

Le projet Prolex de traitement automatique des noms propres

Denis MAUREL
Université François Rabelais Tours
LI (laboratoire d'informatique)

Le projet Prolex

- Des travaux débutés en 1991
- Un nom présenté en 1996... à Grenoble !

Le projet Prolex

- Des travaux débutés en 1991
- Un nom présenté en 1996... à Grenoble !

Maurel D., Belleil C., Eggert E., Piton O. (1996),
Le projet PROLEX, séminaire *Représentations et Outils pour les Bases Lexicales, Morphologie Robuste* de l'action *Lexique* du GDR-PRC CHM,
(Actes p. 164-175), Grenoble, 13-14 novembre

Le projet Prolex

- Des travaux débutés en 1991
- Un nom présenté en 1996... à Grenoble !
- 4 thèses de doctorat et de nombreux stages de maîtrise et de DEA, puis de master...

Le projet Prolex

Friburger N. (2002), Reconnaissance automatique des noms propres ; application à la classification automatique de textes journalistiques

1991

6... à

- 4 thèses de doctorat et de nombreux stages de maîtrise et de DEA, puis de master...

Le projet Prolex

- Des travaux débutés en 1991
- Un nom présenté en 1996... à Grenoble !
- 4 thèses de doctorat et de nombreux stages de maîtrise et de DEA, puis de master...

Tran M. (2006), Prolexbase. Un dictionnaire relationnel multilingue de noms propres : conception implantation et gestion en ligne

Introduction

Trois constats

1. Les noms propres ne figurent que rarement dans les bases de données lexicales utilisées pour le traitement automatique des langues
2. Les noms propres sont souvent la base de dérivations morphologiques

Trois constats

3. Dans un contexte multilingue, ces formes dérivées sont en nombre plus ou moins important suivant les langues et n'ont pas forcément de correspondance dans une langue cible

Exemples

- Les textes journalistiques contiennent 10% de noms propres (Coates-Stephens)
- Dans le *Tour du monde en quatre-vingts jours* de Jules Verne, les quatre noms les plus fréquents sont des noms propres...

Exemples

<i>Nom</i>	<i>Fréquence</i>	<i>Place/Noms</i>	<i>Place/Mots</i>
Fogg	655	1	13
Passepartout	423	2	26
Phileas	301	3	36
Fix	274	4	40
heures	243	5	46



Exemples

Nom lemmatisé	Place/Noms lemmatisés	Fréquence du lemme
Fogg	1	670
Passepartout	2	437
heure	3	317

Une cascade pour la reconnaissance des entités nommées

(thèse de Nathalie Friburger)

Les entités nommées

- MUC 7 (1997)
 - Conférence sur la recherche d'information
 - Définition un nouveau concept à cheval sur la linguistique et la technologie...

Les entités nommées

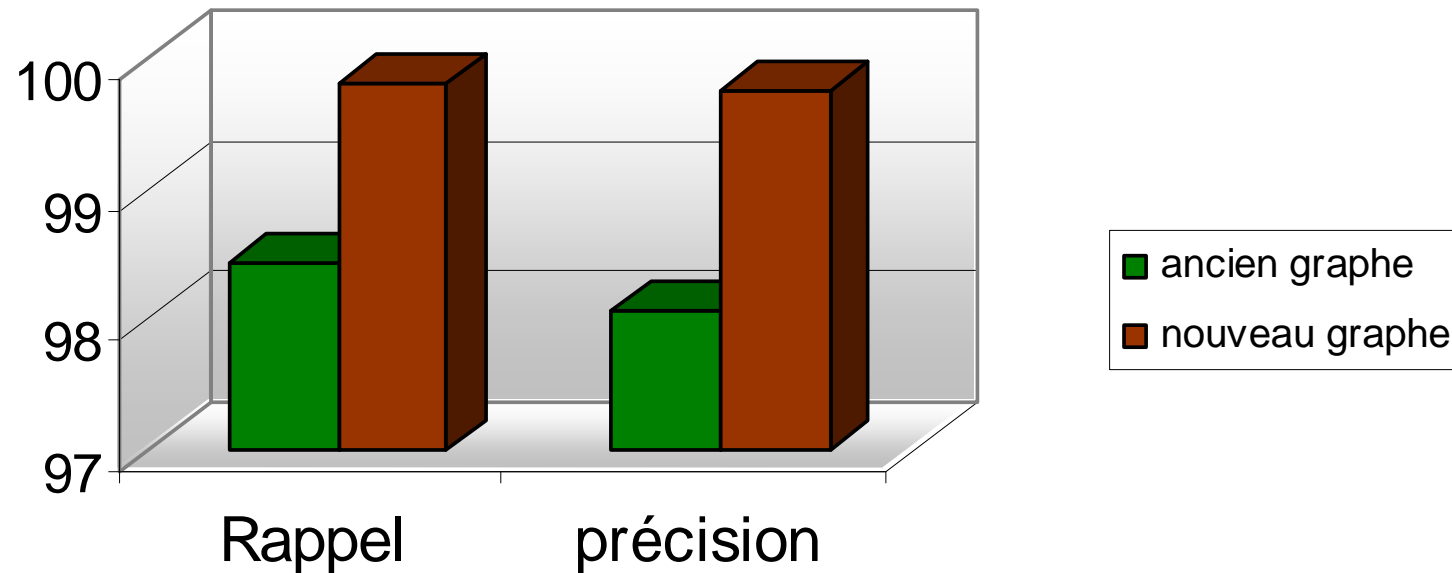
- ENAMEX
Personnes/Lieux/Organisations
- TIMEX
Dates/Heures
- NUMEX
Pourcentages/Valeurs monétaires

Le découpage en phrase

- Une phrase commence par une majuscule et se finit par un point...
 - J'ai demandé à M. A. Dupont. Il m'a répondu...
 - Une manifestation unitaire C.G.T.-C.F.D.T. Voilà l'évènement de ce 1er mai...
 - Cela coûte 100 F. Dupont trouve cela trop cher.
 - Je prends de la vitamine C. Dupont me l'a déconseillé.

Le découpage en phrase

Vérification sur 4,5 Mo de textes variés



Les noms de personne

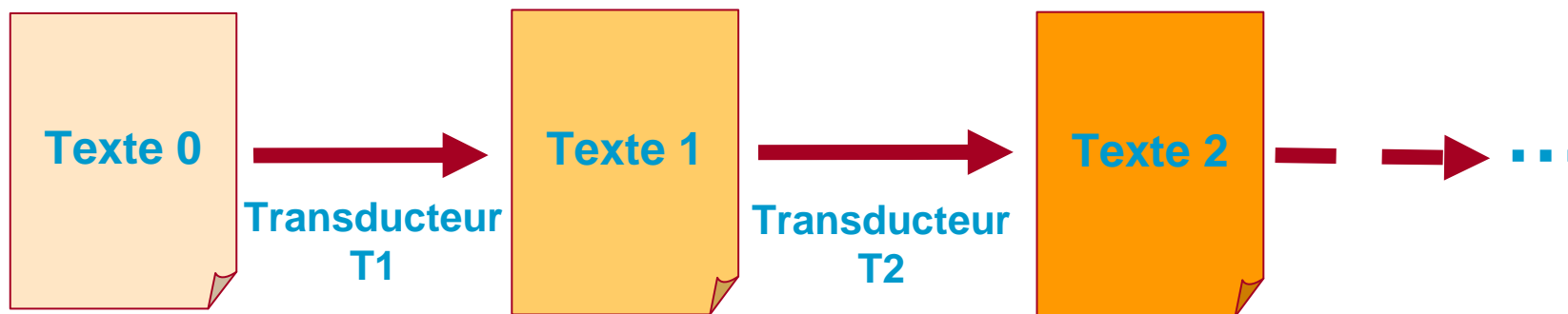
- Preuves externes
 - Les civilités
 - *Madame, Monsieur, Mme, M. etc.*
 - Les titres de toutes sortes
 - *ministre, lieutenant, évêque, etc.*
 - Les noms de professions
 - *juge, architecte, menuisier, etc.*

Les noms de personne

- Preuves internes
 - Les prénoms
 - Dictionnaire de prénoms
 - Description de la structure des prénoms
Jean, Jean-Pierre, J.-P., George W., etc.
 - Les patronymes
 - Patronymes monolexicaux
 - Patronymes polylexicaux
 - français : Dupont de Nemours, de la Fontaine
 - étranger : Mac Donnell-Douglas, O'Ryan, von Bulow, Da Silva, etc.

Cascades de Transducteurs

Une succession de transducteurs appliqués sur un texte



Résultats

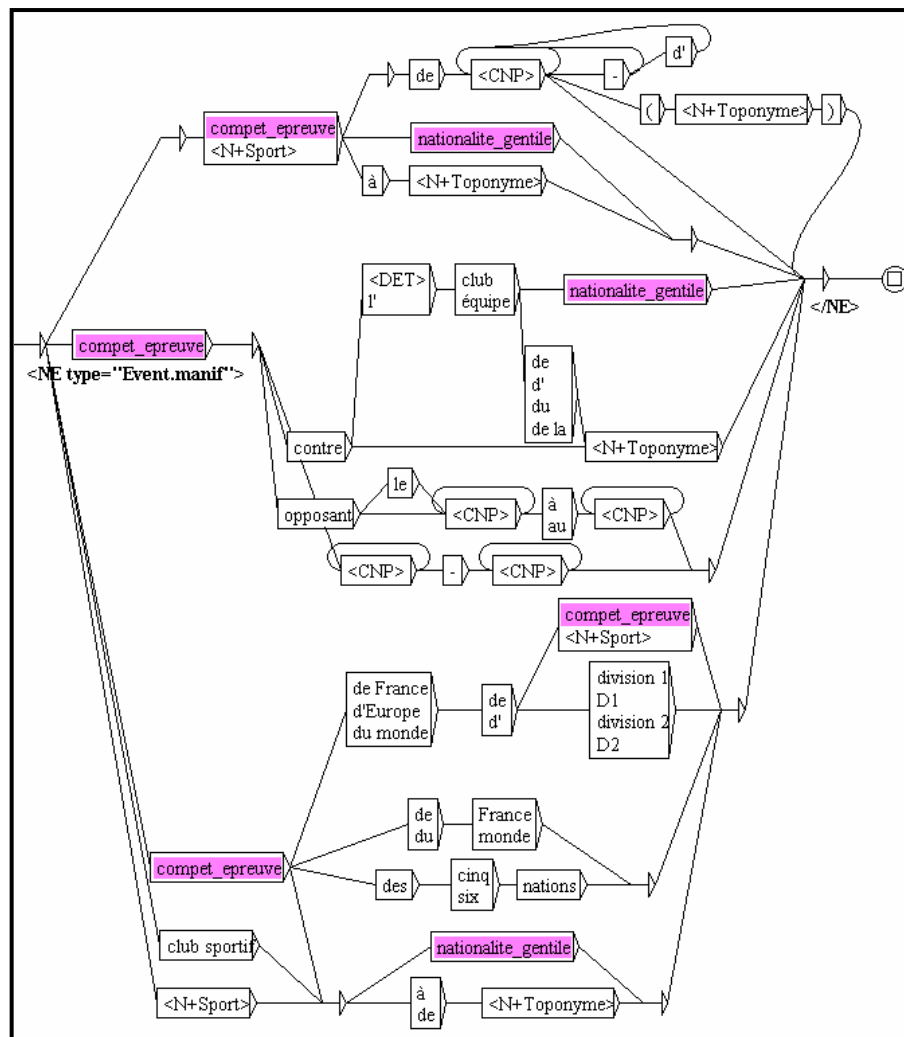
2 numéros du journal *Le Monde*, soit 893 Ko

	Personnes	Organisations	Lieux
Rappel	84,7	78	95,2
Précision	97,5	94,7	95,0

	Personnes	Organisations	Lieux
Rappel	93,7	88,0	96,4
Précision	96,1	92,3	94,5

	<i>Le Monde</i>
Rappel global	93,2
Précision globale	94,4

Graphe "locatif et compétition"



