Lexical Markup Framework: ISO-24613

Gil Francopoulo (INRIA-Loria, France),
Monica Monachini (CNR-ILC,Italy)

LIRICS project

LMF as ISO project

- Work started in Summer 2003 by a new work item proposal issued by the US delegation
- Fall 2003: the French delegation issued a technical proposition for a data model dedicated to NLP lexicons
- Beginning of 2004: ISO-TC37/SC4 decided to form a common ISO project (ISO-24613) with:
  - Nicoletta Calzolari (IT) as convenior
  - two editors:
    Gil Francopoulo (FR)
    Monte George (US)

History & roadmap

- **In 3 years and a half, 13 versions has been written, dispatched (to the National delegations nominated experts), commented and discussed in various ISO technical meetings**
- **+ papers in LREC-2006 + COLING-2004&2006**

- **Situation today:**
  - LMF document is a « committee draft » document (60 pages)
  - In March 2007, the ND allow us to obtain DIS status, provided that we include a couple of comments: we are now currently preparing this DIS version
- **=> target IS (= published standard) in 2008**

Method & motivation

- Try to learn from the past: Eagles, Multext, EDR etc.
- Study current famous lexicons (see « Extended examples of lexicons with LMF » on http://lirics.loria.fr+document area)

- Try to sum up « best practices » of lexicon definition & management
- Work to reach a consensual ISO standard on NLP lexicons

- **Our motivation here today**
- **A) present where we are**
- **B) AND collect your comments**

It's a work on progress. The model is more or less stable: at least, stable enough to be presented and discussed.

Scope

- Range of lexicons, LMF is intended for.
  => MRD + NLP lexicons
  => all MRD and all NLP applications
  => all languages
  => small and large scale lexicons
  => simple and complex lexicons
  => monolingual, bilingual, multilingual

Requirements

- **Multiple orthographies**
- **Morphology**
  -Repr. explicitly all inflected forms
  -Repr. in intension the inflected forms
- **Easily associate spoken form & written form**
- **Repr. complex agglutinating compound words like in German**
- **Repr. fixed, semi-fixed and flexible MWE**
- **Repr. complex syntactic constructions that are mapped onto a semantic representation (as in Eagles)**
- **Allow a semantic organization based on SynSets (like in WordNet) or on semantic predicates (like in FrameNet)**
- **Repr. large scale multilingual resources based on interlingual pivots or on transfer linking**

- **LMF does not address the following topics**
  - general sentence grammar of a language
  - world knowledge representation

General principle

- LMF is a structural data model expressed by a set of Unified Modeling Language (UML) packages.
- LMF is a high level specification based on constants that are defined in other standards

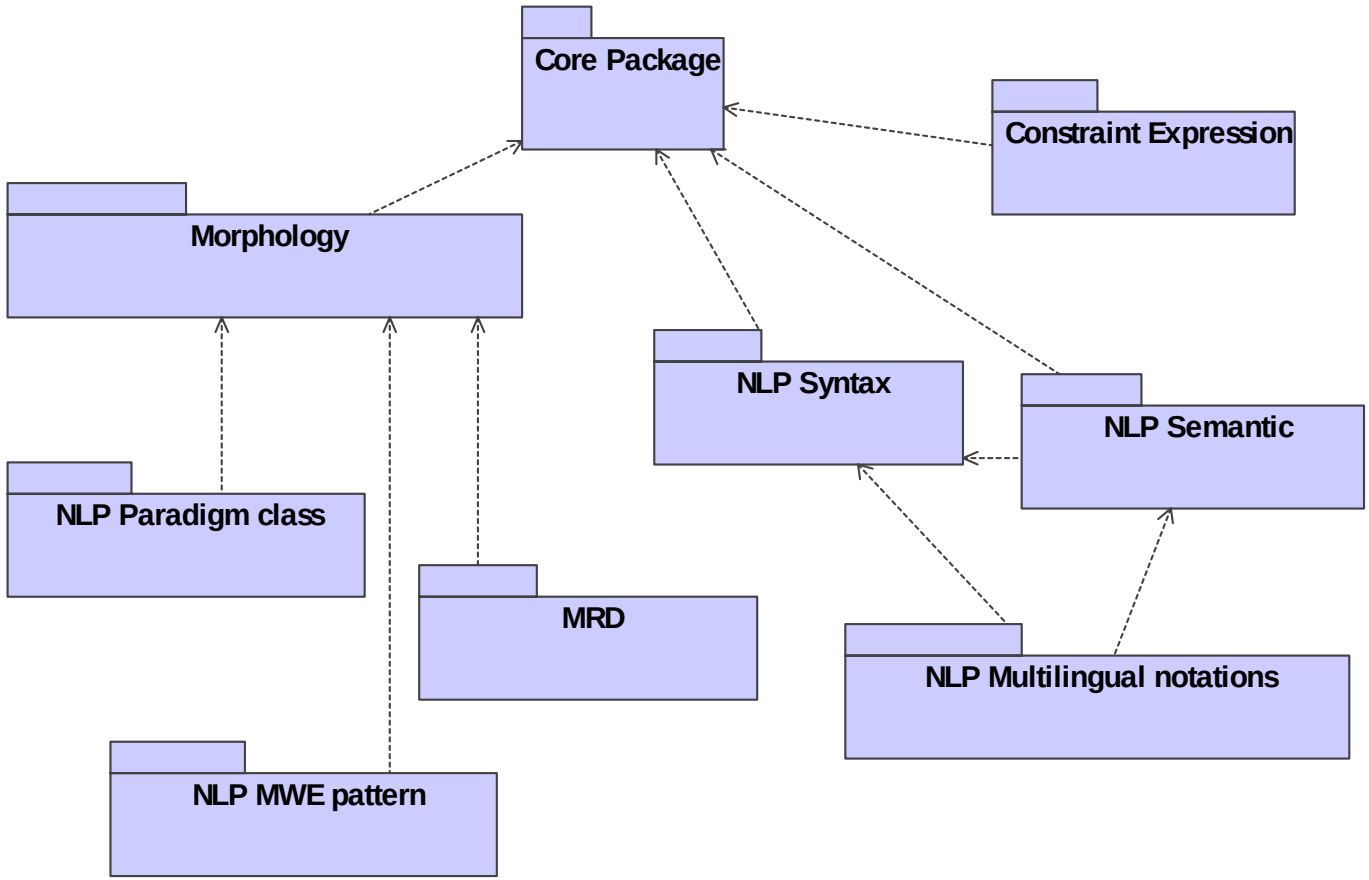
- Each package contains classes
- Each class is specified by:
  - a name
  - an English text describing its usage
  - an UML specification for linking with other classes
- Each class is to be adorned by a set of attribute/value pairs.
- But the attributes are not defined in the LMF specification. Only a list of examples is given. The attributes are to be taken from the data category registry (see next slide).
- The values are either constants or free strings.

