Ontologies and multilingualism

G. Falquet, J. Guyot

March 24, 2015

1.1 Theoretical connections

- the lexicon of each natural language provides a conceptualization of the world.
- to avoid circularity some primitive or basic concepts are not defined within the ontology, a kind of "common" definition is used (the definition in usual dictionaries)
  - in CityGML: Water Bodies sub-model refers to water body classes such as lake, river, ditch, bayou that are not defined in this ontology
- the only way to anchor an ontology in a real domain (meaningful identifiers)

1.2 Practical connections

- domain specific information sources such as dictionaries, reference texts, legal texts, and many other types of documents are expressed in some natural language.
- any usable ontology should be consistent with this terminology and the conceptualization it induces.
- when creating new concepts, their name is constructed from (combinations of) existing linguistic forms.

1.3 Multilingualism

- An ontology may serve as a common reference for an international community
- In ontology driven user interfaces
- In semantic indexing of large multilingual text corpuses
- The information sources required to build an ontology may exist only in some languages
  - the development process must take into account several languages
- — When an ontology needs to be localized

1.4 Ontologies and point of views

- it can be useful to consider each point of view as a different language.
- domain specialists have developed specific vocabularies to exchange information in a precise and non-ambiguous way.
- when a human activity spans several domains, the involved actors may experience communication problems due to this diversity of vocabularies.

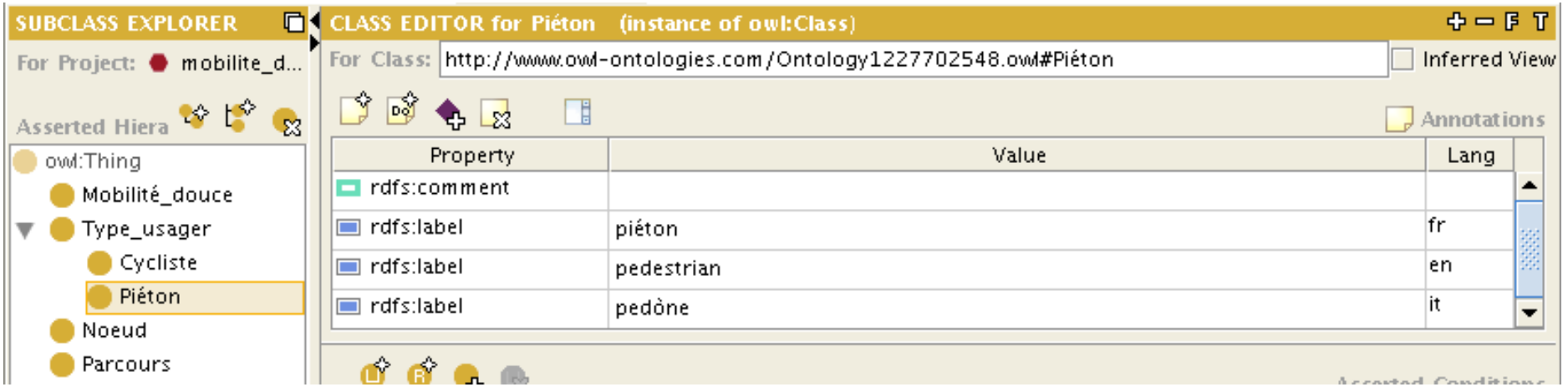
Example

In urbanism related activities: urban engineers, architects, politicians, transportation engineers, or citizen organizations participate in decision processes.

- each one of these groups possesses its own vocabulary and conceptualization of the world,
- improving communication between them cannot rely on the development of a single "monolingual" ontology.
- a situation that is similar to multilingualism or multiculturalism.

2.1 The basic concept-centric approach

- attach linguistic forms to a (neutral) concept



Problems

1. The lexical information attached to a concept is limited to a character string, so there is no possibility to define relationships between lexical forms or to build sophisticated lexical structures
2. The lexical forms (labels) are strictly equivalent, i.e. each label of a concept is supposed to designate exactly this concept. This can be true for very specialized domains but that is rarely the case for wider domains.
3. In RDFS or OWL a string can be qualified with a language name, but not with a region, social class, activity, ...