Exemples d'annotation

- dans les `<NE type="Sigle">PTT</NE>`
- moi je suis native de`<NE type="loc.admi">Pithiviers</NE>` j'aime mieux `<NE type="loc.admi">Orléans</NE>`
- oh j'ai une `<NE typr="prod.art">encyclopédie Quillé</NE>` j'ai le
- `<NE type="time.date.abs">en dix-neuf cent trente-huit</NE>`
- je crois que le `<NE type="org.pol">ministère de l'Education National</NE>`

Exemples d'annotation

- euh passer quelques jours sur la <NE type="loc.geo">Côtes d'Azur</NE>
- euh je suis je travaille à l'<NE type="loc.fac">hôpital d'Orléans</NE> quoi
- parce que nous avons un <NE type="loc.fac">magasin Phildar</NE> juste en face de chez nous
- dans la<NE type="loc.line"> rue Royal</NE> euh

Prolexbase

Un dictionnaire relationnel multilingue de noms propres

Thèse de Mickaël Tran

Prolexbase

- Pour traiter efficacement la question de l'existence de lemmes et de formes en nombre différent d'une langue à une autre, Prolexbase est construite autour:
 - d'un concept multilingue(un pivot)
 - de sa projection sur chaque langue en un ensemble de lemmes (et de formes) morphosémantiquement liés

le *prolexème* (et les *instances*)

Exemple

- *Passepartout*
 - $\text{Prolexème-Fra}_{\text{Passepartout}} = \{\text{Passepartout.N}\}$
 - $\text{Prolexème-Srp}_{\text{Paspартu}} = \{\text{Paspартu.N}, \text{Paspартuov.AP}\}$
 - $\text{Instances-Fra}_{\text{Passepartout}} = \{\text{Passepartout}\}$
 - $\text{Instances-Srp}_{\text{Paspартu}} = \{\text{Paspартu}, \text{Paspартua}, \text{Paspартuom}, \text{Paspартuu}, \text{Paspартuov}, \text{Paspартuova}, \text{Paspартuovih}, \text{Paspартuovim}, \text{Paspартuovo}, \text{Paspартuovu}\}$

La structure de Prolexbase

1. Le niveau indépendant de la langue

Nom commun/nom propre

- Un nom commun se définit par son (ou ses) sens
- Les sens ne sont pas universels, les concepts qu'ils représentent peuvent être raffinés dans une langue et non dans une autre
 - *rivière, fleuve / river*
- Un nom propre désigne son référent
- Le référent ne dépend pas de la langue, pas plus que ses relations avec d'autres référents
- Mais... un même référent peut correspondre à plusieurs noms propres

Exemple

Cité phocéenne

- Si ce terme est supposé suffisamment clair pour un lecteur en langue cible
 - *Phocean city*
- Sinon, il faudra le remplacer par son synonyme
 - *Marseille*
- Éventuellement accompagnée d'une glose
 - *The French city of Marseille*

Le pivot interlingue

- Le pivot n'est pas le référent, mais un *point de vue* sur le référent
- Les différents points de vue sur un même référent sont considérés comme des synonymes

Le pivot interlingue

- Les synonymes sont marqués en suivant la diasystématique de Coseriu
 - Synonymes diachroniques (variété dans le temps)
 - *République populaire de Pologne*
 - *République de Pologne*

Le pivot interlingue

- Les synonymes sont marqués en suivant la diasystématique de Coseriu
 - Synonymes diastratiques (variété relative à la stratification socioculturelle)
 - *Paul-Alain Leclerc*
 - *Julien Clerc*

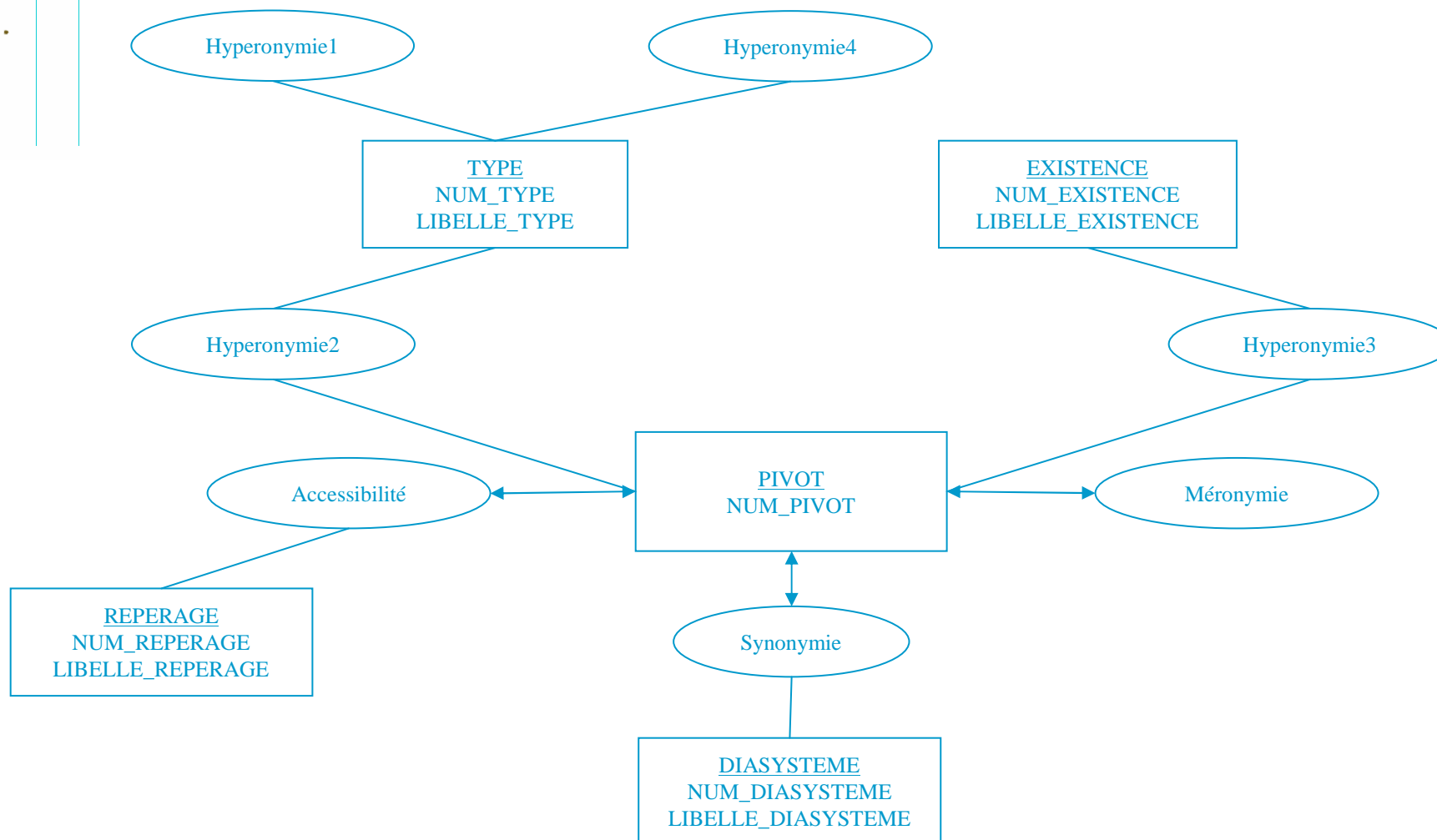
Le pivot interlingue

- Les synonymes sont marqués en suivant la diasystématique de Coseriu
 - Synonymes diaphasiques (variété concernant les finalités de l'emploi)
 - *la Cité phocéenne*
 - *Marseille*

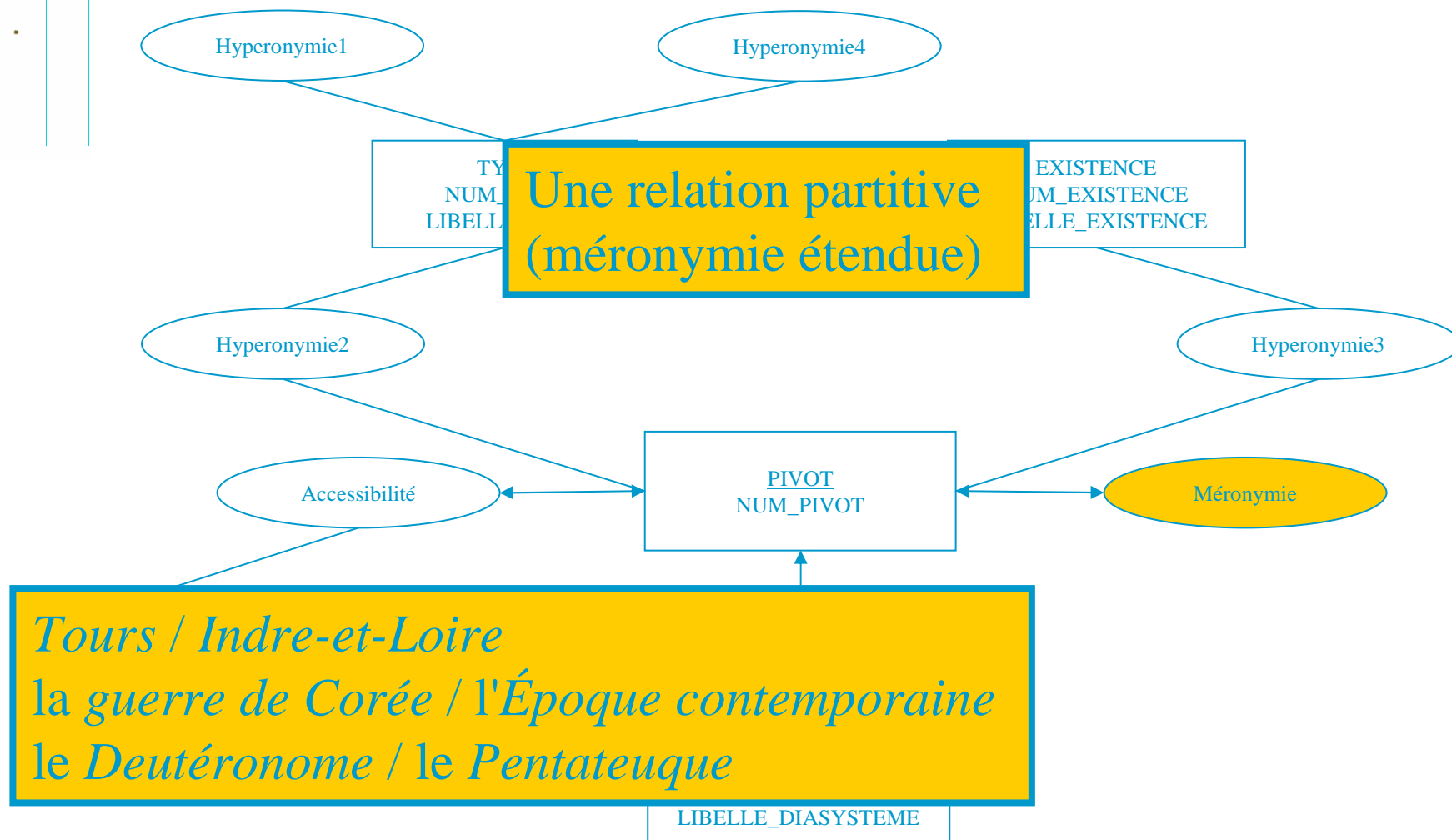
Et le référent?