General principle (cont.)

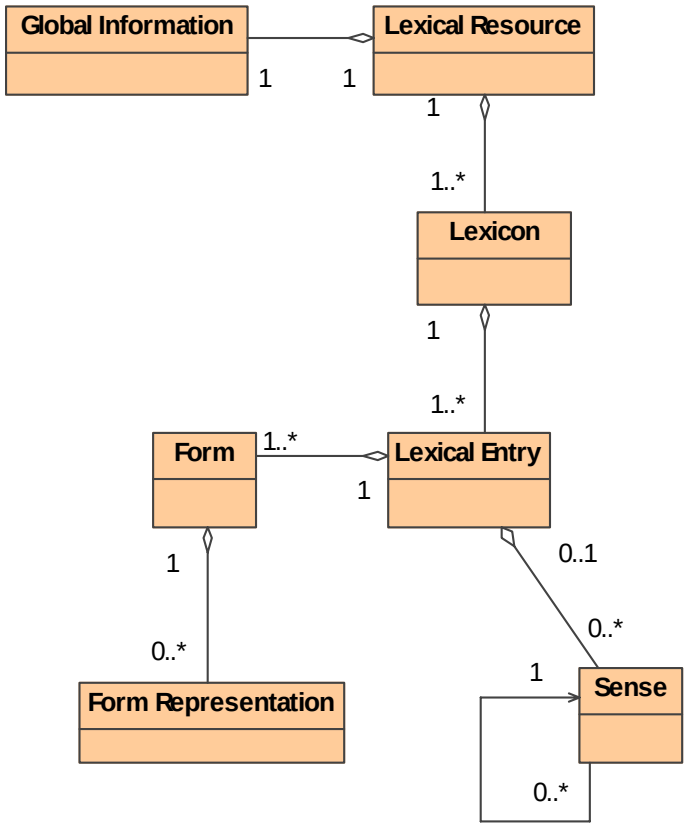
- The free string must conform to Unicode (**ISO/IEC 10646)**
- The constants are to be taken from other standards:
  - language codes **(ISO-639 or IETF BCP-47)**
  - script codes **(ISO-15924)**
  - country codes **(ISO-3166)**
  - dates **(ISO-8601)**
  - data category registry **(rev ISO-12620)** = work in progress to define linguistic constants like /part of speech/, /feminine/ or /transitive/
- The version and name of each of these standards is specified in a class called Global Information
- ISO context= LMF is just one member of a family of standards that are on the way to be defined within TC37/SC4 and all these standards share this particular principle for a good interoperabilty

