# Towards a standardized linguistic annotation of the textual content of labels in Knowledge Representation Systems

**Thierry Declerck[1], Piroska Lendvai[2]**
[1] DFKI GmbH, Language Technology Lab
Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany
[2] Research Institute for Linguistics, Hungarian Academy of Science
Benczúr u. 33, H-1068 Budapest, Hungary
E-mail: declerck@dfki.de, piroska@nytud.hu

**Abstract**

We propose applying standardized linguistic annotation to terms included in labels of knowledge representation schemes (taxonomies or ontologies), hypothesizing that this would help improving ontology-based semantic annotation of texts. We share the view that currently used methods for including lexical and terminological information in such hierarchical networks of concepts are not satisfactory, and thus put forward – as a preliminary step to our annotation goal – a model for modular representation of conceptual, terminological and linguistic information within knowledge representation systems. Our *CTL* model is based on two recent initiatives that describe the representation of terminologies and lexicons in ontologies: the *Terminae* method for building terminological and ontological models from text (Aussenac-Gilles et al., 2008), and the *LexInfo* metamodel for ontology lexica (Buitelaar et al., 2009).

## 1. Introduction

Despite the focused attention of and improvements achieved by the NLP community on various language-related issues of knowledge representation schemes, defining and standardising how natural language expressions should be included in elements of such systems is still not satisfactorily solved. Hierarchically built conceptual networks such as taxonomies and ontologies typically include some non-atomic, i.e. free text, descriptive natural language expression within the conceptual objects they hold. These may serve as definition, comment, or as realization of the terminological content of the concepts. Many taxonomies and ontologies encode terms with the help of an attribute named *label*, which is attached to the concept ID of their classes, as shown in an example ontology written in OWL (Web Ontology Language) in Figure 1.

```
<owl:Class rdf:ID  "SpicyPizza">
    <rdfs:label xml:lang="pt">PizzaTemperada</rdfs:label>
    <rdfs:comment xml:lang="en">Any pizza that has a spicy topping is a
    SpicyPizza</rdfs:comment>
    <owl:equivalentClass>
        <owl:Class>
            <owl:intersectionOf rdf:perseType  "Collection">
             <owl:Class rdf:about="#Pizza"/>
             <owl:Restriction>
                  <owl:onProperty>
                        <owl:ObjectProperty rdf:about="#hasTopping"/>
                  </owl:onProperty>
                  <owl:someValuesFrom rdf:resource  "#SpicyTopping"/>
             </owl:Restriction>
        </owl:intersectionOf>
        </owl:Class>
    </owl:equivalentClass>
</owl:Class>
```

Figure 1: Use of *label* and *comment* attributes for including information in natural language in an ontology[1]

[1] This example is taken from the famous Pizza Ontology delivered with the Protégé Ontology Editor, http://protege.stanford.edu/.

We note though that not all semantic resources (taxonomies and ontologies) comply with this notation.

Two examples, taken from the RadLex ontology and from the XBRL taxonomy, are discussed below.

```
<class>
 <name>RID13218</name>
 <type>anatomy_metaclass</type>
 <own_slot_value>
   <slot_reference>FMAID</slot_reference>
   <value value_type="string">67112</value>
  </own_slot_value>
 <own_slot_value>
   <slot_reference>Synonym</slot_reference>
         <value value_type="string">immaterial physical anatomical
       entity</value>
  </own_slot_value>
 <own_slot_value>
   <slot_reference>Non-English_name</slot_reference>
            <value value_type="string">immaterielles körperliches
       anatomisches Wesen</value>
 </own_slot_value>
 <own_slot_value>
   <slot_reference>Preferred_name</slot_reference>
<value value_type="string">immaterial anatomical entity</value>
 </own_slot_value>
 <own_slot_value>
   <slot_reference>ORIG_Preferred_Name</slot_reference>
 <value value_type="string">immaterial anatomical  entity</value>
 </own_slot_value>
 <own_slot_value>
   <slot_reference>Definition</slot_reference>
   <value value_type="string">Physical anatomical entity  which is
       a three-dimensional space, surface, line or point associated
       with a material anatomical entity. Examples: body space,
       surface of heart, costal margin, apex of right lung, anterior
compartment of right arm.</value>
  </own_slot_value>
 <own_slot_value>
   <slot_reference>Is_A</slot_reference>
   <value value_type="class">RID13441</value>
  </own_slot_value>
 <own_slot_value>
   <slot_reference>Has_Subtype</slot_reference>
   <value value_type="class">RID13221</value>
   <value value_type="class">RID13250</value>
```

```
    <value value_type="class">RID13291</value>
    <value value_type="class">RID13307</value>
    <value value_type="class">RID15845</value>
    <value value_type="class">RID13217</value>
  </own_slot_value>
 <own_slot_value>
    <slot_reference>:ROLE</slot_reference>
    <value value_type="string">Concrete</value>
  </own_slot_value>
 <superclass>RID13441</superclass>
</class>
```