- D'une certaine manière, le référent d'un nom propre est quand même modélisé dans Prolexbase
- C'est un ensemble de synonymes (un *synset*):
l'ensemble de tous
les points de vue le concernant

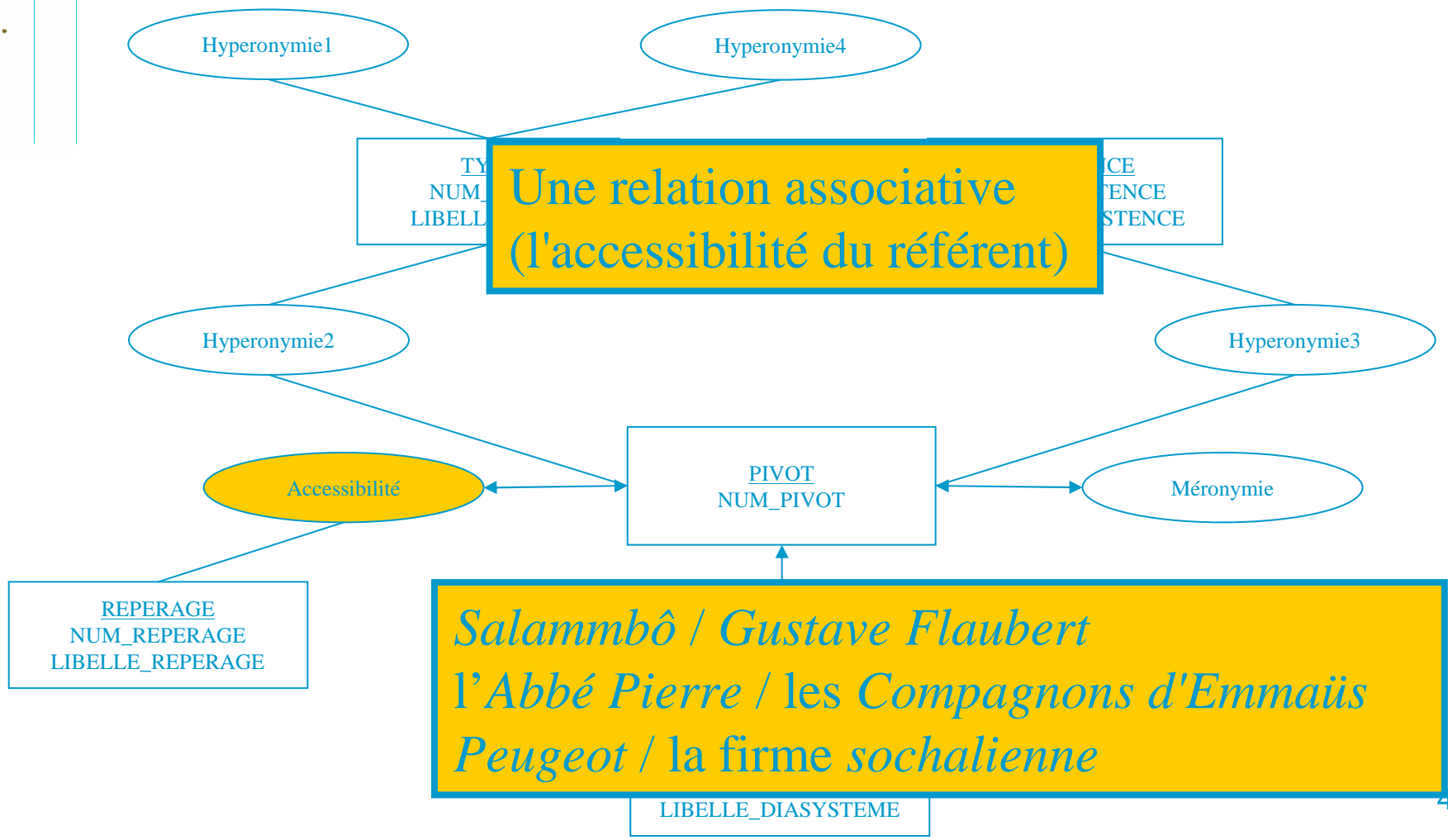
Les relations indépendantes de la langue



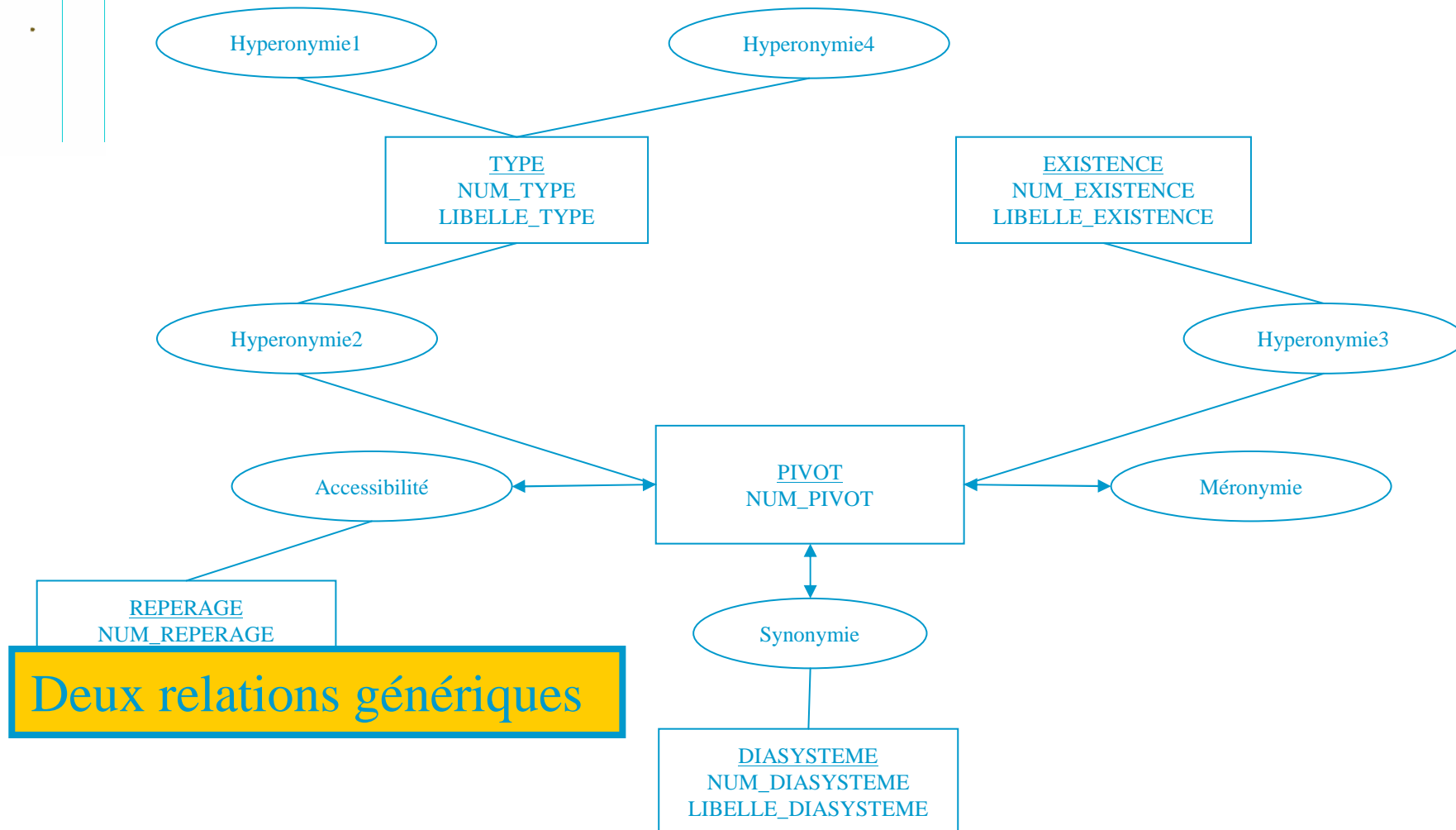
Les relations indépendantes de la langue



Les relations indépendantes de la langue



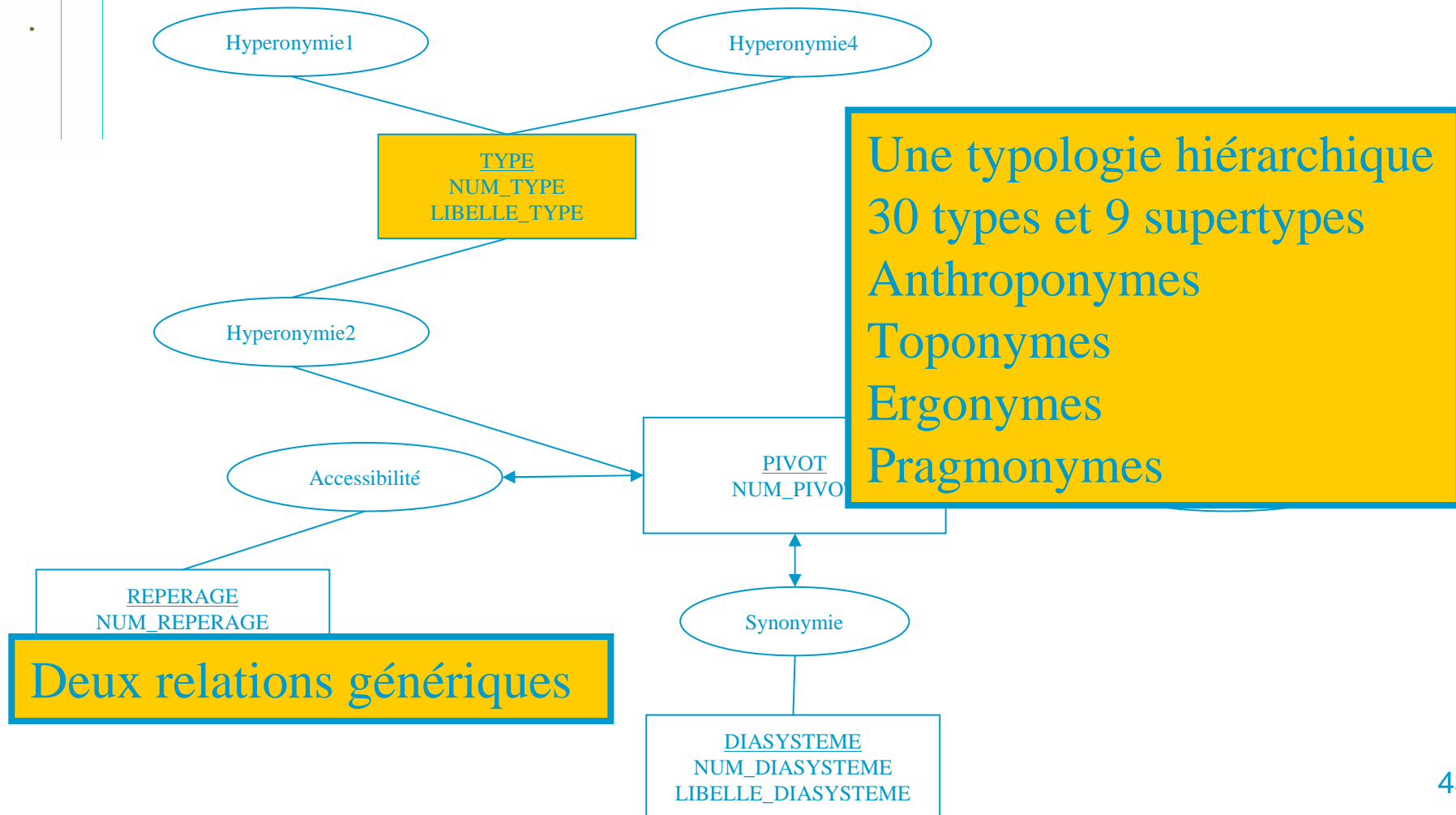
Les relations indépendantes de la langue



