Towards the Generation of Semantically Enriched Multilingual Components of Ontology Labels

Thierry Declerck and Dagmar Gromann

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

Abstract. Ontologies often contain multilingual textual information in annotation properties, such as `rdfs:label` and `rdfs:comment`. While the motivation for using such annotation properties is to provide a human readable description of abstract conceptualization of the domain, we notice that the importance of appropriate natural language use and representation is often neglected. The same can be observed with resources on the Web, such as multilingual taxonomies. Terms often lack consistency and completeness, hampering also an accurate automated natural language processing of such text. We propose a pattern-based transformation of terms in labels, thereby also supporting a multilingual alignment of (sub)components of labels. The source data for our approach is an ontology we derived from an industry classification taxonomy, which we improve as regards consistency and completeness and apply to the process of lexicalization.

Keywords: Ontology Labels, Multilingualism, Terms and Sub-Terms

1 Introduction

Nowadays, it has been increasingly realized that the process of ontology construction is inevitably linked to natural language and related to this development multilingualism is progressively gaining center stage in ontology engineering. There are various possibilities to add natural language strings to ontologies. These strings can be part of RDF URI references, identifying ontological resources (e.g. natural language string used in `rdf:ID`), a fragment (e.g. natural language string in `rdf:about` statements) or marking empty property elements (kind of leaf nodes in a graph, using the `rdf:resource` statement). Natural language strings equally represent the content of the RDF annotation properties `rdfs:label` and `rdfs:comment`, which provide information on ontological resources in a human-readable format.

Herein, we focus on the content of annotation properties. This choice has been partially motivated by the fact that these properties qualify for the inclusion of terminological information, which can be realized in form of longer natural language strings. Additionally, labels and comments locally support multilinguality by means of language tags of RDF literals, i.e., `xml:lang`, whereas this is not the case for RDF URI references.

Analyzing the content of annotation properties in multilingual ontologies, we registered that their realization frequently hampers an accurate automatic linguistic and semantic processing. This type of processing is vital to a large number of ontology-based tasks, such as machine translation, information extraction, cross-lingual ontology mapping. Thus, we investigate if and how cross-lingual preprocessing and linguistic harmonization of terms in ontology labels can be of avail for such processing. At the same time, these initial steps support a multilingual alignment of subcomponents of labels, leading to more fine-grained multilingual resources associated with ontology elements.

Our experimental results are based on the analysis of labels and comments of an ontology we derived from the Global Industry Classification Standard (GICS) taxonomy[1] in English, German, and Spanish. The GICS taxonomy consists of four meta-levels, namely, sector, industry group, industry, sub-industry. These four categories represent the top nodes of the ontology. Each leaf node, i.e., each sub-industry, contains a detailed definition. All classes are indexed by integers, which also indicate the hierarchical structure of the taxonomy: the descending line "10" (Energy), "1010" (Energy), “101010" (Energy Equipment & Services) and "10101010" (Oil & Gas Drilling) represents the first complete branch of the hierarchical tree of the classification scheme[2].

The investigation was triggered by our observation that applying baseline Machine Translation (MT) tools, such as Google Translate, to terms used in GICS produces substantially different terms in target languages than provided by the corresponding languages in GICS. For example, only a partial Spanish translation was obtained for the German compound ellipsis "Eigentums- und Unfallversicherungen", resulting in "Propiedad y accidente", whereas the correct translation should be "Seguro de Propiedad y Accidente" (Property and casualty insurance).

As regards structure, related work will be presented in section 2. Preprocessing steps and corrective patterns for the purpose at hand will be discussed in section 3. Deriving subcomponents of ontology labels for multilingual alignment will be the focus of section 4. Finally, the resulting ontology will be lexicalized by means of *lemon* [7] prior to concluding remarks.

2 Related Work

Research in various areas such as multilingual ontology acquisition [6], cross-lingual ontology mapping [11], ontology lexicalization [7], linguistic enrichment of labels [8], ontology engineering from text [9], and ontology localization [10]

[1] `http://www.standardandpoors.com/indices/gics/en/us`

[2] The definition associated with the leaf concept ID 10101010 is "Drilling contractors or owners of drilling rigs that contract their services for drilling wells."

can be observed. All of these approaches highlight the importance of ontologies labeled in different languages and techniques of acquiring them. While [11] seems to be the closest approach to our investigation, the major difference lies in the fact that [11] (and in fact also [15]) addresses only the language data included in RDF URI reference statements. Consequently, they are not concerned with natural language processing of (possibly lengthy) multilingual natural language strings, but only with finding equivalents of reference expressions in various lexical resources.