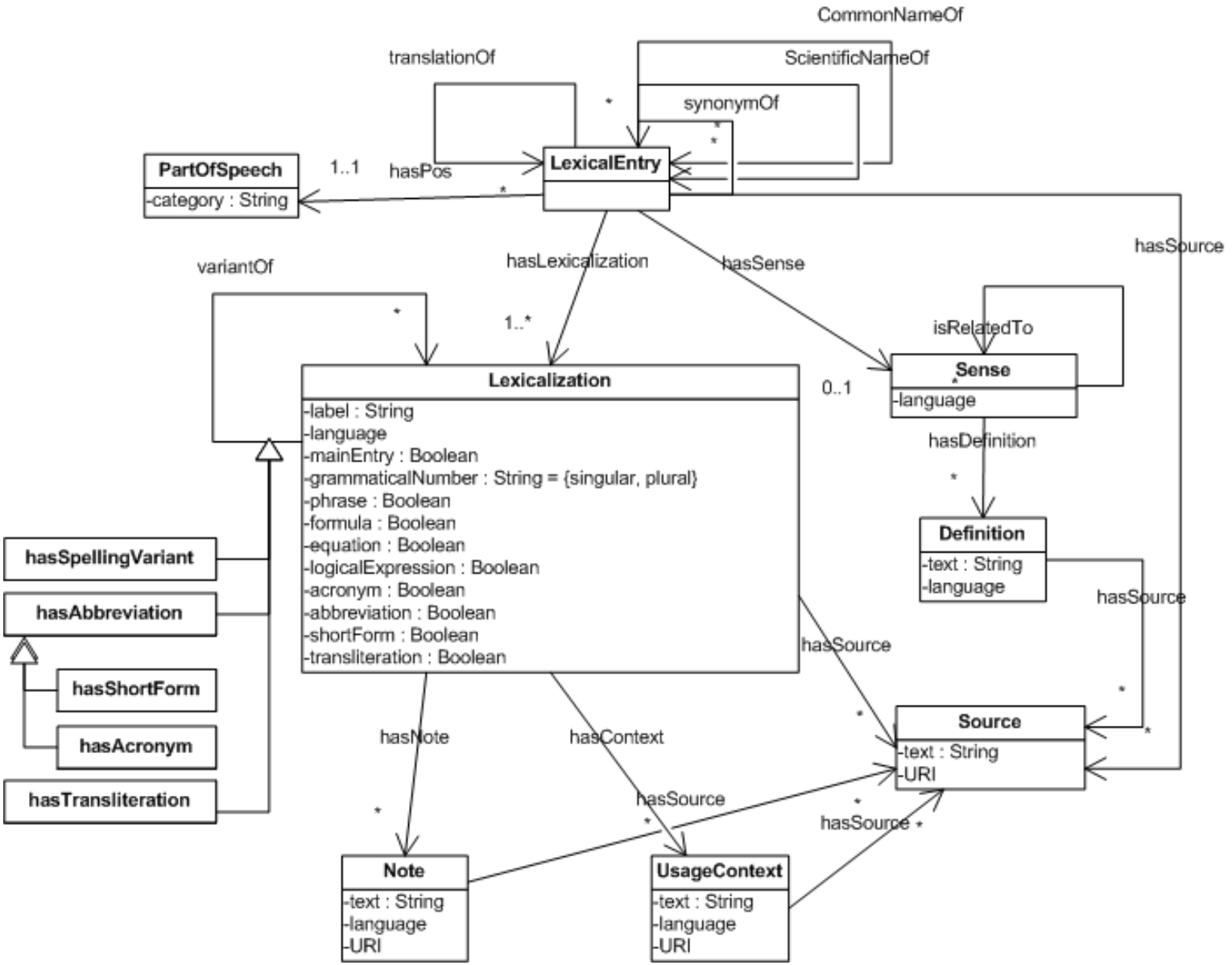
2.2 Concept-centric with structured linguistic elements