Les relations indépendantes de la langue



Les relations indépendantes de la langue



Les relations indépendantes de la langue

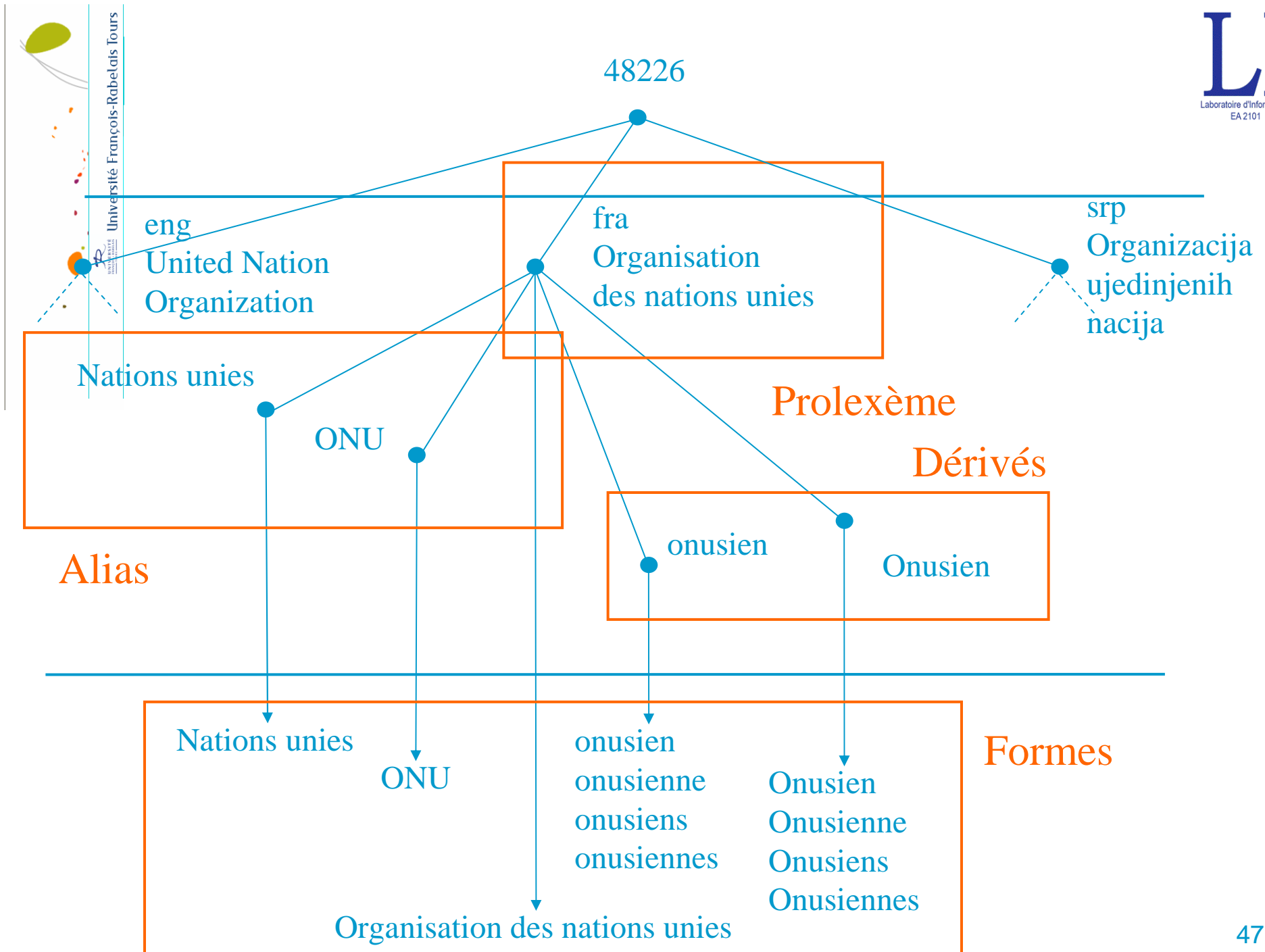
Nom propre						
Anthroponyme			Ergonyme	Pragmonyme	Toponyme	
Individuel	Collectif					
		Groupe				Territoire
Patronyme Personne Prénom Pseudo anthroponyme	Dynastie Ethnonyme	Association Ensemble Entreprise Institution Organisation	Objet Œuvre Pensée Produit Vaisseau	Catastrophe Évènement Fête Histoire Météorologie	Astronyme Édifice Géonyme Hydronyme Ville Voie	Pays Région Supranational

La structure de Prolexbase

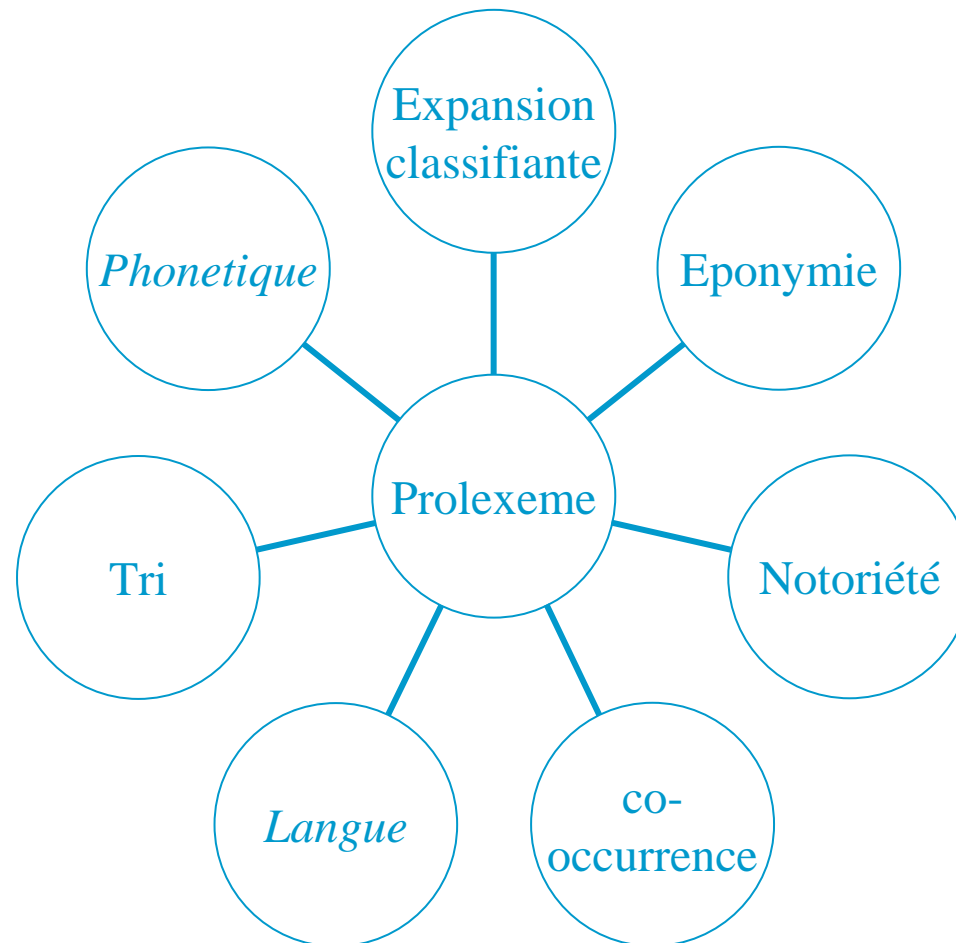
2. Le niveau linguistique

Le prolexème

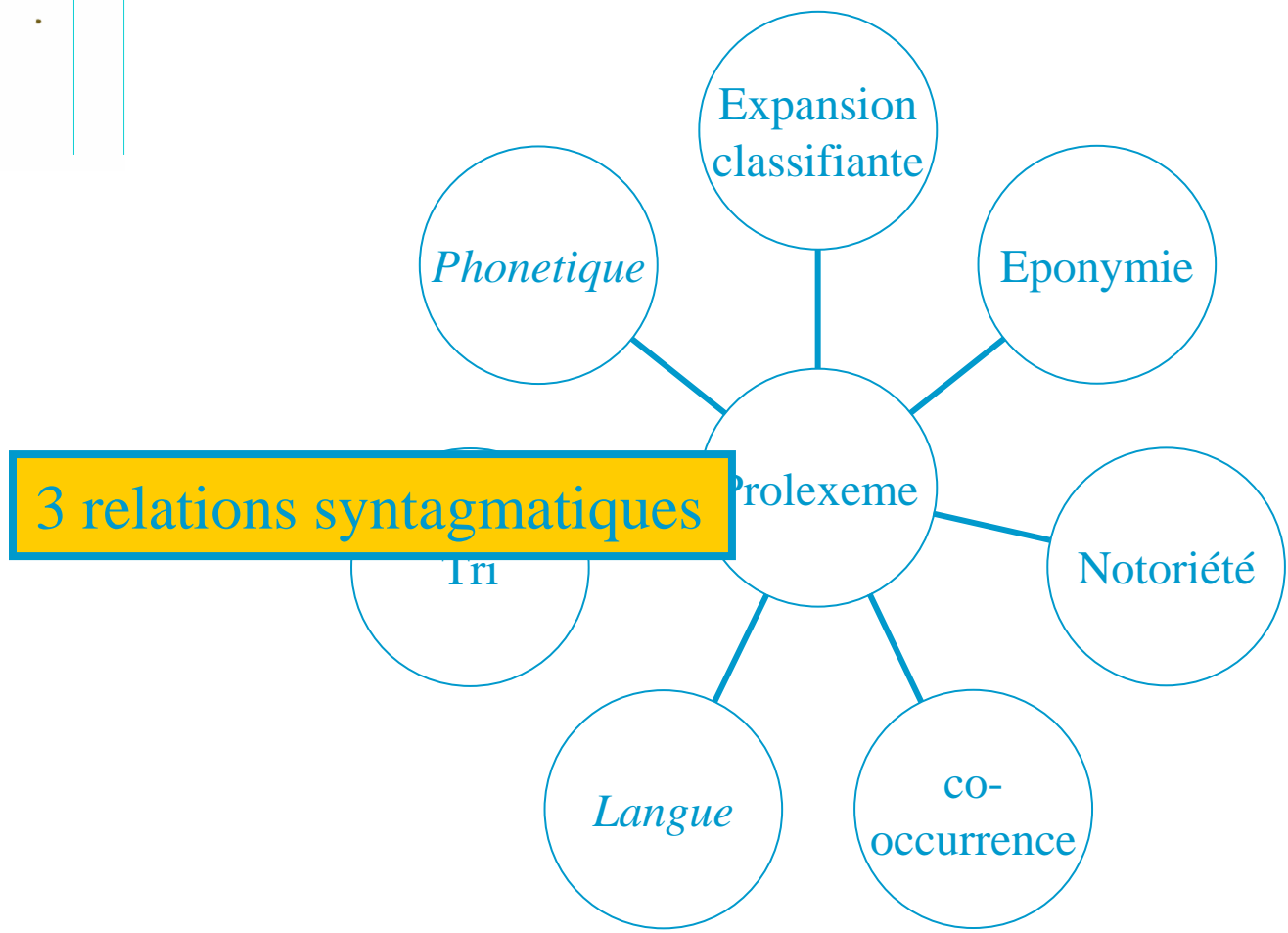
- Un ensemble de lemmes morphosémantiquement liés au nom propre conceptuel
 - Prolexème-Fra_{Finlande} = {Finlande.N, Finlandais.NR, finlandais.AR}
(mais pas *finlandiser*...)
 - Prolexème-Fra_{Organisation des Nations unies} = {Organisation des Nations unies.N, Nations unies.N, Onu.N, Onusien.NR, onusien.AR}



