ProLMF version 1.2. Une ressource libre de noms propres avec des expansions contextuelles

Denis Maurel, Béatrice Bouchou Markhoff
Université François Rabelais Tours
denis.maurel@univ-tours.fr, beatrice.bouchou@univ-tours.fr

RÉSUMÉ

ProLMF est la version LMF de la base lexicale multilingue de noms propres Prolexbase. Disponible librement sur le site du CNRTL, la version 1.2 a été largement améliorée et augmentée par de nouvelles entrées en français, complétées par des expansions contextuelles, et par de petits lexiques en une huitaine de langues.

ABSTRACT

ProLMF 1.2, Proper Names with their Expansions

ProLMF is the LMF version of Prolexbase, a multilingual lexical database of Proper Names. It can be freely downloaded on the CNRTL Website. Version 1.2 had been widely improved and increased, with new French entries whose description includes contextual expansions, and eight small lexica for other languages.

MOTS-CLÉS : ressource libre, base lexicale multilingue, noms propres, expansions contextuelles, schémas de contextualisation, relations sémantiques, alias, point de vue, Prolexbase.
KEYWORDS: free resource, multilingual lexical database, Proper Names, context, semantic relations, alias, point of view, Prolexbase.

1 Les bases de données lexicales

Parmi les ressources utilisées en TAL, les bases de données lexicales occupent une place importante, souvent à l'origine d'applications diverses. Citons entre autres Wordnet (Miller et al., 1990), les dictionnaires Dela (Courtois, Silberztein, 1990), le lexique Morphalou (Romary et al., 2004), le projet Papillon (Mangeot-Lerebours et al., 2003), etc. D'autres ressources libres comme Wikipédia, DBpedia (Auer, Lehmann, 2007), Geonames, Yago 2 (Hoffart et al., 2012), etc., sont structurées autour des entrées lexicales, qu'elles décrivent avec éventuellement des relations paradigmatiques, mais sans informations linguistiques.

Prolexbase (Tran et Maurel, 2006) a la particularité de rassembler des noms propres, en s'intéressant aussi à la morphologie flexionnelle et dérivationnelle de ces noms. Une première version de ProLMF (Bouchou et Maurel, 2008) a été déposée en 2008 sur le site Prolex[1] du CNRTL (Centre national de ressources textuelles et linguistiques), sous une licence libre. Les concepts les plus importants de Prolexbase sont ceux de *point de vue sur un référent* et de *prolexème*. Le premier concept, interlingue, matérialisé par un pivot, signifie que des entrées de Prolexbase peuvent correspondre dans la réalité à un même référent, s'il s'agit de points de vue différents sur celui-ci. Prenons l'exemple récent du pape *François* et

[1] http://www.cnrtl.fr/lexiques/prolex/.

du cardinal *Jorge Bergoglio* : ces deux noms propres correspondraient à deux pivots différents. Le second concept est un ensemble de formes morphosémantiquement (Fellbaum et Miller, 2003) liés à la projection du pivot dans une langue. En français, cet ensemble comprend en général le nom propre lui-même, parfois une forme courte ou un acronyme, souvent un nom et un adjectif relationnels, ces derniers étant les seuls à se fléchir[2]. Dans d'autres langues, où la morphologie flexionnelle et/ou dérivationnelle est plus développée, ce prolexème peut comprendre un grand nombre de lemmes et de formes. Les pivots sont reliés entre eux par trois relations : la synonymie, la méronymie et l'accessibilité (voir section 2.4).

2 Présentation générale de ProLMF

2.1 La norme LMF

La norme LMF (ISO 24613:2008) pour *Lexical Markup Framework* est un meta modèle de représentation des données lexicales (Francopoulo et al., 2006). LMF permet la représentation de bases très différentes dans leurs conceptions, de la simple liste de mots aux bases morphologiques, sémantiques, multilingues, etc. Elle est composée d'un module central (le *core package*) et d'extensions. Le module central, obligatoire, contient les informations générales (par exemple le codage des caractères), le lemme et, facultativement, une ou des formes, un ou des sens. Les extensions permettent de traiter la syntaxe, la sémantique, la morphologie, le multilinguisme, etc. Les attributs de chaque classe respectent, dans la mesure du possible, le registre des *Data categories* (Francopoulo et al., 2008). La Figure 1 présente les classes utilisées par ProLMF ; les classes grisées correspondent à la partie multilingue.

