Towards an ISO standard representation for proper names

Béatrice Bouchou
beatrice.bouchou@univ-tours.fr
LI – Université François Rabelais de Tours
16th April 2007

Summary of the presentation

- ISO/TC 37
  - Data categories
  - Terminological Markup Framework (TMF)
  - Lexical Markup Framework (LMF)
- Prolexbase
  - in TMF
  - in LMF
  - XML (?)

What are ISO standards for?

- Provide a common model for the creation and use of NL resources
- Manage the exchange of data between and among these resources
- Enable the merging of electronic resources

Range

- Monolingual
- Bilingual
- Multilingual
- Linguistic description range from morphology, syntax, semantics to multilingual representation
- Languages are not restricted to European languages
- The range of targeted NLP applications is not restricted

Scalability

- Same specifications for both small and large lexicons

ISO/TC 37

- Terminology *and other language resources*
  - **SC3 -** Computer applications in terminology
    - ISO 12200 - Martif
      - Latest version of TEI Terminology chapter
    - ISO 12620 - Data categories (under revision)
    - ISO 16642 - TMF (Terminological Markup Framework)
  - **SC4 -** Language Resource Management

ISO/TC 37/SC4

The following standards are under preparation:

- ISO/NWI 21829 — Terminology for language resources
- ISO/NP 23679-1 — Word segmentation of written texts – Part 1: General principles and methods
- ISO/NP 23679-2 — Word segmentation of written texts – Part 2: Chinese, Japanese and Korean
- ISO/CD 24610-3 — Language resource management – Feature structures – Part 3: Word segmentation for other languages
- ISO/WD 24611 — Language resource management – Morpho-syntactic annotation framework
- ISO/WD 24612 — Language Resource Management – Linguistic Annotation Framework
- ISO/WD 24613 — Language resource management – LMF (Lexical Markup Framework)

ISO/TC 37/SC 3 & 4

Two-level standards:

- the **high level** specifications provide structural elements, i.e. classes and relations between them: the meta-model
- the **low level** specifications provide standardized constants, i.e. data categories used to “adorn” the classes: ISO 12620

Data categories

- Definition
  - Feature names and values used to describe natural language resources
- Example
  - Features: /part of speech/, /grammatical gender/
  - Values: /feminine/, /plural/, /dual/, /ablative case/
- Role
  - Characterization of structural elements (specification)
  - Linguistic range identification (documentation)

Example of DCR

DCRegistry

version: 1

	Administration Note
DC ID	Administration Note
DC Name	Administration Note
DC Definition	[lrom-en]any general note about the Administered Item;
DC Source Comment	ISO 11179-3
Data Type	basicText (open)
Level(s)	Administration Record

	Administration Status
DC ID	Administration Status
DC Name	Administration Status
DC Definition	[lrom]a designation of the status in the administrative process of a Registration Authority for handling registration requests.
Concept-related Comment	The values and associated meanings of administrative status are determined by each Registration Authority. C.f. registration status
Data Type	basicText (open)
Level(s)	Administration Record

Example of DCR

Conceptual Domain	
DC ID	Conceptual Domain
DC Name	Conceptual Domain
DC Definition	[lrom-en] On the DS level, this field is used to relate the category under description with the set of all its possible values (expressed as a list of data categories). When necessary a datatype (in the sense of XML schemas) may be provided instead of a list of values [lrom-en] On the LS level, to be used when a data category is to be associated to a specific subset of the values declared at DS level
DC Source Comment	ISO 11179-3
Data Type	basicText (open)
Level(s)	Language Section Description Section

Creation Date	
DC ID	Creation Date
DC Name	Creation Date
DC Definition	[lrom-en]the date when the data category has been initially created (for instance in an expertâ€™s working space/private area);
Concept-related Comment	[pseb] must be refine by /Change description/ to show the modifications between the last version and the current one.
Data Type	basicText (open)
Level(s)	Administration Record

Example of DCR

note	
DC ID	ISO12620A-08
DC Name	note
DC Definition	A statement that provides further information on any part of a terminological entry. [lrom - DS]additional information associated with the DS level, excluding technical information that would normally be described within /Explanation/ [lrom - LS]additional information associated with the LS level, excluding technical information that would normally be described within /explanation/ [lrom]may refine /definition/ to indicate approval, acceptability, or applicability in a given context. In particular, it should be used to record alternative definitions, or older definitions that one mat want to keep for documentary purposes. The /status/ field should not be used alone at DS level
DC Source Comment	For definition of related term, see ISO 1087-1, 3.8.5. ISO12620-2
Data Type	noteText (open)
Level(s)	Language Section Description Section Name Section Used only for refinement

Data categories

- Data Category Registery (DCR): http://syntax.inist.fr
  - NLP resource developers can refer to this DCR, while building their proper DCR…
- Motivation
  - Reference framework for comparing models and structures for NLP
  - Towards a better interoperability between information systems

Example : Morphalou

- Dictionary of inflected forms for French (www.cnrtl.fr)
- Only forms yet (no senses)
  - 60 940 common nouns
  - 8790 verbs
  - 22790 adjectives
  - 1579 adverbs
  - etc.

Data Categories in Morphalou

- <orthography>
- <grammaticalCategory>
- <grammaticalGender>
- <grammaticalNumber>
- <spellingVariantOf>
- <feminineVariantOf>
- <originatingEntry target="68340">
- …

Meta-models

- TMF
  - standards and guidelines for creating and using terminological data collections
  - Core model + data categories
- LMF
  - creation and use of electronic lexical resources
  - generic skeleton + data categories
  - More components in the skeleton…

TMF meta-model

- All following slides on TMF are taken from:

Laurent Romary

Laboratoire Loria-INRIA

TMF Meta-model

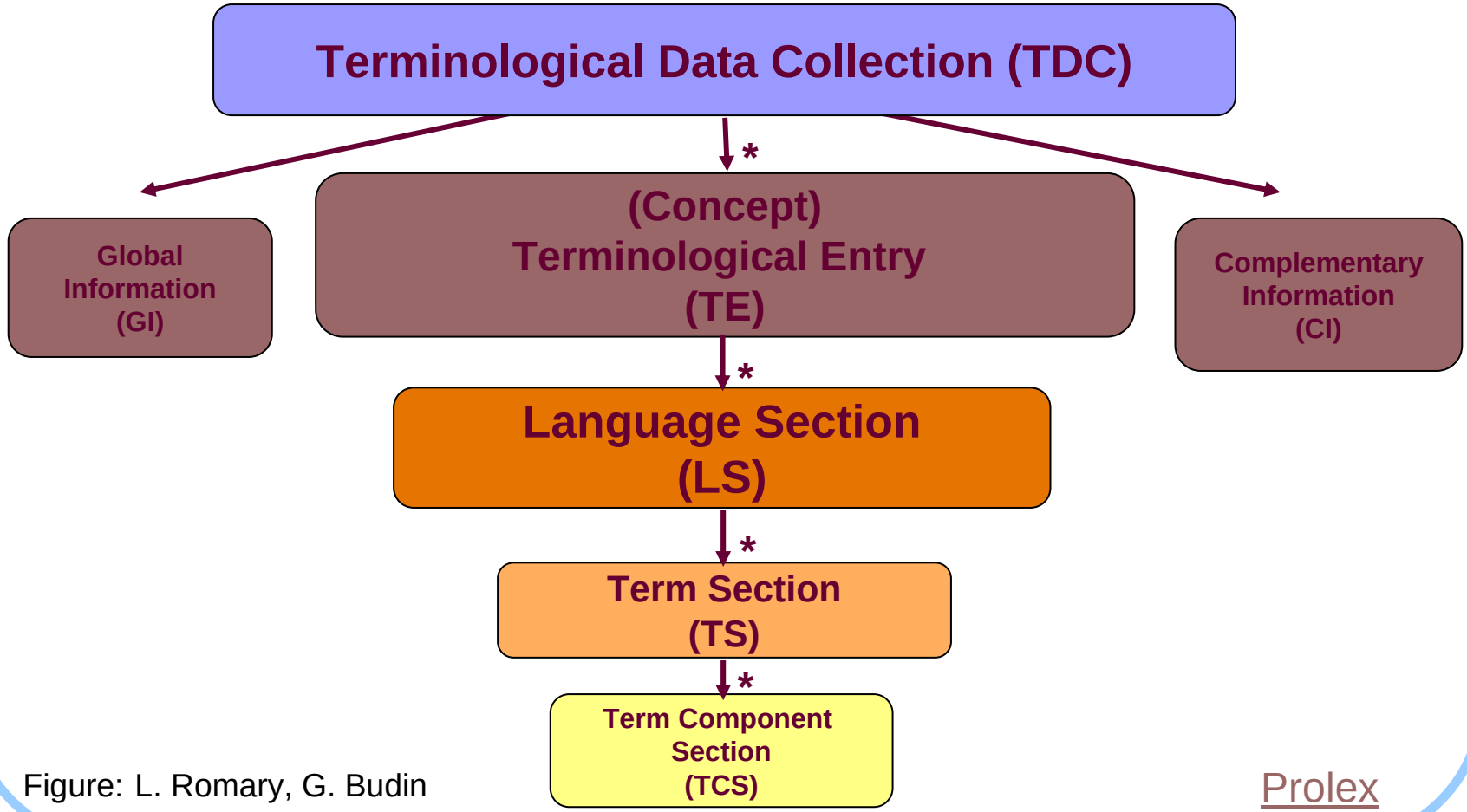


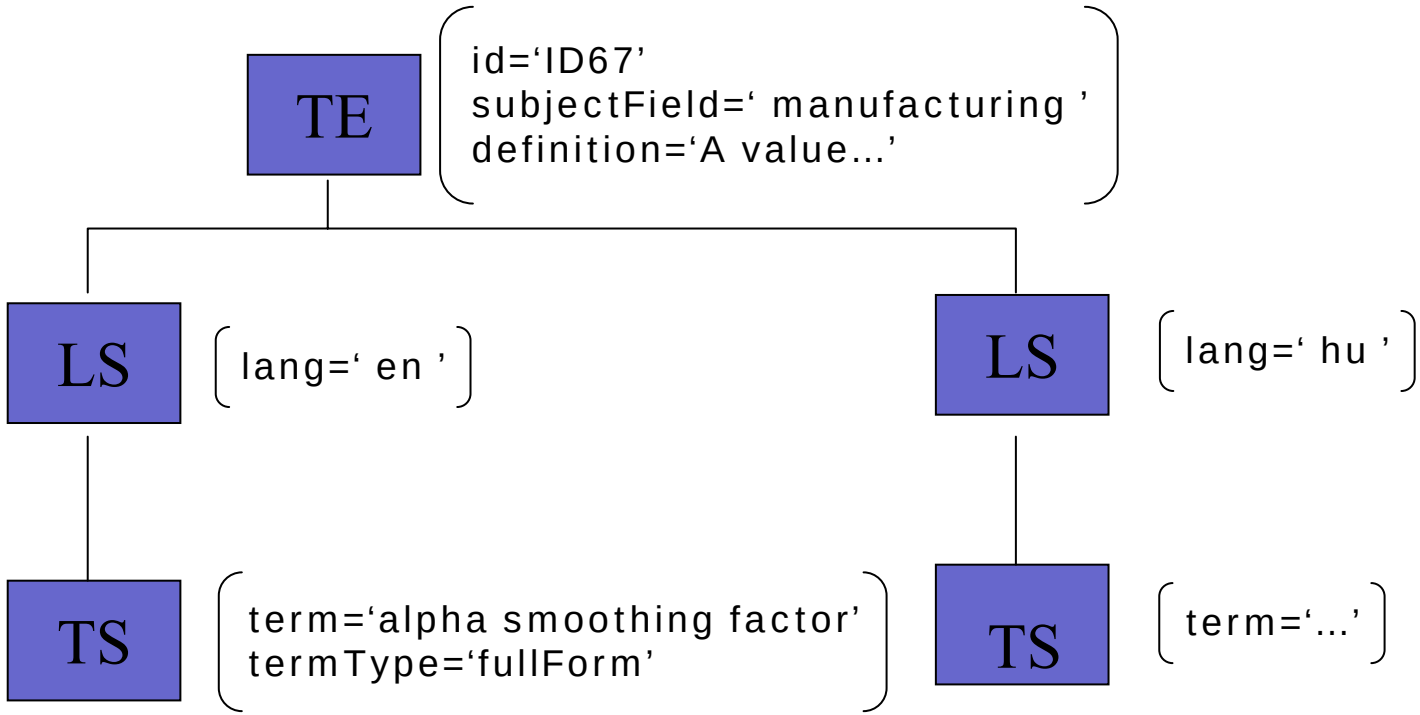
Figure: L. Romary, G. Budin

Prolex

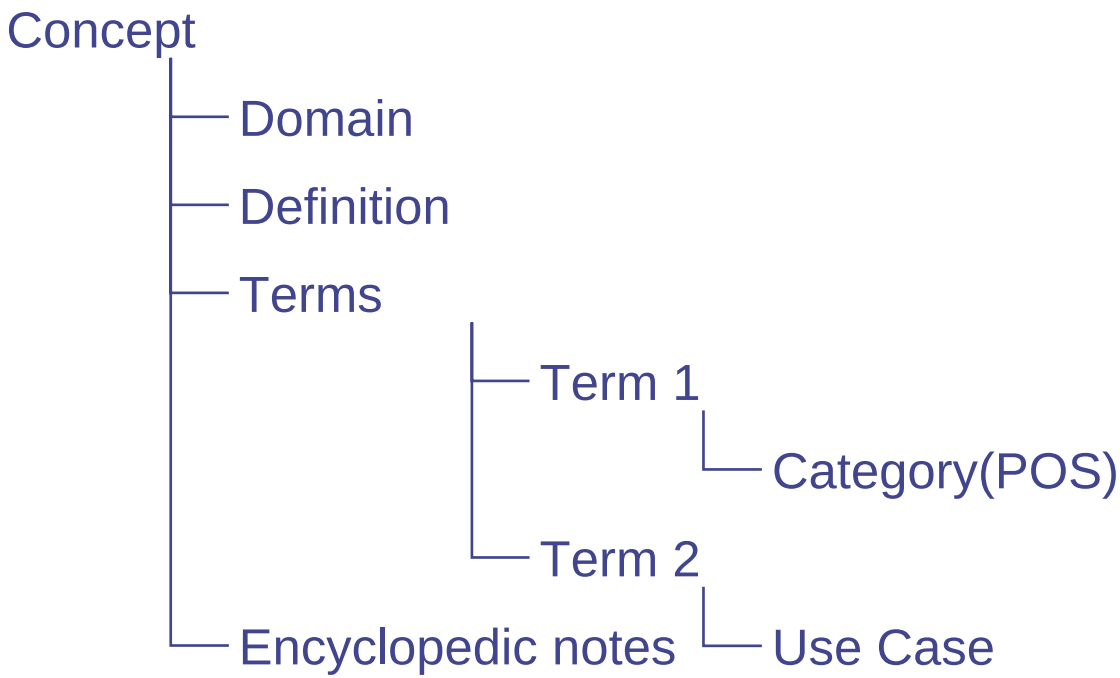
Example in TBX www.lisa.org

```
<termEntry id="ID67">
    <descrip type="subjectField">manufacturing</descrip>
    <descrip type="definition">A value between 0 and 1 used in ...
    </descrip>
    <langSet lang="en">
        <tig>
            <term>alpha smoothing factor</term>
            <termNote type="termType">fullForm</termNote>
        </tig>
    </langSet>
    <langSet lang="hu">
        <tig>
            <term>Alfa ... </term>
        </tig>
    </langSet>
</termEntry>
```