Current and future results of our work might best be compared to state-of-the-art research in the field of lexico-syntactic patterns, which are part of ontology design patterns[3] and mostly used for learning ontologies from natural language text (e.g. [5]). For this purpose and for the approach we apply to the analysis of the content of ontology labels, many different linguistic processes, such as tokenization, lemmatization, shallow parsing are used, also often combined with statistical machine learning techniques to learn ontologies from large sets of documents, e.g. Text2Onto [4]. The major problem of such patterns is low precision and over-generalization, which [3] try to overcome by restricting their main approach to three sets of patterns.

The creation of ontologies from text (e.g. [12, 2]) or other resources such as thesauri (e.g. [1]) and taxonomies has been a thriving research topic as of late. However, the use of multilingual information as a means of coherence and consistency check of ontology labels calls for further investigation. Our work seems to open the possibility to offer better proposals for the use of more consistent terminology in labels associated with ontology elements in a cross-lingual setting.

3 Initial Processing Steps and Cross-Lingual Corrective Patterns

We concentrate in this experiment on multilingual aspects in the GICS ontology we derived from the original taxonomy, having in mind the potential for an improved translation base for terms in this domain and for Information Extraction in documents describing among others activities of companies. Initially, we focused on labels in the three languages English, German, and Spanish, but have already experimented with Russian labels.

To remedy the deficient translatability of GICS labels, we investigated the transformation of the surface realization of the contained terms. In order to achieve a better readability of the ontology by engineers and users and better prepare labels to automatic processing, we transform non-lexical symbols to lexical correspondents, apply lexico-syntactic patterns to resolve compound ellipses, and complement labels based on constituency discrepancies across languages, i.e., missing constituents in one or more languages.

Replacing non-lexical items by their lexical correspondents refers to punctuation and ampersands. Duplicate occurrences of punctuation such as ",." are

[3] http://ontologydesignpatterns.org

corrected. Ampersands occur 159 times in the English taxonomy, the coordination word "and" not being used at all, while the German version features 117 occurrences and the Spanish only uses the coordination marker "y". The ampersand character serves to represent coordination, but automated linguistic decomposition of terms containing ampersands is not supported by off-the shelf NLP tools. As a rather straight-forward step the ampersand was replaced by "and" and "und" (DE).

At a more complex level we transform so called compound ellipses in GICS labels in fully lexicalized strings. Elliptical compounds represent the outcome of a deletion process of identical constituents in either the right or the left part of the coordination. For instance, the hyphenated German compound "Erdöl- und Erdgasförderung" (Oil and Gas Drilling) is transformed to "Erdölförderung und Erdgasförderung" (Oil Drilling and Gas Drilling). This transformation is not trivial as it requires both the analysis of the compounds and the resolution of the ellipsis, attaching the constituent "Förderung" to "Erdöl" in the example above. This process necessitated the use and adaptation of a morphological analysis component and the generation of ellipsis grammars, which are both implemented in the NooJ[4] finite state framework. Examples of the lexico-syntactic patterns implemented in NooJ are provided below.

[Examples of Resolution Patterns of Elliptical Coordinations]

```
DE: <NN1>hyphen und <NN2+NN3> resolved to <NN1+NN3> und <NN2+NN3>
EN: <NN1> and <NN2> <NN3>  resolved to <NN1> <NN3> and <NN2> <NN3>
ES: <NN1> <ADJA1> y <ADJA2> resolved to <NN1> <ADJA1> y <NN1> <ADJA2>

DE: <NN1+NN2> und hyphen<NN3> resolved to <NN1+NN2> und <NN1+NN3>
EN: <NN1> <NN2> and <NN3>  resolved to <NN1> <NN2> and <NN1> <NN3>
ES: <NN1> y <NN2> de <NN3> resolved to <NN1> de <NN3> y <NN2> de <NN3>
```

The presence of the German hyphen compound triggers the resolution of ellipses into coordinated structures in labels for other languages attached to the same concept. For instance, the German example above triggers the transformation of the English label "Oil and Gas Drilling" to "Oil Drilling and Gas Drilling" and of the Spanish label "Perforación de Pozos Petrolíferos y Gasíferos" to "Perforación de Pozos Petrolíferos y Perforación de Pozos Gasíferos". The resolution not only concerns single nouns, but also nominal phrases, e.g. "Perforación de Pozos", and adjectival phrases. As our algorithm requires the presence of a German hyphen, terms such as "Commercial Services and Supplies" (related to the German "Gewerbliche Dienste und Betriebsstoffe") are not resolved and are also not supposed to be resolved. All definitions attached to GICS terms confirm our approach to ellipsis resolution. Further examples of resolution in all three languages are as follows.