ProLMF 1.2 comporte :

- un lexique français avec lemme, forme et sens pour chaque entrée lexicale, ainsi que des schémas de contextualisation ;
- quelques petits lexiques (allemand, anglais, italien, néerlandais, polonais, portugais et serbe) avec uniquement lemme et sens ;
- et une description au niveau multilingue avec des informations typologiques et, surtout, des relations entre pivots (synonymie, méronymie et accessibilité).

2.2 Les informations globales

Les informations globales indiquent que ProLMF respecte la norme ISO 639 pour le codage des noms de langues sur trois lettres[3] et la norme ISO 15924 pour le codage des noms d'écriture sur quatre lettres[4] ; elles indiquent aussi que le codage des caractères est implanté en UTF-8.

[2] Par exemple le pivot 38558 correspond en français à cinq lemmes et à un ensemble de dix-sept formes {Paris, Parisien, Parisienne, Parisiens, Parisiennes, parisien, parisienne, parisiens, parisiennes, Parigot, Parigote, Parigots, parigotes, parigot, parigote, parigots, parigotes}, qui ne contient pas *parisianisme*, pourtant bien dérivé morphologiquement de Paris, mais dont le sens est lexicalisé.

[3] C'est-à-dire respectivement : *deu*, *eng*, *fra*, *ita*, *nld*, *pol*, *por*, *spa* et *srp*.

[4] C'est-à-dire *latn* pour *latin* et *cyrl* pour *cyrillique*.

```
<LexicalRessource>
  <GlobalInformation languageCoding="ISO 639" scriptCoding="ISO 15924" characterCoding="UTF-8"
entrySource="Prolexbase" resourceName="ProLMF" version="1.2"/>
```

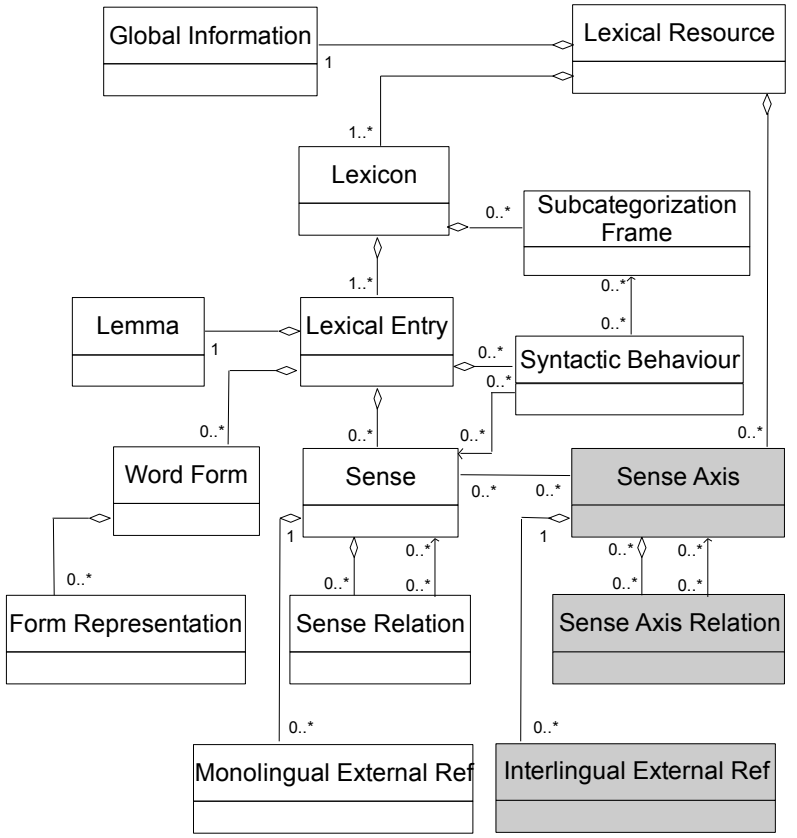

FIGURE 1 – Les classes utilisées par ProLMF.