Figure 2: Representation of a class in the RadLex ontology

In the RadLex case depicted in Figure 2, the established use of the *label* element in XML and OWL standards, as depicted in Figure 1, is performed by the relation element *preferred name*. The canonical medical term "immaterial anatomical entity" is additionally encoded in a non-intuitive manner, using the element name value_type, because it is not distinctive from references to other attributes of the class (e.g. to relations such as has_subtype).

The next example shows an entry in the XBRL taxonomy. Here, the element *label* is also used in a non-standard way, since it denotes in fact a class, while its attribute xlink:label is equivalent to the *Class ID* in Figure 1 (which equals to *class name* in RadLex). In case the standard use of *label* would be applied, the label's value would be the term Participating interests and shares in associated enterprises - Uncalled amounts - Movements during the period. These kinds of expressions are encountered in balances of annual company reports.

```
<label
xlink:label="ParticipatingInterestsSharesAssociatedEnterprisesUncalledAm
ountsMovements_lab"
xlink:type="resource"
xlink:role="http://www.xbrl.org/2003/role/documentation" xml:lang="en">
Participating interests and shares in associated enterprises - Uncalled
amounts - Movements during the period
</label>
```

Figure 3: Representation of a concept in the XBRL taxonomy

We showed the examples from RadLex and XBRL not only because of their non-standard solutions for the inclusion of information in natural language, but also in order to illustrate the cognitive and linguistic complexity of terms that needs to be formalized for optimal structuring of semantic resources used in knowledge representation systems.

In particular, an important NLP task, semantic annotation, often takes place on the basis of the mapping of labels of ontologies or taxonomies and natural language expressions occurring in documents that are external to structured resources. Overlapping in style with the natural language expressions in Figures 1-3, external documents contain lexical units featuring similar morphological and syntactic properties. Our study elaborates on these closely related issues as well.

## 2. Linguistic complexity of terms in ontologies

Linguistic markup is typically added to textual documents to represent morphological, syntactic, etc. information. Given the complexity of natural language expressions realizing terms in ontological resources, we assume that mere lexical treatment of these is insufficient, and thus propose a incorporating a fully-fledged linguistic annotation of labelling terms. An example of the possible linguistic annotation of a term taken from the German version of RadLex „Skelettmuskel des medialen Oberschenkels" (*skeletal muscle of medial thigh* is listed below in an informal way.

1 Categorial Information for the whole term
   - „hasCat" => "NP",

2 Dependency Information for the whole term
   - "hasHead" => "Skelettmuskel",
   - "hasModifier" => "des medialen Oberschenkels",
   - "hasModifierType" => "PostModGen",

3 Recursive dependency Information
   - "hasModHead[2]" => "Oberschenkel",
   - "hasModMod[3]" => "medialen",

4 Recursive constituency and morpho-syntactic Information
   - "hasHeadPos" => "Noun",
   - "hasHeadCase"=> "Nominative|Accusative",
   - "hasHeadCompound" => "Skelett Muskel",
   - "hasHeadLemma" => "skelett muskel",
   - "hasModCat"[4] => „NP",
   - "hasModHeadLemma"=> "oberschenkel",
   - "hasModHeadPoS" => "Noun",
   - "hasModHeadCase" => „Gen",
   - "hasModModPoS" => "Adj",

Figure 4: List of possible linguistic annotation for an ontology label

In Figure 4 we can see the kind of linguistic objects we would use for annotating terms in ontology labels. In order to standardize the representation of this information, we will adopt the developing standards at ISO TC37/SC4, including the multi–layered annotation approach suggested there. Our main source of inspiration for the linguistic annotation is given in (Ide et al., 2006) and (Ide et al., 2007).

Given the potential complexity of such annotation, we shall not add the markup directly to the ontology classes, but rather suggest integrating the annotation in the ontology within a specific, separate representation layer. We draw on existing models for the integration of terminological and linguistic information in ontologies.