1. **conceptual level:** intended to represent the concepts (or meanings) and their definitions. ontological elements
2. terminological level: made of terms, which are associations between concepts and lexical forms.
   - the chemical term acid associates the linguistic form "acid" to the concept defined as a *compound which donates a hydrogen ion to another compound in a reaction.*
   - terminological relationships such as antonymy.
3. lexical level: forms, which are character strings used in written language.
   - lexical relationships such as plural or other inflectional variants. .

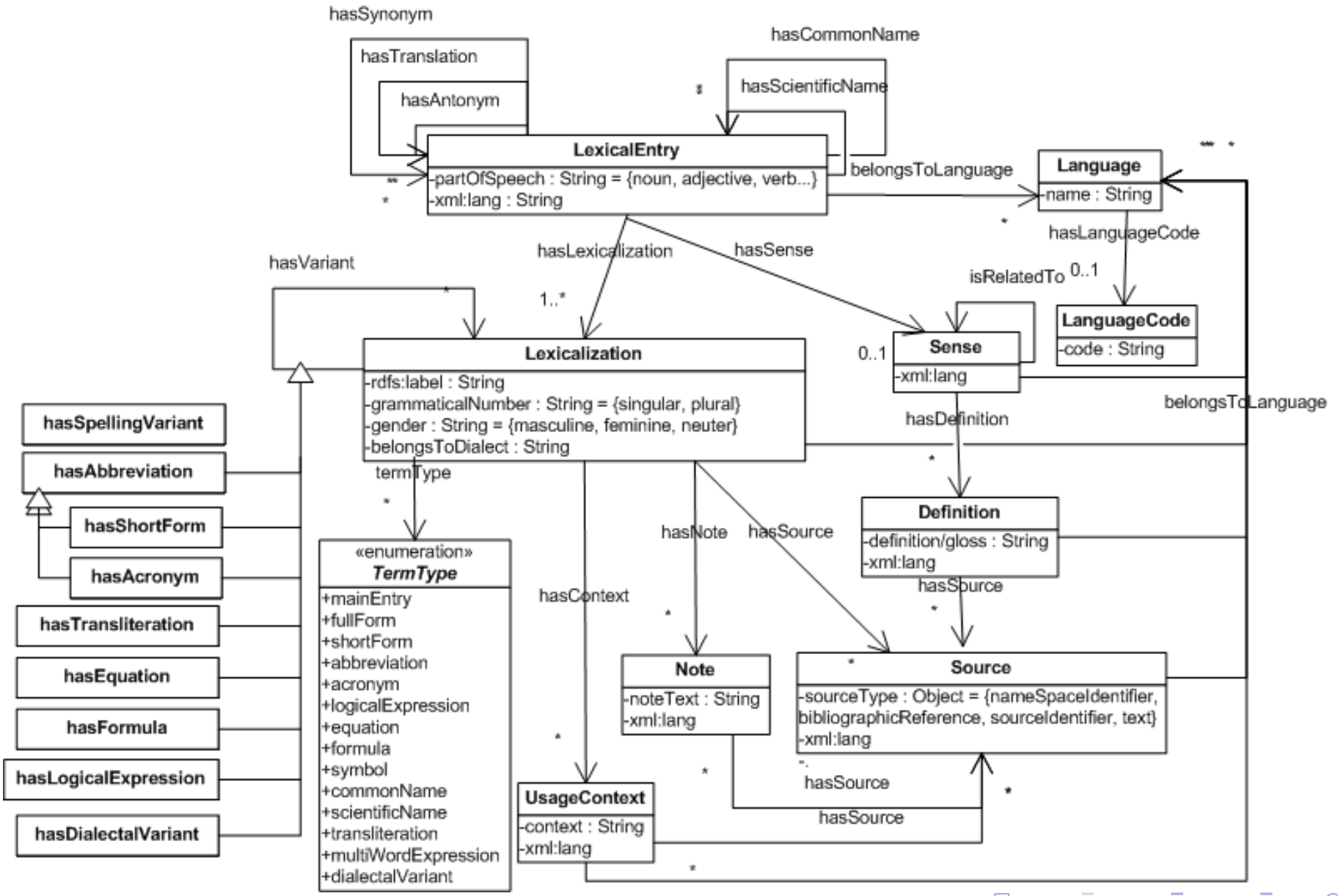
Another model (Aguado de Cea, 2009)

