

Standardised Stand-off Linguistic Annotation of Ontology Labels

Thierry Declerck (DFKI, LT-Lab)

Piroska Lendvai (Uni Tilburg)

*And Questions to the speakers by
Thierry Declerck*



Motivation

- Ontology/Taxonomy classes are often introduced with natural language expressions (terms), that reflect the „meaning“ of the class in an human readable way or one way this class can be realised in textual documents
 - Those terms are ideally encoded within a „label“ feature associated with the ontology class. A (modified) ex. From Radlex
 - Class = RID1382
 - Label: lang=„en“ string=„right inferior pulmonary ligament“
 - Label: lang=„de“ string=„Rechtes inferiores Lungenligament“
- In most of the cases no linguistic information is associated with the natural language expressions used in the labels
 - But models for combining lexical/linguistic and ontological information within (domain) ontologies have been recently proposed (see for example the LexInfo approach by Buitelaar et al, ESWC 2009)



An approach for the Linguistic Annotation of Ontology Labels

- Stand-off annotation of ontology labels with linguistic information (but searching for compatibility with LexInfo or other models)
 - No overload of the ontology with additional linguistic information, which can be quite substantial in case the content of the label is a larger phrase or even a full sentence
 - The linguistic information externally associated with the labels can be organised in an ontological structure, which is mirroring the original domain ontology.



Possible Benefits of the Approach

- Possibility to link constitutive parts – linguistically speaking – of the string content of the label, which might have themselves no corresponding concepts in the actual ontology, to possibly related classes in other ontologies, or to suggest the addition of such classes in the actual ontology
 - If the noun/noun compound „pulmonary ligament“ is being used in other labels, internal or external to the actual ontology, one can investigate if the corresponding classes might be related
 - The compound noun „lymph node“ is occurring as the linguistic head of noun phrases in many labels in the Radlex ontology, but never as the only term of the label of a class
 - Suggest a new class, and/or;
 - Search if this class is not part of another (to be related) ontology..



Possible Benefits of the Approach

- Linguistic annotation of ontology/taxonomy labels can ease the corresponding semantic annotation of text
 - One can better model the variation of surface realisation of the concept labels: Not searching for a 100% match but for compatible linguistic annotation (terms in the class label and in the text sharing for example the same lemmas, even in different word orders)
- Use of a standardised linguistic annotation strategy for the labels of all available ontologies.



Relation to ISO TC37/SC4

- Apply the ISO strategy for linguistic annotation (Linguistic Annotation Framework, LAF), including feature structures (in XML) and a multi-layered stand-off annotation approach
 - Point from the feature structure representing an multilayered annotation graph to the label of the ontology class
- Mapping of the tagset of the linguistic annotation to the morpho-syntactic and syntactic data categories defined in ISOcat, or direct use of the data categories of ISO



An entry in the RadLex Ontology

```
<class>
  <name>RID2694</name>
  <type>radlex_metaclass</type>
  <own_slot_value>
    <slot_reference>Preferred_Name</slot_reference>
    <value value_type="string">Skelettmuskel des medialen Oberschenkels</value>
  </own_slot_value>
  <own_slot_value>
    <slot_reference>Part_Of</slot_reference>
    <value value_type="class">RID2660</value>
  </own_slot_value>
  <own_slot_value>
    <slot_reference>:ROLE</slot_reference>
    <value value_type="string">Concrete</value>
  </own_slot_value>
  <superclass>RID2660</superclass>
</class>
```

- 1) In blue font: the class label in German language (English: *skeletal muscle of medial thigh*)
- 2) The super-class RID2660 is the class marked by the label „Oberschenkel“ (English: *thigh*)