[Annotation Results of NooJ Processing applied to German, English and Spanish]

[4] http://www.nooj4nlp.net/pages/nooj.html

```
<EL TYPE="Energiezubehör#und#Energiedienst">Energiezubehör und -dienste</EL>
<ELLLL TYPE="Grosshandel#und#Einzelhandel">Gross- und Einzelhandel</ELLLL>

<EL TYPE="Energy#Equipment#and#Energy#Services">Energy Equipment and
Services</EL>
<EFOURD TYPE="Oil#:#Exploration#and#Oil#:#Production#and#Gas#:#Exploration#and
#Gas#:#Production">Oil and Gas Exploration and Production</EFOURD>

<EL TYPE="Equipos#de#Energía#y#Servicios#de#Energía">Equipos y Servicios de
 Energía</EL>.
<ELLL TYPE="Productos#Madereros#y#Productos#Papeleros">Productos Madereros y
Papeleros</ELLL>
```

At times the authors of the industry classification apply a colon to structure terms, such as *Metalle & Bergbau: Diverse* (Diversified Metals and Mining). Frequently, these constructs can only be resolved using prepositions instead of compounding, because terms such as *Heiwerkerausrüstungseinzelhandel* (Home Improvement Retail) do not exist. Structures using colons could only be observed in German labels of GICS.

As a final preprocessing step we evaluated complementing labels on the basis of a cross-lingual comparison. The German "Integrierte Erdöl- und Erdgasbetriebe" lacks any equivalent of "betrieb" (company) in the English or Spanish version. Despite the fact that the taxonomy is about business activities, the word company does virtually not occur in the English or Spanish designations of concepts, only in definitions. For the sake of completeness, we decided to complement the English and Spanish label with the equivalent of the missing term taken from sibling concepts in the same sector or definitions. In this case, we add "companies" and "empresas" on the basis of the assumption that multilingual labels associated with concepts should, where feasible, have the same amount and quality of information.

The presented algorithm ports all terms to a shared surface realization and depicts the different but aligned language specific realizations. While the patterns for resolving general ellipsis can be applied to other sources, such as the Industry Classification Benchmark (ICB)[5], the second case of terms separated by colon seems to be specific to GICS. Currently the algorithm has been implemented for the indicated languages, however, we have performed experiments with their utilization for other not closely related languages, such as Russian. Many lexico-syntactic patterns can be applied directly to the Russian designations, such as the compound "Хранение и транспортировка нефти и газа" (Storage and Transportation of Oil and Gas) can be resolved to "Хранение нефти и транспортировка нефти и Хранение газа и транспортировка газа" (Oil Storage and Oil Transportation and Gas Storage and Gas Transportation).

The representation of the fact that we modified the original terms (or labels) remains to be an issue. Indicating the modification is important to the authors of the taxonomy as well as people analyzing data. As a tentative step, for this

[5] http://www.icbenchmark.com/

purpose we have introduced the annotation property "preprocessed" to clarify that we have adapted the original content of labels and definitions.

4 Multilingual Alignment and Sub-Term Structures

Performing initial preprocessing steps facilitates the multilingual alignment of terms and components of terms. For the purpose of multilingual alignment, we have extensively analyzed and utilized existing hierarchical relations and definitions. In a second step we create relations to indicate sub-term relations in the actual ontology. By creating an additional terminological resource, we derive a second subsumption hierarchy focusing on sub-term relations, which is supposed to facilitate Information Extraction based on the ontology we created.

4.1 Term Alignment

Within the taxonomic structure of GICS we are able to establish relations between (sub)terms along the line of class hierarchies. GICS is structured along four major meta-categories in sector, industry group, industry, and sub-industry. Terms used in a super-class can thus be used for comparing a term in one language with the terms of other languages. Not only the line hierarchy is interesting for us but equally siblings in the hierarchy provide vital information.

