepts. Thus, we are not only able to deliver the combined xEBR and Belgian National Bank terminology, but we can also automatically link to other national legislations. Our current work focused on the relation between the Belgian and Spanish taxonomies as mediated by xEBR. Our tool also delivers the Spanish correspondences for the "IntangibleFixedAssets" example, both at a conceptual and terminological level, as can be seen in Table 3:

```
"concept" => "pgc-07-c-bs_ActivoNoCorrienteInmovilizadoIntangible"
"prefLabel" => "I. Inmovilizado intangible"
"altLabel" => "Activo no corriente inmovilizado intangible"
```

[The concept in the Spanish taxonomy corresponding to the xEBR concept *xebr_ IntangibleFixedAssetsTotal* with two associated labels in the Spanish language.]

Analysts can submit an instance XBRL document encoded in the Spanish taxonomy to our tool and receive both the xEBR concepts with the associated

[6] Due to limited space, we do no display all labels here. We just mention that the national taxonomies distinguish between labels and verbose labels, which we encode then as *prefLabel* vs *altLabel*, using RDF and SKOS for encoding this information:

<http://www.xbrl.org/xbrl_be.owl#pfs_hasIntangibleFixedAssets>
<http://www.xbrl.org/skos.owl#exactMatch>
<http://www.xbrl.org/xebr.owl#hasIntangibleFixedAssetsTotal> .
<http://www.xbrl.org/xbrl_be.owl#pfs_hasIntangibleFixedAssets>
<http://www.xbrl.org/skos.owl#prefLabel> "Immaterielle Anlagewerte"@de .

English labels as well as the Belgian concepts with the associated labels in four languages. Consequently, we have built an integrated terminological repository, generated on the basis of officially accepted terminologies in different business reporting legislations in Europe. This multilingual term base allows for a semantic processing of instance documents generated by national banks or by business registers, which use these taxonomies as their primary source of knowledge.

3 Extracting Multilingual Term Variants from Web Sources

Our second experiment is dedicated to the extension of the term base we generated from the official taxonomies with automatically detected term variants in on-line sources, which have been automatically extracted as structured or semi-structured data. For the time being, we consult company information on the bilingual web presence of the DAX Index of the German Stock Exchange (deutsche-boerse.com)[7], on the monolingual page of the Bundesanzeiger[8], and in annual reports published directly by companies. The annual report published by the company BASF SE serves as an example herein. In this case, we consult the bilingual, i.e., English and German, PDF reports of BASF manually, contrary to the other sources, from which the data has been extracted automatically.

Concentrating on various reports in various languages for one company for a specific year allows for the additional use of a simple heuristics in order to detect multilingual term correspondences: the financial positions associated with terms have the same values. We are well aware of the fact that this heuristics cannot be applied to all financial positions in reports. For example, the monetary value of *Total assets* and *Total equity and liabilities* should be identical, as can be seen in Table 1, however no equivalence relation can be established as they are no variants of each other. Nevertheless, the taxonomy indicates possible positions of terms in specific parts of tables, which provides us with a precise context for the application of our heuristics.

Some results for the BASF example are summarized in Table 1, which exemplifies that equivalences among monolingual business reporting concepts can be established on the basis of previously normalized financial figures. Thus, a synonymy relation between *Langfristiges Fremdkapital* and *Langfristige Verbindlichkeiten* in German or between *short term assets* and *current assets* in English can be established. As regards the bilingual level, a relation can be established between, for example, the German terms *Langfristiges Fremdkapital* and *Langfristige Verbindlichkeiten* and the English term *Longterm Liabilities*.

[7] See for example the bilingual DAX pages on the company BASF: `http://www.boerse-frankfurt.de/de/aktien/basf+se+DE000BASF111/kennzahlen` and `http://www.boerse-frankfurt.de/en/equities/basf+se+DE000BASF111/key+figures`.

[8] The "Bundesanzeiger" is the official institution for company reporting in Germany. `https://www.bundesanzeiger.de/ebanzwww/wexsservlet`.