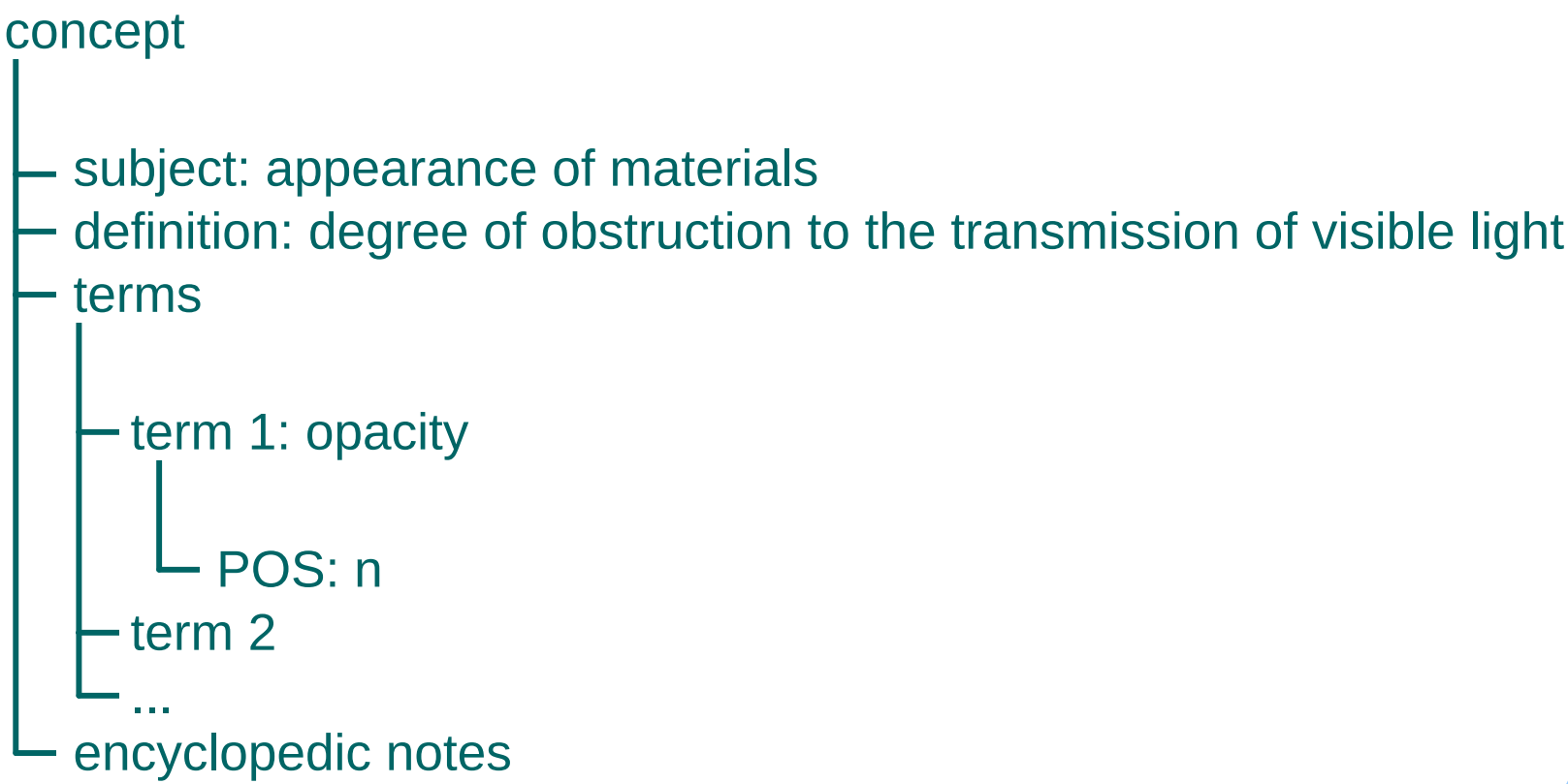
TMF model of previous TBX example



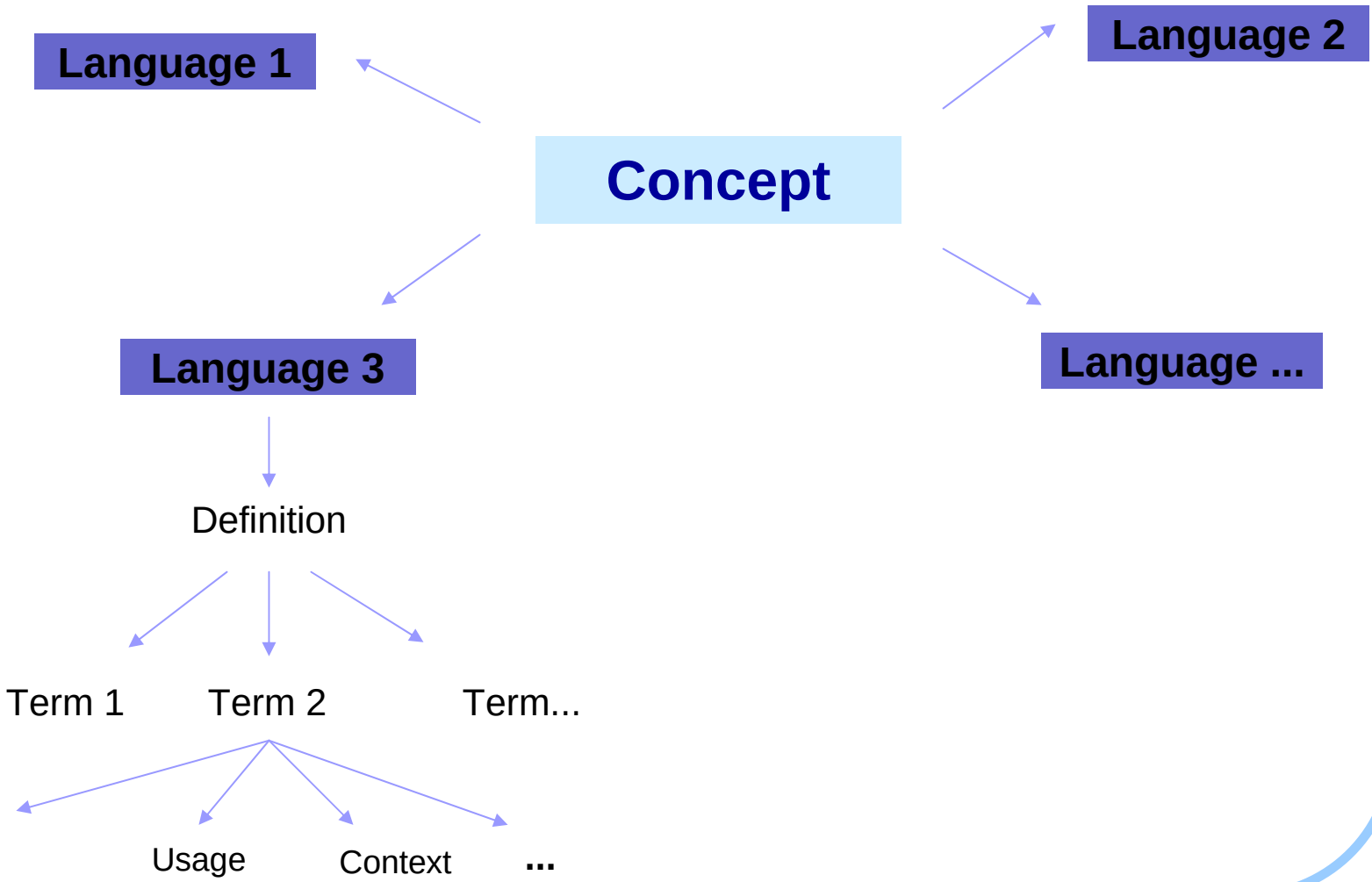
Concept view followed by TMF



Concept view: example



TMF multilingual



LMF skeleton

- All following slides on LMF are taken from:

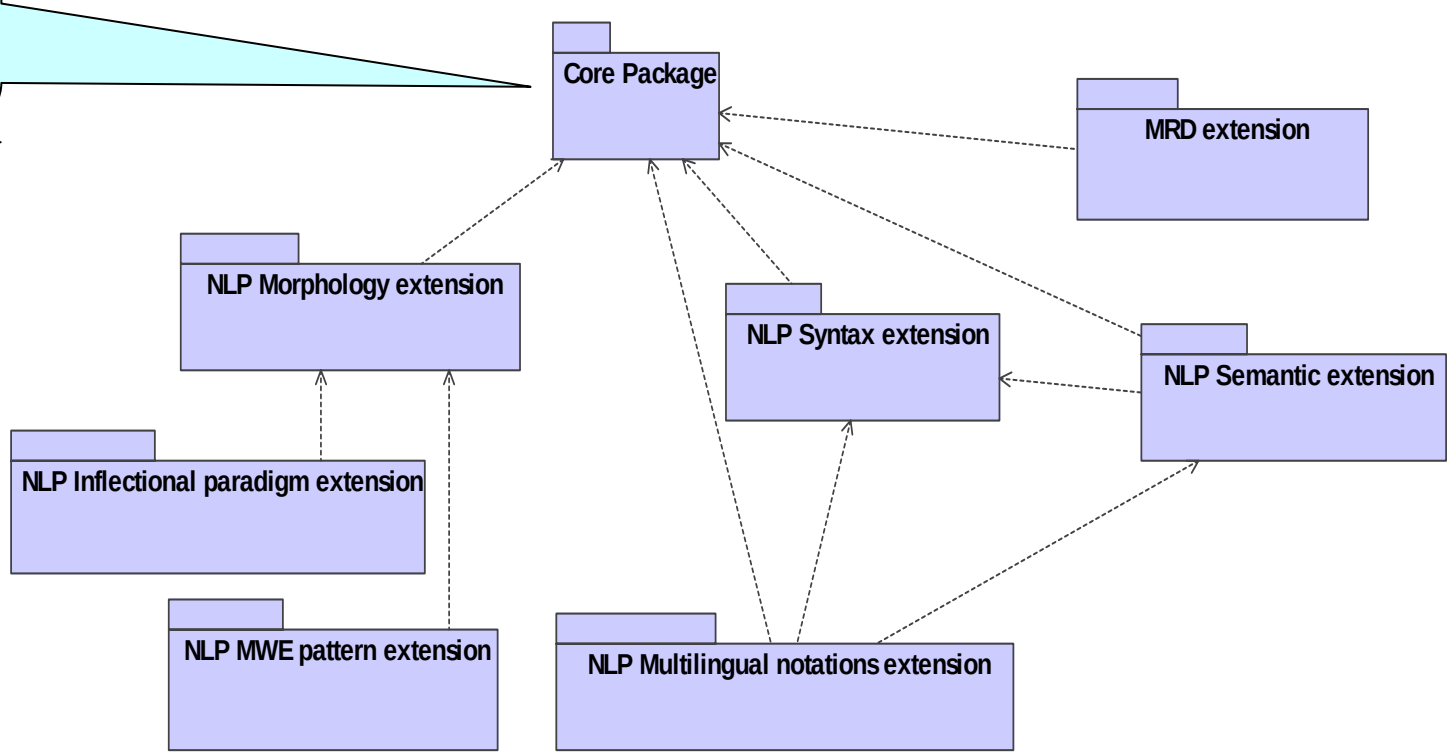
Monica Monachini

CNR-ILC - Pisa

Structure of LMF

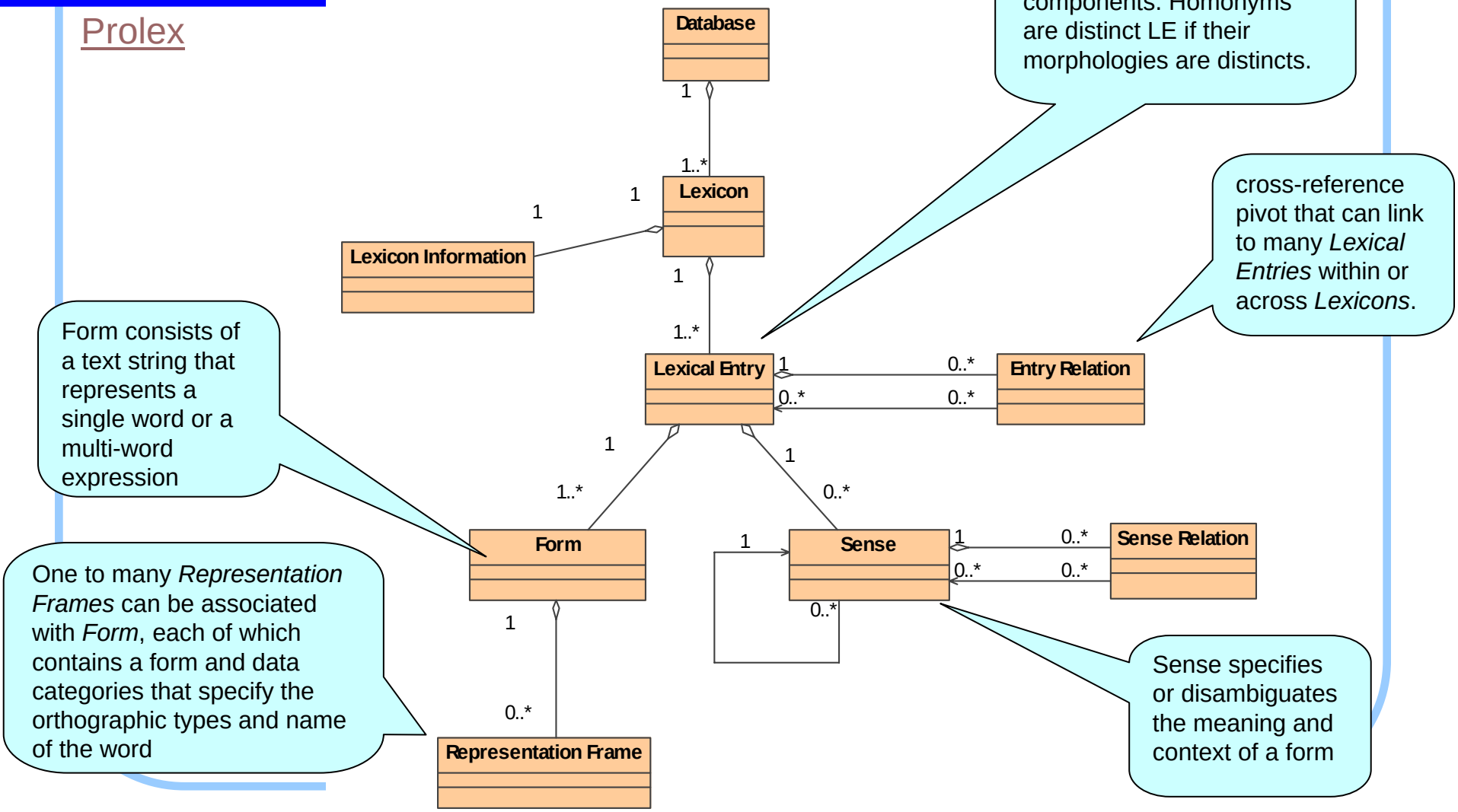
Structural skeleton, with the basic hierarchy of information in a lexical entry