Chaque lexique comporte comme attribut son code de langue et son code d'écriture. Lorsque deux écritures sont utilisées, comme par exemple en serbe (latin et cyrillique), il faudrait en principe distinguer ces écritures à l'intérieur de chaque forme. La version 1.2 de ProLMF n'implante pas cette solution. À titre transitoire, nous avons créé deux lexiques serbes, un en latin et l'autre en cyrillique.

```
<Lexicon languageIdentifier="deu" script="latn">
...
</Lexicon>
...
<Lexicon languageIdentifier="srp" script="cyrl">
...
</Lexicon>
```

2.3 Les entrées lexicales

Une entrée lexicale correspond à une seule partie du discours et comporte un lemme, une ou

des formes et un ou des sens. Comme cela a été dit plus haut, pour les langues autres que le français, la partie forme n'est pas renseignée. Sinon, chaque forme comprend son genre et son nombre.

Par exemple, la ville de Paris en serbe a pour lemme *Pariz* :

```
<LexicalEntry partOfSpeech="noun">
  <Lemma>Pariz</Lemma>
  <Sense idSense="P400" refSenseAxis="38558" termProvenance="fullForm" label="properName"/>
</LexicalEntry>
```

Dans la norme LMF, les entrées sont regroupées par catégories du discours, lemmes et formes ; les sens peuvent différer. Pour ProLMF, lorsqu'il y a plusieurs sens, il s'agit d'homographes. Par exemple, l'adjectif relationnel *neuilléen* correspond dans la base à deux villes : *Neuilly-l'Évêque* (*rarelyUsed* !) –pivot : 13346- et, bien sûr, *Neuilly-sur-Seine* –pivot : 18220-.

```
<LexicalEntry partOfSpeech="adjective">
  <Lemma>neuilléen</Lemma>
  <WordForm grammaticalGender="masculine" grammaticalNumber="singular">neuilléen</WordForm>
  <WordForm grammaticalGender="masculine" grammaticalNumber="plural">neuilléens</WordForm>
  <WordForm grammaticalGender="feminine" grammaticalNumber="singular">neuilléenne</WordForm>
  <WordForm grammaticalGender="feminine" grammaticalNumber="plural">neuilléennes</WordForm>
  <Sense idSense="D8233" refSenseAxis="13346" termProvenance="relationalAdjective"
frequency="rarelyUsed" label="derivative" refSense="P13346"/>
  <Sense idSense="D11433" refSenseAxis="18220" termProvenance="relationalAdjective"
frequency="infrequentlyUsed" label="derivative" refSense="P18220"/>
</LexicalEntry>
```

Le sens comprend jusqu'à six attributs : un identifiant (*idSense*), la référence au pivot multilingue (*refSenseAxis*), éventuellement la catégorie d'alias [5] ou de dérivé [6] (*termProvenance*), la notoriété[7] (*frequency*), l'indication s'il s'agit d'un nom propre ou d'un dérivé (*label*) et, dans ce dernier cas, la référence au sens dont il est dérivé (*refSense*)[8].

Dans le cas d'un nom propre, chaque sens peut être associé à un ou des contextes ; ces contextes sont décrits dans le lexique correspondant, au même niveau que les entrées lexicales (voir section 2.5).