Lexica and lexical resources created in the initial processing are now utilized to create multilingual alignments of the terminology contained in the taxonomy. We utilize lemmas of the normalized labels to facilitate the multilingual alignment as represented by the NooJ output illustrated below.

[Example of NooJ Annotation Result]

```
<TYPE="Integrierte#Erdoelbetriebe#und#Integrierte#Erdgasbetriebe">
Integrierte Erdoel- und Erdgasbetriebe</>
```

The associated lexical information in NooJ tells us in this case that "Integrierte" is the adjectival form derived from the verb "integrieren" (to integrate). The lemma of the head of the compound noun "betriebe" being then "Betrieb" (company). Thereby, we are able to establish term relations on the basis of the hierarchy, such as depicted below for the GICS class "101020".

[Example of Term Alignment]

```
"de"  => "Erdoel, Erdgas und nicht erneuerbare Brennstoffe",
"en"  => "Oil, Gas and Consumable Fuels",
"es"  => "Petroleo, Gas y Combustibles",

"trans"  =>
"Erdoel@de = Oil@en = Petroleo@es ::
 Erdgas@de = Gas@en = Gas@es ::
 Nicht erneuerbare Brennstoffe@de = Consumable Fuels@en = Combustibles@es"
```

Term pairs may vary strongly across different sectors within one classification. For instance, "Leisure products" equals "Freizeitartikel" in German, while "Agricultural Products" corresponds to "Landwirtschaftliche Produkte". Once "product" is aligned with "Artikel", in a different sector it maps to "Produkte". Nevertheless, this fact does not hamper automating the alignment process, which has been done on the basis of a Java tool, porting the preprocessed labels to the subsumption hierarchy of the ontology. At times, this initial alignment can lead to multiple mappings of terms depicted in Fig. 1.

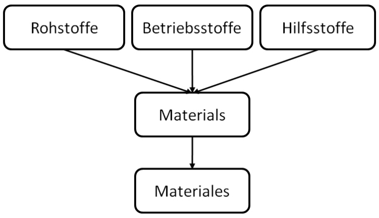

Fig. 1. Different Conceptualization of Cross-Lingual Designations

The interesting point about the example in Fig. 1 is the different conceptualization across languages. The German multi-word term corresponds to the single word expression "Material" in English and Spanish, which constitutes a challenge for cross-lingual alignment as there seems to be no equivalent for the three German expressions in the other languages.

In such cases other term pairs within the same sector are analyzed as regards re-occurrence of terms. If no equivalence can be detected, the definition has to be searched. Should the terms be only contained in one label, then additional resources, such as bilingual dictionaries or other multilingual industry classifications, might be consulted. However, in other cases clear misalignments occur, such as "Betriebsstoffe" in German being aligned to "Professional Services" (en) and "Servicios Profesionales" (es). As the same designation is part of another sub-industry in the sector, the incorrect alignment can be corrected on the basis of the existing correct alignment to "Professionelle Dienste". The definition in each language further confirms this alignment.

Our special focus is on terminal nodes in the original taxonomy as they contain detailed definitions, which further facilitates the cross-lingual alignment and validation of alignment correctness. As a tentative approach, we use lexico-syntactic patterns again to extract some basic information contained in definitions, exemplified by one pattern in German below. The extracted information as well as manually derived alignments from definitions are both used to validate the previously described alignments of designations of taxonomic concepts.

[Pattern for Extraction of Information in Definitions]

```
German:
<NP1>, die sich mit <NP2>, <NP3>und OR oder <NP4>
von <NP5> tätig sind OR beschäftigen.
Definition "Pharmazeutika": "Unternehmen, die in der Erforschung,
Entwicklung oder Herstellung von Pharmazeutika tätig sind."
```

Definitions provide further information to facilitate the construction of proper terms and term alignments. <NP1> represents a synonym of the word company, e.g. manufacturer, producer, provider, whereas the other noun phrases relate to business activities. One Spanish example is the industry of *Transportes*, which has the industry group *Transporto Aereo* and sub-industry *Lineas Aereas* referring to the former term explicitly in its definition.

Analyzing siblings creates relations that would otherwise not be evident. For instance, "Building Products" might not be related to "Aerospace and Defense" in any other domain. Within this sector, however, they are related as it regards the manufacturing of aerospace and defense equipment. The extracted term pairs of the definitions allow us to add these additional information to the label to strive towards completeness of information.