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

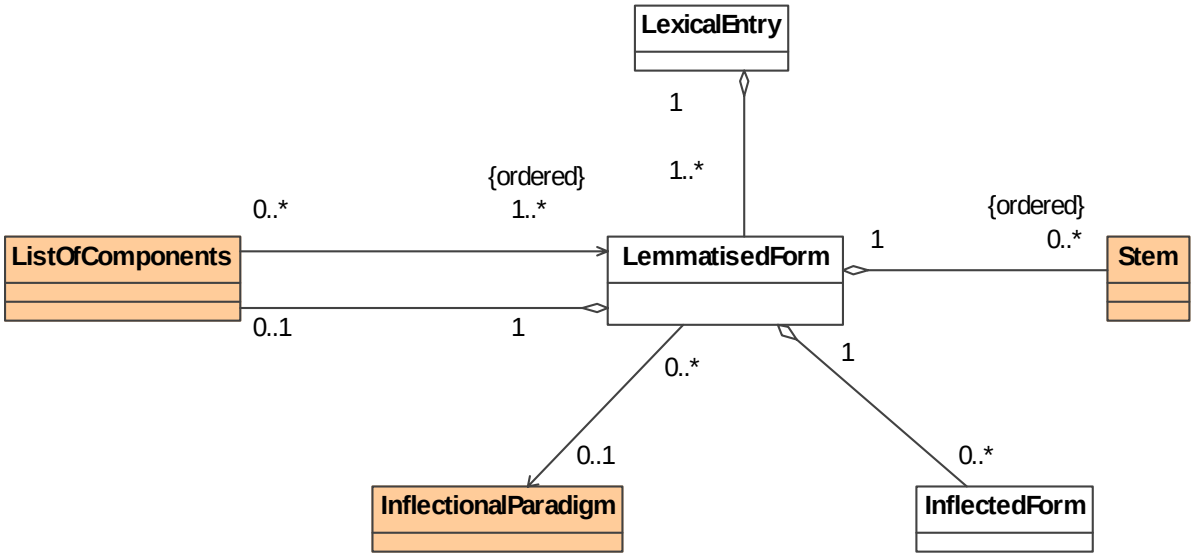


Core package

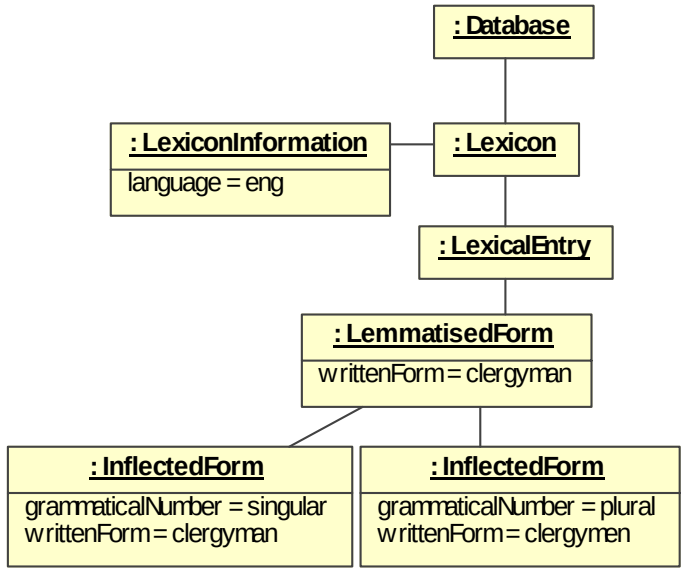
Prolex



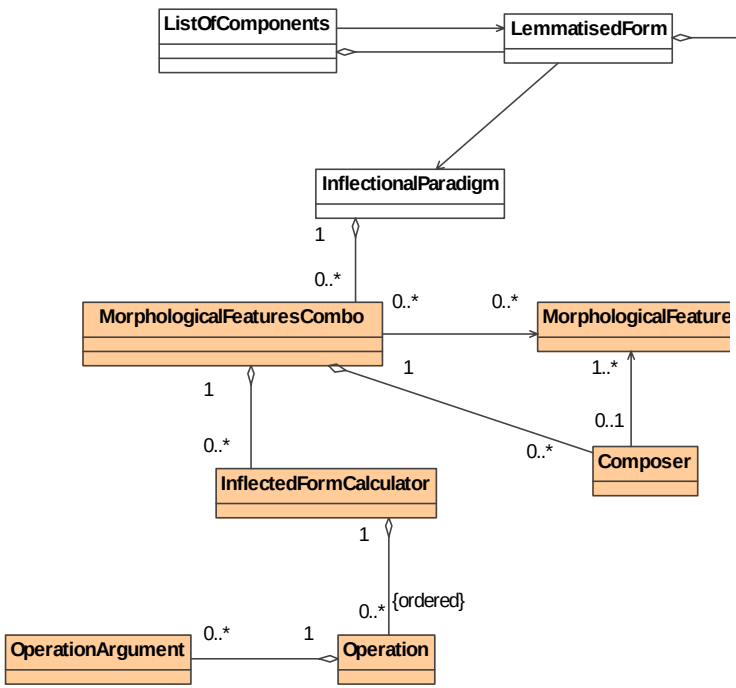
Package for extensional morphology



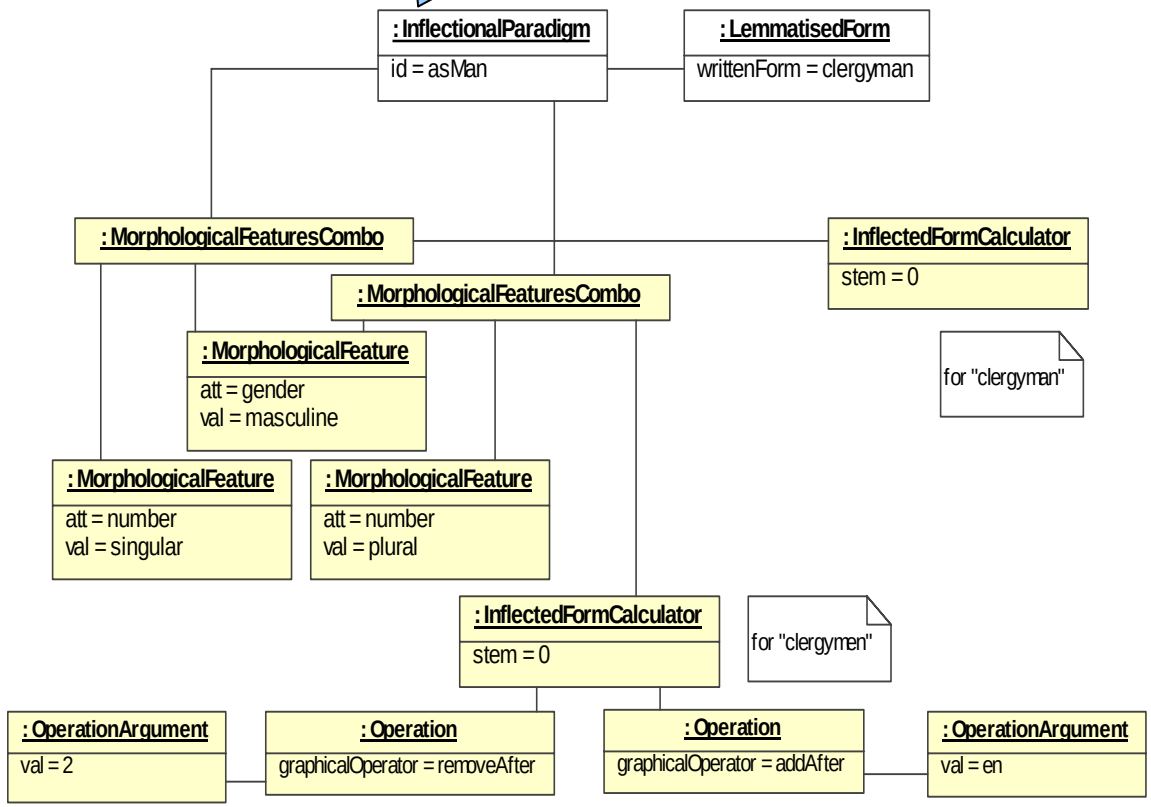
1st strategy:describe the morphology representing explicitly all inflections



Package for inflectional paradigm



2nd strategy:
declare an inflectional paradigm;
use the inflectional paradigm
extension for defining it



Package for NLP syntax

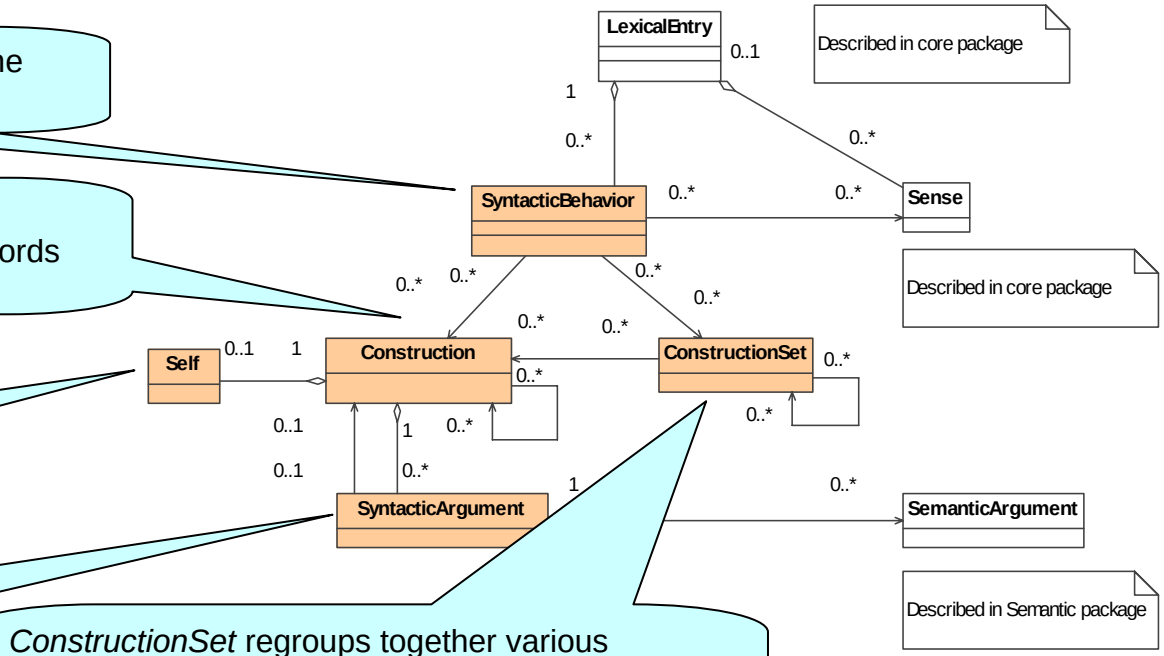
Syntactic behavior represents one of the behaviors of one (or more) senses

