LIRICS WP2 – NLP Lexica

Monica Monachini
monica.monachini@ilc.cnr.it
CNR-ILC – Pisa
+ *LIRICS partners*

8th October 2007

Summary

- WP2 Main Objectives
- External Relations
- WP2 Outstanding Results
  - Deliverables
  - Synergies within and outside the project
- Impact
- Future activities

WP2 Main Objectives

- Define a “family” of standards for NLP lexicons on an international scale, not limited to research institutions but with industrial support
  - **Two-level** standards:
  - provide the structural elements (lexical classes and relations between them), i.e. the meta-model;
  - provide standardized constants, i.e. data categories used to “adorn” the lexical classes
- Build a test suite of lexical entries

WP2 Methodology

- Build on the past, by endorsing major standardization activities and *best-practices* in the field
- Propose the ISO binomial,
  - high-level specification structure,
  - low-level specification adornment
- Identify lexical information to combine with the lexical model: input to the ISO Data Category Registry to enable specific implementations.
- Accompany the model with a set of examples

External relations

- UNI Italian Delegates for ISO/TC37
- ISO TC 37/SC 4/WG4
- ISO TC 37/SC 4/TDG2 Morphosyntax
- ISO TC 37/SC 4/TDG4 Syntax
- ISO TC 37/SC 4/TDG3 Semantic

- NEDO grant

- BOOTStrep

- Language-Grid

- Senso Com

forum for discussion

Results: Data Categor

D 2.1 Evaluation of existing Standards for NLP Lexica

+ Appendices .xml
constraints in datcat combinations as XML features-structures**,** *GrammGender [feminine, masculine, neuter]*)

- Maximum unified set of candic categories subdivided along th linguistic description:
  - ISO **Morphosyntactic profile** (TDG2 – *leader*: *Gil Francopoulo*): 366 data categories
    - DatCats coming from D2.1 and D3.1 relevant to Mo-sy description
  - ISO **Syntactic profile** (TDG4 – *leader*: *Thierry Declerck*)
    - DatCats coming from D2.1 and D3.1 relevant for syntactic description
  - ISO **Semantic profile** (TDG3 – *leader*: *Harry Bunt*)
    - DatCats coming from D2.1 relevant to lexical description are being contributed together with datcats relevant for time annotation

Results: LMF

- The Lexical Markup Framework: a high-level lexical meta-model, a flexible environment for user-defined mark-up languages
- Two revisions:
  - rev.9 M18 CD version (submitt.03-06; results on 06-06)
  - rev.14 M30 DIS version (submitt.11-06; results on 02-07; FDIS ballot started in Aug'07)

D 2.2a Interim Report WD of NLP lexical standard for CD Ballot.

D 2.2b Lexical standard for ISO Ballot.

What is LMF for?

- provide a common model for the creation and use of lexical resources
- manage the exchange of data between and among these resources
- enable the merging of electronic resources to form extensive global resources.

Range of topics:

- monolingual,
- bilingual
- multilingual lexical resources

Scalability

- the same specifications are to be used for both small and large lexicons

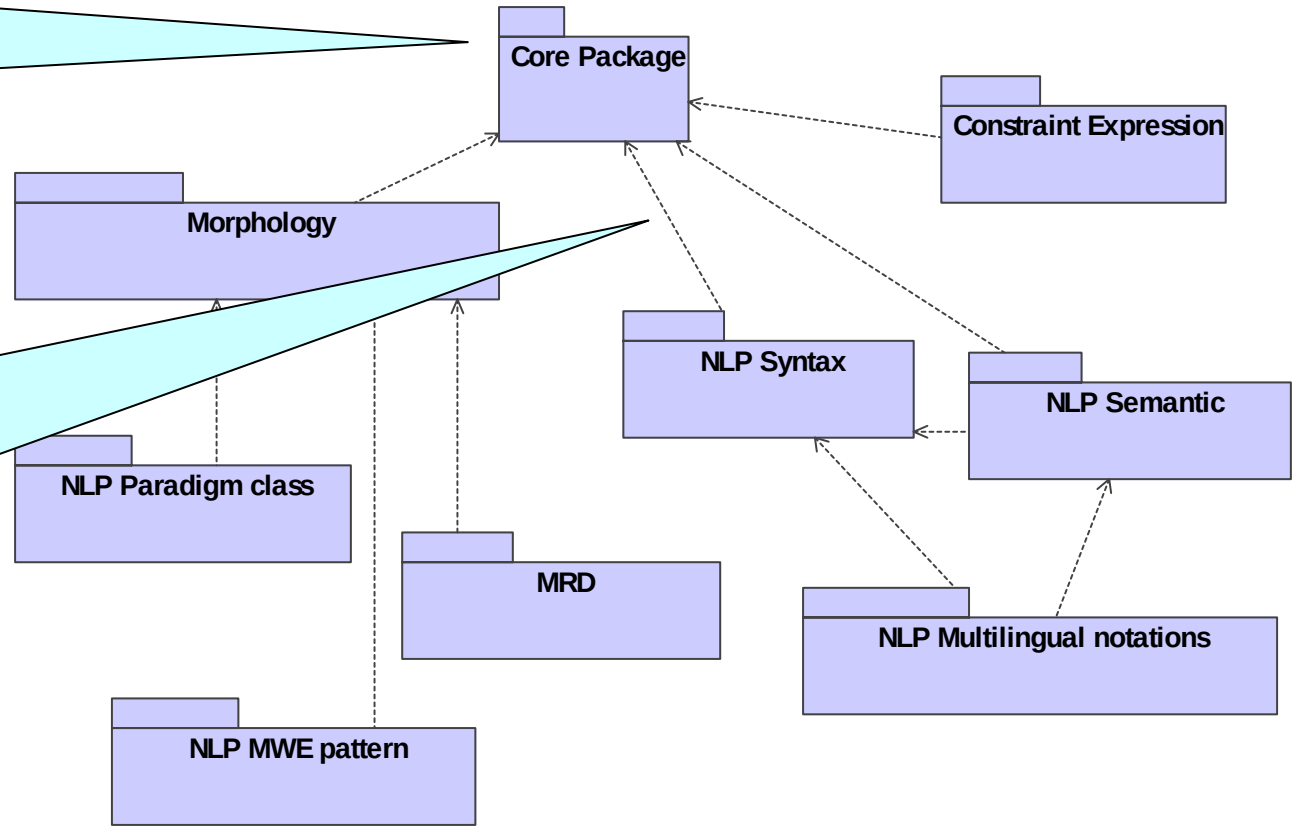
Coverage

- linguistic description range from morphology, syntax, semantic to multilingual representation
- languages are not restricted to European languages
- the range of targeted NLP applications is not restricted.

Structure of LMF

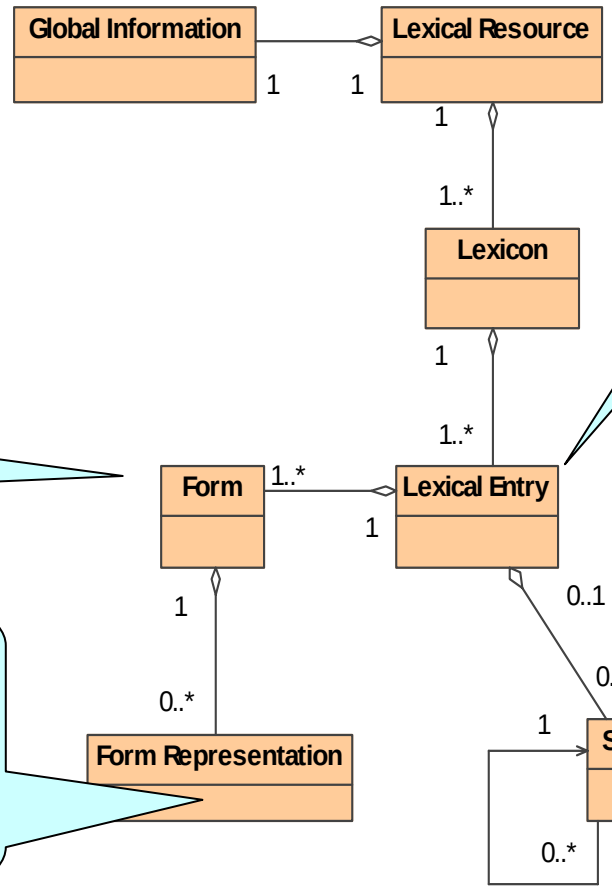
Structural skeleton, with the basic hierarchy of information in a lexical entry

extend a subset of core-model classes; are conformant to the core model; cannot be used regardless to the core model



LMF specifications comply with modeling UML principles;
an XML DTD allows implementation

Core package



Container for managing the top level language components. The number of words or MWe of the lexicon is equal to the number of lexical entries in a given lexicon.

Form consists of a text string that represents a single word or a multi-word expression

One to many *Form Representation* can be associated with *Form*, each of which contains a form and data categories that specify the orthographic types and name of the word

Sense specifies or disambiguates the meaning and context of a form

Results: Lexical Test Suites

- A *mapping of well known NLP lexicon practices* against the model accompanies LMF rev.9 (ISO Auxiliary working document).
- A set of lexical entries conformant to the model accompanies LMF rev.14 (LMF XML DTD), facilitating its acceptance and implementation and promoting the development of LMF conformant lexicons.
  - Principles: *relevance* of linguistic concepts, *conciseness* and *precision* of what is represented, *conformity* to ISO.
  - As a by-product ☾ e.g. for IT development of linguistic services, API allowing to perform queries to the MySql lexical database from Java applications.

D 2.3 Test Suite of ISO conformant lexical entries

Lexical Test suites: coverage

Extension/Info represented	*Lang*	**lexicon/application**
(part of) the syntactic/semantic extensions; correspondence layer	*Italian*	NLP gen-purpose Lex (pluri-lingual)
Morphological extension (different PoS, DatCat from Morphosyntactic Profile)	*Spanish*	Full-form dictionary tagging applications
MWs extension, Constrain expression, paradigm pattern extension	*French*	TagDico
Morphological and semantic extension: pronunciation info and synonymy rel.s	*English*	LMF Demo Lexicon of the Lexus tool
morphological extension: FormRepresentation, orthographic and synonymic variants)	*English (BioLexicon)*	Domain Lexicon for text mining appl.s in Bio
Syntactic and semantic extension; correspondence	*Japanese (NEDO)*	Multilingual Lex for semantic web appl.s