Table 1. Monolingual term variants and bilingual term correspondences established by comparing different financial reports for the same company in the same period

German	Figure	English	Source
Umsatzerlöse	63.873	Sales	BASF
Umsatz	63.873		Bundesanzeiger
Umsatz	63.873	Sales	DAX
Langfristige Vermögenswerte	34.532	Long-term assets	BASF
Langfristiges Vermögen	34.532		Bundesanzeiger
Anlagevermögen insgesamt	34.532	Total Capital Assets	DAX
Kurzfristige Vermögenswerte	24.861	Short-term assets	BASF
Kurzfristiges Vermögen	24.861		Bundesanzeiger
Umlaufvermögen	24.861	Total Current Assets	DAX
Langfristiges Fremdkapital	21.168	Long-term liabilities	BASF
Langfristiges Fremdkapital	21.168		Bundesanzeiger
Langfristige Verbindlichkeiten	21.168	Total Longterm Liabilities	DAX
Gesamtkapital (Passiva)	59.393	Total equity and liabilities	BASF
Gesamtvermgen (Aktiva)	59.393	Total assets	BASF
Gesamtkapital (Passiva)	59.393		Bundesanzeiger
Gesamtvermgen (Aktiva)	59.393		Bundesanzeiger
Bilanzsumme	59.393	Total Liabilities and Equity	DAX

Mediated by the corresponding xEBR concepts, these German and English term variants can also be linked to other languages, re-using the mechanisms described in section 2, so that the German term variants *Kurzfristige Vermögenswerte*, *Kurzfristiges Vermögen* and *Umlaufvermögen* can be linked – via the xEBR concept *xebr_CurrentAssetsTotal* – to the Spanish labels *B) ACTIVO CORRIENTE* and *Activo corriente*, which are associated to the concept *pgc-07-c-bs_ActivoCorriente* of the Spanish taxonomy.

4 Conclusion and Future Work

We have described completed and ongoing work in the building of an integrated term base in the domain of standardized business reporting. The starting point is a core taxonomy that maps reporting and financial concepts from various European taxonomies. We integrated this taxonomy in our set of financial and reporting ontologies, proposing at the same time a multilingual extension of the labels with all the terms officially introduced in the national taxonomies. As a second step, we automatically extract term variants for the extended labels of xEBR concepts from on-line sources, also in a multilingual fashion, thus augmenting the term base that supports the information extraction task applied to financial reporting documents.

Acknowledgements. This work has been supported under the Seventh Framework Programme of the European Commission through the Monnet project

(Multilingual ONtologies for NETworked knowledge) co-funded by the European Commission with Grant No. 248458.

References

1. Declerck,T., Krieger, H.U., Thomas, S.M., Buitelaar, P., O'Riain, S., Wunnder, T., Maguet, G., McCrae, J., Spohr, D., Montiel-Ponsoda, E.: Ontology-based Multilingual Access to Financial Reports for Sharing Business Knowledge across Europe. In: Proceedings of MONTIFIC/ECQA Conference, September, Budapest (2012)
2. Wunner, T., Buitelaar, P., O'Riain, S.: Semantic, Terminological and Linguistic Interpretation of XBRL. In: Proceeding of EKAW, October, Lisbon (2010)
3. Declerck, T., Lendvai, P., Wunner, T.: Linguistic and Semantic Features of Textual Labels in Knowledge Representation Systems. In: Proceedings of ISA6, ISO/ACL-SIGSEM Workshop, January, Oxford (2011)
4. McCrae, J., Espinoza, M., Montiel-Ponsoda, E., Aguado-de- Cea, G., Cimiano, P.: Combining statistical and semantic approaches to the translation of ontologies and taxonomies. In: Proceedings of the Fifth workshop on Syntax, Structure and Semantics in Statistical Translation (SSST-5), Portland (2011)
5. Aguado-de-Cea, G., Montiel-Ponsoda, E.: Term variants in ontologies. In: Proceedings of the 30th International Conference of AESLA, pp. 19-21 April, Spain, Lleida (2012)

Consolidating European Multilingual Terminology across Languages and Domains

Tatiana Gornostay, Roberts Rozis, Andrejs Vasiļjevs, Inguna Skadiņa

Tilde, Riga, Latvia
{tatiana.gornostay, andrejs, roberts.rozis,
inguna.skadina}@tilde.lv

Abstract. This short paper addresses the problem of consolidating European multilingual terminology resources across languages and domains. In the introduction section we identify the task and the necessity of a consolidated interface for different, usually dispersed, multilingual terminology resources. In the second section we give examples of state-of-the-art approaches to the solution of this task – the Quest tool and the EuroTermBank portal. The third section presents a brief overview and reports on midterm results (achieved during the first year of the project) of an on-going research on consolidating European multilingual terminology resources as part of the ICT PSP EU project META-NORD within the META-NET initiative and the META-SHARE open linguistic infrastructure. Finally, we make conclusions and outline future work.