LMF structural data model

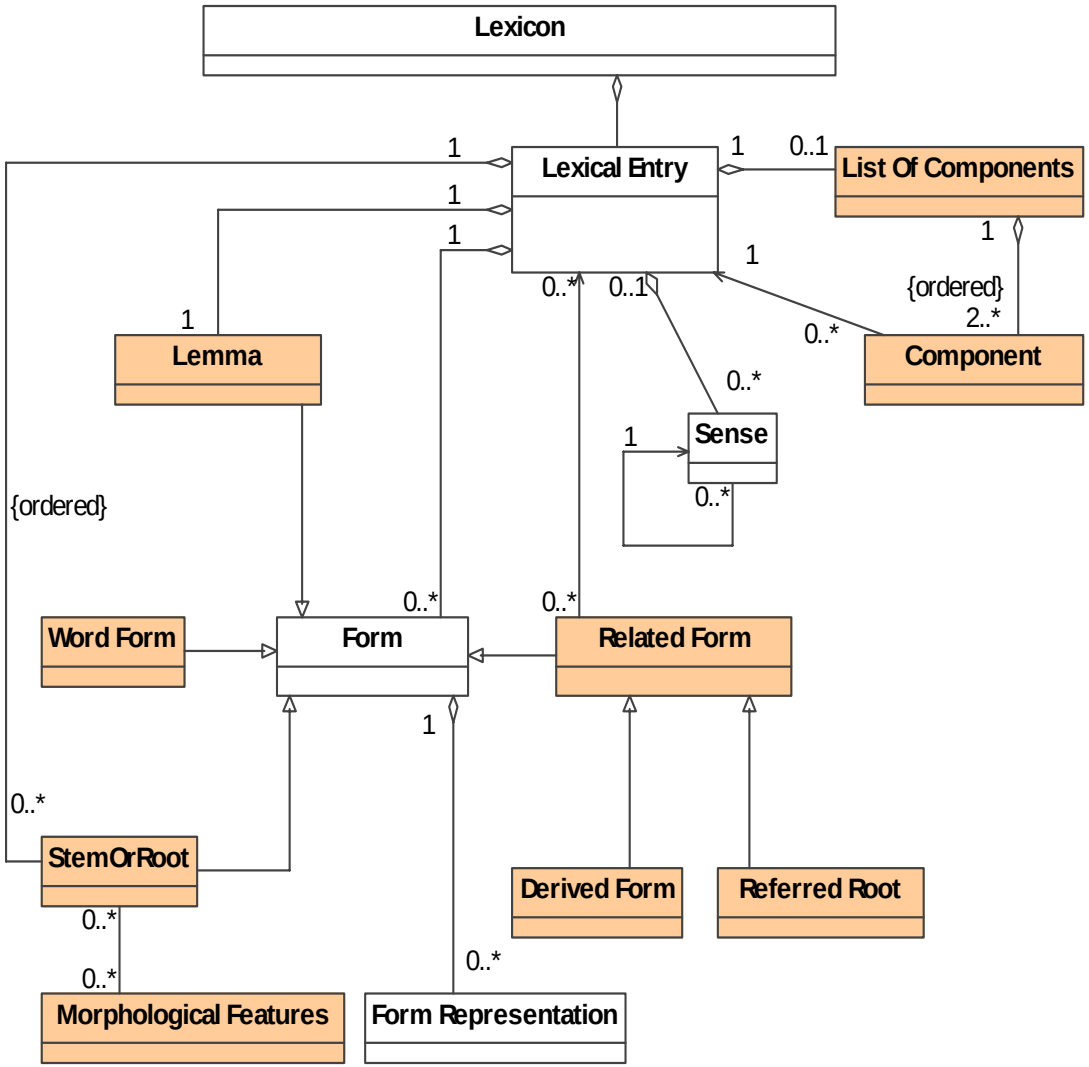
- One core package and 8 packages for extensions