[2] head of the modifying phrase

[3] modifier within the modifying phrase

4 category of the modifying phrase

## 3. Improved models for the integration of terminological and lexical information in ontologies

Commonly used inclusion methods of terminological and lexical information in taxonomies or ontologies, as described in the Introduction, are not satisfying. There exist two representation proposals that are improving this situation. One is concentrating on the terminological aspect (Aussenac-Gilles et al., 2007; Raymonet et al., 2009) and one on the lexical aspect (Buitelaar et al., 2009). (Buitelaar et al., 2009) propose a model called *LexInfo* and suggest adding lexical, morpho-syntactic and chunking information to the labels of ontology classes. The authors design an OWL representation scheme for this set of linguistic information and its linking to ontology classes. LexInfo supports in this among others the ontology-based semantic annotation of text.

(Aussenac-Gilles et al., 2007) describe a model called *Terminae* and suggest having within ontologies two distinct, but interlinked high levels of classes: one for the hierarchy of concepts (and associated relations), and one for (a list of) terms that point to the concepts they denote. In this way the concept level world gets cleaner, and we can avoid for example the very cumbersome manner of encoding synonyms and other related terms as this is done in RadLex (see Figure 2): synonyms are now encoded within the terminology level of the ontology. An advantage of this approach lies in the fact that a subset of a terminology can more easily be identified and re-used in other (domain) ontologies. (Raymonet et al., 2009) give an example of the application of Terminae in the automotive domain. We note that in Terminae the lemma and part-of-speech information is encoded within the term classes.

We suggest the merging of LexInfo and Terminae, whereas we would apply the full model of LexInfo to each word in a term. In doing so we take lexical information completely out of the descriptions of both domain and term classes. We suggest thus to have three layers of description within the ontology, where a meta- class has three main subclasses describing domain-class, terminology and linguistic hierarchies. The linguistic layer is based on and extends LexInfo.

## 4. A model for the integration of conceptual, terminological and linguistic objects in ontologies (CTL)

Building on the Terminae model, we add a conceptual level in the ontology dealing with linguistic objects, which themselves are modeled on the base of the LexInfo model. The layer of linguistic objects is then pointing to the terminology only through the representation of the tokenized terms, and via this layer to the class hierarchy. All other linguistic information is to be considered as building an abstract object. It is clear that we do not want to include in the linguistic layer all the XML code (which is in fact representing feature structures), since this would introduce verbosity in the ontological description. We rather use the name of the actual feature structure (or its *type*), which is then linked to the token list of a term. We plan to register those names in the ISO data category registration infrastructure (called ISOcat)[5]. This linking mechanism is represented in an informal way below, taking again RadLex as our basis example:

```
Domain_Class:
        hasId: RID2694
        hasREL: Part_Of
        hasSuperclass: RID2660
Term_Class:
        hasId: Term:1767
        hasString: Skelettmuskel des medialen Oberschenkels
        hasTokens: [t1 Skelettmuskel] [t2 des] [t3 medialen] [t4
Oberschenkels]
        hasClass: Class:RID2694

Linguistic_Class:
        hasId: LO:14
        hasName: Ling:postNominalGentiveModification
    hasTerm: Term:1767_hasTokens[t1-t4]
Linguistic_Class:
        hasId: LO:215
        hasName: NP_Genitive
        hasTermTokens: Term:1767_has_Tokens[t2-t4]
Linguistic_Class:
        hasId: LO:213
        hasName: NOUN_Nominative
        hasTermTokens: Term:1767_Tokens[t1]
        …
```

The reader can see how the linguistic objects are pointing to the tokenized terms, and how the terms point then to the classes. On the basis of this model, we can obtain a matrix of linguistic objects, terms, and classes (including attributes and relations). This matrix can then deliver interesting insights on the use of natural language in knowledge representation systems. In the longer term, this can lead to proposal for a normalization of natural language expressions that fit best for building a terminology representing most adequately a formal representation of a domain.

## 5. Expected improvements for ontology population and ontology learning tasks

We expect our model to support improvements both in the ontology-based semantic annotation of textual documents and in the semi-automatic generation of ontologies from text. In the concrete example of applying RadLex for the semantic annotation of radiology reports, we notice that there are linguistic discrepancies between the terms encoded in the ontology and the way the concepts are expressed in the reports. So for example in the ontology we have the terms: *Axillärer Lymphknoten*, *Mediastinaler Lymphknoten* or *Hilärer Lymphknoten* . Those terms are implementing the feature structure: ADJ_modified_NP.