2.4 La partie multilingue

Comme cela a été rappelé en section 1, la partie multilingue comprend un pivot par « point de vue » sur un nom propre. Par exemple, on distinguera un pivot pour *Paris* –pivot : 38558- et un autre pour *Ville lumière* –pivot : 55120-, même si le référent, la ville de Paris est le même. Ces deux pivots seront en relation de synonymie diaphasique[9]. Deux autres

[5] Par exemple : *fullForm*, *shortForm*, *acronym*...
[6] Par exemple : *relationalAdjective*, *relationalName*, *quasiRelationalName*...
[7] Suivant les *data categories*, l'attribut de notoriété prend trois valeurs : *rarelyUsed*, *infrequentlyUsed* et *commonlyUsed*.
[8] En français, c'est en général le nom propre ou un alias, comme *Onusien*, dérivé de *Onu* et non d'*Organisation des Nations Unis*. Dans des langues à forte productivité dérivationnelle, comme le serbe, cet attribut est beaucoup plus diversifié.
[9] Nous utilisons les traits définis par (Coseriu, 1998) : *diaphasique* (variation d'emploi), *diachronique* (variation dans le temps), *diatopique* (variation dans l'espace) et *diastratique* (variation socio-culturelle).

relations[10] existent : la méronymie[11] (les Champs-Élysées –pivot : 49215- est une avenue parisienne) et l'accessibilité[12] (Paris est la capitale de la France –pivot : 27-). Ces pivots sont associés à la typologie des noms propres du projet Prolex et à un paradigme d'existence[13] :

```
<SenseAxis id="38558">
  <InterlingualExternalRef externalSystem="typology" externalReference="city"/>
  <InterlingualExternalRef externalSystem="existence" externalReference="historical"/>
  <SenseAxisRelation label="partitiveRelation" refSenseAxis="49215"/>
  <SenseAxisRelation label="quasiSynonym" refSenseAxis="55120" usageNote="diaphasic"/>
  <SenseAxisRelation id="1" label="associativeRelation" refSenseAxis="27" subjectField="capital"/>
</SenseAxis>
```

2.5 Les règles d'aliasisation

Dans la version 1.2, un grand nombre d'alias ont été ajoutés par des règles d'aliasisation, pour permettre la création automatique de formes courtes[14]. Par exemple, le prolexème *Wolfgang Amadeus Mozart* est complété par l'alias *Wolfgang Mozart* et les noms de ville construits avec une préposition et un toponyme, comme *Neuilly sur Seine*, sont pour une grande part associés à une forme courte sans complément prépositionnel, comme *Neuilly*.

3 Les expansions contextuelles

La nouveauté la plus importante de ProLMF 1.2 est l'introduction de règles d'expansion contextuelle. Celles-ci peuvent se diviser en trois catégories :

- La présence éventuelle d'un déterminant (*la France*) et le choix de la préposition locative pour les noms de pays (*en France*).
- l'expansion classifiante (*la ville de Paris*)
- l'expansion d'accessibilité (*Paris, la capitale de la France*)

Certains sens sont aussi complétés par un lien vers Wikipédia (classe MonolingualExternalRef).

3.1 Déterminants et prépositions locatives

Les noms propres en français sont parfois précédés d'un déterminant. Nous avons noté cette information en indiquant de quel déterminant il s'agit. Dans le cas particulier des noms de pays, nous avons indiqué aussi la préposition locative à utiliser. Par exemple *France* est en général précédé de l'article *la* et s'utilise avec la préposition *en* (contrairement à *Portugal*, par exemple). Cette indication se trouve dans le sens associé à l'entrée lexicale France :