IT sample entry 1/2

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE LexicalResource SYSTEM "LMFNLP.dtd">
<LexicalResource dtdVersion="14">
<GlobalInformation>
   <feat att="source" val="ILC-CNR test suites number 1 for Italian"/>
</GlobalInformation>
<Lexicon>
   <feat att="language" val="ita"/>
   <LexicalEntry id="LE_abbandonare">
      <feat att="POS" val="V"/>
      <Lemma>
         <feat att="writtenform" val="abbandonare"/>
      </Lemma>
      <Sense id="USem73115abbandonare
         <PredicativeRepresentation predicate="PREDabbandonare_2" correspondences="ISObivalent">
            <feat att="link" val="Master"/>
         </PredicativeRepresentation>
         <SenseRelation targets=" USem67171mollare USem79703lasciare">
            <feat att="relation_type" val="Synonym"/>
         </SenseRelation>
         <SenseRelation targets=" USemD5371agire">
            <feat att="relation_type" val="Isa"/>
         </SenseRelation>
      </Sense>
      <SyntacticBehaviour id="SB_SYNUabbandonare"
      senses=" USem59592a ndonare USem73115abbandonare"
      subcategorizationFrames="regularSVO"/>
   </LexicalEntry>
```

IT sample entry 2/2

Activities, Meetings, Conferences

LIRICS WPs BI- TRI- LATERAL Working Meetings:

- Pisa, April 07; Paris, 10.5.07 : “topical” WP2 ☾ WP5 meeting

LIRICS Meetings

- Paris, 11.05.2007: Industrial Advisory Board Meeting
- Provo, Aug 2007

ISO Meetings

- Beijing, 20-26 August 2006; Provo, August 2007: ISO TC37 Annual meeting.

LMF related Meetings

- April, September, December 2006 LMF presented at Bootstrep
- Kyoto, Jan 2007: LMF discussed in the framework of the NEDO grant

Conferences

- COLING’06, Sydney: LMF multilingual
- GLDV’07: Tuebingen, April: LMF multilingual
- GL’07 Paris, 10-11 May 2007: LMF and Bootstrep
- LTC’07 Poznan, October 2007: LMF and BootStrep

Impact of LMF

- BioLexicon: a term lexicon for text-mining applications in the bio domain
  - Normalization of nomenclature in bio terms; link with the ontology
- NEDO: a harmonized multi-Asian lexicon for advanced industrial technologies