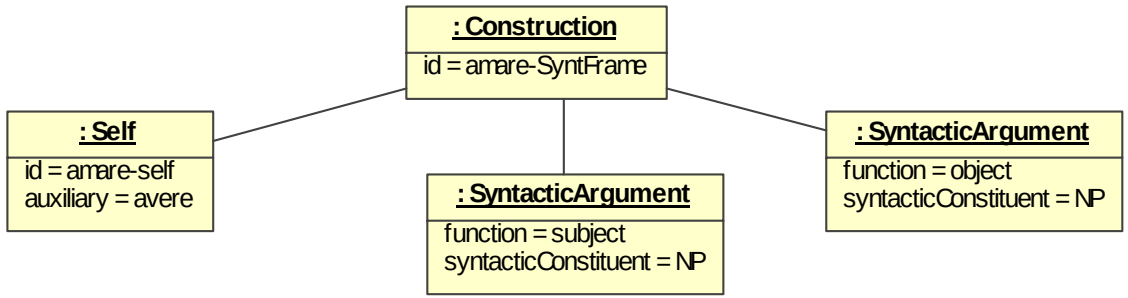
Construction describes one syntactic construction and can be shared by all words with the same syntactic behavior

Self refers to the head lexical entry and describes syntactic properties

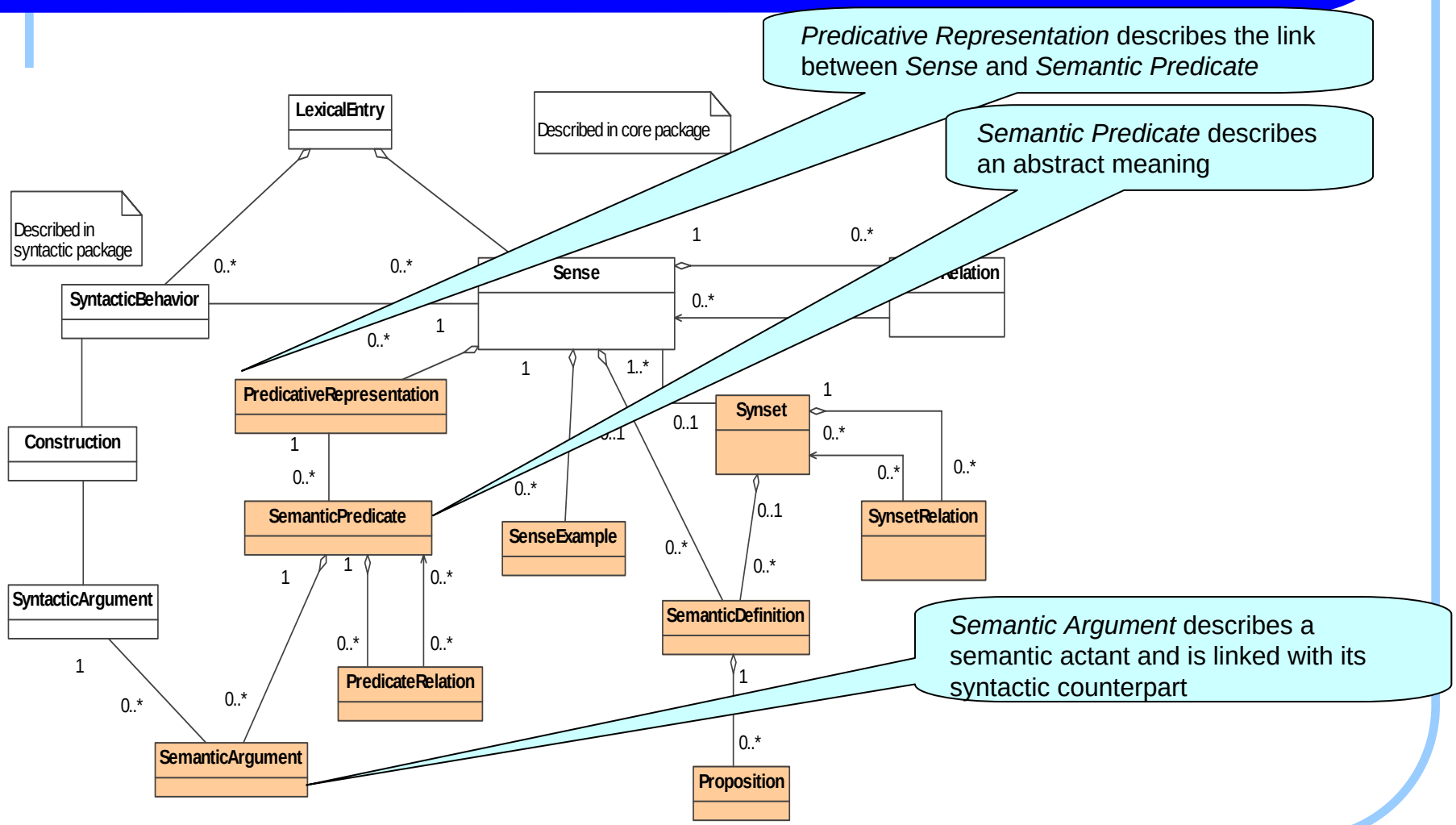
Syntactic Argument describes a syntactic actant