Linguistic Annotation that can be associated to the label

- Label: [Skelettmuskel des medialen Oberschenkels](#)
- Categorial Information for the whole term
 - „hasCat" => "NP",
- Dependency Information for the whole term
 - "hasHead" => "Skelettmuskel",
 - "hasModifier" => "des medialen Oberschenkels",
 - "hasModifierType" => "PostModGen",
- Recursive dependency Information
 - "hasModHead" => "Oberschenkel", # head of the modifying phrase
 - "hasModMod" = "medialen", # modifier within the modifying phrase
- Recursive constituency and morpho-syntactic Information
 - "hasHeadPos" => "Noun",
 - "hasHeadCase" => "Nominative|Accusative",
 - "hasHeadCompound" => "Skelett Muskel",
 - "hasHeadLemma" => "skelett muskel",
 - "hasModCat" => „NP", # cat of the modifying phrase
 - "hasModHeadCompound" => "Ober Schenkels",
 - "hasModHeadLemma" => "ober schenkel",
 - "hasModHeadPoS" => "Noun",
 - "hasModHeadCase" => „Gen",
 - "hasModModPoS" => "Adj",



Another example in Radlex

```
<class>
  <name>RID3600</name>
  <type>radlex_metaclass</type>
  <own_slot_value>
    <slot_reference>Preferred_Name</slot_reference>
    <value value_type="string">Arthritis durch Cholesterol-Kristallablagerung</value>
  </own_slot_value>
  <own_slot_value>
    <slot_reference>Is_A</slot_reference>
    <value value_type="class">RID3565</value>
  </own_slot_value>
  <own_slot_value>
    <slot_reference>:ROLE</slot_reference>
    <value value_type="string">Concrete</value>
  </own_slot_value>
  <superclass>RID3565</superclass>
</class>
```

- 1) **In blue font:** the class label in German language (English: *arthritis of cholesterol crystal deposition*)
- 2) The super-class RID2660 is the class marked by the label „Kristallarthropathie “ (English: *crystal induced arthritis*)



Linguistic Annotation that can be associated to the second label

- Label: **Arthritis durch Cholesterol-Kristallablagerung**
- Categorical Information for the whole term
 - „hasCat" => "NP",
- Dependency Information for the whole term
 - "hasHead" => „ArthritisI",
 - "hasModifier" => "durch Cholesterol-Kristallablagerung",
 - "hasModifierType" => "PostModPrep-Durch",
- Recursive dependency Information
 - „hasModHead“ => „durch“,
 - "hasModCompHead" => "Cholesterol-Kristallablagerung", # „Comp“ stays for Complement
- Recursive constituency and morpho-syntactic Information
 - "hasHeadPos" => "Noun",
 - "hasHeadCase" => "Nominative",
 - "hasModHead" => " Cholesterol-Kristallablagerung",
 - "hasModCat" => „PP",
 - "hasModCompHeadCompound" => " Cholesterol - Kristall Ablagerung",
 - "hasModCompHeadLemma" => "cholesterol - kristall ablagerung",
 - "hasModCompHeadPoS" => "Noun",
 - "hasModCompHeadCase" => „Accusative",



Strategy for mapping ontology labels to surface realisation

- German Labels for „Lymphknoten“ (*Lymph Node*)
 - Pre-nominal adjectival modification
 - Abdomineller Lymphknoten
 - Aortocavaler Lymphknoten
 - Retrocavaler Lymphknoten
 - Paraaortaler Lymphknoten
 - Mesenterialer Lymphknoten
 - Axillärer Lymphknoten
 - Hilärer Lymphknoten
 - non-ATS thorakaler Lymphknoten
 - (Noun/NounCompounds)
 - Leberpfortenlymphknoten
 - Lymphknotendissektion
 - Post-nominal genitive modification
 - Lymphknoten des hinteren Beckenkamms
 - Lymphknoten der kleinen Kurvatur/Ligamentum gastrohepaticum
- Realisation in the German corpus
 - Zahlreiche Lymphknoten mesenterial, paraaortal und im kleinen Becken
 - Keine pathologisch vergrößerten Lymphknoten axillär, mediastinal und hilär.
 - Zahlreiche, bis zu 0,9 cm messende Lymphknoten im mesenterialen Fettgewebe, retroperitoneal sowie im kleinen Becken
 - Keine Lymphknotenfiliae (the word „filiae“ is not in the ontology)
 - Keine vergrößerten Lymphknoten bds axillär, mediastinal und hilär.
 - Nachweis der kleinen Lymphknoten beidseits in der Halsgefäßnervenscheide ohne Größenzunahme.