  - OWL.RDF conceptual model; XML core set lexical entries
- NICT Language-Grid Service Ontology: composite services to connect existing language services on users' request
  - Ontologized version of LMF
- *Senso-Comune*: a linguistic a knowledge base (wiki modalities) for It
- The CLARIN infrastructure will get inspired from LIRICS results
- LMF & FLaReNet *Fostering Language Resources Network* (if accepted)

Future activities

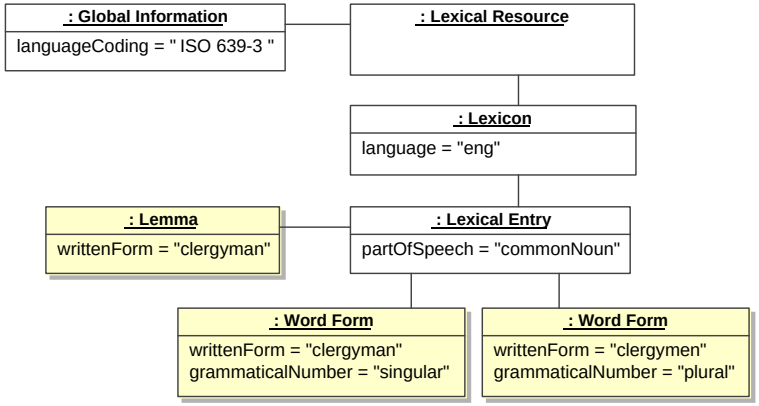
- A successful ISO standard: FDIS Feb'08:
  - define the specification
  - communicate
  - make LMF usable and operational
- Provide a space in a WIKI for LMF to continue working, store guidelines/examples, ease dissemination
- Build a simple, usable, standalone/web-based, free and open tool for building LMF lexicons (LEXUS)
- LMF User Guide with examples
- Map all commonly used lexicons into LMF
- Publish the DatCats that are needed for an LMF lexicon.
- Address the ontological layer in the lexicon (internal task force to provide input to the ontological task force)

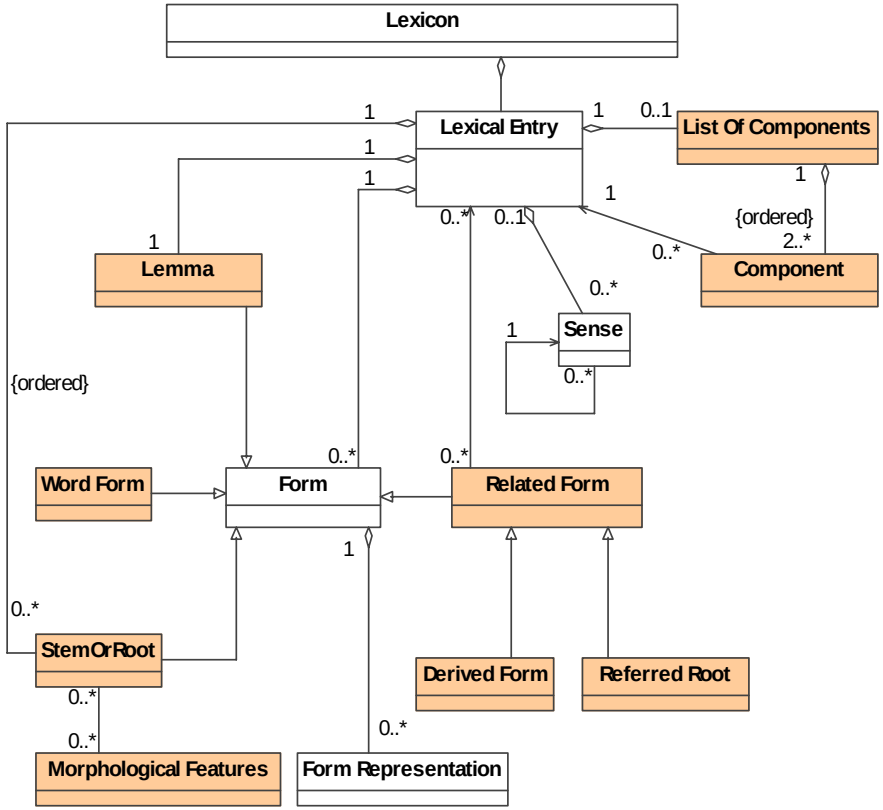
THE END

LMF is ...

- a structural data model expressed by a set of Unified Modeling Language (UML) packages
- a high level specification based on constants 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.
  - attributes are not defined in the LMF specification, to be taken from the data category registry; values are either constants or free strings.

Package for Morphology





Package for NLP syntax

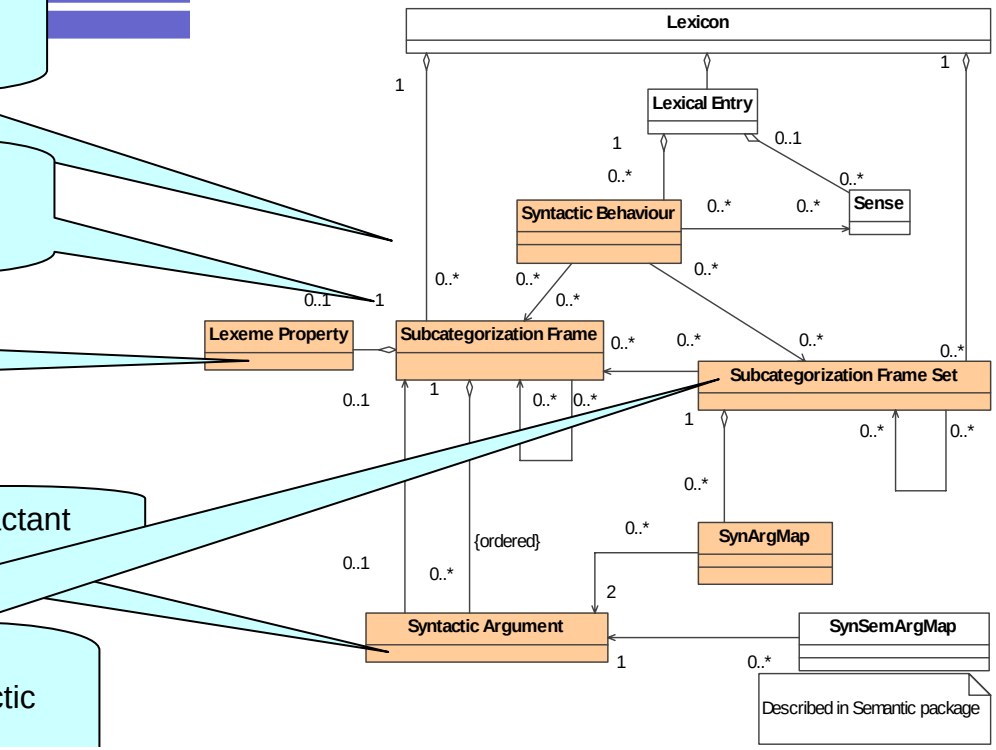
SyntacticBehavior represents one of the behaviors of one (or more) senses

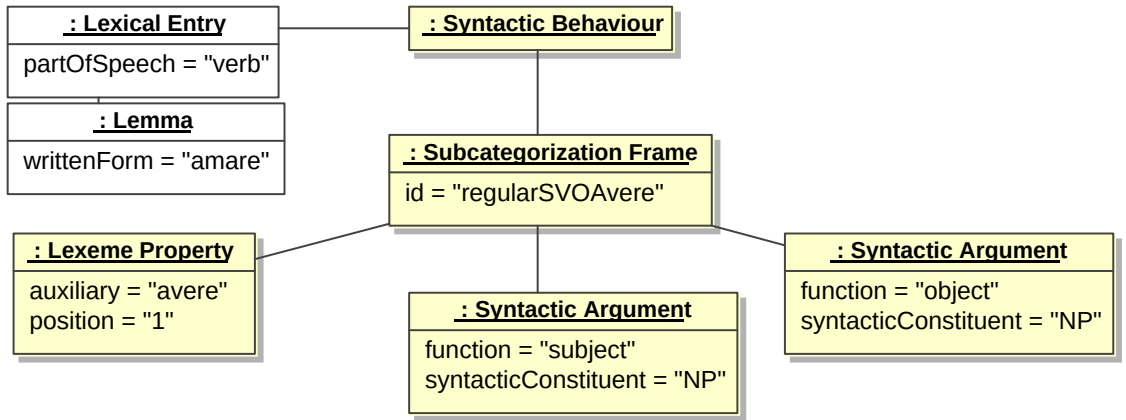
SubcategorizationFrame describes one syntactic construction and can be shared by all words with the same syntactic behavior

LexemePoperty refers to the head lexical entry and describes syntactic properties

Syntactic Argument describes a syntactic actant

SubcatFrameSet regroups together various *Syntactic Constructions* and factorizes syntactic descriptions to have a minimum of syntactic behavior elements in the lexicon.





