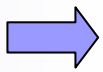
Lexical Markup Framework (iso 24613)

Gil FRANCOPOULO
INRIA-loria
gil.francopoulo@wanadoo.fr



not a one way presentation, interrupt when you want

Introduction

- One of the crucial aspects impacting HLT is the need to optimize the production, maintenance and extension of linguistic resources.
  => management + interchange
  A very frequent mentioned need is: « how to merge resources? ».

- A second crucial aspect involves optimizing the process leading to integration of linguistic resources in applications

Scope

- The scope is: all natural languages, MRD & NLP lexicons.
- The goal is not to represent all lexicons in the world. The goal is to represent the best practices of the 2 fields: MRD & NLP, and so to target a maximum of lexicons
- LMF must provide a good linguistic solution for notions like word (single word + multi-word expressions), morphology, syntactic behavior, sense, semantic relations, syntactic-semantic mapping, translation…

Principles

- In these complex domains, it's not possible to anticipate everything
- Hypothesis: common best practices can be specified as structural elements
- These structural elements are:
  a) specified in the LMF document
  b) will be decorated as the user convenience by ISO 12620 datcats
- Separately from LMF, a registry of datcats is provided (see rev ISO 12620)

Organization

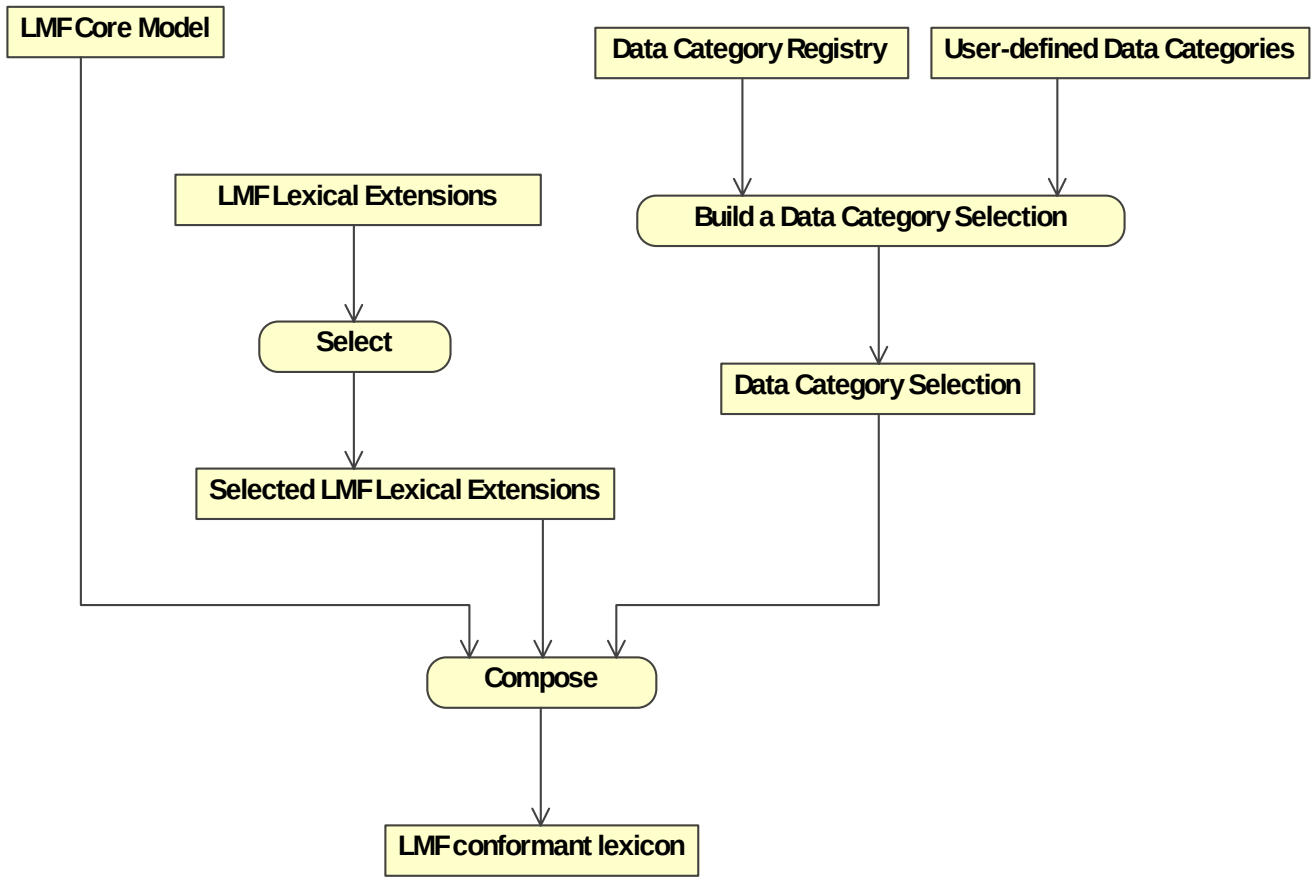
- All languages, MRD+NLP
  => the scope is rather broad