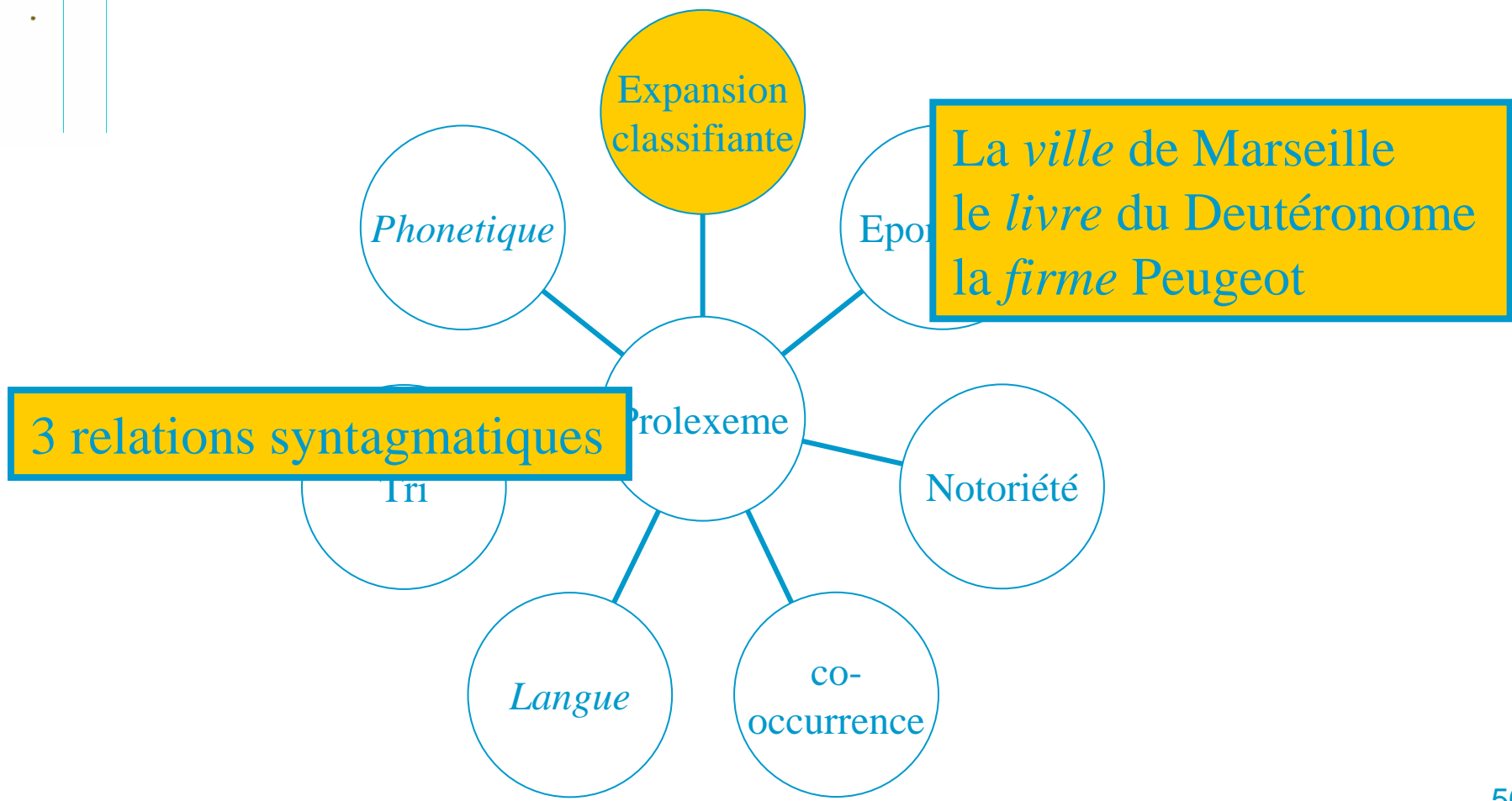
Les relations dépendantes de la langue



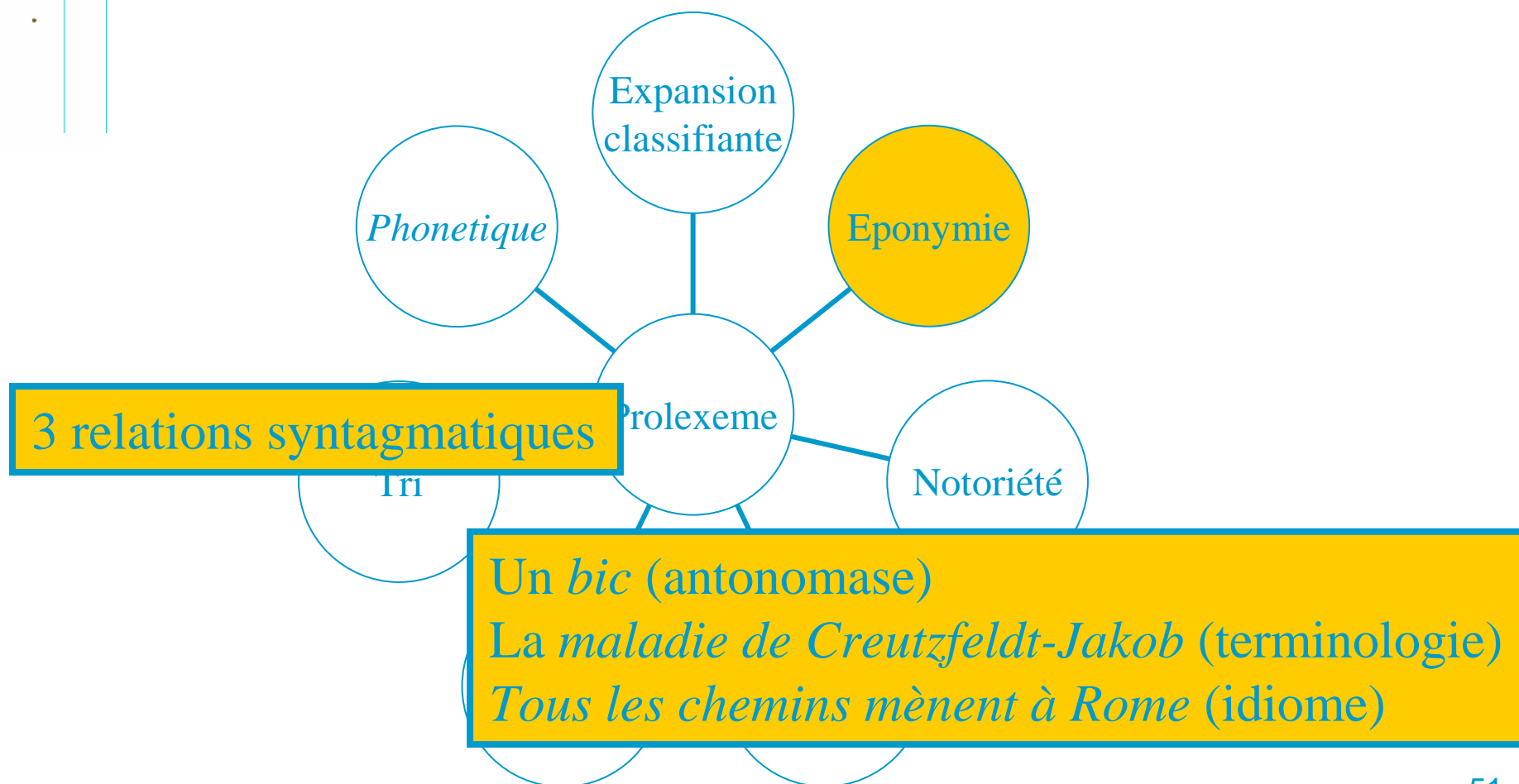
Les relations dépendantes de la langue



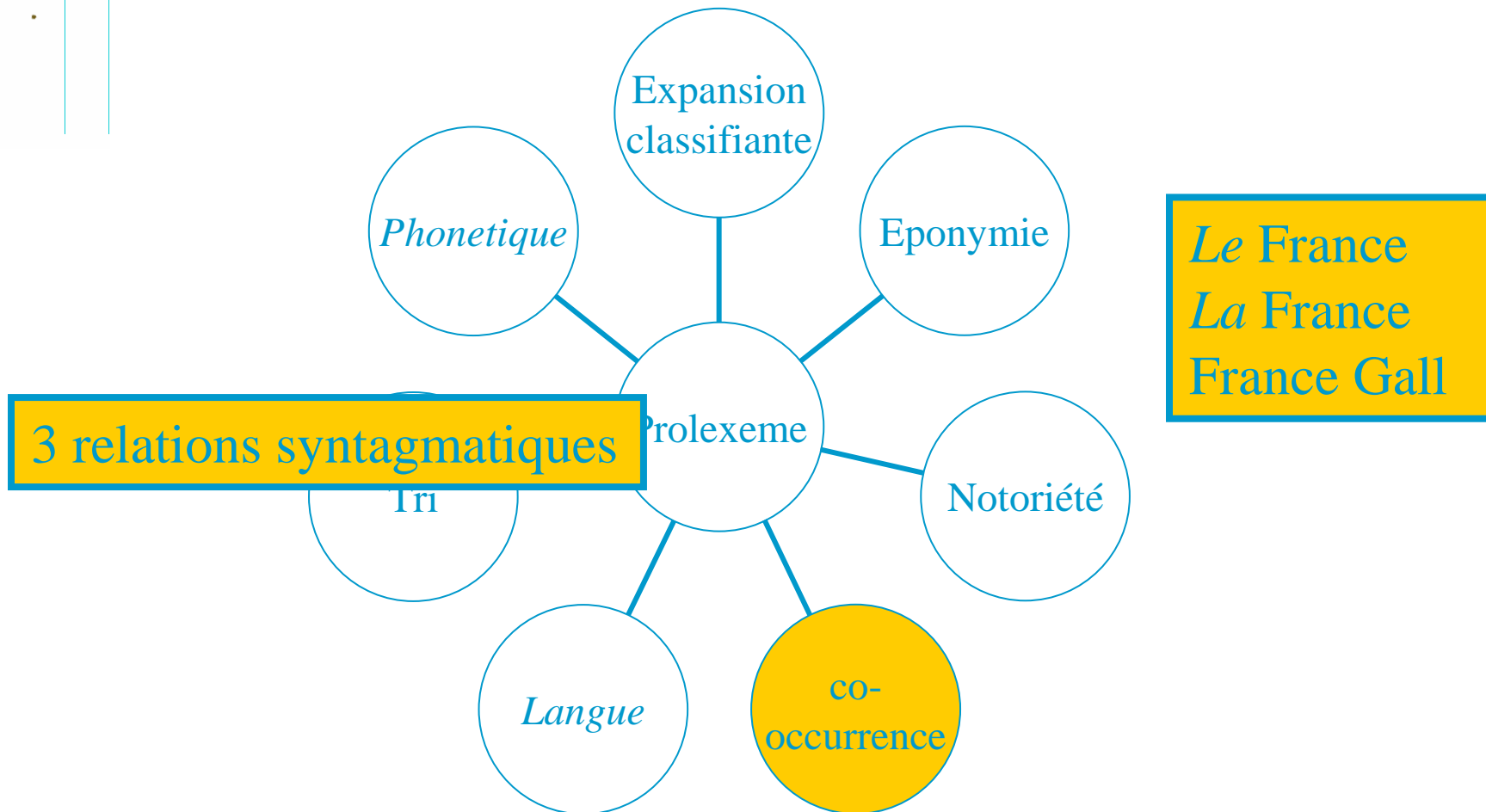
Les relations dépendantes de la langue



Les relations dépendantes de la langue