XML representation

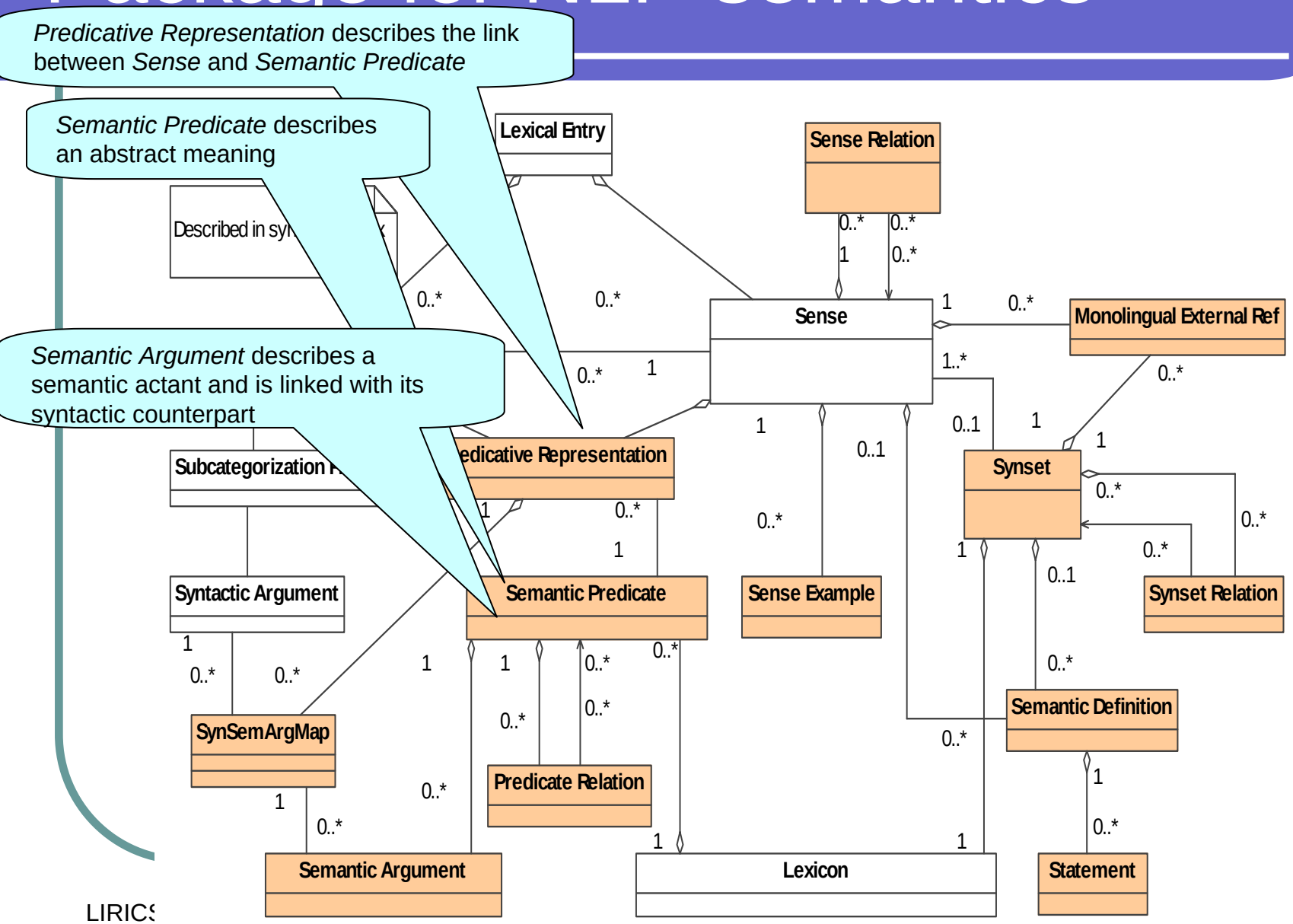
```
<LexicalResource dtdVersion="14">
<GlobalInformation>
  <feat att="source" val="clips"/>
</GlobalInformation>
<Lexicon>
  <feat att="language" val="ita"/>
<LexicalEntry>
  <feat att="partOfSpeech" val="verb"/>
  <Lemma>
    <feat att="writtenForm" val="amare"/>
  </Lemma>
  <SyntacticBehaviour subcategorizationFrames="regularSVOAvere"/>
</LexicalEntry>
<SubcategorizationFrame id="regularSVOAvere">
  <LexemeProperty>
    <feat att="auxiliary" val="avere"/>
    <feat att="position" val="1"/>
  </LexemeProperty>
  <SyntacticArgument>
    <feat att="function" val="subject"/>
    <feat att="syntacticConstituent" val="NP"/>
  </SyntacticArgument>
  <SyntacticArgument>
    <feat att="function" val="object"/>
    <feat att="syntacticConstituent" val="NP"/>
  </SyntacticArgument>
</SubcategorizationFrame>
</Lexicon>
</LexicalResource>
```

Package for NLP semantics

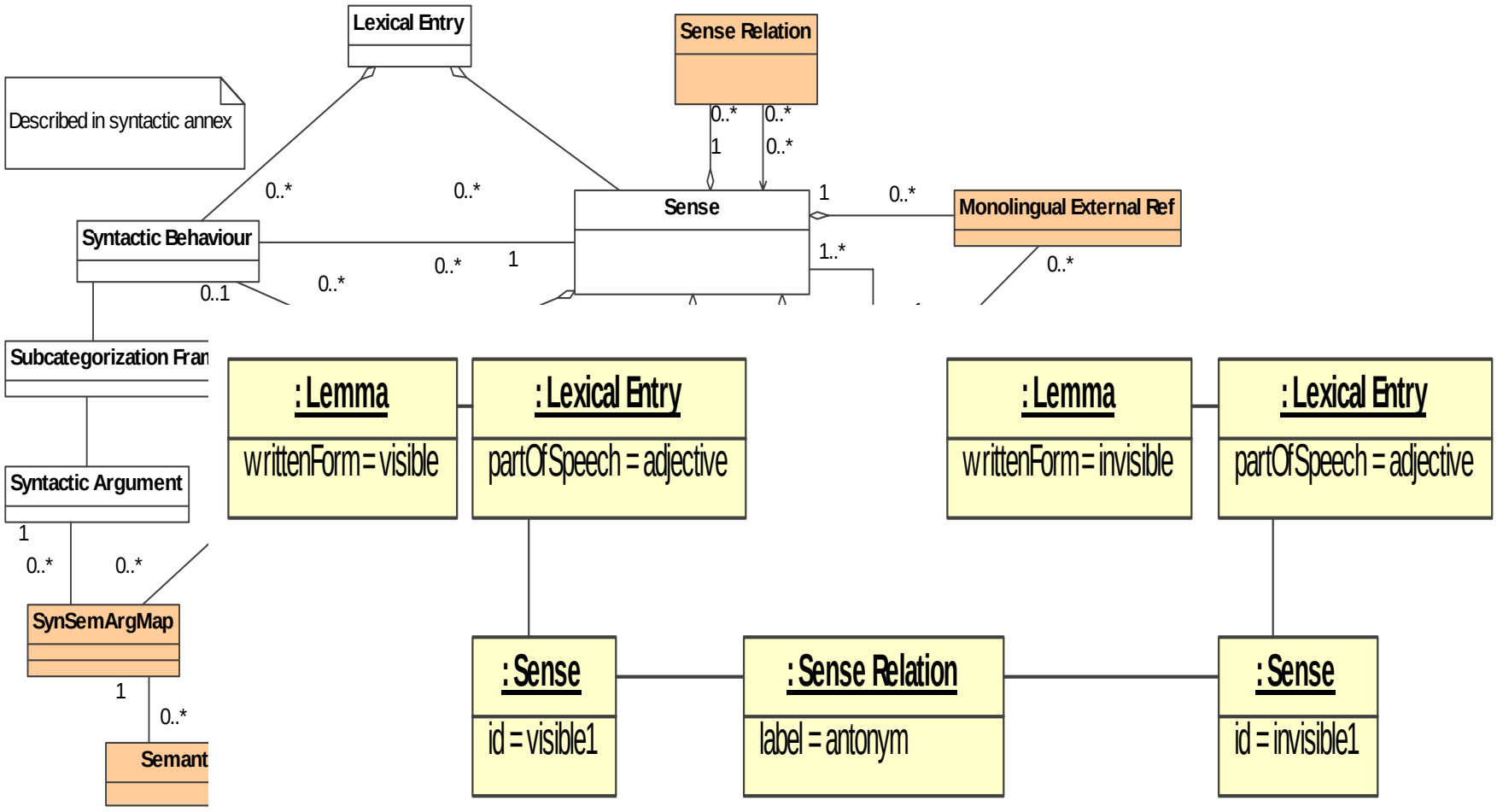
Predicative Representation describes the link between *Sense* and *Semantic Predicate*

Semantic Predicate describes an abstract meaning

Semantic Argument describes a semantic actant and is linked with its syntactic counterpart



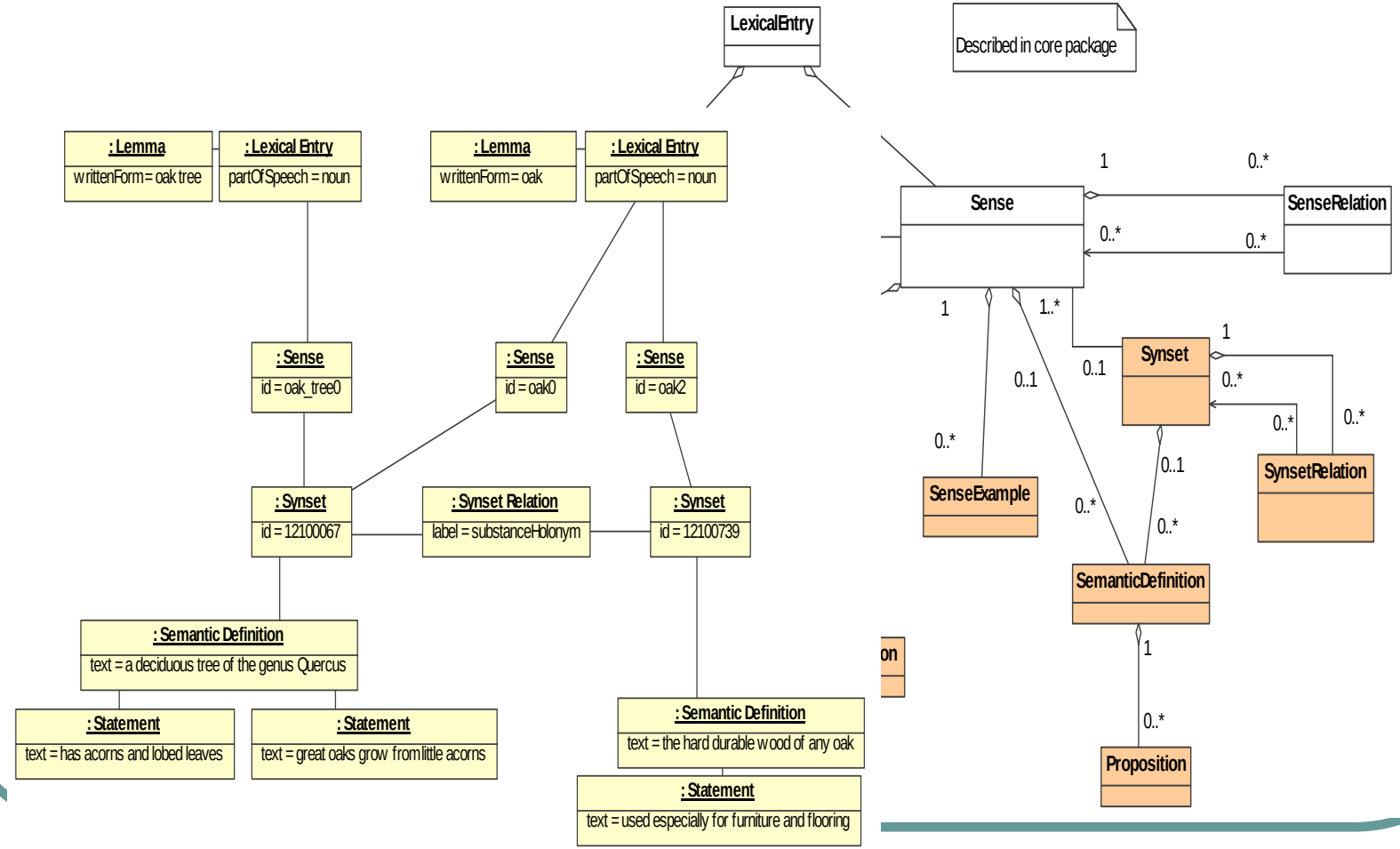
Package for NLP semantics (cont.)



XML representation

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE LexicalResource SYSTEM "LMFNLP.dtd">
<LexicalResource dtdVersion="14">
<GlobalInformation>
   <feat att="label" val="ILC-CNR test suites number 1 for Italian"/>
</GlobalInformation>
<Lexicon>
   <feat att="language" val="fra"/>
   <LexicalEntry>
   <feat att="partOfSpeech" val="adjective"/>
   <Lemma>
      <feat att="writtenForm" val="visible"/>
   </Lemma>
   <Sense id="visible1">
      <SenseRelation targets="invisible1">
         <feat att="label" val="antonym"/>
      </SenseRelation>
   </Sense>
</LexicalEntry>
<LexicalEntry>
   <feat att="partOfSpeech" val="adjective"/>
   <Lemma>
      <feat att="writtenForm" val="invisible"/>
   </Lemma>
   <Sense id="invisible1"/>
</LexicalEntry>
</Lexicon>
</LexicalResource>
```