Decision: a modular organization with:

  - a core model for common elements
  - a series of 5 optional modules (called extensions) to be combined together in order to address various types of lexicons

- No specific extension for a particular family of languages

Process

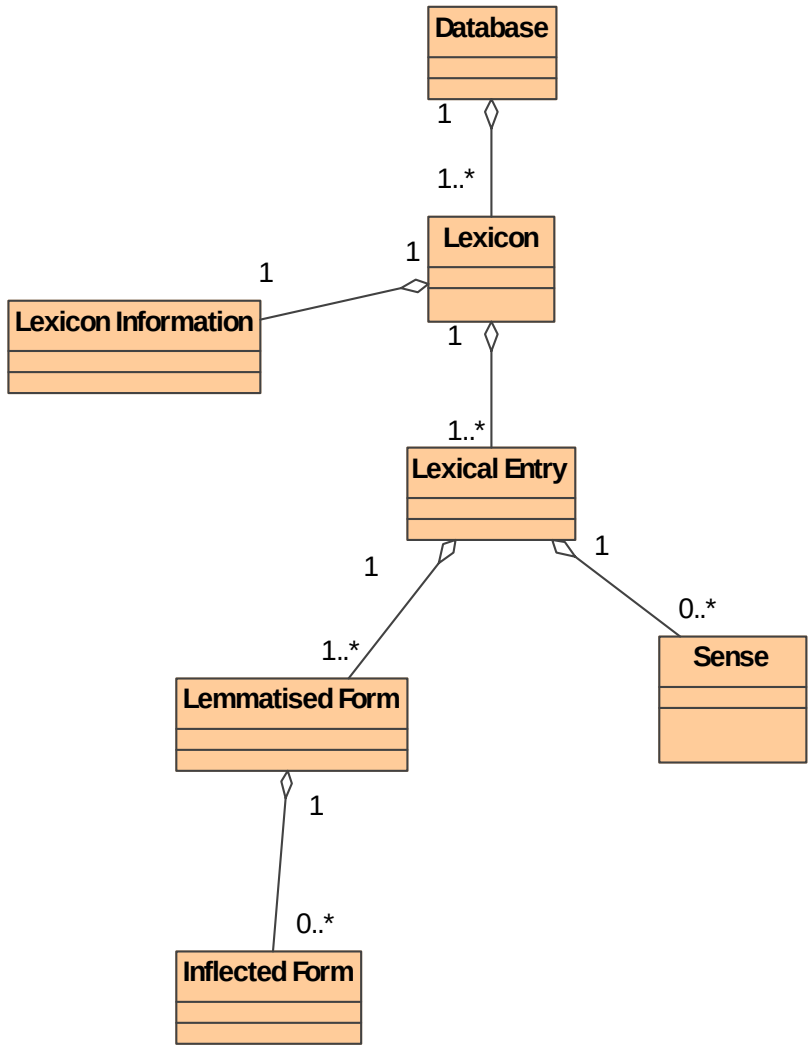


UML activity diagram

- So, LMF is a meta-model in the sense that the standard specifies a mechanism that allows the user to define his own lexical model
- According to two distinct criteria:
  - The DCS
  - The selected extensions