[10] Toutes les relations ne sont bien sûr notée qu'une fois, sur un seul des deux pivots.
[11] Celle-ci comprend les relations classiques (lieux et évènements), mais nous l'avons aussi étendue aux filiales d'entreprises, à la nationalité des personnes, etc.
[12] Dans un « dictionnaire de noms propres », un nom propre est accessible via un autre nom propre et non via une définition. L'accessibilité comporte volontairement dans Prolexbase douze repérages (*subjectFile*) très larges : *relative*, *creator*, *capital*... Ces repérages seront démultipliés dans chaque langue par les contextes d'accessibilité (section 3).
[13] La typologie Prolex est volontairement réduite à trente types et supertypes. Celui-ci comprend trois valeurs.
[14] L'application de ces règles est bien sûr supervisée, comme toute la base. Dans ProLMF, ces alias ne sont pas distingués des alias entrés manuellement, mais cette information est dans Prolexbase, ainsi que la règle appliquée.

```
<LexicalEntry partOfSpeech="noun">
  <Lemma>France</Lemma>
  <WordForm grammaticalGender="feminine" grammaticalNumber="singular">France</WordForm>
  <Sense idSense="P27" refSenseAxis="27" termProvenance="fullForm" frequency="commonlyUsed"
label="properName">
    <SyntacticBehaviour refSubcategorizationFrame="CO3"/>
    <SyntacticBehaviour refSubcategorizationFrame="CO7"/>
    <MonolingualExternalRef externalSystem="Wikipedia"
externalReference="http://fr.wikipedia.org/wiki/France"/>
  </Sense>
</LexicalEntry>
```

Les balises *SyntacticBehaviour* font référence à des règles de sous-catégorisation, elles aussi décrites dans le lexique, après les entrées lexicales :

```
<SubcategorizationFrame id="CO3" introducer="Determiner">la</SubcategorizationFrame>
<SubcategorizationFrame id="CO7" introducer="locativePreposition">en</SubcategorizationFrame>
```

Cette relation s'applique à tous les noms propres de la base, prolexèmes et alias.

3.2 Les expansions

La relation d'expansion classifiante associe à un prolexème une expansion qui peut apparaître, soit à sa gauche, soit à sa droite. Toutes les expansions qui existent dans une langue ne se retrouvent pas forcément dans une autre langue. Si l'expansion d'un nom propre est omise dans un texte, il est parfois nécessaire de la rétablir lors de la traduction de celui-ci, afin d'apporter un complément d'information au lecteur. Ainsi, le nom propre *la Loire* devient en anglais *the Loire River*. Dans la version 1.2, un grand nombre de prolexèmes sont liés à des expansions, comme des précisions toponymiques (ville, rivière, aéroport…), des professions (acteur, industriel, compositeur…) ou autres (archange, cité légendaire, fête…).

La relation d'expansion d'accessibilité est la projection dans une langue de la relation d'accessibilité sur les pivots interlingues. Comme cela a été rappelé ci-dessus, la relation d'accessibilité comprend une indication très large de repérage (capitale, parent, créateur…) qui correspond à diverses informations (père/frère, auteur/compositeur…). Ces formes sont parfois différentes d'une langue à l'autre et d'un mot à l'autre (par exemple, le repérage *capitale* donnera en français *capitale* ou *chef lieu*, suivant le nom propre considéré).

Par exemple, *Paris* –pivot : 38558- a pour expansion classifiante *la ville de* et pour expansion d'accessibilité *la capitale de*, toutes deux féminin et singulier :

```
<LexicalEntry partOfSpeech="noun">
  <Lemma>Paris</Lemma>
  <WordForm grammaticalGender="masculineFeminine"
grammaticalNumber="singular">Paris</WordForm>
  <Sense idSense="P38558" refSenseAxis="38558" termProvenance="fullForm" frequency="commonlyUsed"
label="properName">
    <SyntacticBehaviour refSubcategorizationFrame="CC222"/>
    <SyntacticBehaviour refSenseAxisRelation="1" refSubcategorizationFrame="AC4"/>
    <SyntacticBehaviour refSubcategorizationFrame="CO1"/>
    <MonolingualExternalRef externalSystem="Wikipedia"
externalReference="http://fr.wikipedia.org/wiki/Paris"/>
  </Sense>
</LexicalEntry>
```

Avec les descriptions suivantes :

```
<SubcategorizationFrame id="CO1" introducer="Determiner">zero</SubcategorizationFrame>
<SubcategorizationFrame id="CC222" introducer="classifyingContext" grammaticalGender="feminine" grammaticalNumber="singular">la ville de </SubcategorizationFrame>
<SubcategorizationFrame id="AC4" introducer="accessibilityContext" grammaticalGender="feminine" grammaticalNumber="singular">la capitale de la </SubcategorizationFrame>
```

4 Conclusion

Nous avons présenté dans cet article ProLMF 1.2 en détaillant les différences significatives avec la version 1.1. Cette ressource est disponible sur le site Prolex[15] du CNRTL (Centre national de ressources textuelles et linguistiques) sous une licence LGPL-LR, accompagnée d'un schéma XML. Le tableau 1 indique le nombre d'entrées pour chaque langue.