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

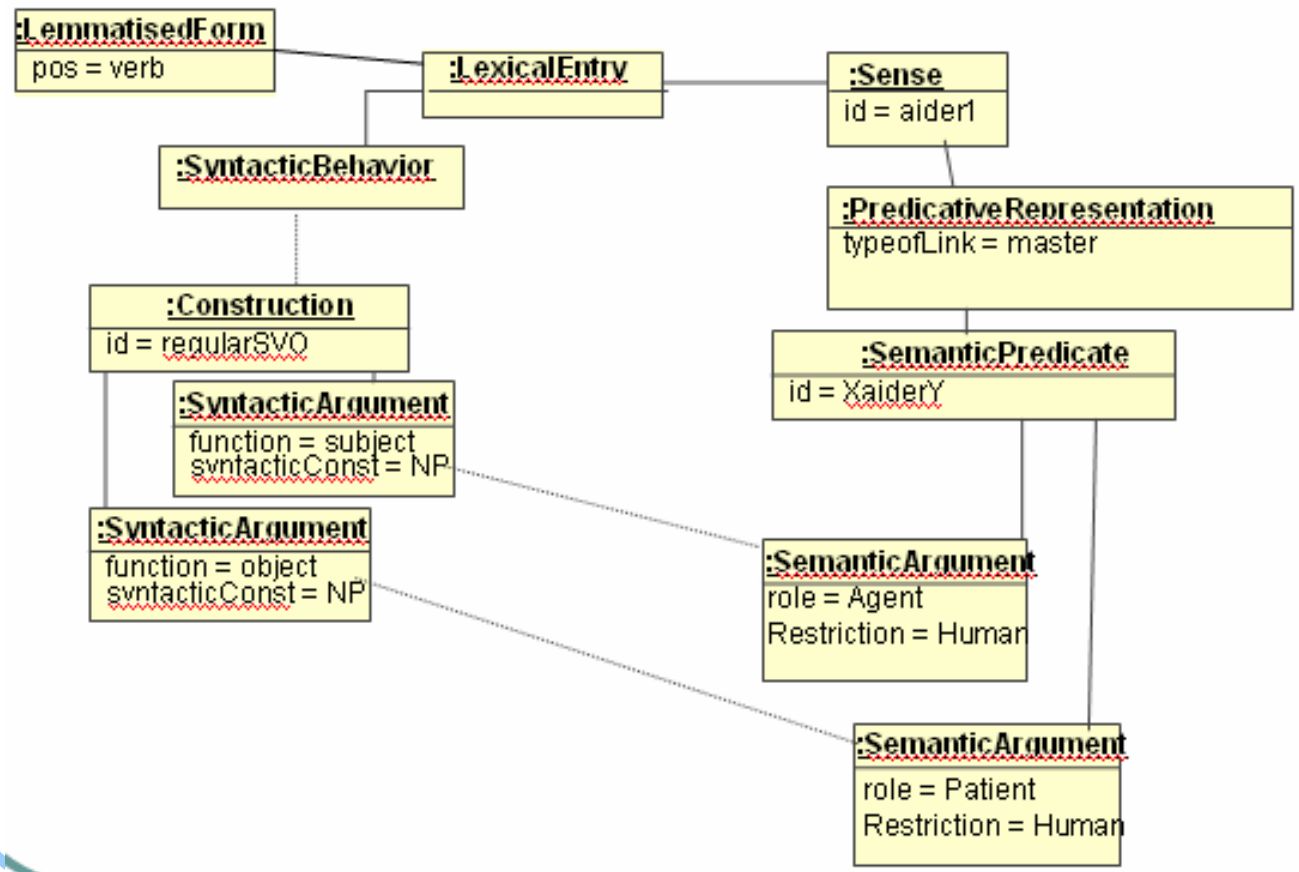




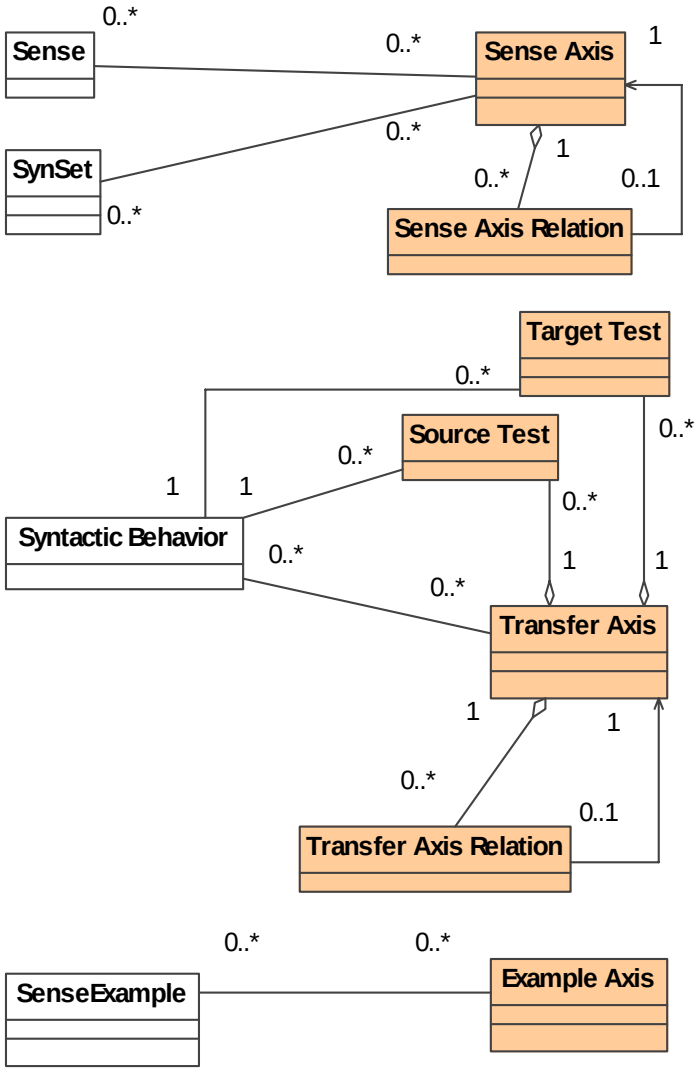
Package for NLP semantics

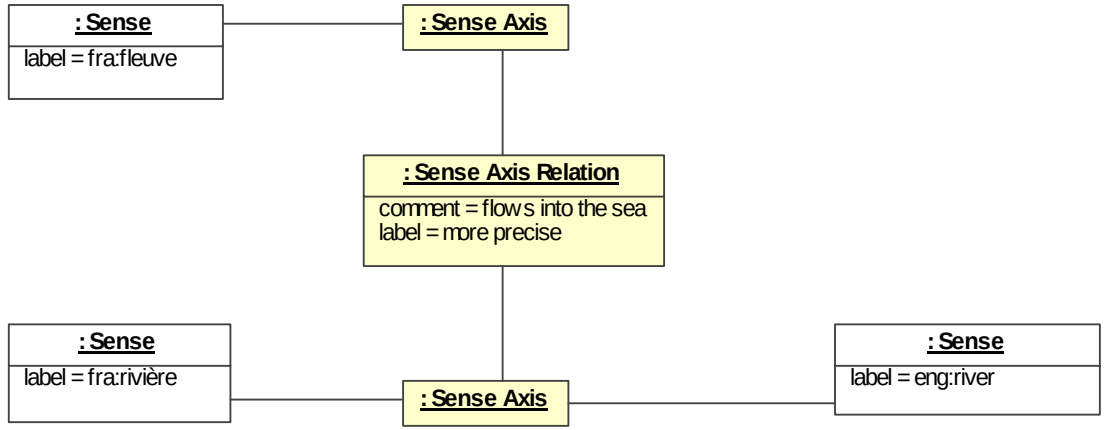


Example (package for semantics)



Package for Multilingual representation

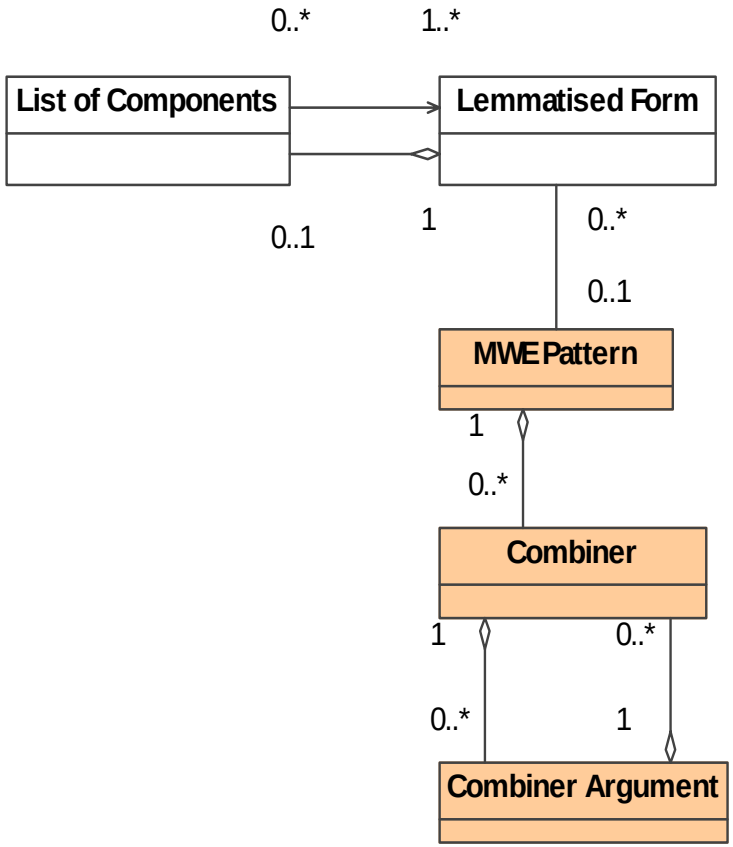




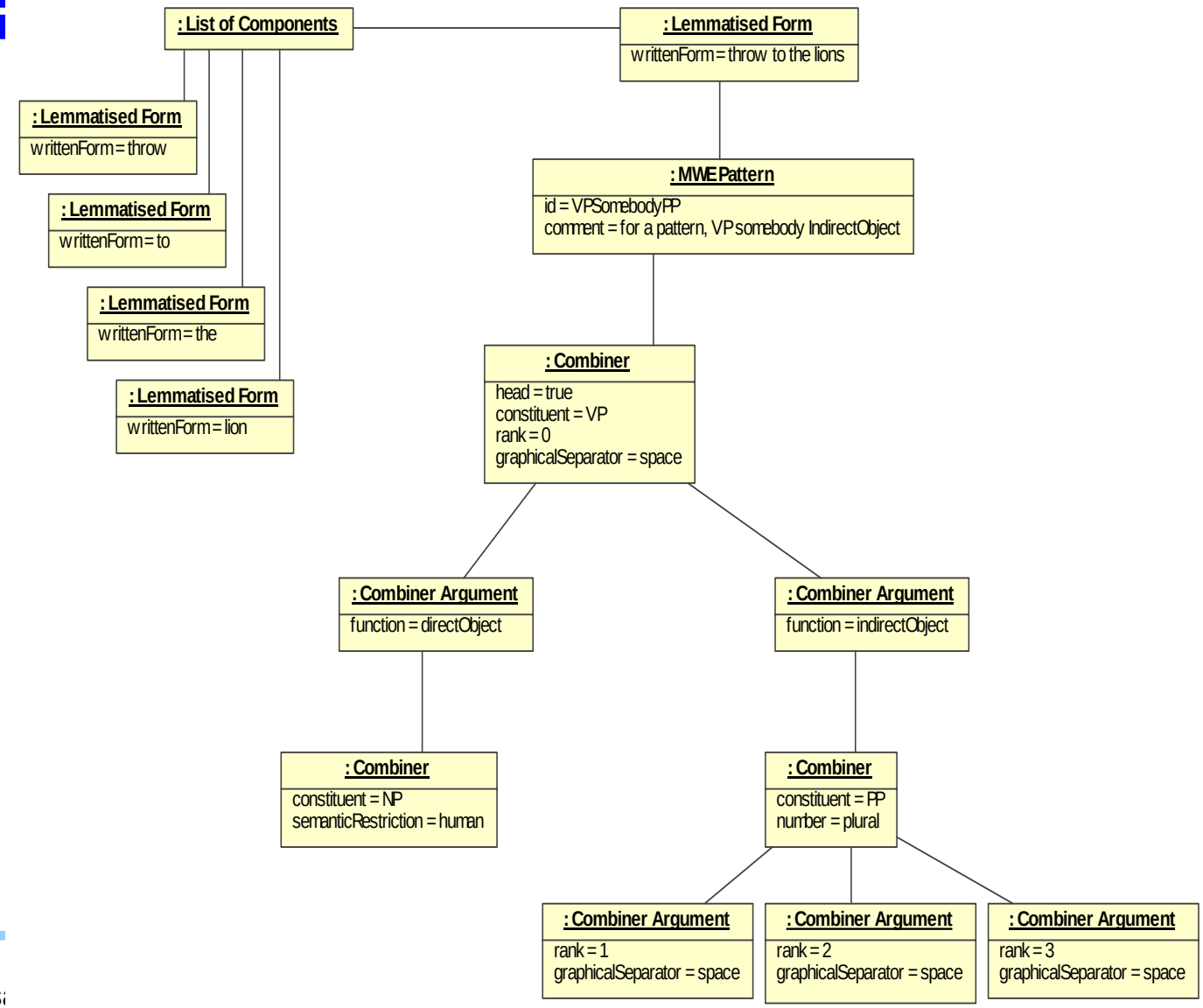
Sense Axis Relation describes the linking between two different *Sense Axis*

Source and *TargetTest* permit to express conditions about the translation on the source/target language side

Package for Multiword expressions



Example (multiword expression)



TMF and LMF (L. Romary)

- Terminology (and ontology)
  - Term: corresponds to a concept in a specific domain
  - Concept → several possible linguistic forms
    - Onomasiological view
- Lexicography
  - Lexical entry + genericity
  - LE → one or several senses
    - Semasiological view

TMF or LMF? (G. Budin)

- In general, a terminological resource is organized into **concept entries**, each of which includes one or more terms designating a particular concept.
- In general, a lexical resource is organized into **lexical entries**, each of which includes one or more senses of a particular lexical item (a word or phrase).
- A concept entry containing multiple terms can be split into multiple lexical entries, one per term...
- Multiple lexical entries associated with the same concept can be combined into one concept entry...

Prolexbase

- A database of proper names and their relationships
- Four levels in the conceptual model:
  - Instances
  - Linguistic

  For each language
  - Conceptual
  - Meta conceptual

  Multilingual (common)
- http://tln.li.univ-tours.fr/tln_prolex/prolex.php

Details on Prolex model

- See the presentation of Denis Maurel in the same seminar…

Prolex features

- Proper name “sense”: a point of view on a referent
- Semantic description is at multilingual level
- The multilingual level is important (it is false that “a proper name do not need to be translated”)
- The prolexeme groups several terms, these terms are related to the same pivot (“sense”)

Prolex features

- Importance of derivations performed from the prolexeme:
  - derivatives are related to the same point of view on the same referent as their prolexeme
- The complete linguistic description (including morphology, syntax, and semantics):
  - extensional (the instance layer)
  - intensional (inflection paradigms, rules for aliases and derivatives).