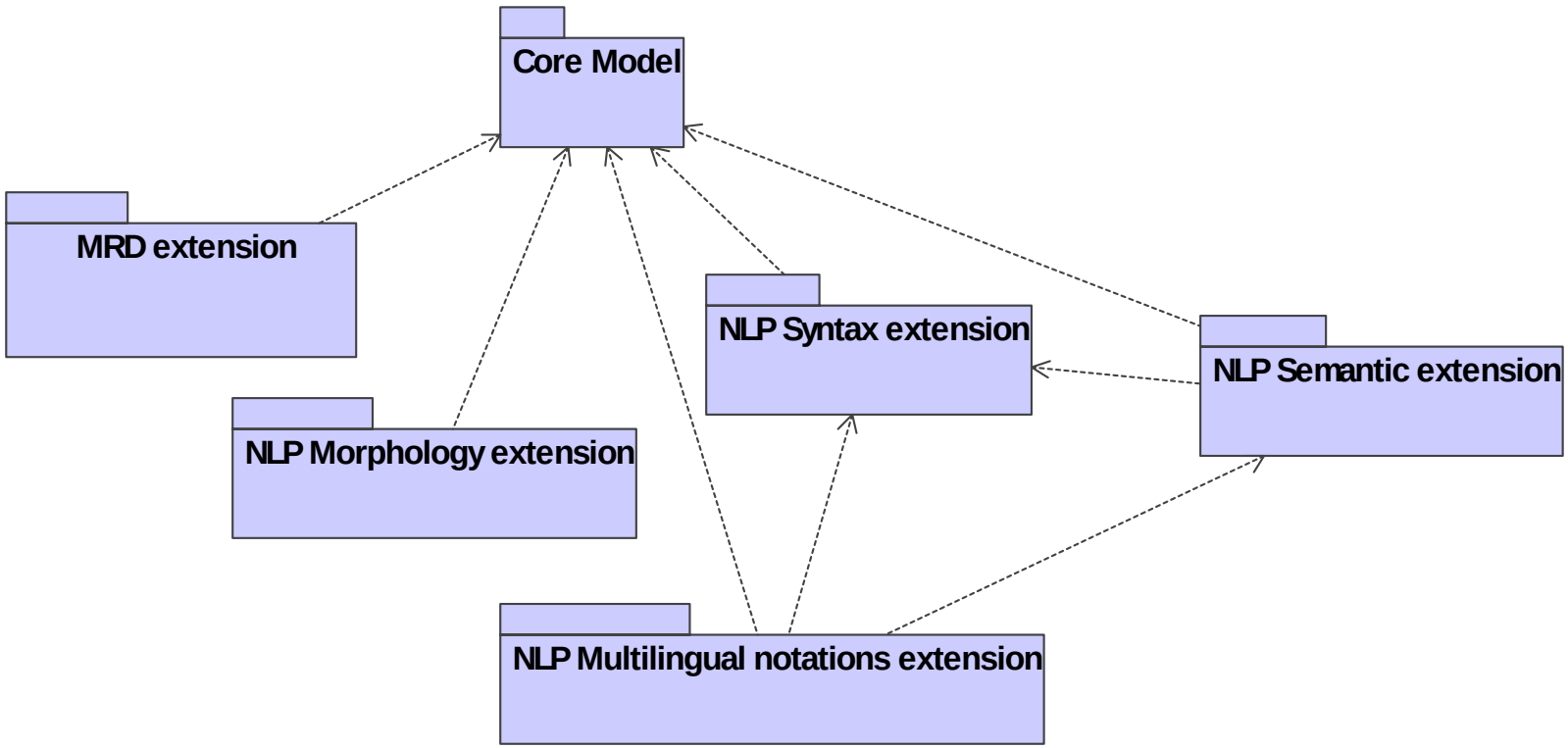
LMF is not a model for a specific lexicon

Core Model



UML class model diagram

Extensions


Core Model
MRD extension
NLP Syntax extension
NLP Semantic extension
NLP Morphology extension
NLP Multilingual notations extension

NLP morphology extension

1) From the linguistic point of view

- Mono vs multi-orthographic languages
- Internal morphology
  - Isolating languages
  - Agglutinating languages
  - Inflectional languages
- Polysynthesis
- Single words & MWE