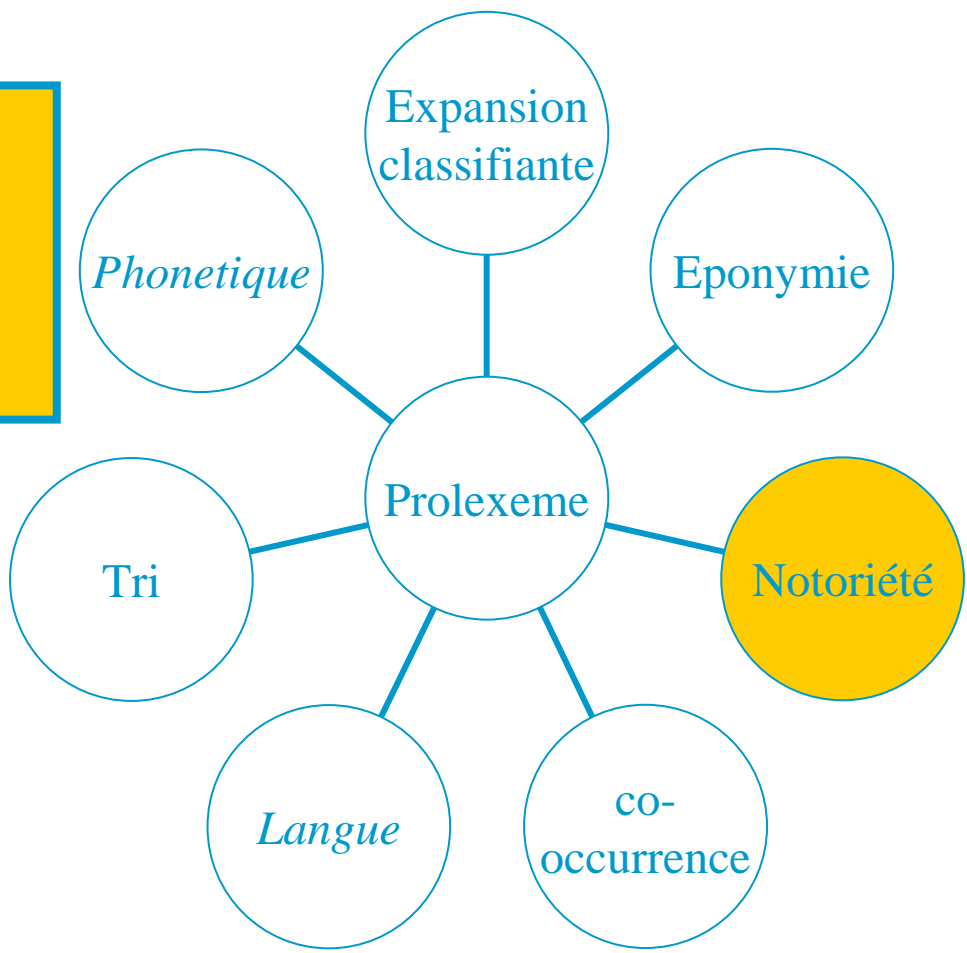


Les relations dépendantes de la langue



Les relations dépendantes de la langue

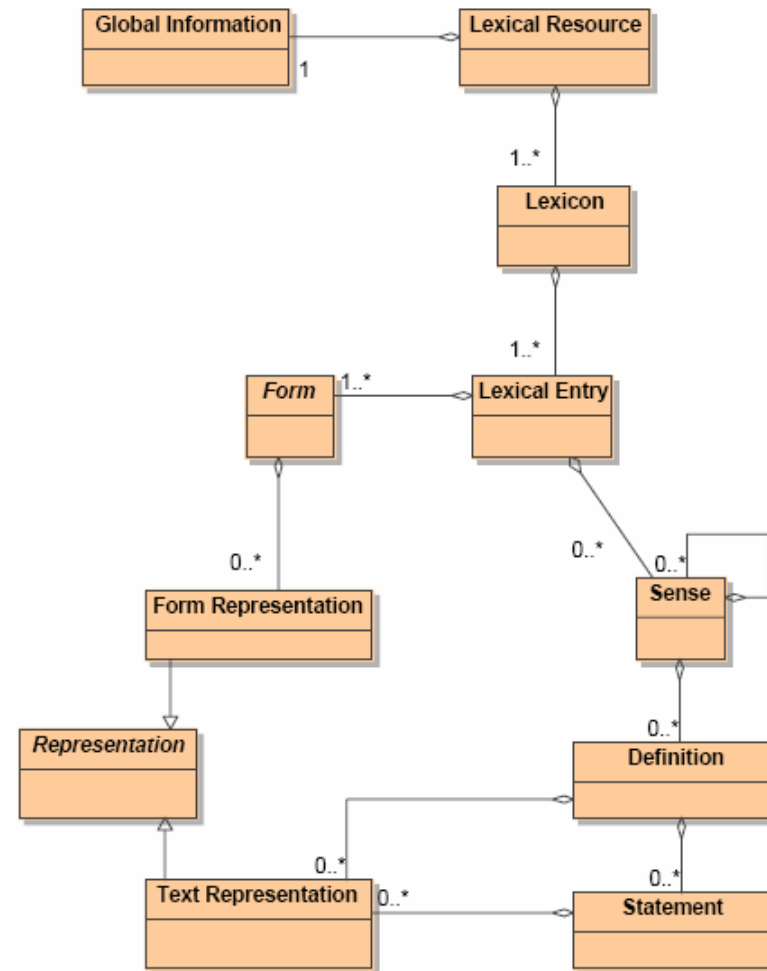
Trois usages:
rare
peu fréquent
fréquent



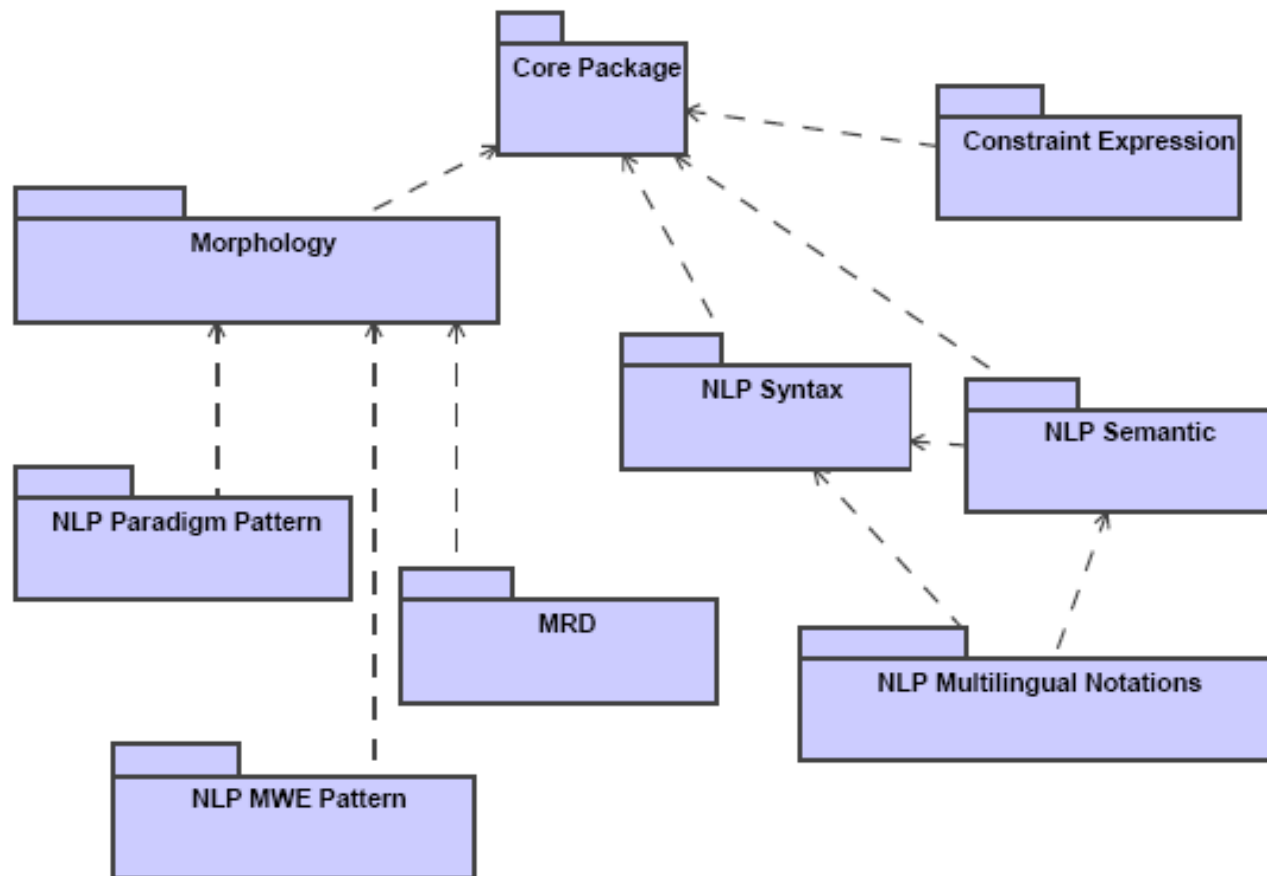
Une version LMF de Prolexbase

(Lexical Markup Framework)