1. Lexical layer: characters and symbols that make up the syntax (ASCII encoding, UNICODE, etc.)
2. Syntactic layer: structure of characters and symbols, i.e., the grammar. It embraces different representation languages (e.g. RDF(S), OWL, etc.)
3. Representation paradigm layer: paradigm followed in the representation of the ontology (frames, semantic networks, DL, etc.) that allows a certain way of expressing and structuring knowledge
4. Terminological layer: terms or labels selected to name ontology elements
5. Conceptual layer: related to conceptualization decisions, such as granularity, expressiveness, perspective, etc.
6. Pragmatic layer: final layout of the model according to user's needs

The LIR model 1 [3]

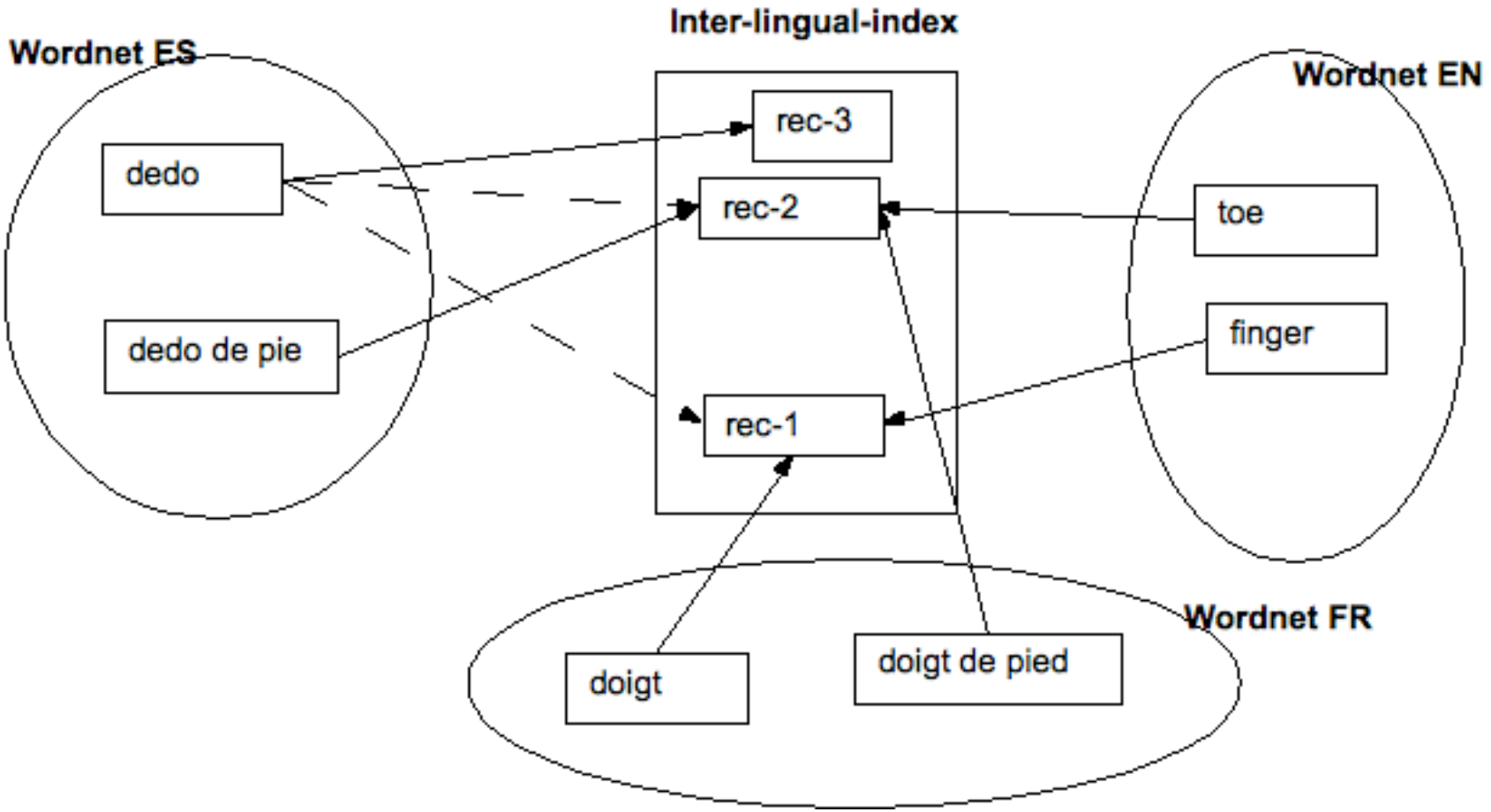


LIR model 2 [2]