Keywords: language resource, terminology resource, multilingualism, consolidation, linguistic infrastructure

1 Introduction

Terminology is multidisciplinary and comprises primarily such tasks as the analysis of concepts and conceptual systems; creation of new terms; identification, recognition, extraction of existing terms; compilation of terminology resources, for example, dictionaries, banks, databases, i.e., terminography; application of terminology resources, for example, in translation, including computer-assisted and machine translation; management of terms.

After a long history of terminology work and as the number of terminology resources grow every year, the task of the consolidation of different, usually dispersed, terminology resources becomes more urgent.[1] According to recent research surveys on user practice in terminology work, one of the most required functionality of a terminology resource is a consolidated interface [2, 4].

[1] See, for example, a PhD thesis on consolidation of heterogeneous terminology resources [8].

2 Dispersed Terminology and Its Consolidation

There have been several efforts to provide reasonable solutions in accessing multilingual terminology resources for language workers. For example, the *Quest* tool, a one-stop access to a series of general-interest terminology databases, brings consolidated terminology content to translators in the Directorate-General for Translation of the European Commission [6]. Quest is not a terminology database but a metasearch interface which translators can use to query several databases simultaneously [Ibid.: 9].

One of the major efforts in the consolidation of terminology resources is EuroTermBank[2] – a centralized online publicly available term bank for the EU languages which provides a federated access to 5 interlinked external term banks [1, 7, 10] (Fig. 1).

Moreover, a free innovative multilingual terminology translation tool *EuroTermBank Terminology Add-in*[3] was developed. The tool integrates terminology resources from EuroTermBank into the most widely exploited working environment among language workers – *Microsoft Word 2003/2007* [3].

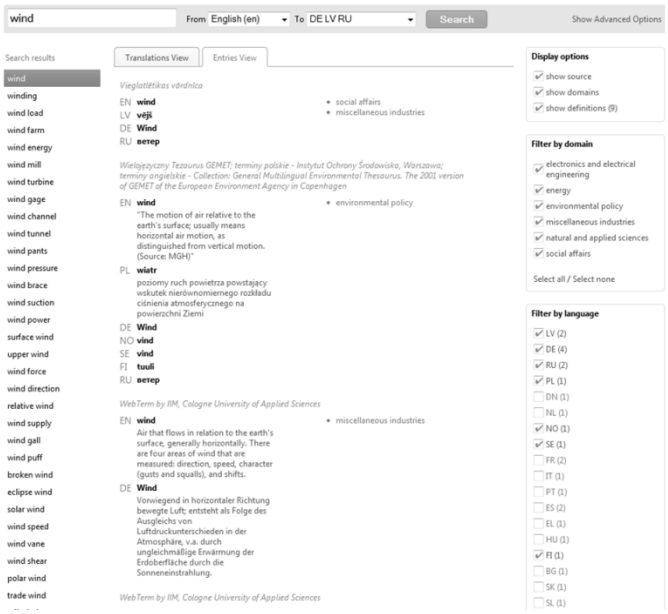


Fig. 1. Consolidated representation of terminology entries in EuroTermBank from different bilingual and multilingual resources

[2] www.eurotermbank.eu

[3] The tool can be downloaded under the following link: http://www.eurotermbank.com/downloads.aspx

3 Consolidation of Terminology in META-NORD

The ICT PSP project META-NORD[4] contributes to building an open linguistic infrastructure for language resources by identifying and describing language resources in the Baltic and Nordic countries [5, 11] and by populating language resources (after IRP issues are cleared) and their metadata into the open distributed META-SHARE platform.[5] The first batch of META-NORD language resources was released in November 2011, two more batches are planned in July, 2012 and January, 2012.

META-NORD addresses a growing demand for consolidating dispersed terminology resources. Within the task of consolidating European multilingual terminology across languages and domains, META-NORD aims at:

- extending META-SHARE with monolingual and bilingual multilingual terminology resources across Europe;
- integrating the EuroTermBank platform into META-SHARE by adapting EuroTermBank to relevant data access and sharing mechanisms;
- populating EuroTermBank with additional terminology resources and thus broadening the language coverage of EuroTermBank [9].

EuroTermBank will be integrated into the META-SHARE platform as a distributed terminology repository (Fig. 2).