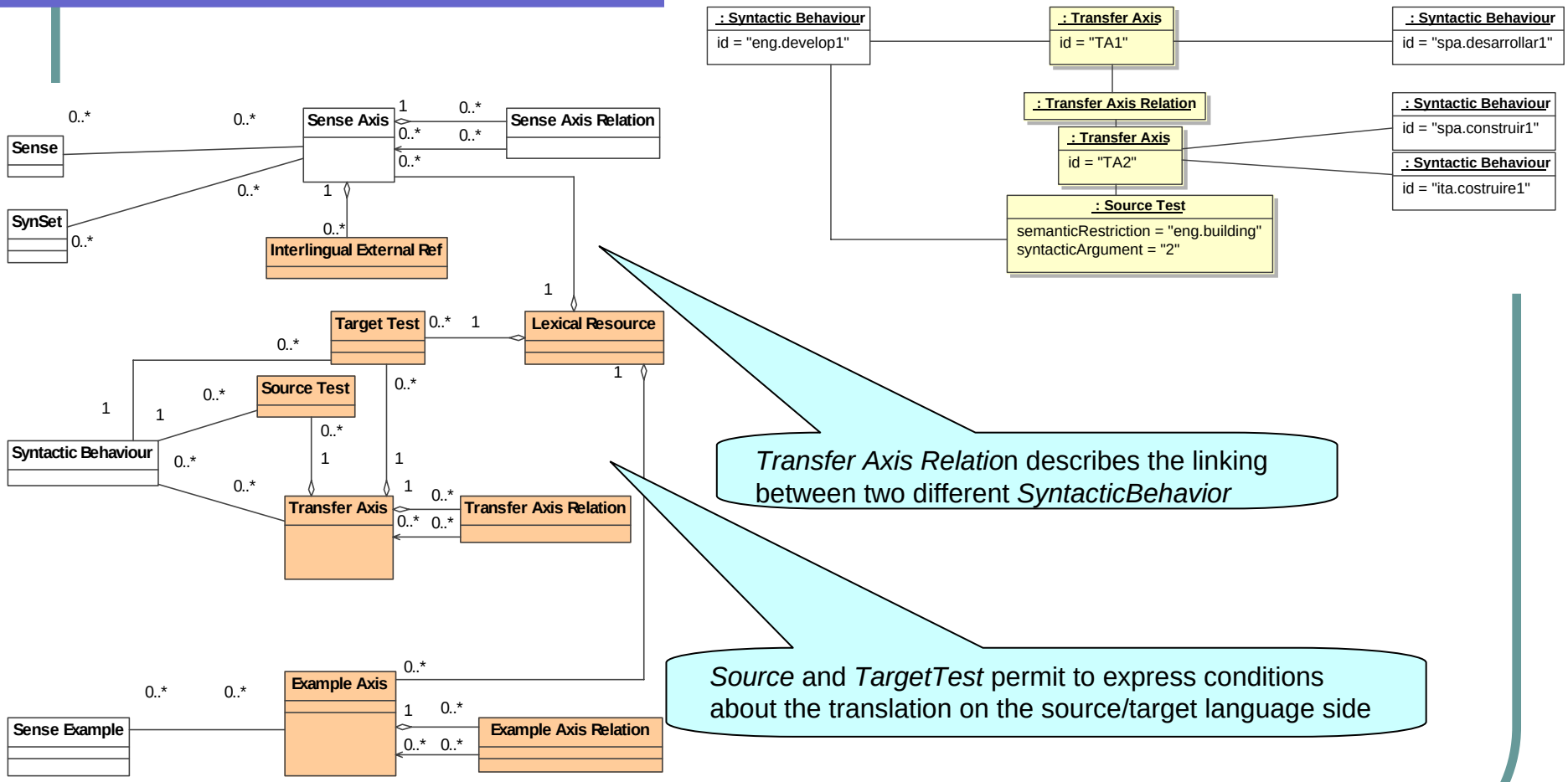
Package for NLP semantics (cont.)