ISO Standard for Prolex

- LMF ?
  - Proper names are not a « specialized language » (as a terminology)
  - Complete linguistic description
  - But special « sense », no semantic description at lexical entry level
  - But derivative hierarchy not easy to represent in LMF

ISO Standard Prolex: Meta-model

- TMF ?
  - Proper names are special citizens of languages: pivots resemble concepts
  - Semantics defined at concept level
  - Multilingual links via concepts
  - But: complete lexical information not easy to express
  - But: derivative hierarchy difficult to represent in TMF

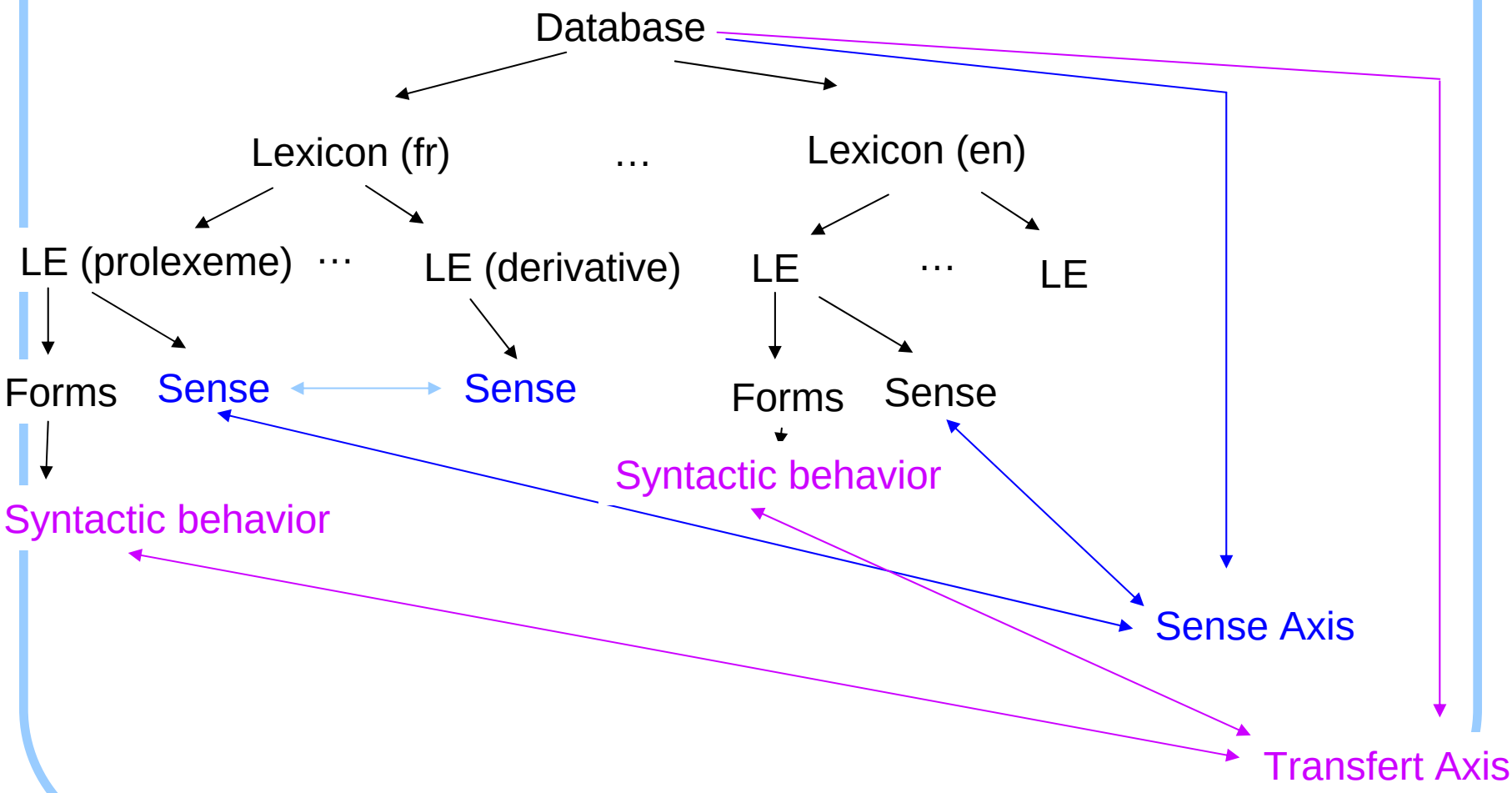
Prolex in LMF (linguistic – instance)

- Prolexeme → Lexical Entry
  - the Sense part of this LE is « prolexeme » (+ number)
- Alias → Form of the LE that represents the corresponding prolexeme
  - with the Related Form class of Morphology extension
- Derivative → Lexical Entry
  - the Sense part of this LE is the relation name (e.g. « relational adjective ») + its « parent » number
- Links of « derivation » (from « parent » to « child »): Entry Relation (?)
- Instance → Form of the LE that represents the corresponding prolexeme or derivative
  - with the subclasses Lemmatised Form and Inflected Form

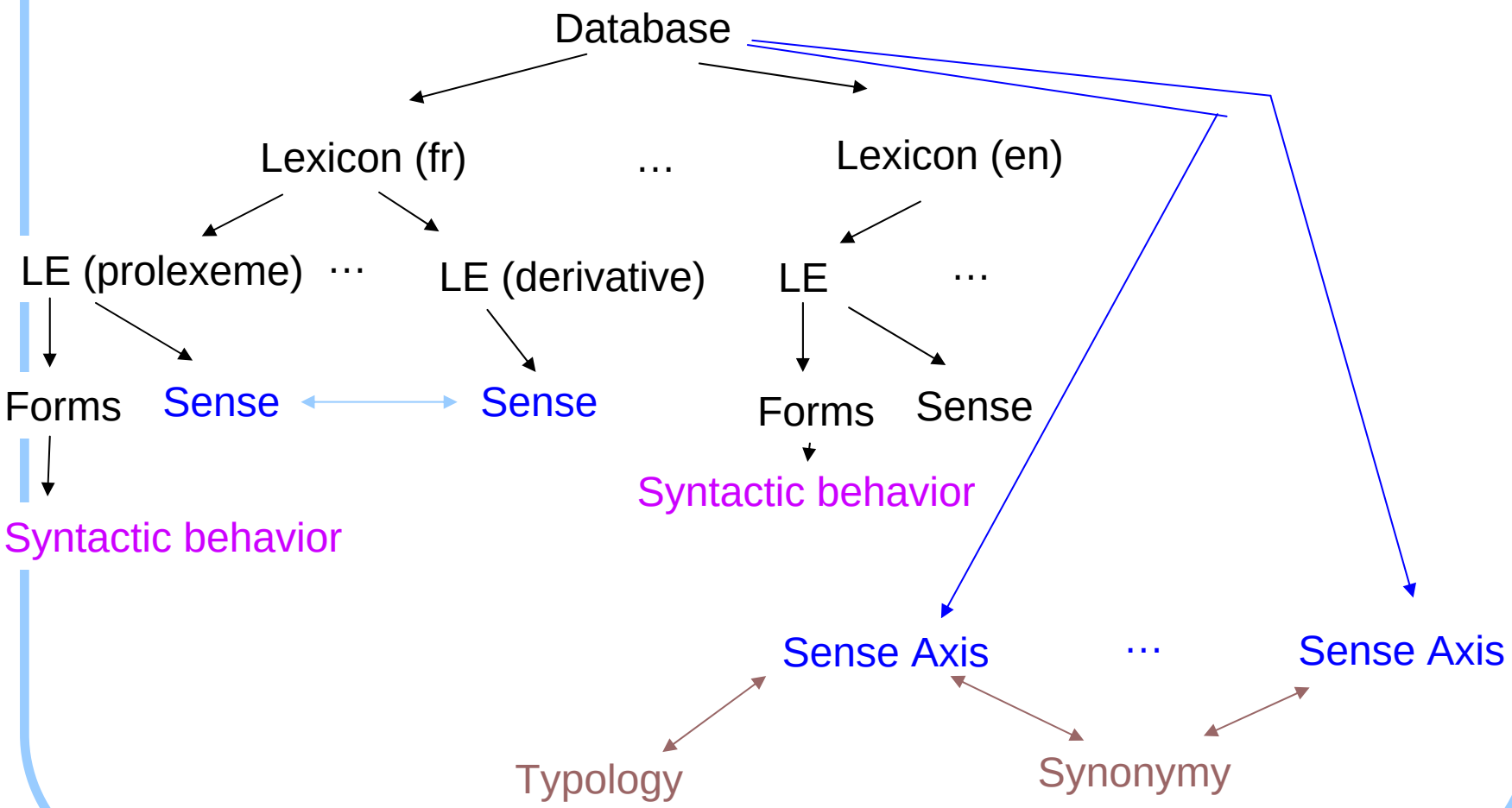
Prolex in LMF (conceptual – meta)

- Pivot → Sense Axis of Multilingual extension
  - Links prolexemes (via their sense part)
- Conceptual relations → Sense Axis Relation of Multilingual extension + data categories
- Type → Sense Axis Relation + Interlingual External Ref of Multilingual extension.
- Language name → Lexicon Information part
- Multilingual links → Axis of Multilingual extension

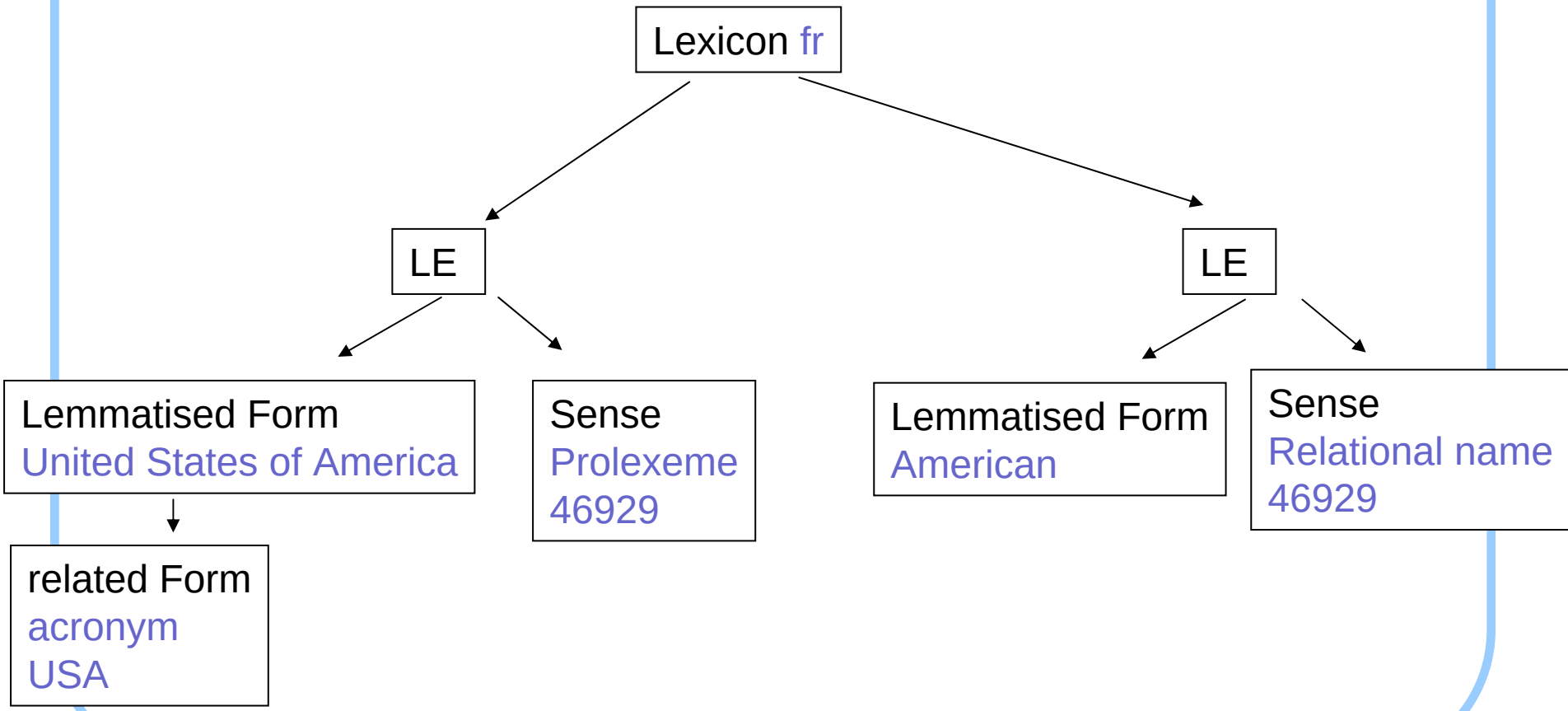
Prolex LMF Abstract Model



Prolex LMF Abstract Model



Prolex LMF Example



Prolex in TMF (linguistic – instance)