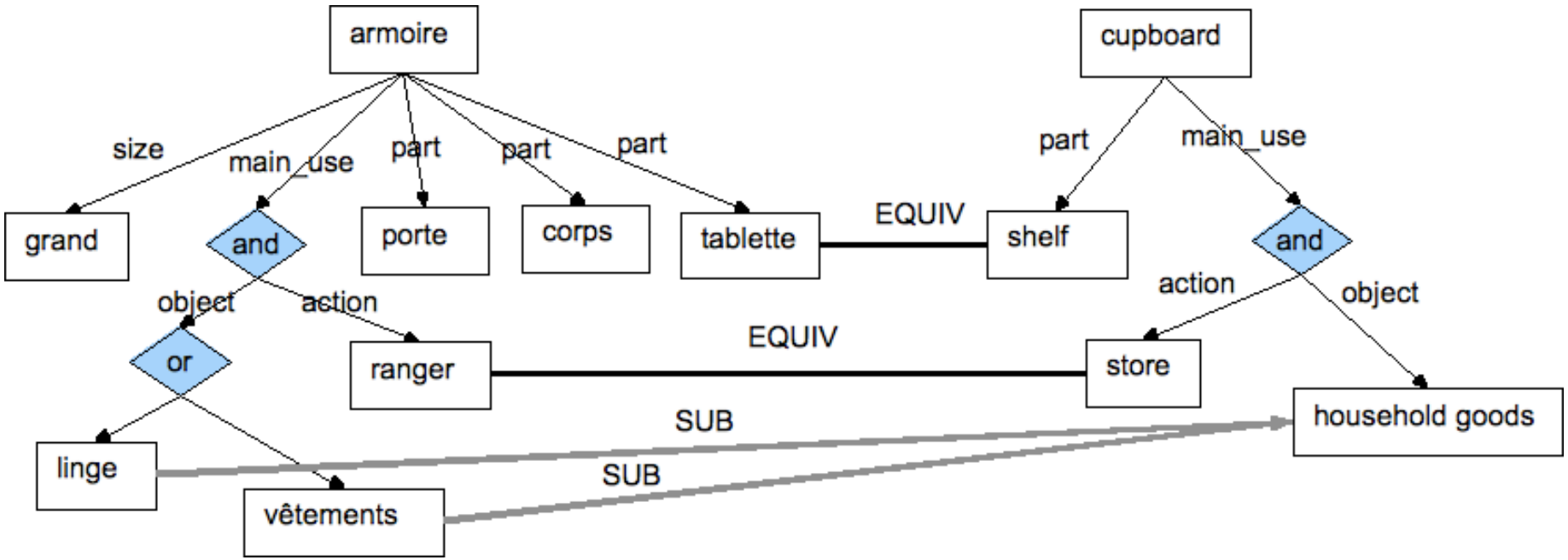

hasSynonym
hasTranslation
hasAntonym
hasCommonName
hasScientificName
LexicalEntry
-partOfSpeech : String = {noun, adjective, verb...}
-xml:lang : String
belongsToLanguage
Language
-name : String
hasLanguageCode
hasVariant
hasLexicalization
hasSense
1..*
isRelatedTo
0..1
LanguageCode
-code : String
Lexicalization
-rdfs:label : String
-grammaticalNumber : String = {singular, plural}
-gender : String = {masculine, feminine, neuter}
-belongsToDialect : String
0..1
Sense
-xml:lang
belongsToLanguage
hasDefinition
termType
hasSpellingVariant
hasAbbreviation
hasShortForm
hasAcronym
hasTransliteration
hasEquation
hasFormula
hasLogicalExpression
hasDialectalVariant
«enumeration»
TermType
+mainEntry
+fullForm
+shortForm
+abbreviation
+acronym
+logicalExpression
+equation
+formula
+symbol
+commonName
+scientificName
+transliteration
+multiWordExpression
+dialectalVariant
hasContext
hasNote
hasSource
Definition
-definition/gloss : String
-xml:lang
hasSource
Note
-noteText : String
-xml:lang
Source
-sourceType : Object = {nameSpaceIdentifier, bibliographicReference, sourceIdentifier, text}
-xml:lang
hasSource
UsageContext
-context : String
-xml:lang
hasSource

2.3 Interconnection and alignment approach

- maintain separate ontologies, one for each language,
- an interconnection structure links equivalent concepts



Applications: finding translations



Application: Multilingual information retrieval

1. the information need is expressed by a set of keywords or key phrases or sen-tences in the user's own languages
2. the document corpus contains documents written in different natural languages

The basic idea is to replace each term that appears in a document or in the query by a concept identifier. Then it becoms possible to apply mono-lingual IR techniques, simply replacing the word space by the concept identifier space.

Tasks in MLIR

Disambiguation. find the concept that corresponds to an ambiguous lexical form