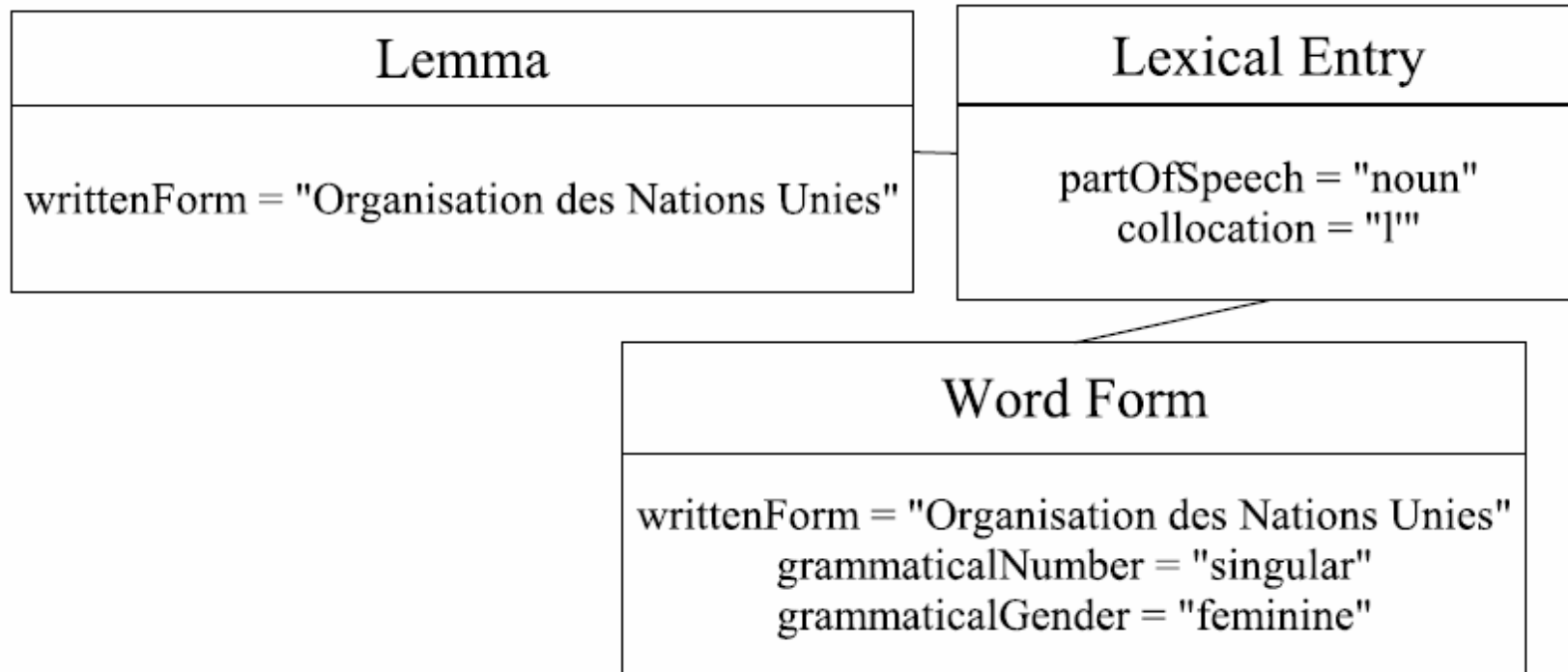
LMF core package



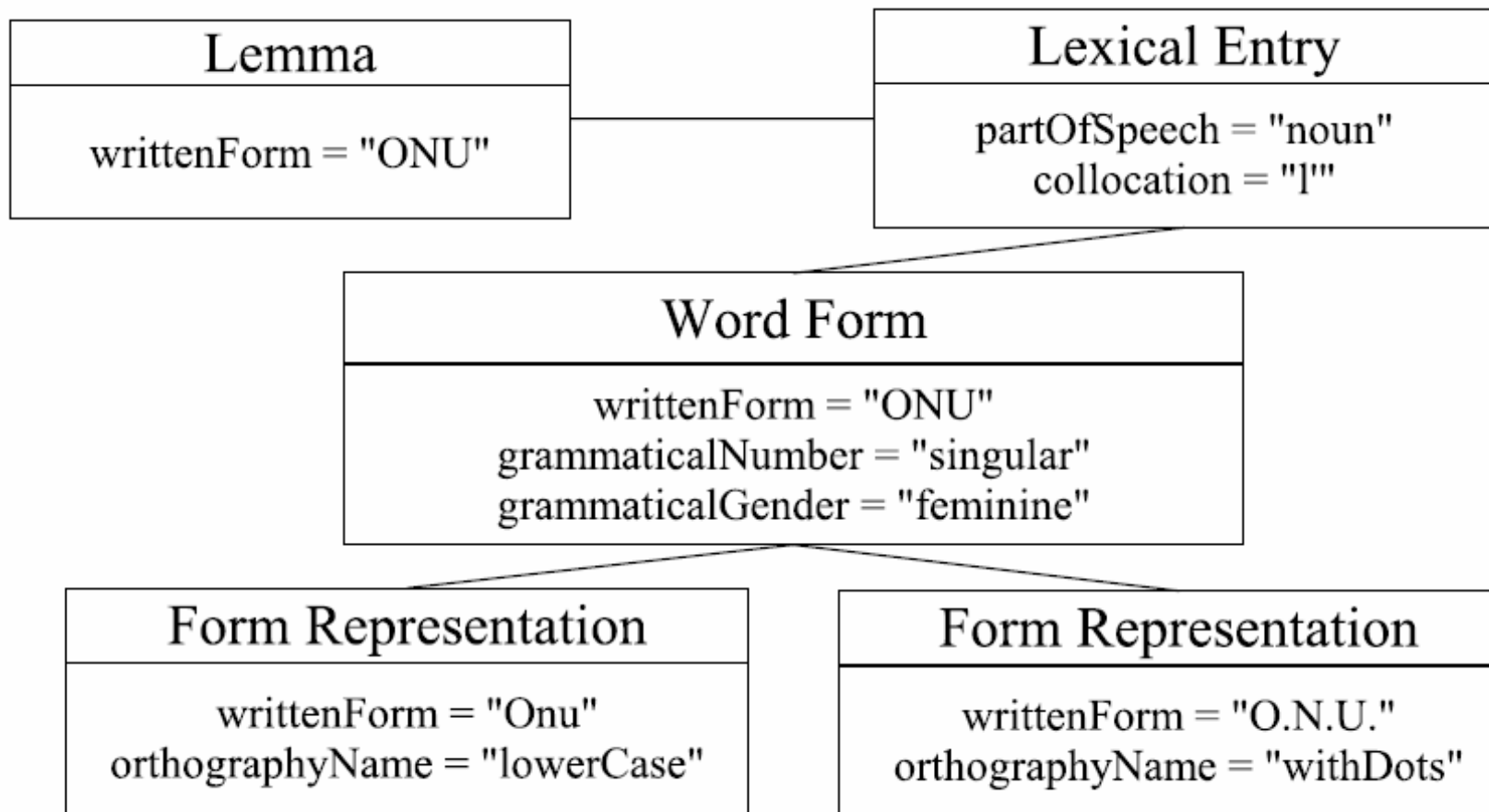
LMF extension packages



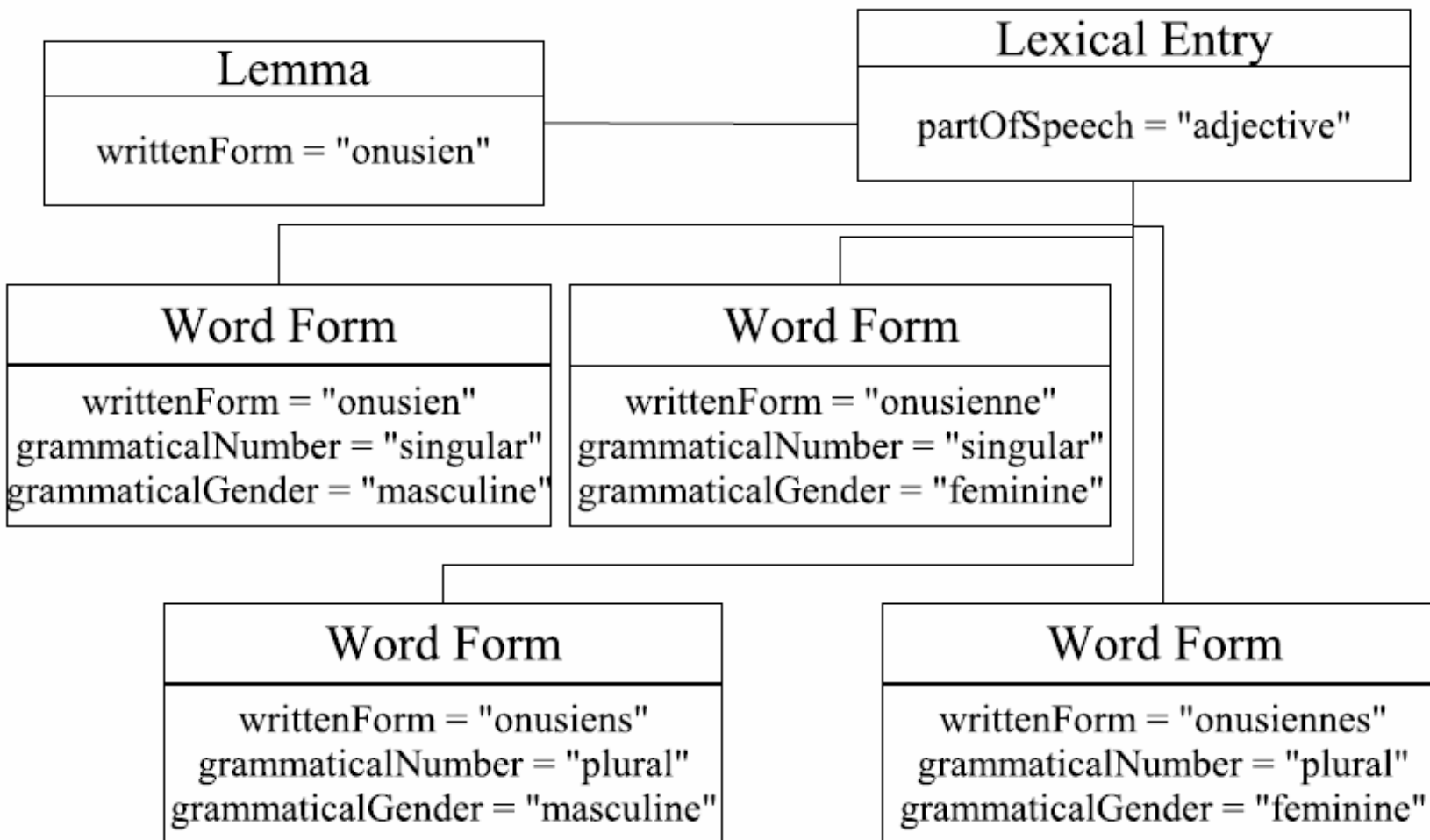
Prolex - LMF



Prolex - LMF

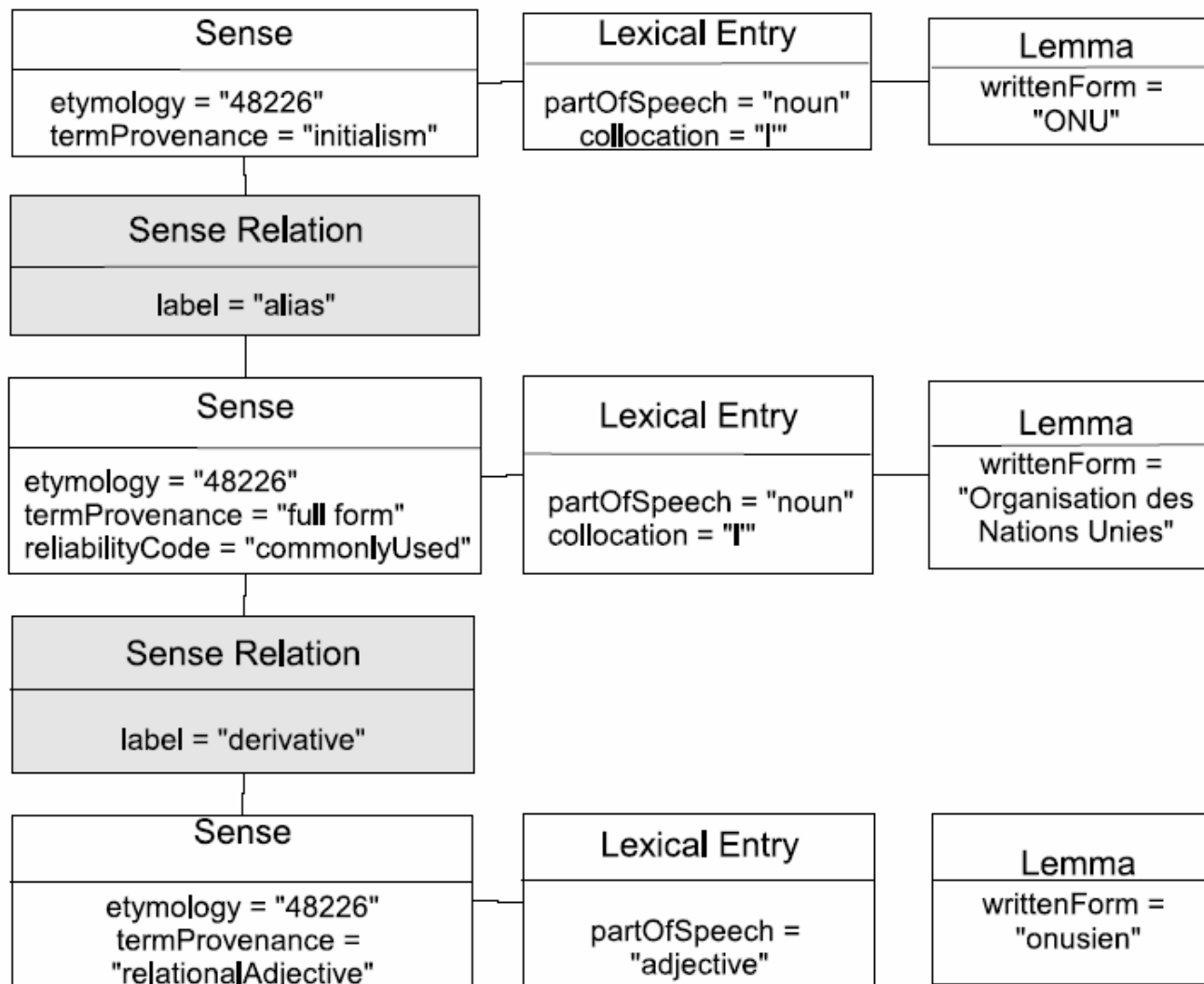


Prolex - LMF

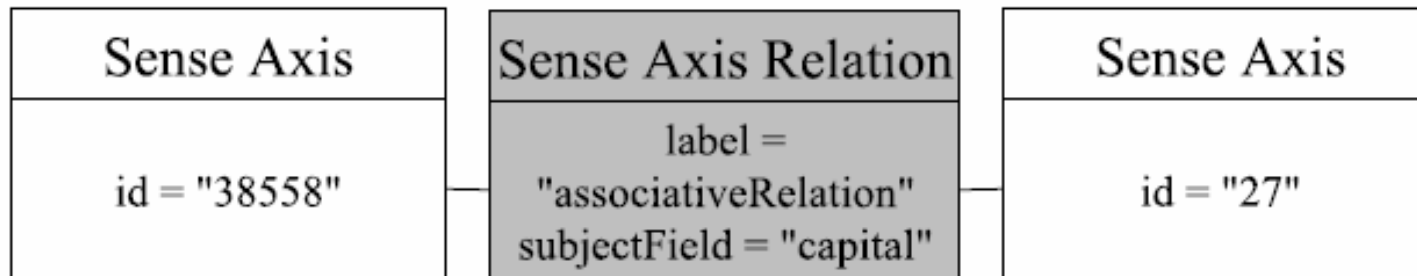




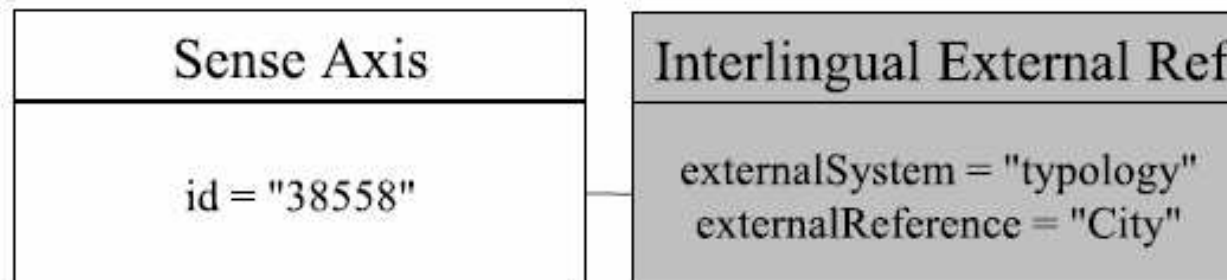
Prolex - LMF



Prolex - LMF



Prolex - LMF



Prolex - LMF

```

<?xml version="1.0" encoding="UTF-8" ?>
- <LexicalRessource>
  <GlobalInformation entrySource="Prolex" />
  - <Lexicon languageSymbol="fra">
    ...
    - <LexicalEntry partOfSpeech="noun">
      <Collocation determiner="zero" />
      <Lemma writtenForm="Paris" />
      <WordForm writtenForm="Paris" grammaticalGender="masculineFeminine" grammaticalNumber="singular" />
      <Sense id="38558" etymology="38558" termProvenance="fullForm" reliabilityCode="commonlyUsed" />
    </LexicalEntry>
    - <LexicalEntry partOfSpeech="noun">
      <Lemma writtenForm="Parisien" />
      <WordForm writtenForm="Parisiennes" grammaticalGender="feminine" grammaticalNumber="plural" />
      <WordForm writtenForm="Parisiens" grammaticalGender="masculine" grammaticalNumber="plural" />
      <WordForm writtenForm="Parisienne" grammaticalGender="feminine" grammaticalNumber="singular" />
      <WordForm writtenForm="Parisien" grammaticalGender="masculine" grammaticalNumber="singular" />
      - <Sense id="18666" etymology="38558" termProvenance="relationalName" reliabilityCode="commonlyUsed">
        <SenseRelation label="derivative" targets="38558" />
      </Sense>
    </LexicalEntry>
    - <LexicalEntry partOfSpeech="adjective">
      <Lemma writtenForm="parisien" />
      <WordForm writtenForm="parisiennes" grammaticalGender="feminine" grammaticalNumber="plural" />
      <WordForm writtenForm="parisiens" grammaticalGender="masculine" grammaticalNumber="plural" />
      <WordForm writtenForm="parisienne" grammaticalGender="feminine" grammaticalNumber="singular" />
      <WordForm writtenForm="parisien" grammaticalGender="masculine" grammaticalNumber="singular" />
      - <Sense id="18667" etymology="38558" termProvenance="relationalAdjective" reliabilityCode="commonlyUsed">
        <SenseRelation label="derivative" targets="38558" />
      </Sense>
    </LexicalEntry>
    - <LexicalEntry partOfSpeech="noun">
      <Lemma writtenForm="Parigot" />
      <WordForm writtenForm="Parigotes" grammaticalGender="feminine" grammaticalNumber="plural" />
      <WordForm writtenForm="Parigots" grammaticalGender="masculine" grammaticalNumber="plural" />
      <WordForm writtenForm="Parigote" grammaticalGender="feminine" grammaticalNumber="singular" />
      <WordForm writtenForm="Parigot" grammaticalGender="masculine" grammaticalNumber="singular" />
      - <Sense id="20799" etymology="38558" termProvenance="quasiRelationalName" reliabilityCode="commonlyUsed">
        <SenseRelation label="derivative" targets="38558" />
      </Sense>
    </LexicalEntry>
    ...
  </Lexicon>
  ...
- <SenseAxis id="38558">

```

Prolex - LMF

```
<?xml version="1.0" encoding="UTF-8" ?>
```

```
<?xml version="1.0" encoding="UTF-8" ?>
- <LexicalRessource>
  <GlobalInformation entrySource="Prolex" />
  - <Lexicon languageSymbol="fra">
```

```
...
```

```
    nine" grammaticalNumber="singular" />
    reliabilityCode="commonlyUsed" />
```

```

- <LexicalEntry partOfSpeech="noun">
  <Lemma writtenForm="Parisien" />
  <WordForm writtenForm="Parisiennes" grammaticalGender="feminine" grammaticalNumber="plural" />
  <WordForm writtenForm="Parisiens" grammaticalGender="masculine" grammaticalNumber="plural" />
  <WordForm writtenForm="Parisienne" grammaticalGender="feminine" grammaticalNumber="singular" />
  <WordForm writtenForm="Parisien" grammaticalGender="masculine" grammaticalNumber="singular" />
  - <Sense id="18666" etymology="38558" termProvenance="relationalName" reliabilityCode="commonlyUsed">
    <SenseRelation label="derivative" targets="38558" />
  </Sense>
</LexicalEntry>
- <LexicalEntry partOfSpeech="adjective">
  <Lemma writtenForm="parisien" />
  <WordForm writtenForm="parisiennes" grammaticalGender="feminine" grammaticalNumber="plural" />
  <WordForm writtenForm="parisiens" grammaticalGender="masculine" grammaticalNumber="plural" />
  <WordForm writtenForm="parisienne" grammaticalGender="feminine" grammaticalNumber="singular" />
  <WordForm writtenForm="parisien" grammaticalGender="masculine" grammaticalNumber="singular" />
  - <Sense id="18667" etymology="38558" termProvenance="relationalAdjective" reliabilityCode="commonlyUsed">
    <SenseRelation label="derivative" targets="38558" />
  </Sense>
</LexicalEntry>
- <LexicalEntry partOfSpeech="noun">
  <Lemma writtenForm="Parigot" />
  <WordForm writtenForm="Parigotes" grammaticalGender="feminine" grammaticalNumber="plural" />
  <WordForm writtenForm="Parigots" grammaticalGender="masculine" grammaticalNumber="plural" />
  <WordForm writtenForm="Parigote" grammaticalGender="feminine" grammaticalNumber="singular" />
  <WordForm writtenForm="Parigot" grammaticalGender="masculine" grammaticalNumber="singular" />
  - <Sense id="20799" etymology="38558" termProvenance="quasiRelationalName" reliabilityCode="commonlyUsed">