- Each instance of prolexeme → Term Section (TS)
- Each instance of derivative → Term Section (TS)

  under the same TE as their prolexeme

  A data category (\eg "associative relation") can link the prolexeme with its derivatives (and derivatives with their derivatives)

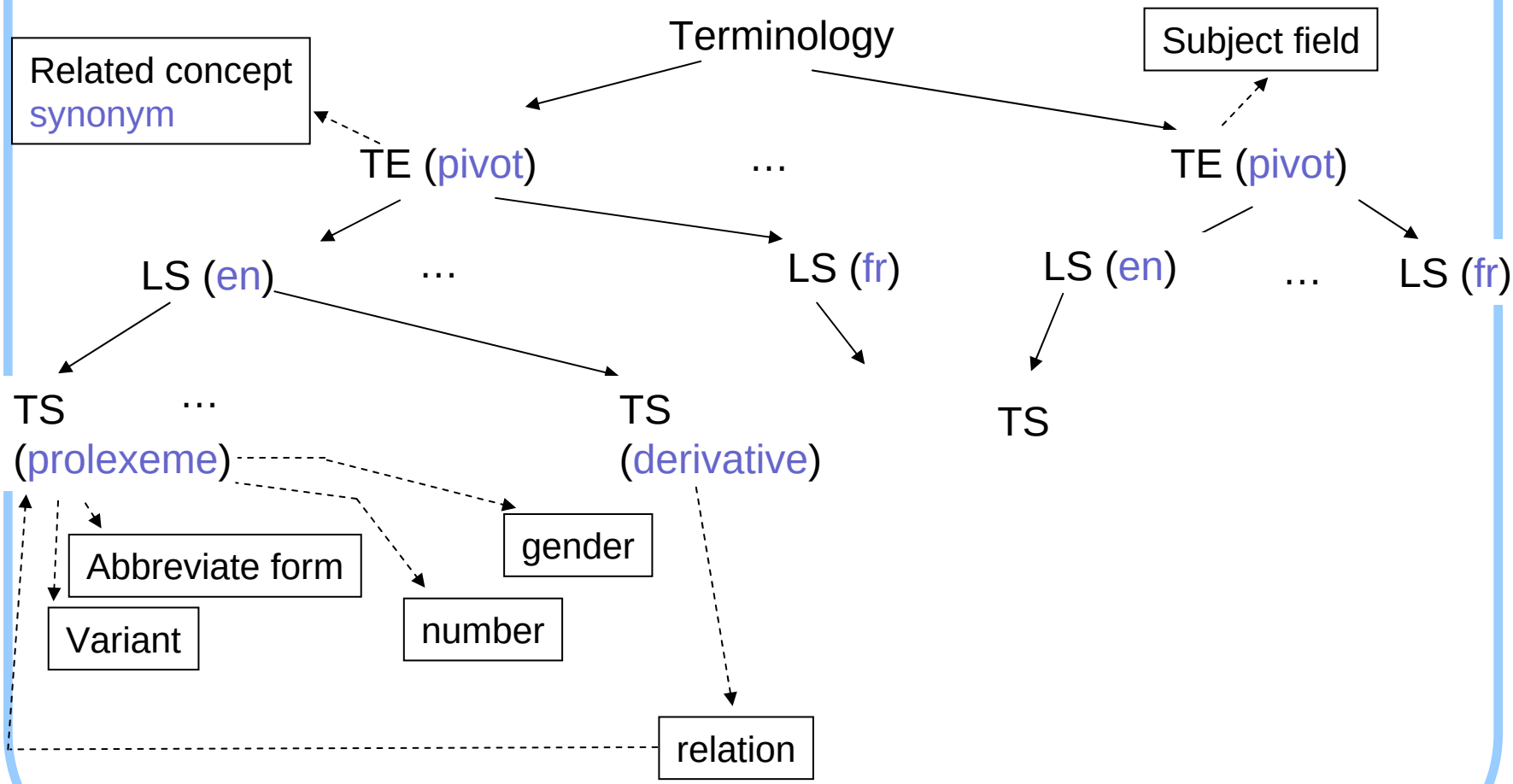
  Data categories: "grammatical gender", "grammatical number" are associated to TS

- Alias → Data category attached to the TS of the prolexeme: "abbreviated form", "acronym", "variant", etc.

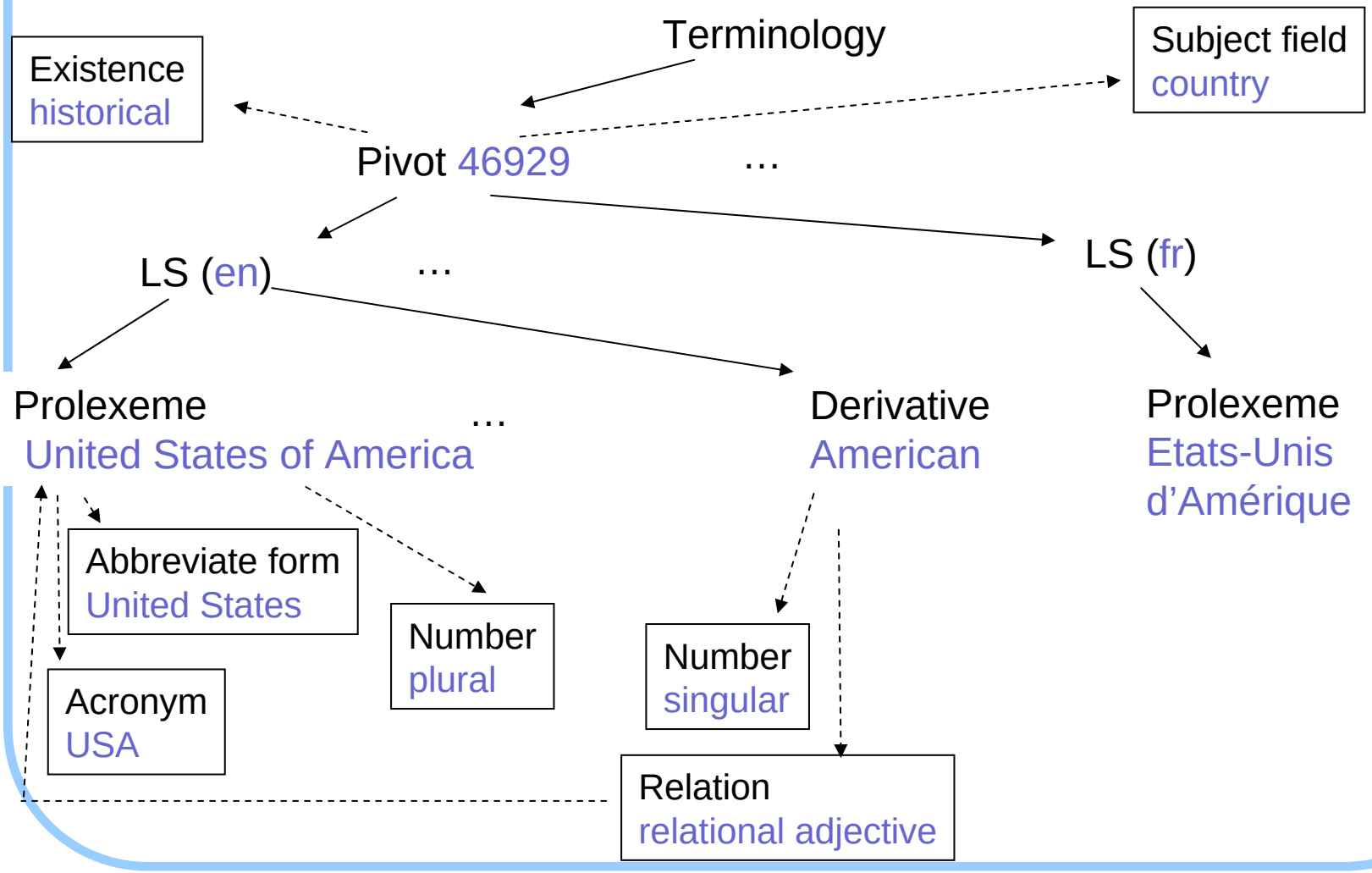
Prolex in TMF (conceptual – meta)

- Pivot → Terminological Entry (TE)
- Conceptual relations → Data categories: "related concept", "concept relation" …
- Type → Data categories: "subject field", "broader concept"…
- Language → Language section (LS)
- Multilingual links → Via the TEs, plus data categories attached to the LS instances

Prolex TMF abstract model



Prolex TMF example



What about XML?

- XML design is not so simple
- XML design is not mature
- Proposal in NLDB 2005
  - Database is (i) a multilingual part and (ii) a set of languages (not so far from LMF…)
  - Main entries in multilingual part : pivots
  - Main entries in language part : prolexemes
- From conceptual (abstract) model to a logical model : many solutions for XML…

Representing relations*:

- Polysemia
  - LMF : several senses of a lexical entry
  - TMF : not considered (a term is related to only one concept…)
  - Prolex : as TMF, one pivot for each point of view on a referent
- Synonymy
  - LMF : Links from one LE (one sense) to another one
  - TMF : in concept structure*
  - Prolex : conceptual relations between pivots

(*L. Romary)

XML Instances of ISO models

- I did not find any report on discussions
- XML « instances » of TMF: TML
  - MSC (MARTIF with Specified Constraints)
  - Geneter
  - GMT: a tool for comparing two TML…
- XML « instances » of LMF:
  - XML schemas: a DTD + a RelaxNG specification
  - To be analysed

Conclusion

- A lot of work remains to be done…