ProLMF 1.2		
fra	73 029	Entrées lexicales[16]
	10	Déterminants et prépositions locatives
	228	Expansions classifiantes
	101	Expansions d'accessibilité
deu	1 124	Entrées lexicales (Lemmes)
eng	790	
ita	751	
nld	683	
pol	8 236	
por	523	
spa	741	
srp-latn	1355	
srp-cyrl	980	

TABLE 1 – ProLMF 1.2 en chiffres

[15] http://www.cnrtl.fr/lexiques/prolex/.
[16] Dont 3 267 entrées lexicales obtenues par des règles d'aliasisation.

Nous avons comme perspective pour la poursuite de ce travail :

- le complément des liens vers Wikipédia et l'introduction de liens vers Geonames ;
- la mise en ligne sur le site du CNRTL d'une version 2.1 de ProLMF avec un nombre important d'entrées lexicales en anglais et en polonais ;
- l'ajout de la langue arabe (à plus long terme).

Remerciements

Nous remercions vivement ceux qui, après téléchargement de ProLMF 1.1, ont pris la peine de nous signaler des erreurs et des suggestions. Tout particulièrement Pascal Malaise, Karen Fort et Benoît Sagot. Nous remercions aussi Małgorzata Spędzia qui a créé le module polonais de la version 1.2.

Références

AUER S., LEHMANN J. (2007). What have Innsbruck and Leipzig in common? Extracting Semantics from Wiki Content. ESWC 2007. LNCS 4519:503-517.

BOUCHOU B., MAUREL D. (2008), Prolexbase et LMF: vers un standard pour les ressources lexicales sur les noms propres, *Traitement automatique des langues*, 49(1):61-88.

COSERIU E. (1998), Le double problème des unités dia-s, *Les Cahiers δια. Etudes sur la diachronie et la variation linguistique* 1:9-16.

COURTOIS B., SILBERZTEIN M. (1990), Dictionnaires électroniques du français, *Langues française*, 87:11-22.

FELLBAUM C., MILLER G. A. (2003), Morphosemantic Links in WordNet, TAL, 44(2):69–80.

FRANCOPOULO G., MONTE G., CALZOLARI N., MONACHINI M., BEL N., PET M., SORIA C. (2006), Lexical Markup Framework (LMF), LREC 2006.

FRANCOPOULO G., DECLERCK T., SORNLERTLAMVANICH V., DE LA CLERGERIE E., MONACHINI M. (2008), Data Category Registry: morpho-syntactic and syntactic profiles, Workshop *Uses and usage of language resource-related standards* (*LREC'2008*), 31-39.

HOFFART J., SUCHANEK F. M., BERBERICH K., WEIKUM G. (2012). YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. Artificial Intelligence Journal, Special Issue on Artificial Intelligence, Wikipedia and Semi-Structured Resources.

MANGEOT-LEREBOURS M., SÉRASSET G., LAFOURCADE M. (2003), Construction collaborative d'une base lexicale multilingue, le projet Papillon, TAL, 44-2:151-176.

MILLER G., BECKWITH R., FELLBAUM C., GROSS D., MILLER K. (1990), Introduction to WordNet : an on-line lexical database, *International Journal of Lexicography*, 3:235-244.

ROMARY L., SALMON-ALT S. FRANCOPOULO G. (2004), Standards going concrete: from LMF to Morphalou, in Workshop on Electronic Dictionaries, COLING-04.

TRAN M., MAUREL D. (2006), Prolexbase : Un dictionnaire relationnel multilingue de noms propres, *Traitement automatique des langues*, Vol. 47(3):115-139.