Terms aligned in this section are represented in the GICS ontology as annotation properties with the respective `xml:lang` property. Initial preprocessing and the correct alignment of terms serve to improve the overall quality of the natural language representation of the ontology. The alignment of terms equally helps to reveal inconsistencies or in other words improve the consistency of ontology labels.

4.2 Sub-Term Relations

At this point our ontology consists of five main classes according to the taxonomic structure, the four meta-levels and an additional class "Company". The latter features a "hasBusinessActivitiy" object property to the main class "Sub-Industry" so that upon instantiating a company various activities can be added. In addition, all taxonomic categories have a `subClass` relationship to the respective meta-category.

Creating sub-term relations introduces an additional structure not originally part of the GICS taxonomy, which is why we have decided to create an additional OWL-DL resource dedicated to terminology and terminological relations. For "isSubTermOf" relations it might be worth considering a transitive characteristic, that is: "P(x,y) and P(y,z) implies P(x,z)"[6], so each term y isSubTermOf x, z isSubTermOf y, which implies that z isSubTermOf x. This allows us to state that "Trucks" is a subterm of "Heavy Trucks" and at the same time of "Farm Machinery and Heavy Trucks". This type of decomposition abides by the terminological principles presented in ISO704:2009.

[6] `http://www.w3.org/TR/2004/REC-owl-guide-20040210/#PropertyCharacteristics`

In order to account for the terminological relations and levels, pseudo-categories, i.e., categories not originally part of the taxonomy and generated for terminological reasons, have to be introduced to the original hierarchy. This is due to the fact that terminological relations focus on hypernymic, meronymic relations. For example, the subcategories of *Energy* all refer to either Energy, Oil, Gas, or Consumable Fuels, all of which have to be introduced to the terminological structure.

The decomposition of e.g. "Oil Equipment and Gas Equipment and Oil Services and Gas Services" centers around the constituent and divides the term at the second "and". Accordingly, the definition of sub-industries has to be adapted to the changed concept and added to the terminological entry. Information extracted from definitions in the previous step are added to the terminology in order to enlarge the contained vocabulary.

A terminological representation of these natural language labels of an ontology provides a highly beneficial overview of contained terms, their sub-terms and relations between them. This facilitates duplicity and consistency evaluations of labels. In combination with part of speech, morphological, and syntactic information represented in *lemon*, there are various application scenarios from facilitating the creation of new labels to machine translation.

5 Lexicalizing Ontology Labels

Several approaches and models seek to provide a lexicon-ontology interface to reduce the complexity of the ontology, while at the same time providing full lexical information on the natural language representation of ontologies.

The *lemon* model [7] was developed within the Monnet project[7] and represents textual and linguistic information contained in ontologies as external RDF resource and establishes semantics by means of relating entries to the ontology, i.e., the relation represents a means to disambiguate words. It adapts the main principles of the Lexical Markup Framework (LMF) standardized in ISO 24613 and unites it with the core features of *LexInfo* in order to elaborate a specific ontology-lexicon model. Lexicon objects describe syntactic and morpho-syntactic properties, which are related to entities of the ontology via sense objects. Subsequent to applying state labels to the entry, i.e., preferred, alternative, hidden reference, the lexical sense links to the lexical entry, which might be decomposed to its individual elements.

Lexicons based on *lemon* can be created automatically by means of the *lemon* generator[8]. The following lexicon was created on the basis of the seed ontology, without any preprocessing and term alignment. As can be seen, decomposition of the term "Energy Equipment & Services" fails due to the ampersand and the ellipsis.

[7] `http://www.monnet-project.eu`

[8] `http://monnetproject.deri.ie/lemonsource/`

[*lemon* decomposition of "Energy Equipment & Services"]

```
<lemon:decomposition xmlns:ns0=
"http://www.w3.org/1999/02/22-rdf-syntax-ns#" ns0:parseType="Collection">
     <lemon:Component  rdf:about="unknown:/GICS__en/
               Energy%2BEquipment%2B%26%2BServices#comp">
               <lemon:element rdf:resource="unknown:/GICS__en/Energy"/>
     </lemon:Component>
     <lemon:Component rdf:about="unknown:/GICS__en/
               Energy%2BEquipment%2B%26%2BServices#comp2">
               <lemon:element rdf:resource="unknown:/GICS__en/Equipment"/>
     </lemon:Component>
     <lemon:Component rdf:about="unknown:/GICS__en/
              Energy%2BEquipment%2B%26%2BServices#comp3">
             <lemon:element rdf:resource="unknown:/GICS__en/Services"/>
     </lemon:Component>
</lemon:decomposition>
```