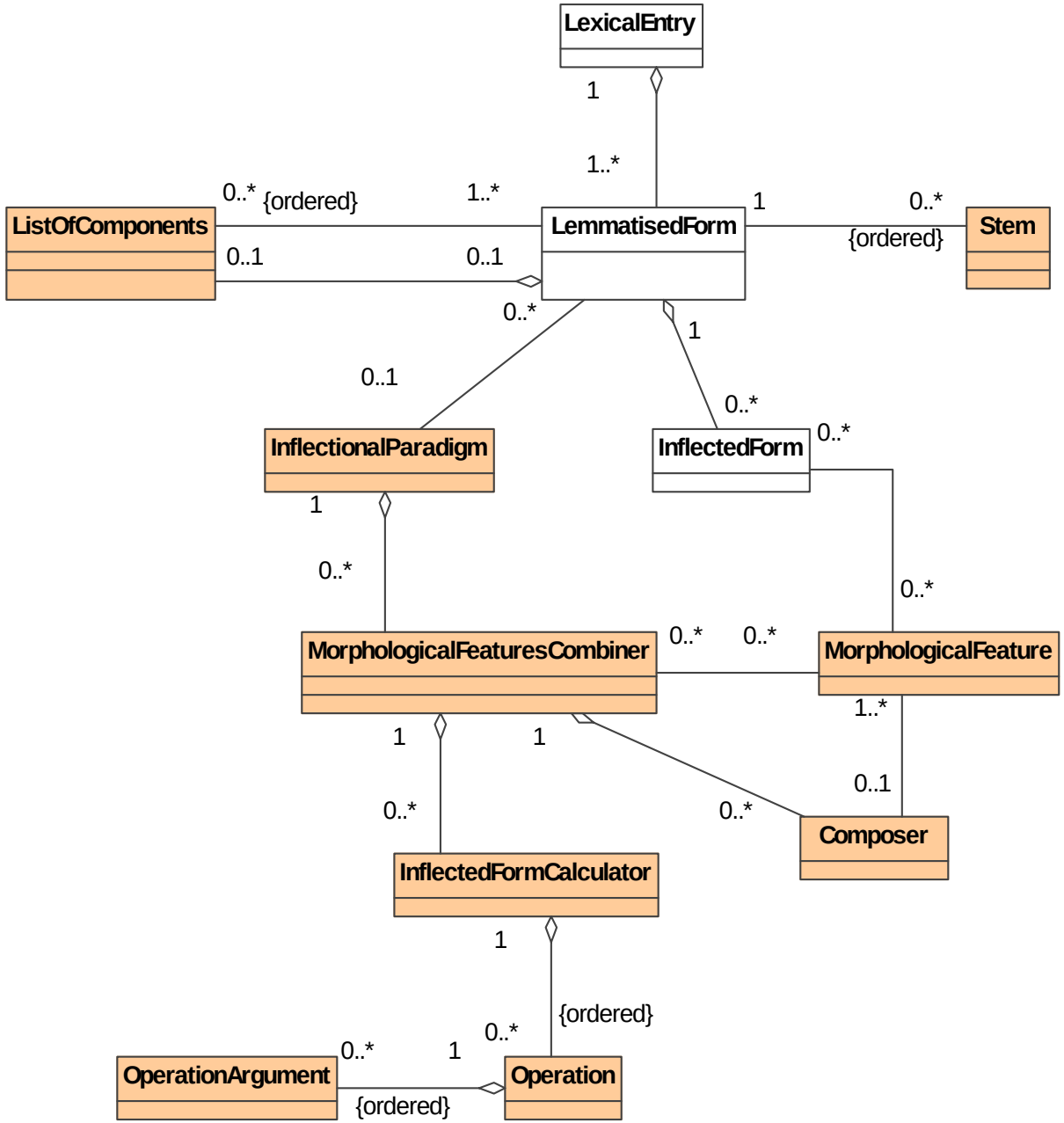
Language universals & linguistic typology (B. Comrie)

Morphology (cont.)