    <SenseRelation label="derivative" targets="38558" />
  </Sense>
</LexicalEntry>
...
</Lexicon>
...
- <SenseAxis id="38558">
```

Prolex - LMF

```
<?xml version="1.0" encoding="UTF-8" ?>
- <LexicalRessource>
  <GlobalInformation entrySource="Prolex" />
  <Lexicon languageSymbol="fra">
    ...
  - <LexicalEntry partOfSpeech="noun">
```

```
...
- <LexicalEntry partOfSpeech="noun">
  <Collocation determiner="zero" />
  <Lemma writtenForm="Paris" />
  <WordForm writtenForm="Paris" grammaticalGender="masculineFeminine" grammaticalNumber="singular" />
  <Sense id="38558" etymology="38558" termProvenance="fullForm" reliabilityCode="commonlyUsed" />
</LexicalEntry>
```

```
  <Sense id="18666" etymology="38558" termProvenance="relationalName" reliabilityCode="commonlyUsed">
    <SenseRelation label="derivative" targets="38558" />
  </Sense>
</LexicalEntry>
- <LexicalEntry partOfSpeech="adjective">
  <Lemma writtenForm="parisien" />
  <WordForm writtenForm="parisiennes" grammaticalGender="feminine" grammaticalNumber="plural" />
  <WordForm writtenForm="parisiens" grammaticalGender="masculine" grammaticalNumber="plural" />
  <WordForm writtenForm="parisienne" grammaticalGender="feminine" grammaticalNumber="singular" />
  <WordForm writtenForm="parisien" grammaticalGender="masculine" grammaticalNumber="singular" />
  <Sense id="18667" etymology="38558" termProvenance="relationalAdjective" reliabilityCode="commonlyUsed">
    <SenseRelation label="derivative" targets="38558" />
  </Sense>
</LexicalEntry>
- <LexicalEntry partOfSpeech="noun">
  <Lemma writtenForm="Parigot" />
  <WordForm writtenForm="Parigotes" grammaticalGender="feminine" grammaticalNumber="plural" />
  <WordForm writtenForm="Parigots" grammaticalGender="masculine" grammaticalNumber="plural" />
  <WordForm writtenForm="Parigote" grammaticalGender="feminine" grammaticalNumber="singular" />
  <WordForm writtenForm="Parigot" grammaticalGender="masculine" grammaticalNumber="singular" />
  <Sense id="20799" etymology="38558" termProvenance="quasiRelationalName" reliabilityCode="commonlyUsed">
    <SenseRelation label="derivative" targets="38558" />
  </Sense>
</LexicalEntry>
...
</Lexicon>
...
- <SenseAxis id="38558">
```

Prolex - LMF

```
<?xml version="1.0" encoding="UTF-8" ?>
- <LexicalRessource>
  <GlobalInformation entrySource="Prolex" />
  - <Lexicon languageSymbol="fra">
    ...
    - <LexicalEntry partOfSpeech="noun">
      <Collocation determiner="zero" />
      <Lemma writtenForm="Paris" />
      <WordForm writtenForm="Paris" grammaticalGender="masculineFeminine" grammaticalNumber="singular" />
      <Sense id="38558" etymology="38558" termProvenance="fullForm" reliabilityCode="commonlyUsed" />
    </LexicalEntry>
    - <LexicalEntry partOfSpeech="noun">
      <Lemma writtenForm="Parisien" />
      <WordForm writtenForm="Parisiennes" grammaticalGender="feminine" grammaticalNumber="plural" />
      <WordForm writtenForm="Parisiens" grammaticalGender="masculine" grammaticalNumber="plural" />
      <WordForm writtenForm="Parisienne" grammaticalGender="feminine" grammaticalNumber="singular" />
      <WordForm writtenForm="Parisien" grammaticalGender="masculine" grammaticalNumber="singular" />
      - <Sense id="18666" etymology="38558" termProvenance="relationalName" reliabilityCode="commonlyUsed">
        <SenseRelation label="derivative" targets="38558" />
      </Sense>
    </LexicalEntry>
    - <LexicalEntry partOfSpeech="adjective">
      <Lemma writtenForm="parisien" />
      <WordForm writtenForm="parisiennes" grammaticalGender="feminine" grammaticalNumber="plural" />
```

```
- <LexicalEntry partOfSpeech="noun">
  <Lemma writtenForm="Parigot" />
  <WordForm writtenForm="Parigotes" grammaticalGender="feminine" grammaticalNumber="plural" />
  <WordForm writtenForm="Parigots" grammaticalGender="masculine" grammaticalNumber="plural" />
  <WordForm writtenForm="Parigote" grammaticalGender="feminine" grammaticalNumber="singular" />
  <WordForm writtenForm="Parigot" grammaticalGender="masculine" grammaticalNumber="singular" />
  - <Sense id="20799" etymology="38558" termProvenance="quasiRelationalName" reliabilityCode="commonlyUsed">
    <SenseRelation label="derivative" targets="38558" />
  </Sense>
</LexicalEntry>
...
</Lexicon>
...
- <SenseAxis id="38558">
```

Prolex - LMF

```

...
- <SenseAxis id="38558">
  <InterlingualExternalRef externalSystem="typology" externalReference="city" />
  <InterlingualExternalRef externalSystem="existence" externalReference="historical" />
  <SenseAxisRelation label="quasiSynonym" targets="55120" subjectField="diaphasic" />
  <SenseAxisRelation label="partitiveRelation" targets="53865" />
  <SenseAxisRelation label="partitiveRelation" targets="53687" />
  <SenseAxisRelation label="partitiveRelation" targets="53042" />
  ...
  <SenseAxisRelation label="associativeRelation" targets="27" subjectField="capital" />
  <SenseAxisRelation label="associativeRelation" targets="5" subjectField="capital" />
  <SenseAxisRelation label="associativeRelation" targets="54453" subjectField="headquarters" />
  <SenseAxisRelation label="associativeRelation" targets="54454" subjectField="headquarters" />
  ...
</SenseAxis>
...
</LexicalResource>

```