The application of off-the-shelf NLP tools to labels in fact negatively influences the efficiency of an automated *lemon* based lexicalization process of labels, as most commonly used tools are not in the position to handle such types of (mainly nominal) ellipsis. Considering the fact that ontology labels to a large extend only consist of nouns and noun compounds, the issue is a vital one. We apply the process of lexicalization to the annotation property `rdfs:label` available in all languages covered in the GICS ontology, namely German, English, Spanish. For this purpose we use *lemon* for the representation of linguistic information added to these labels and linking to the original ontology elements.

Lexicalization supports the decomposition of terms into sub-terms, that is it facilitates the application of patterns to detect cross-lingual alignments at the level of components of terms/labels. The linguistic information in the *lemon* representation is being used for consolidation. However, we consider the decomposition of terms to be part of the terminological level, thus, introducing the terminological resource for GICS in section 4. The example below shows the encoding of constituency and part-of-speech information subsequent to our initial preprocessing and term alignment process.

[Constituency and Part-Of-Speech Information of "Energy Equipment and Energy Services" in *lemon*]

```
<lemon:entry>
     <lemon:LexicalEntry rdf:about="unknown:/lexicon__en/Energy+Equipment+and+Energy+Services">
        <lemon:sense>
          <lemon:LexicalSense rdf:about="unknown:/lexicon__en/Energy%2BEquipment%2Band%
             2BEnergy%2BServices#sense">
            <lemon:reference rdf:resource="http://www.semanticweb.org/ontologies/2012/8/GICS.owl#GICS101010"/>
          </lemon:LexicalSense>
        </lemon:sense>
        <lemon:canonicalForm>
          <lemon:Form rdf:about="unknown:/lexicon__en/Energy+Equipment+and+Energy+Services#form">
            <lemon:writtenRep xml:lang="en">Energy Equipment and Energy Services</lemon:writtenRep>
          </lemon:Form>
        </lemon:canonicalForm>
        <lemon:phraseRoot>
          ...
          <lemon:constituent rdf:resource="http://monnetproject.deri.ie/tags/penn/node/NN"/>
          ...
          <lemon:constituent rdf:resource="http://monnetproject.deri.ie/tags/penn/node/NNS"/>
          ...
          <lemon:constituent rdf:resource="http://monnetproject.deri.ie/tags/penn/node/NP"/>
```

```
         ...
         <lemon:constituent rdf:resource="http://monnetproject.deri.ie/tags/penn/node/CC"/>
         ...
         <lemon:constituent rdf:resource="http://monnetproject.deri.ie/tags/penn/node/NN"/>
         ...
         <lemon:constituent rdf:resource="http://monnetproject.deri.ie/tags/penn/node/NP"/>
         ...
         <lemon:constituent rdf:resource="http://monnetproject.deri.ie/tags/penn/node/NP"/>
...
</lemon:entry>
```

Due to space constraints the example only provides an English version, however, the same improved results can be observed in German and Spanish. The above example provides that *lemon* was in the position to decompose the term and provide part-of-speech information, using the Penn Treebank Notation. The lexical sense contains the link to the ontology and the original label as "writtenRep", followed by information on individual elements of the term. This use case is supposed to show that that such type of preprocessing and term alignment has beneficial effects on ontology labels.

6 Concluding Remarks and Future Work

We have preprocessed the labels of an ontology we derived from the GICS taxonomy, for the time being in English, German, and Spanish. We showed a pattern-based approach to resolving compound ellipses, which can be generalized across resources, such as the Industry Classification Benchmark (ICB). Thereby, we created terms initially not contained in the resource and thus, inaccessible to ontology-based tasks, such as Information Extraction. We aligned the terms across all three languages. Terms contained in definitions were extracted and additionally aligned to increase the overall quality and validate existing alignments. Furthermore, the normalized and aligned terms were included in a terminological resource in OWL-DL to provide explicit sub-term relations and decompose complex, long labels. Lexicalizing the derived ontology with its processed labels as opposed to the initial ontology served to exemplify the usefulness of such (pre)processing of labels.