XML representation

```
<LexicalEntry>
  <feat att="partOfSpeech" val="noun"/>
  <Lemma>
    <feat att="writtenForm" val="oak tree"/>
  </Lemma>
  <Sense id="oak_tree0" synset="s12100067"/>
</LexicalEntry>
<LexicalEntry>
  <feat att="partOfSpeech" val="noun"/>
  <Lemma>
    <feat att="writtenForm" val="oak"/>
  </Lemma>
  <Sense id="oak0" synset="s12100067"/>
  <Sense id="oak2" synset="s12100739"/>
</LexicalEntry>
<Synset id="s12100067">
  <SemanticDefinition>
    <feat att="text" val="a deciduous tree of the genus Quercus"/>
    <Statement>
      <feat att="text" val="has acorns and lobed leaves"/>
    </Statement>
    <Statement>
      <feat att="text" val="great oaks grow from little acorns"/>
    </Statement>
  </SemanticDefinition>
  <SynsetRelation targets="s12100739">
    <feat att="label" val="substanceHolonym"/>
  </SynsetRelation>
</Synset>
<Synset id="s12100739">
  <SemanticDefinition>
    <feat att="text" val="the hard durable wood of any oak"/>
    <Statement>
      <feat att="text" val="used especially for furniture and flooring"/>
    </Statement>
  </SemanticDefinition>
</Synset>
```