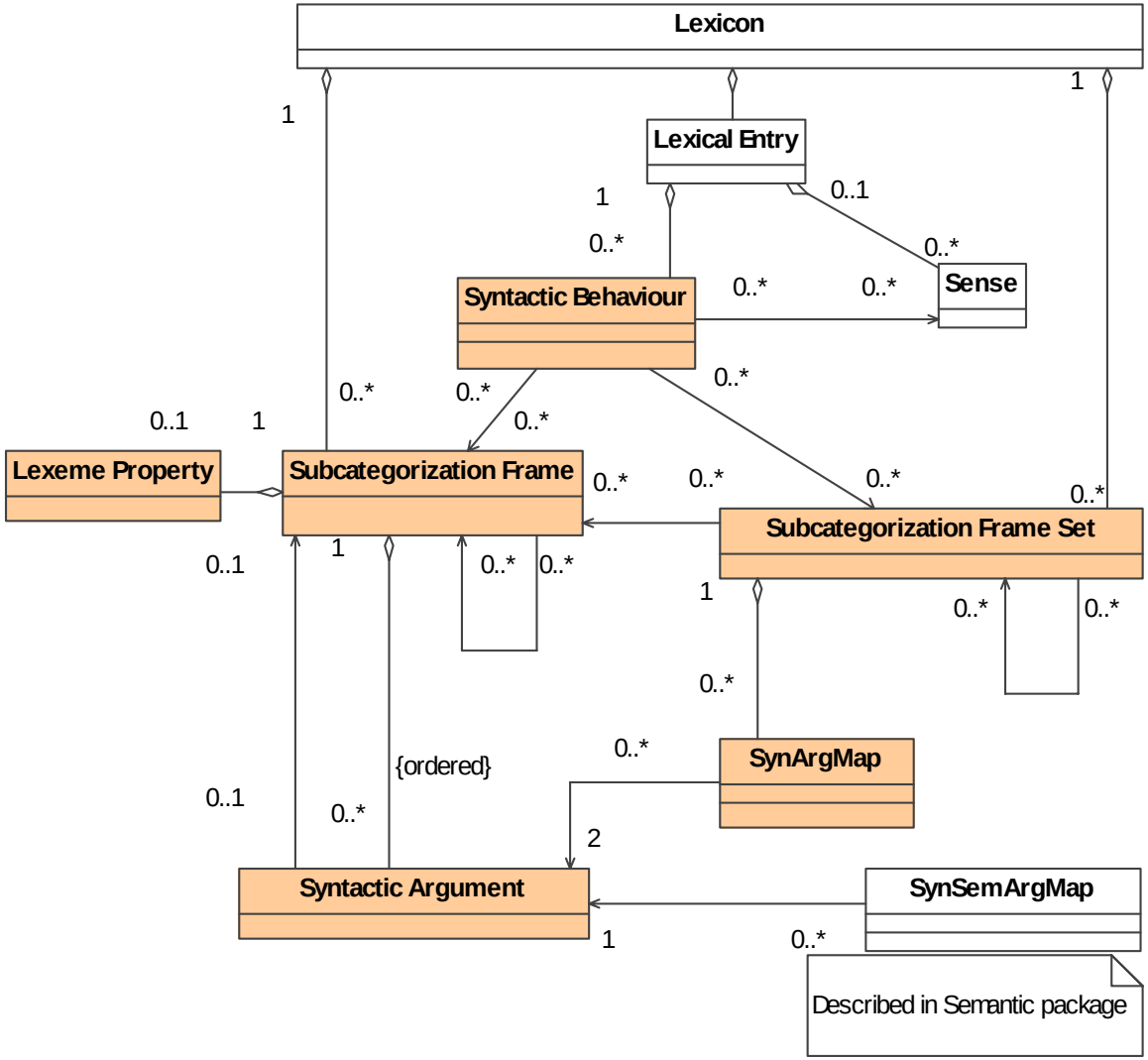
data model: core package



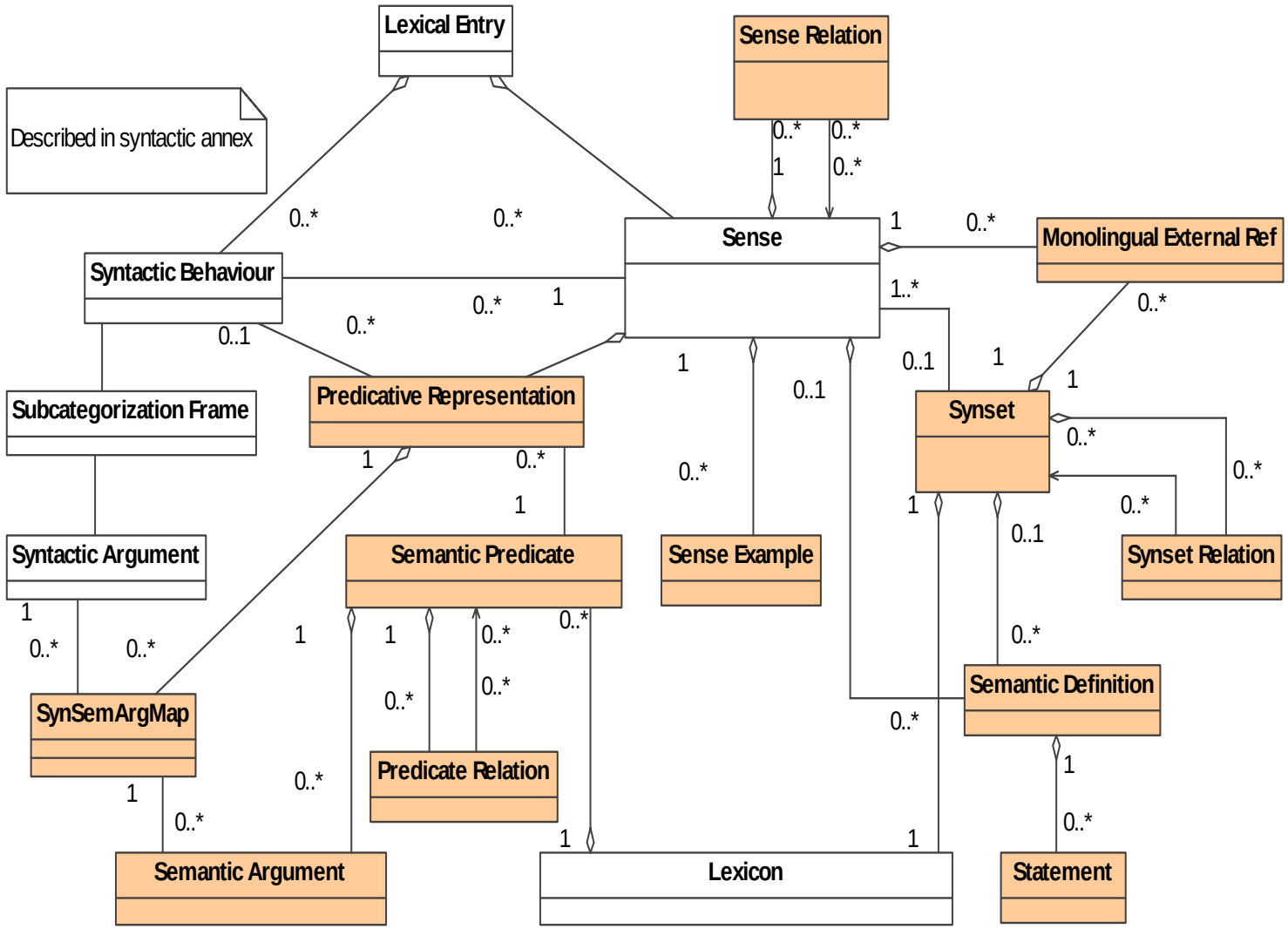
data model: Morphology



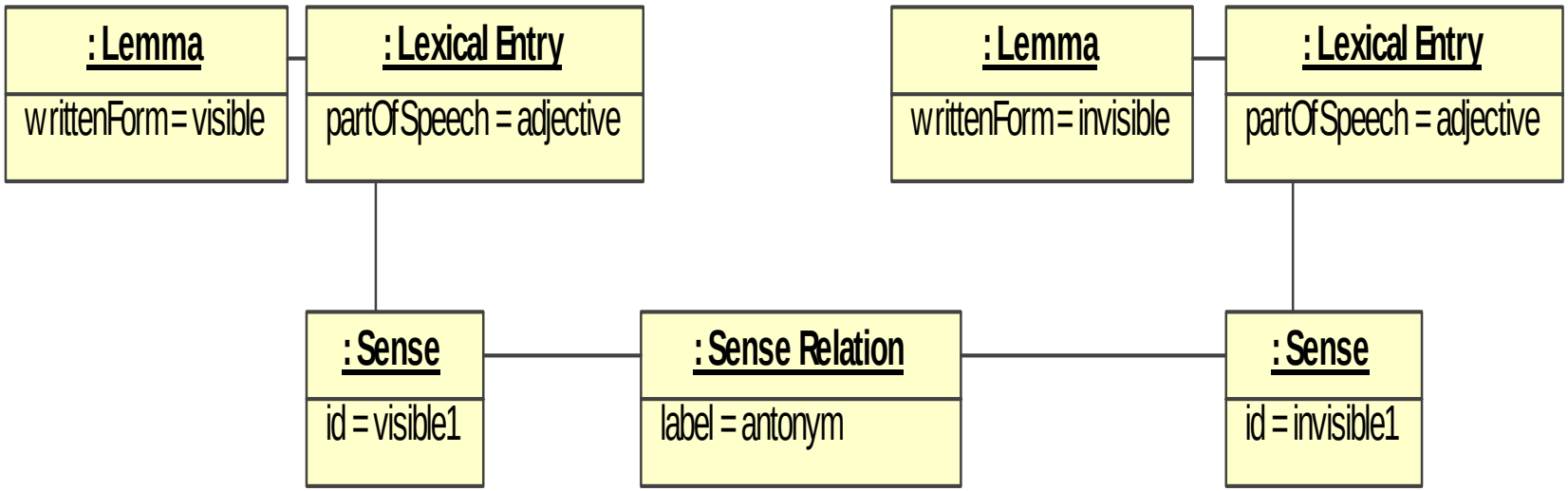