As regards future work, we are currently investigating the applicability of our pattern-based approach to other language families than Romance languages. One further approach that might be interesting is the automation of the creation of a terminological resource for the ontology, similar to the idea of the *lemon* generator.

References

1. Kless, D., Jansen, L., Lindenthal, J., Wiebensohn, J.:A Method of Re-Engineering a Thesaurus into an Ontology. In: Donelly, M., Guizzardi, G. (eds): Formal Ontology in Information Systems - Proceedings of the Seventh International Conference (FOIS 2012), pp.133–146. IOS Press, Amsterdam (2012)
2. Serra, I., Girardi, R.: A Process for Extracting Non-Taxonomic Relationships of Ontologies from Text. Intelligent Information Management 3, 119–124 (2011)

3. Maynard, D. F. A., Peters, W.: Using lexicosyntactic Ontology Design Patterns for Ontology Creation and Population. In Proceecdings of the Workshop on Ontology Patterns (2009)
4. Cimiano, P., Voelker, J.: Text2Onto - A Framework for Ontology Learning and Data-driven Change Discovery. In: Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB), Alicante, Spain (2005)
5. Klaussner, C., Zhekova, D.: Lexico-Syntactic Patterns for Automatic Ontology Building. In: Proceedings of the International Conference Recent Advances in Natural Language Processing (2011)
6. Nichols, E., Bond, F., Tanaka, T., Fujita, S., Flickinger, D.: Multilingual Ontology Acquisition from Multiple MRDs. In Proceedings of the 2nd Workshop on Ontology Learning and Population, pp. 10–17 (2006)
7. Buitelaar, P., Cimiano, P., McCrae, J., Montlie-Ponsoda, E., Declerck, T.: Ontology Lexicalization: The *lemon* Perspective. In: Slodzian, M., Valette, M., Aussenac-Gilles, N., Condamines, A., Hernandez, N., Rothenburger, B. (eds.): Workshop Proceedings of the 9th International Conference on Terminology and Artificial Intelligence, pp- 33–36, Paris, France, INALCO, Paris (2011)
8. Declerck, T., Lendvai P.: Towards a standardized linguistic annotation of the textual content of labels in knowledge representation systems. In: LREC 2010- The seventh international conference on Language Resources and Evaluation. Interna- tional Conference on Language Resources and Evaluation (LREC-10), Malta (2010)
9. 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, (2008)
10. Mejía, M.E., Montiel-Ponsoda, E., Aguado de Cea, G., Gómez-Pérez, A.: Ontology Localization. In: Suárez-Figueroa, M.C.,Gómez-Pérez, A., Motta, E., Gangemi, A. (eds): Ontology Engineering in a Networked World. pp. 171–191, Springer Berlin Heidelberg (2012)
11. Fu, B., Brennan, R., O'Sullivan, D.: Using Pseudo Feedback to Improve Cross-Lingual Ontology Mapping. In: Proceedings of the 8th Extended Semantic Web Conference (ESWC 2011), LNCS 6643, pp. 336–351 (2011)
12. de Cea, G.A., Gómez-Pérez, A., Ponsoda, E.M., Suárez-Figueroa, M.C.: Natural Language-Based Approach for Helping in the Reuse of Ontology Design Patterns. In: Proceedings of the 16th International Conference on Knowledge Engineering and Knowledge Management Knowledge Patterns (EKAW 2008), Acitrezza, Italy (2008)
13. Suárez-Figueroa, M.C., Gómez-Pérez, A., Fernández-López, M.: The NeOn Methodology for Ontology Engineering. In: Suárez-Figueroa, M.C.,Gómez-Pérez, A., Motta, E., Gangemi, A. (eds): Ontology Engineering in a Networked World. pp. 9–34, Springer Berlin Heidelberg (2012)
14. Cimiano P., Buitelaar P., McCrae J., Sintek M.: LexInfo: A declarative model for the lexiconontology interface. Journal of Web Semanics, Vol. 9, No. 1, pp. 29–51 (2011)
15. Vertan, C., v.Hahn, W. Challenges fort he Multilingual Semantic Web. In Proceedings of the International Workshop on Semantic web Technologies for Machine Translation, in conjunction with MT-Summit X, Phuket, Thailand (2005)