[5] see http://www.isocat.org/

In the radiology reports we find then text like: „Lymphknoten axillär, mediastinal und hilär“, with the following (simplified) linguistic structure:

[NOUN Lymphknoten] [ADVP [ADV axillär] [PUNCT ,] [ADV mediastinal] [CONJ und] [ADV hilär] ]

We can associate to this segment of text a feature structure named Noun_mod_by_Enum_Coord_Adverbs. In our actual work, we manage to unify the feature structures in the ontology and the feature structure in the text with the following algorithm:

Find in labels of the ontology the head noun corresponding to the head noun of the feature structure detected in the text:

– „Lymphknoten“ as head of Nominal phrases

Search for lemma of modifiers in the labels that correspond to the lemmas of the (adverbial) modifiers found in the text.

– hilär (lemma of adv) = hilär (lemma of adj „hilärer“), etc.

Distribute then the head noun into the feature structures of the coordinated ADVP, lacking such an head, and generate the semantic annotation for the textual segment.

Concerning the possibility to improve the generation/extension of ontologies from text, the linguistic analysis of terms associated with classes in existing ontologies can give fruitful insights and suggests that similar linguistic constructions in external text are providing for candidate for new classes in existing ontologies or as a the building blocks for new (domain) ontologies).

We additionally discovered the potential of our model for the task of ontology consistency checking. By heuristics, similar feature structures associated to terms should point to similar conceptual constructs in the ontology. In the RadLex Ontology (v2.0 for German), we see that postnominal gentive modification is very frequently associated with a “IS_A” relation between two concepts. But in the case of the two following terms: 1. "Ligamentum des Handgelenks" (*ligament of wrist joint*) and 2. "Ligamentum des Ellenbogengelenks" (*ligament of elbow joint*), we find that 1. is in a “Is_A” relation to Handgelenk" (*wrist joint*), whereas 2. is in a “Part_Of” relation to "Ellenbogengelenk" (*elbow joint*). But we have in 1. and 2.nearly identical linguistic objects, the only difference being in the first part of the compounds Handgelenks and Ellenbogengelenks (hand vs elbow). We notified the discrepancy in the naming of the relations to the domain expert, and he confirmed our findings: Both terms denote the same type of relation to their head nouns (“Handgelenks” and “Ellenbogengelenks” respectively).

## 6. Conclusion

We have presented a proposal for combining two models for the integration of terminological and lexical information in ontologies: *Terminae* and *LexInfo*. Our proposal, called CTL, implements a three layers representation model of class, terminology and linguistic objects, whereas the latter are no longer limited to lexical information but are covering the full range of linguistic phenomena, including constituency and dependency. We are currently working on formalizing our approach, taking into account also standardization work for linguistic annotation at ISO TC37/SC4. We also show that the approach benefits linguistic and semantic analysis of external documents that are often to be linked to semantic resources for semantic enrichment with concepts, and that new concepts can be extracted or inferred on the base of the linguistic and semantic analysis of the documents.

## 7. Acknowledgements

The research presented in this paper is partially funded by the European Commission in the context of the FP7 project MONNET - Multilingual Ontologies for Networked Knowledge, with grant agreement number 248458.

## 8. References

Buitelaar, P., Cimiano, P. Haase, P., Sintek, M. (2009). Towards linguistically grounded ontologies. In *Proceedings of the 6th European Semantic Web Conference.* Springer Berlin/Heidelberg, pp. 111--125.

Buitelaar, P., Declerck, T. (2003) Linguistic Annotation for the Semantic Web. In S. Handschuh & S. Staab (Eds.), *Annotation for the Semantic Web*. IOS Press, pp. 93--111.

Aussenac-Gilles, N., Szulman, S., Despres, S. (2008). 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.

Reymonet, A., Thomas, J., Aussenac-Gilles, N. (2009) Ontology based information retrieval: an application to automotive diagnosis. In *Proceedings of International Workshop on Principles of Diagnosis*.

Gruber, T.R (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, 43 (5-6):907 928.

Ide, N., and Romary, L. (2006). Representing linguistic corpora and their annotations. In *Proceedings of LREC2006*.

Ide, N., and Suderman, K. (2007). GrAF: A graph-based format for linguistic annotations. In *Proceedings of the ACL2007 Workshop on Linguistic Annotation Workshop*.