2) From the point of view of the strategy of description

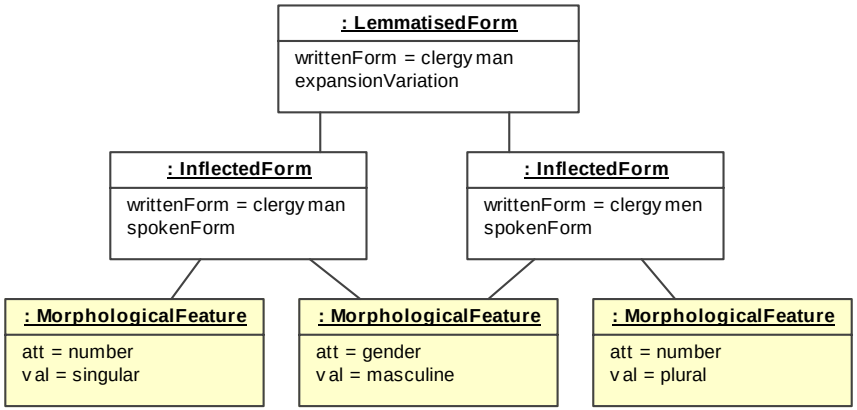
- In extension vs intension (depends on the language)
- For intension, the lexicon manager has the choice between different levels of precision: under-specified description or full description