data model: Syntax



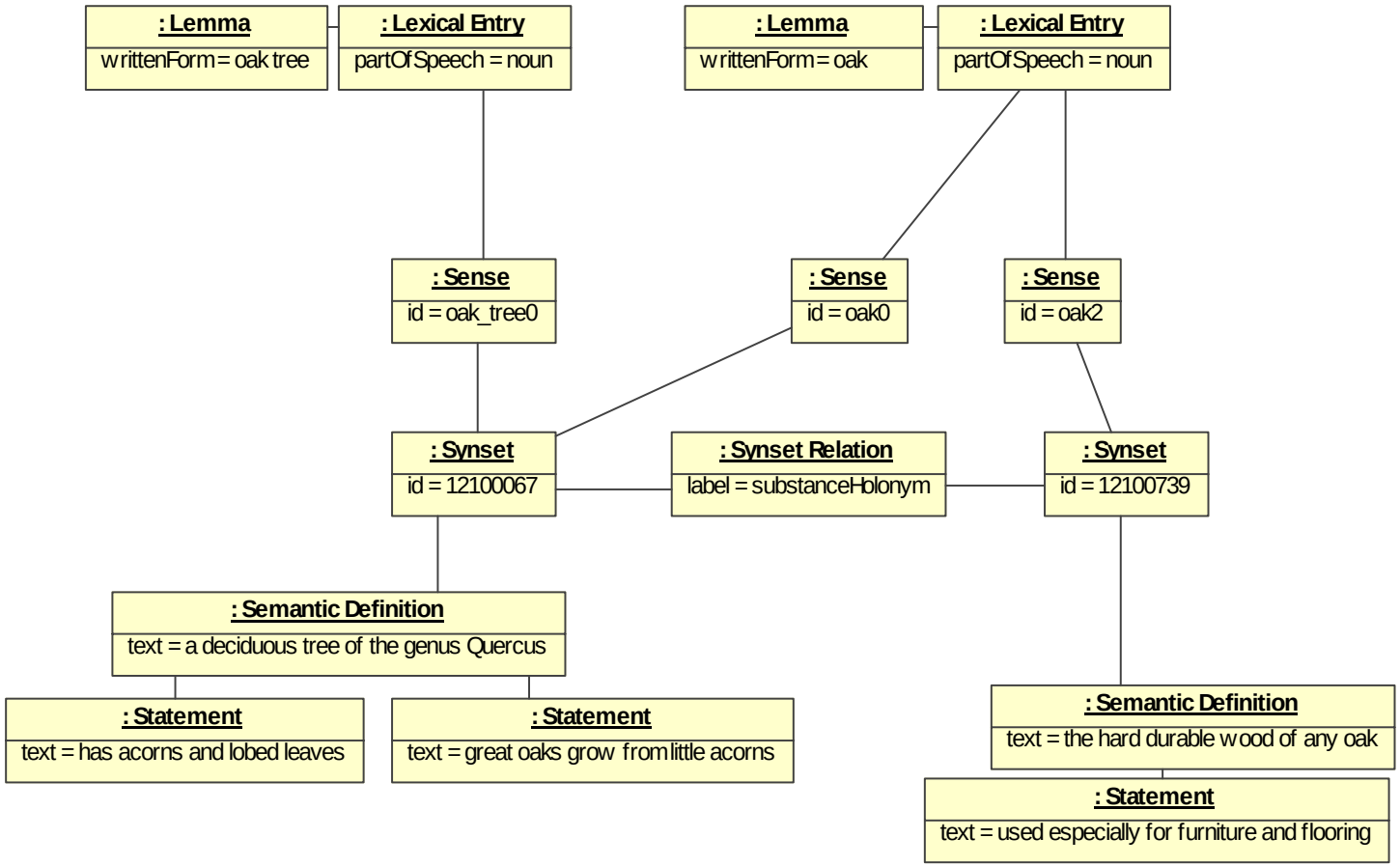
data model: Semantics



Semantics: example#1



Semantics: example#2 (WordNet 2.1)