Feature Structure for Pre-nominal adjectival Modification

- For labels like „Axillärer Lymphknoten“

```
„hasCat" => "NP",  
  
"hasHead" => „Lymphknoten",  
"hasHeadPoS" => „N",  
"hasHeadCompound" =>  
    „Lymph Knote",  
"hasHeadLemma" => „knoten",  
  
"hasModifier" => " Axillärer",  
"hasModifierPoS" => " Adj",  
"hasModifierLemma" => " axillär",
```

PreModAdjectival



Parsing Problem for certain Surface Realisation in Corpus

- For example, expressions like
„Lymphknoten axillär, mediastinal und
hilär.“
 - Incomplete sentence
 - Sequence „Noun + Coord Adverbs“ normally
not a chunk

Feature Structures for Surface Realisation in Corpus

- „Lymphknoten axillär, mediastinal und hilär.“

Noun ["hasString" => „Lymphknoten“,
"hasCompound" => „Lymph Knoten“,
"hasLemma" => „knoten“,]

EnumAdvCoord [„hasCoord" => „und“,
„hasAdvStrings“ =>
<axillär, mediastinal, hilär>
„hasAdvLemmas“ =>
<axillär, mediastinal, hilär>]

Need for a rule for grouping/unifying the coordinated adverbs and the noun, and for transforming the generated structure in 3 equivalents PreModAdjectival feature structures that can unify with the corresponding labels in the ontology.



Questions addressed to the Speakers of the Workshop

FLaReNet-Silt Workshop, Pisa
19.09.2009



Questions to Paul

- Principled way of adding language information to ontology labels?
 - Who does this?
 - Terminology as the base?
 - Danger of overload of ontologies?
- Syntactic information is needed, not only lexicon (labels can include full sentences). Extensions of LexInfo (beyond subcat information)
- Differences between linguistic ontologies (like „Gold“) and Lexinfo?
- Can the work on linguistically analysing labels of ontologies lead to a reorganization of the ontologies?
- I like the idea of pointing from LMF to Ontologies. Extend to other ISO Standards (libraries of feature structures, data categories)
 - Mechanisms for doing this?
- Uses of ontology labels for improving lexicon work?



Questions to Chu Ren

- Ontologies are not necessarily based on the search for causes. ISA and Part-Of relations as basic ontological information
 - Description of anatomy
 - Multimedia ontologies
- I like the conclusion: ontologies to be grounded in human experiences, but there is still a need for formalization. It seems to be that there is a gap between knowledge conveyed in natural and in artificial languages. Is LexInfo a good possible „bridge“?
- Natural language for human communication, Ontologies for machine communication?



Questions to Massimo

- Question in fact to all: How does it look like in the human mind. Conceptual knowledge different from language knowledge?
- Is knowledge organized around words?
 - Language generation seems to organized from other linguistic units as from word-boundaries. Is the lexicon really the basic building blok of conceptual knowledge
- Wordnet is subjective (probably) but ontologies probably too (even if modelled on the base of scientific consensus)



Questions to Kyrill

- Talk addresses the applicability of ontologies in the annotation of text
 - Is grammar always a representation of relations between lexicon items?
 - Many constructions in text are using word with a not-lexicalized meaning.
 - Where is the middle in the middle Layer Ontology. User-based evaluation?
 - The presented middle layer ontology seems very similar to upper-level ontology. Introducing a new layer is not introducing additional complexity?
 - For manual or automatic annotation?



Questions to Piek

- Talk has good questions to other contributors
 - From WordNet to „PhraseNet“ or even „SentenceNet“? In ontologies we deal not only with words but more complex terms.
 - Relation between wordNet and syntactic structures?
 - Vocabularies used in ontologies are de facto disambiguated. Simplification of vocabularies. We should not aim at representing vocabularies in ontologies but at a linguistic representation of the vocabularies used in Ontologies (agreement on this)