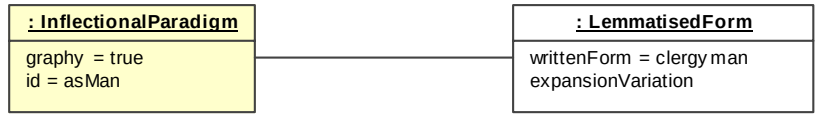
NLP Morphology

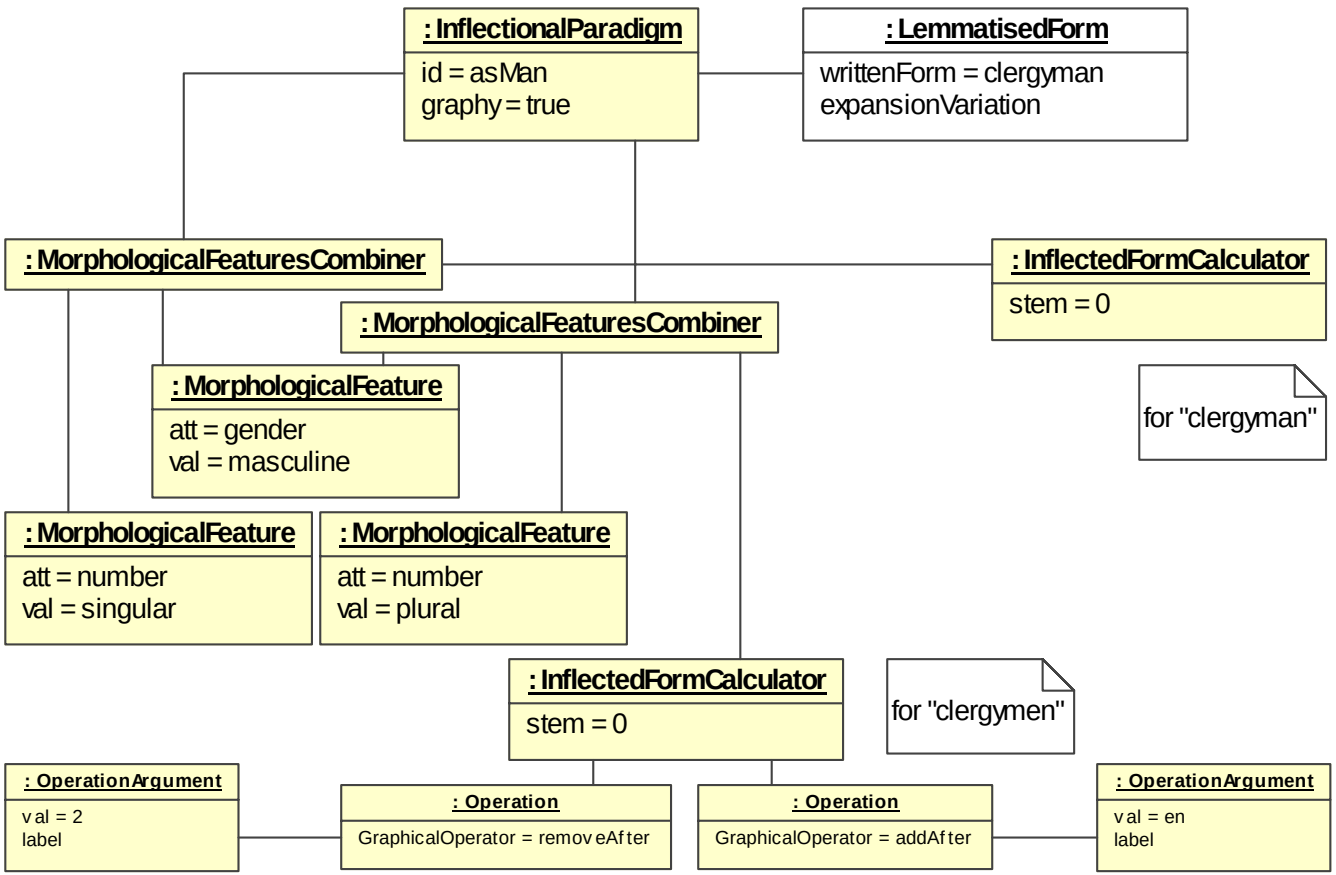


UML class model

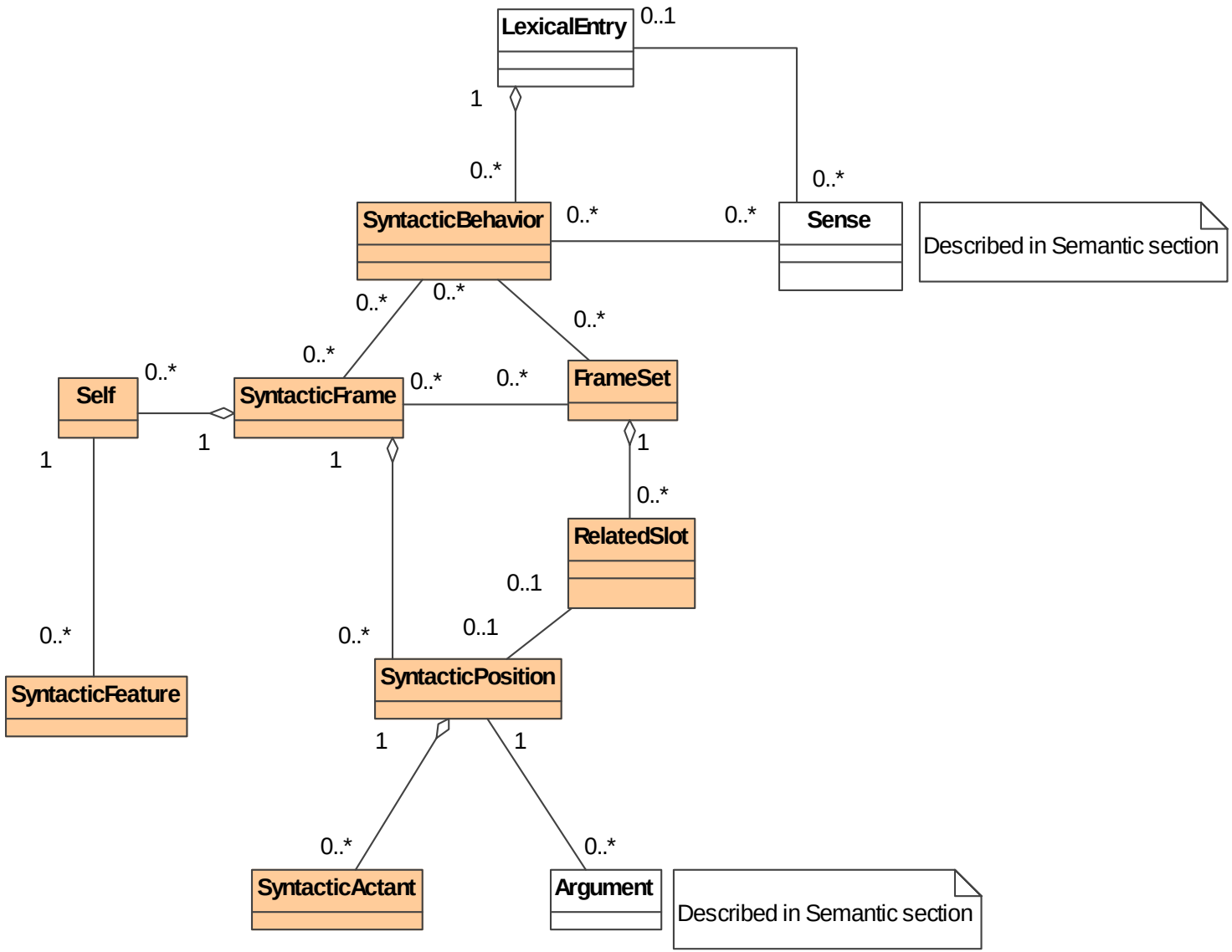
UML object model example:
Three ways to describe « clergyman »