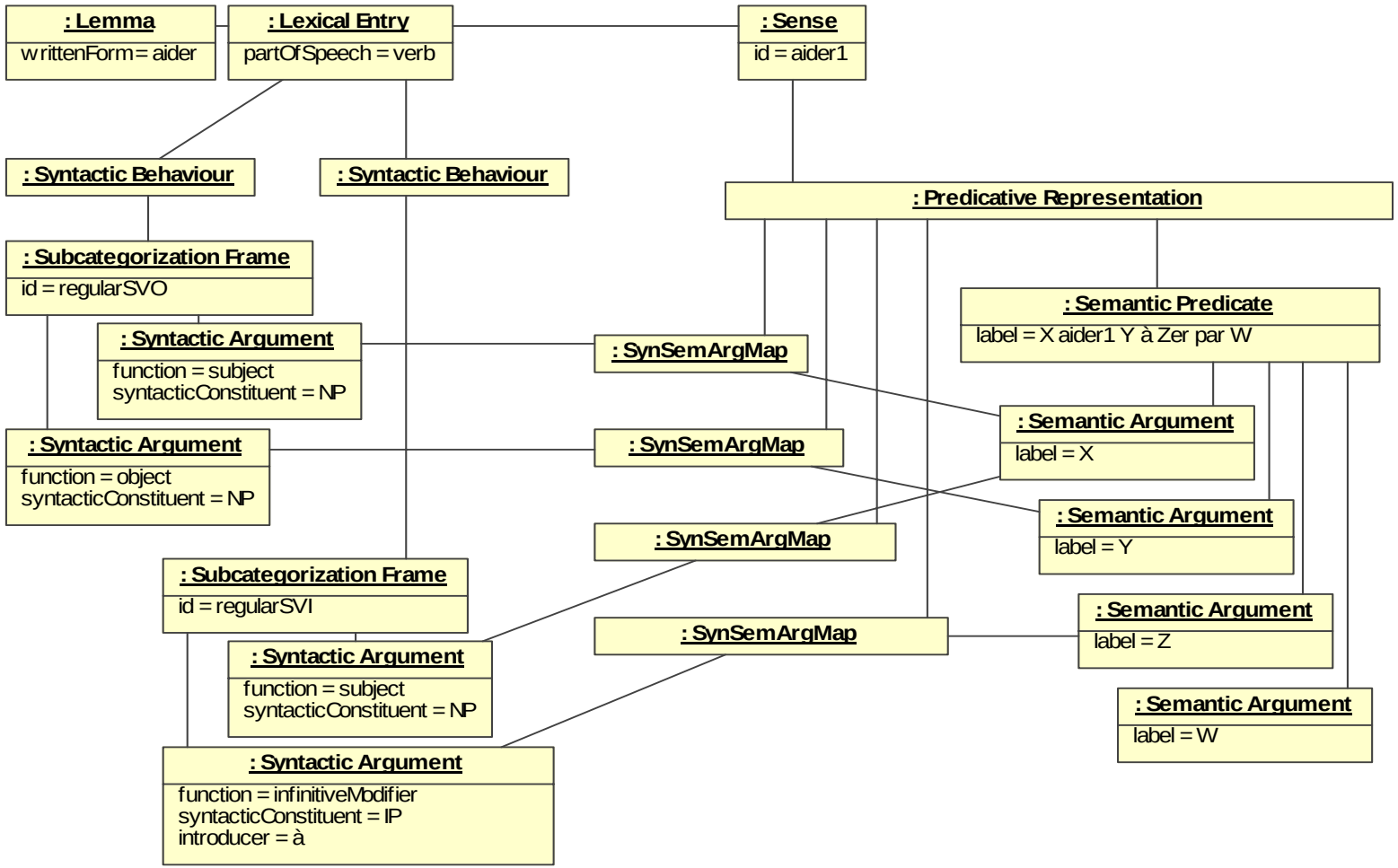


The same data can be expressed by the XML fragment (DTD in annex)

```
<LexicalEntry>
   <DC att="partOfSpeech" val="noun"/>
   <Lemma>
      <DC att="writtenForm" val="oak tree"/>
   </Lemma>
   <Sense id="oak_tree0" synset="12100067"/>
</LexicalEntry>
<LexicalEntry>
   <DC att="partOfSpeech" val="noun"/>
   <Lemma>
      <DC att=writtenForm" val="oak"/>
   </Lemma>
   <Sense id="oak0" synset="12100067"/>
   <Sense id="oak2" synset="12100739"/>
</LexicalEntry>
<Synset id="12100067">
   <SemanticDefinition>
      <DC att="text" val="a deciduous tree of the genus Quercus"/>
      <Statement>
         <DC att="text" val="has acorns and lobed leaves"/>
      </Statement>
      <Statement>
         <DC att="text" val="great oaks grow from little acorns"/>
      </Statement>
   </SemanticDefinition>
   <SynsetRelation targets="12100739"
      <DC att="label" val="substanceHolonym"/>
   </SynsetRelation>
</Synset>

...
```

Semantics: example#3 (DEC)