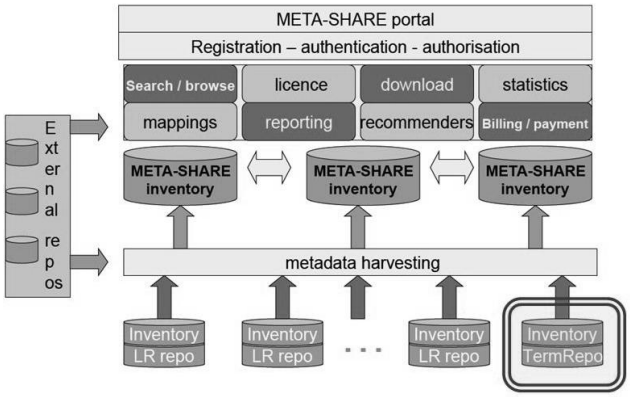


Fig. 2. Terminology repository within the META-SHARE network

META-SHARE will contain among others: (1) EuroTermBank as a local terminology repository consisting of its own terminology resources and metadata that will follow the META-SHARE schema; (2) a local inventory consisting of metadata for the terminology resources stored at the local repository; (2) a META-SHARE inventory consisting of the metadata for the terminology resources stored in the repository.[6]

[4] www.meta-nord.eu

[5] www.meta-share.eu

[6] See more about the overall architecture of META-SHARE at: www.meta-net.eu/meta-share/architecture.

4 Conclusions and Future Work

META-NORD lays the ground for fruitful cooperation in identifying, consolidating, and sharing terminology resources across Europe. During the first year of the project 10 terminology resources have been identified to be interlinked with the META-SHARE portal via EuroTermBank. The work on integrating EuroTermBank into META-SHARE has started recently and the number of interlinked terminology resources can be increased during the second year of the project. It is also anticipated that the META-NORD initiative can be further extended to other European countries by other projects within the META-NET network – CESAR, METANET4U, and T4ME.

References

1. Auksoriūtė A., Belogrīvs I., Bielevičienė A. [et al.].Towards Consolidation of European Terminology Resources: Experience and Recommendations from the EuroTermBank Project. Signe Rirdance and Andrejs Vasiļjevs (eds.). Tilde, Riga, Latvia, 2006.
2. Blancafort, H. and Gornostay, T. Calling Professionals: Help Us to Understand Your Needs! A questionnaire-based online survey about terminology and corpora practices. Electronic resource: http://www.ttc-project.eu/images/stories/TTC_Survey_2010.pdf.
3. Gornostay T., Vasiljevs A., Rirdance S., and Rozis R. Bridging the Gap – EuroTermBank Terminology Delivered to Users' Environment. In Proceedings of the 14th Annual Conference of the European Association for Machine Translation, Saint-Raphaël, France, 2010.
4. Gornostay, T. Terminology management in real use. In: Proceedings of the 5th International Conference "Applied Linguistics in Science and Education", Saint-Petersburg, Russia, 2010.
5. Skadiņa, I., Vasiļjevs, A., Borin, L., Smedt, De K., Lindén, K., Rögnvaldsson, E. META-NORD: Towards Sharing of Language Resources in Nordic and Baltic Countries. In Proceedings of the workshop IJCNLP 2011 on Language Resources, Technology and Services in the Sharing Paradigm, Chiang Mai, Thailand, 2011.
6. Translation tools and workflow: European Commission Directorate-General for Translation. Luxembourg, Office for Official Publications of the European Communities, 2007.
7. Vasiljevs A. and Rirdance S. Consolidation and unification of dispersed multilingual terminology data. In Proceedings of the International Conference "Recent Advances in Natural Language Processing", Borovets, Bulgaria, 2007.
8. Vasiļjevs, A. Consolidation of Heterogeneous Terminology Resources. PhD thesis, University of Latvia, Riga, Latvia, 2011.
9. Vasiļjevs, A., Forsberg, M., Gornostay, T. [et al.]. Creation of an Open Shared Language Resources Repository in the Baltic and Nordic Countries. In Proceedings of the 8th International Conference on Language Resources and Evaluation, Istanbul, Turkey, 2012.
10. Vasiljevs, A., Rirdance, S. and Liedskalnins. A. EuroTermBank: Towards Greater Interoperability of Dispersed Multilingual Terminology Data. In Proceedings of the 1st International Conference on Global Interoperability for Language Resources, Hong Kong, 2008.
11. Vasiļjevs, A., Skadiņa, I., Sandford, B.P., Smedt, De K., Borin, L. META-NORD: Baltic and Nordic Branch of the European Open Linguistic Infrastructure. In Proceedings of the 18th Nordic Conference on Computational Linguistics, Riga, Latvia, 2011.