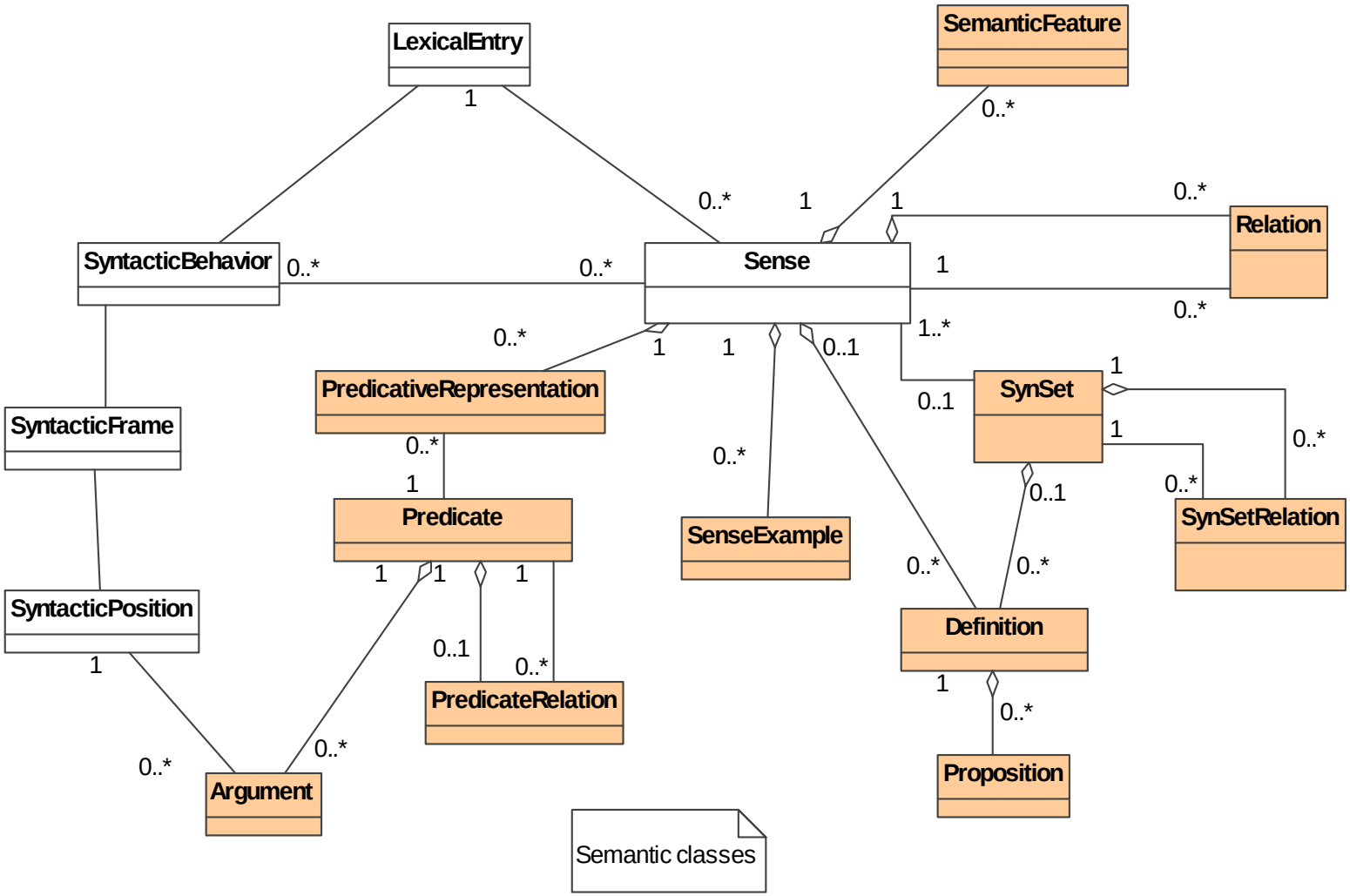


NLP Syntax

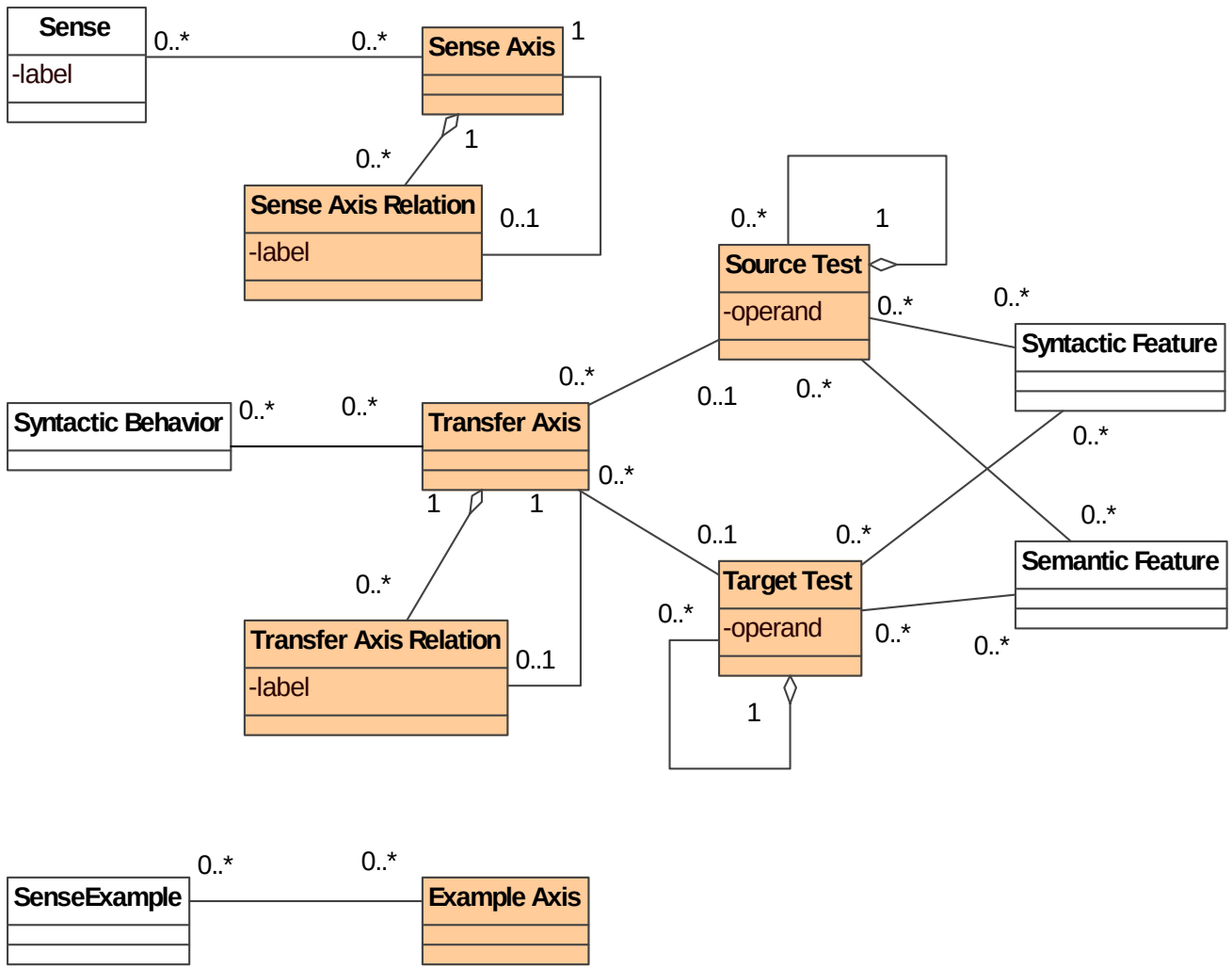


UML class model

NLP Semantic



NLP multilingual notations



UML class model

Question #1: decision to take

- Until now, we worked only on the conceptual model with UML
- No real decision concerning the external physical model: just some tests.
- Do you think we need XML specifications in the document?

Question #2: decision to take

- What is the status of the study « Extended examples of LMF » ?
- Do we consider this document as an informative annex or a separate technical report?

Question #3: strategy to adopt concerning the ISO CS

- Do we schedule one CD then a CD ballot?
- Or do we produce a series of CD to be sent to the ISO experts for comments and then a CD ballot?