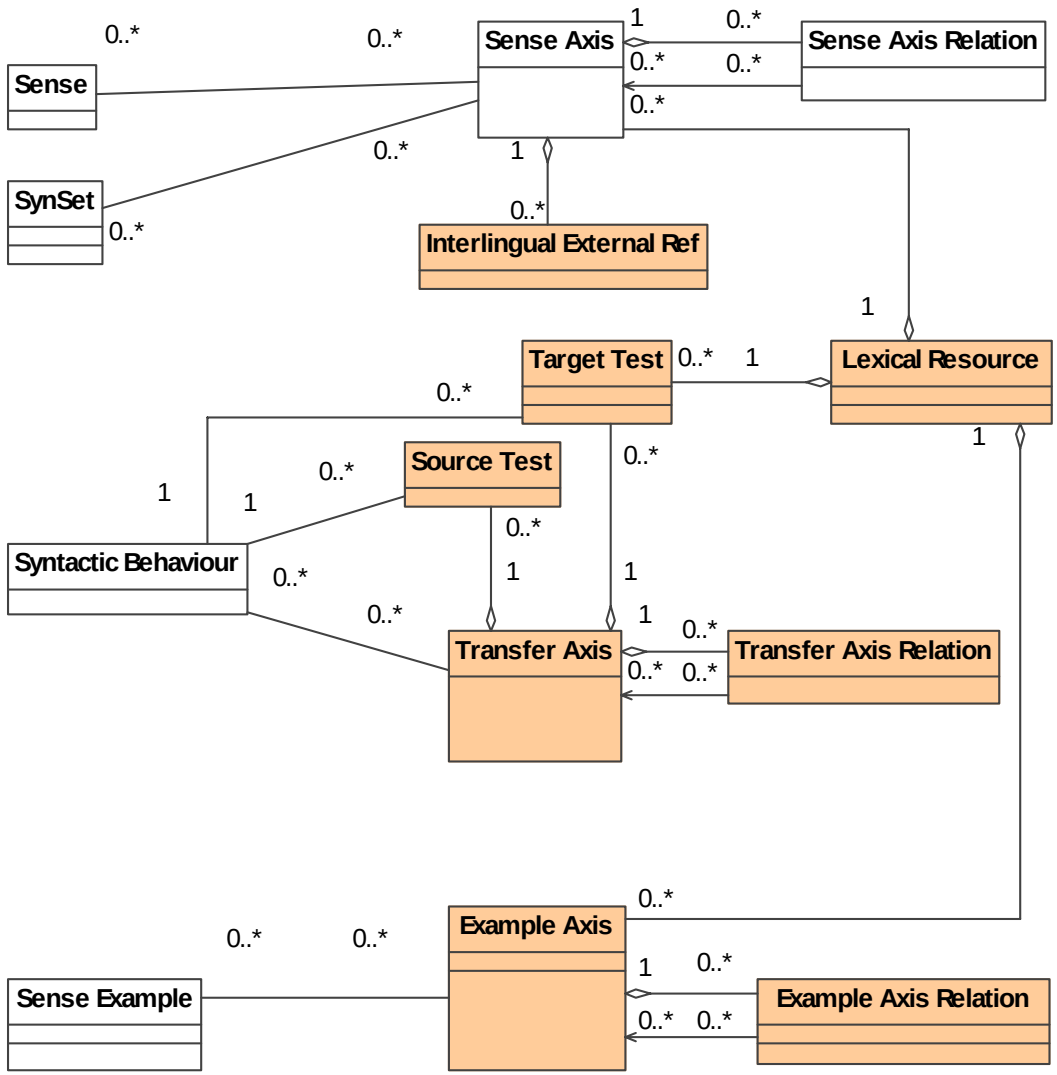


Package for multilingual notations

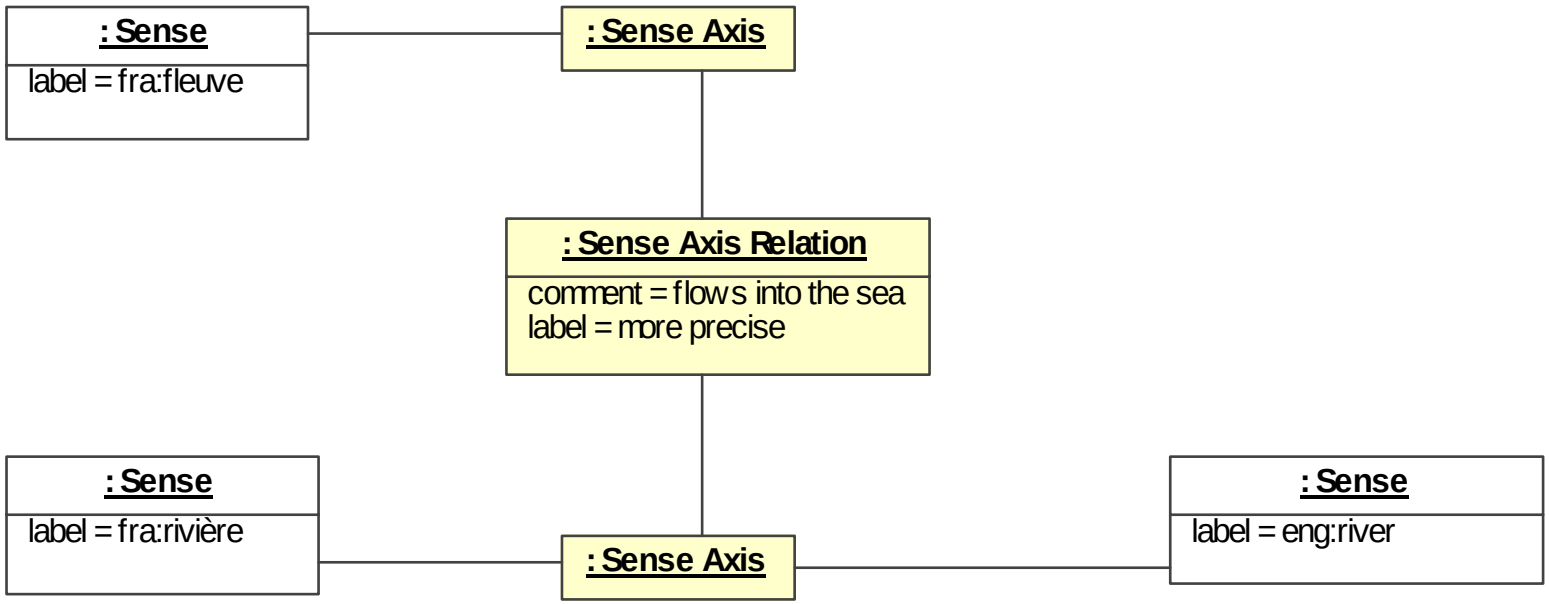
- For both interlingual pivots and transfer approach:
  - Sense+Synset (of different languages) may be linked by a SenseAxis
  - SyntacticBehavior (of different languages) may be linked by a TransferAxis