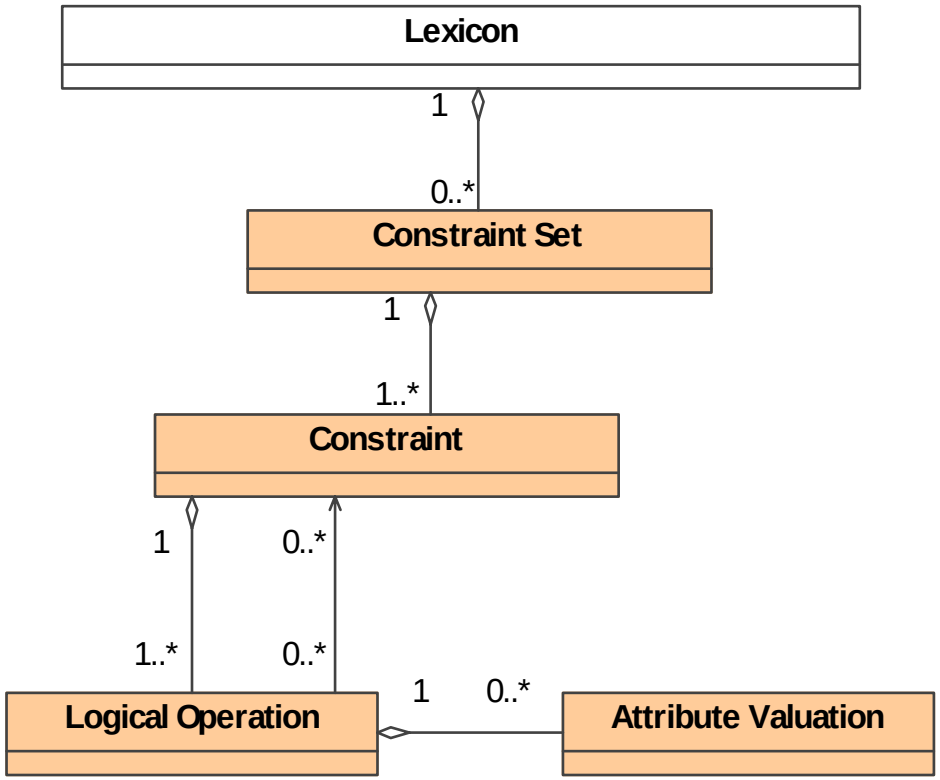
Package for Multilingual representation



XML representation

```
<TransferAxis id="TA1" synBehaviours="eng.develop1 esp.desarrollar1">
  <TransferAxisRelation targets="TA2"/>
</TransferAxis>
<TransferAxis id="TA2" synBehaviours="esp.construir1 ita.costruir1">
  <SourceTest>
    <feat att="semanticRestriction" att="eng.building"/>
    <feat att="syntacticArgument" att="2"/>
  </SourceTest>
</TransferAxis>
```

Package for Constraint Expression



XML representation

```
            <feat att="partOfSpeech" val="adjective"/>
        </AttributeValuation>
        <Constraint>
            <feat att="label" val="genderNumber"/>
            <LogicalOperation>
                <feat att="operator" val="logicalOr"/>
                <AttributeValuation>
                    <feat att="grammaticalGender" val="masculine"/>
                    <feat att="grammaticalNumber" val="singular"/>
                </AttributeValuation>
                <AttributeValuation>
                    <feat att="grammaticalGender" val="masculine"/>
                    <feat att="grammaticalNumber" val="plural"/>
                </AttributeValuation>
                <AttributeValuation>
                    <feat att="grammaticalGender" val="feminine"/>
                    <feat att="grammaticalNumber" val="singular"/>
                </AttributeValuation>
                <AttributeValuation>
                    <feat att="grammaticalGender" val="feminine"/>
                    <feat att="grammaticalNumber" val="plural"/>
                </AttributeValuation>
            </LogicalOperation>
        </Constraint>
    </LogicalOperation>
</Constraint>
</ConstraintSet>
</Lexicon>
</LexicalResource>
```