Reasoning. documents that do not match the query at the keyword level are nevertheless relevant.

Interactive search.

faceted search, proposes to build the user query by navigating within (subsets of) the domain ontology.

Application: Semantic Annotation of Documents

A semantic annotation, in its simplest form, is a list of concepts belonging to a domain ontology.
A more precise kind of annotation consists in semantic graphs, for instance RDF graphs. In this case the graph nodes correspond to individuals that are concept instances and the labeled edges represent semantic relations between these individuals.

Ontologies for semantic annotation

Terminologically rich and multilingual ontologies play a key role to enable semantic annotation.

1. They serve as references for labeling the graph nodes (with concept identifiers) and the graph edges (with relation identifiers).
2. Automatically annotating large collections of documents requires natural language processing tools (in particular parsers) to recognize the lexical forms corresponding to concepts and concept instances. These tools must be provided with adequate lexical information.
3. Natural language processing tools can take advantage of ontological knowledge to solve syntax analysis problems. For instance, ambiguous sentences may be disambiguated if some domain knowledge is available.

References


Guadalupe Aguado de Cea, Mauricio Espinoza, Asunción Gómez-Pérez, Margherita Sini. (2009). Multilingual and Localization Support for Ontologies. Retrieved February 17, 2009, from http://www.neon-project.org/web-content/images/Publications/neon_2008_d242.pdf


E. Montiel-Ponsoda, G. Aguado de Cea, A. Gómez-Pérez, and W. Peters. Modelling Multilinguality in ontologies. In Proceedings of the Coling08 Conference, in Manchester, UK. August 2008.


W. Peters, E. Montiel-Ponsoda, G. Aguado de Cea. Localizing Ontologies in OWL. In Proceedings of the OntoLex07 Workshop, co-located at the ISWC 2007 Conference in Busan, South Corea. 2007.