- Possibility to share or to duplicate Axis
- Possibility to add sourceTest or targetTest
- Possibility to link Examples (from different languages)

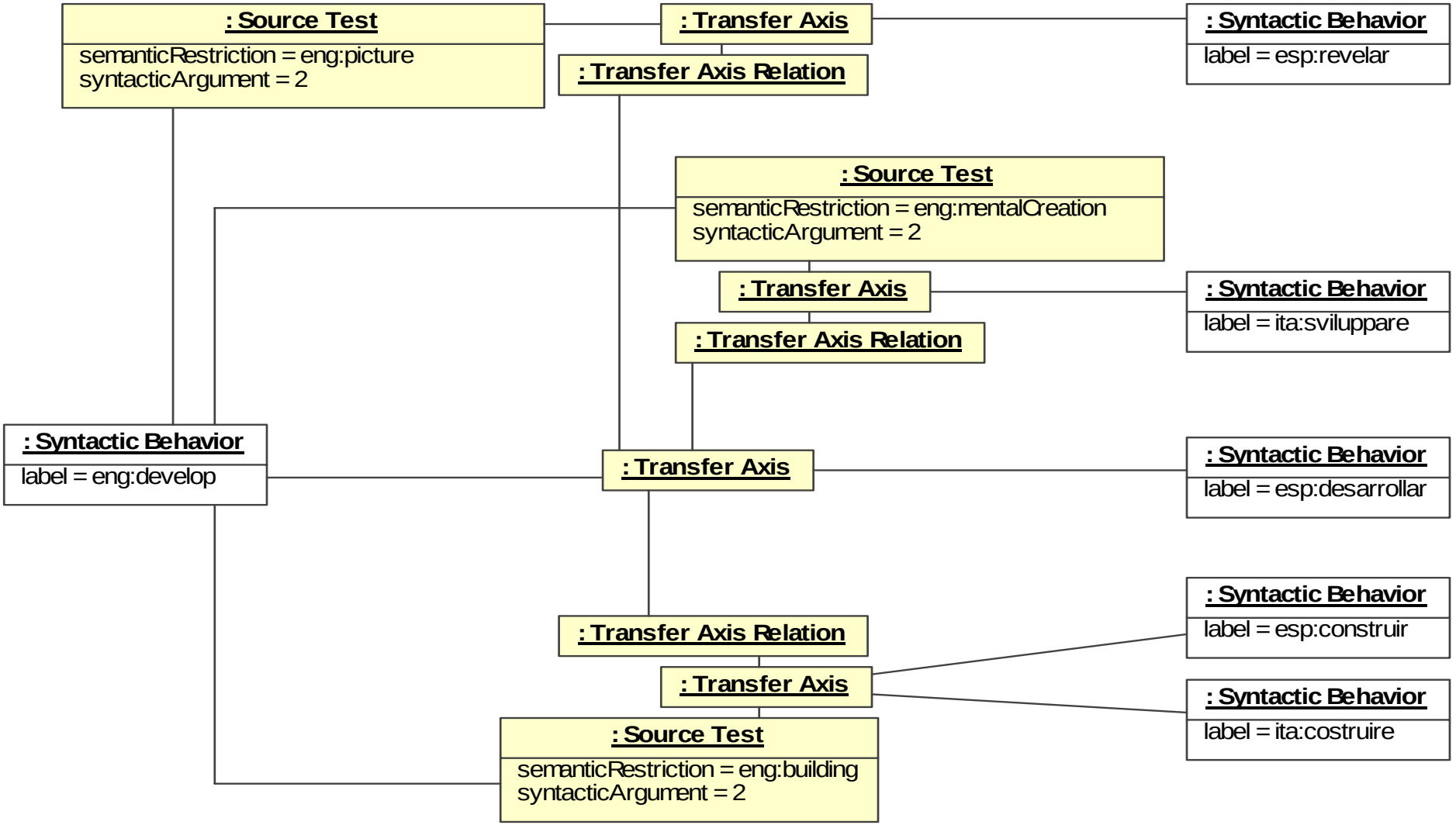
Data model: Multilingual notations



Multilingual notation example#1



Multilingual notation example#2



Connection with external systems like ontologies

- It's not the purpose of the semantic and the multilingual packages to provide a complex knowledge organization system
- LMF focus is NLP lexicons as required by user needs expressed through the channels of the National Delegations
- But we must provide to our users a clear linking with these external systems

Differences

- A semantic node in LMF is a data structure representing the meaning of a word in a particular language
- A node in a knowledge representation system is a data structure representing an elementary piece of what ‘exists’
- What ‘exists’ can be examined by separating issues of concept definition (ontology) and facts (concrete or imaginary facts), but from an LMF perspective, we stop where the meaning of a word stops
- Ontologies and fact data bases are considered as external systems

Provided mechanisms

- The mechanism cannot be a naive attribute adornment because the cardinality is one to many: intermediate classes must be designed for this purpose
- The connection is provided by two classes MonolingualExternalRef and MultilingualExternalRef
- These classes are adorned by /externalSystem/ and /externalReference, resp. to the name of the external system and to the relevant node in this given system

Last slide

- **Acknowledgements:**
  The work presented here is partially funded by the EU eContent-22236 LIRICS project
- **Future readings:**
  - LMF-revision-13, see http://lirics.loria.fr
  Don’t hesitate to contact us (gil.francopoulo@wanadoo.fr)
- **Thank you**