

**Prolexbase :**  
**Un lexique syntaxique et sémantique**  
**de noms propres**

Denis Maurel, Mickaël Tran, LI (Laboratoire d'Informatique de l'Université François-Rabelais de Tours)

Après avoir présenté rapidement notre ontologie des noms propres, nous détaillerons le regroupement sémantique de différents lemmes en un *nom propre conceptuel* et les relations sémantiques entre ces différents noms propres conceptuels. Puis nous introduirons les liens syntaxiques entre un ou plusieurs noms propres conceptuels et des grammaires locales.

## **1 Le projet Prolex**

Ce dont on a le plus besoin en TAL, c'est de disposer de lexiques à large couverture qui expliquent suffisamment de propriétés linguistiques et de relations entre les entrées pour qu'on puisse les utiliser comme instruments de traitement non seulement de la phrase simple, mais aussi de la phrase complexe et du discours. Il nous faut donc disposer d'information sur les relations entre noms propres (communes aux langues traitées) et sur leurs expansions (dans une langue particulière). Ces relations sont souvent à l'origine d'anaphores, pronominales ou lexicales, à l'intérieur d'un même texte.

Pour traiter correctement les noms propres dans le cadre du TAL, il est nécessaire, dans un premier temps, de définir un modèle de représentation pour permettre ensuite une création et une gestion cohérente d'un dictionnaire relationnel. Il s'agit d'étudier les différents aspects du domaine des noms propres et d'identifier les concepts, les relations et les attributs. Parmi les différentes approches ou modèles de représentation de l'information, nous avons décidé d'utiliser une approche ontologique [Gruber 1995].

Nous présenterons rapidement notre ontologie des noms propres (Figure 1). Les quatre niveaux de notre ontologie peuvent être regroupés en deux parties :

- Une partie supérieure commune aux langues traitées : les niveaux conceptuel (le référent suivant différents points de vue) et méta-conceptuel (la typologie et l'essence), hiérarchisés par la relation d'hyponymie. Elle contient les relations entre noms propres (Prédication, Méronymie, Synonymies synchronique et diachronique).
- Une partie inférieure particulière à une langue donnée : le niveau linguistique (morphologie dérivationnelle) et celui des instances (morphologie flexionnelle). Chaque langue possèdera sa propre arborescence à partir de la forme canonique d'un nom propre, forme que nous appellerons dans la suite *prolexème*. Il n'existe pas d'arborescence générale s'appliquant pour toutes les langues, étant donné les grandes divergences et la complexité des mécanismes s'appliquant sur les noms propres. Selon la langue, cette arborescence pourra être plus ou moins complexe.

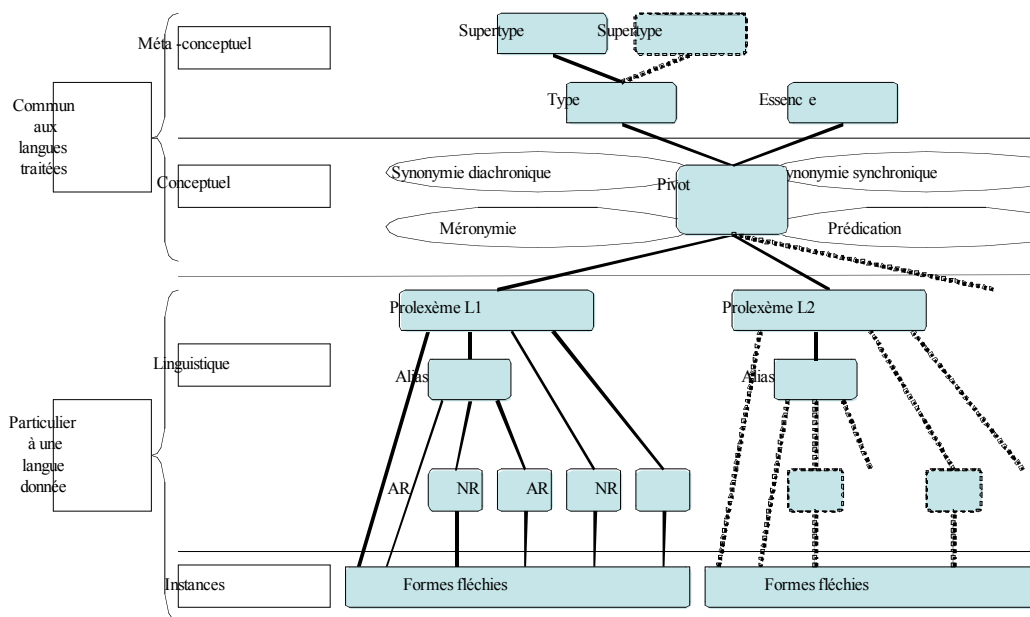


Figure 1 : L'ontologie Prolex

## 2 Une unité sémantique pour les noms propres : le prolexème

Le prolexème français regroupe sous une même unité sémantique plusieurs lemmes (et donc, au final, un grand nombre d'instances) : le lemme du nom propre lui-même, ceux de ses alias (variantes, abréviations, acronymes, sigles, transcriptions différentes...), de ses dérivés (même supplétifs) et ceux des dérivés de ses alias. La Figure 2 présente l'exemple du nom propre *Organisation des nations unies*.

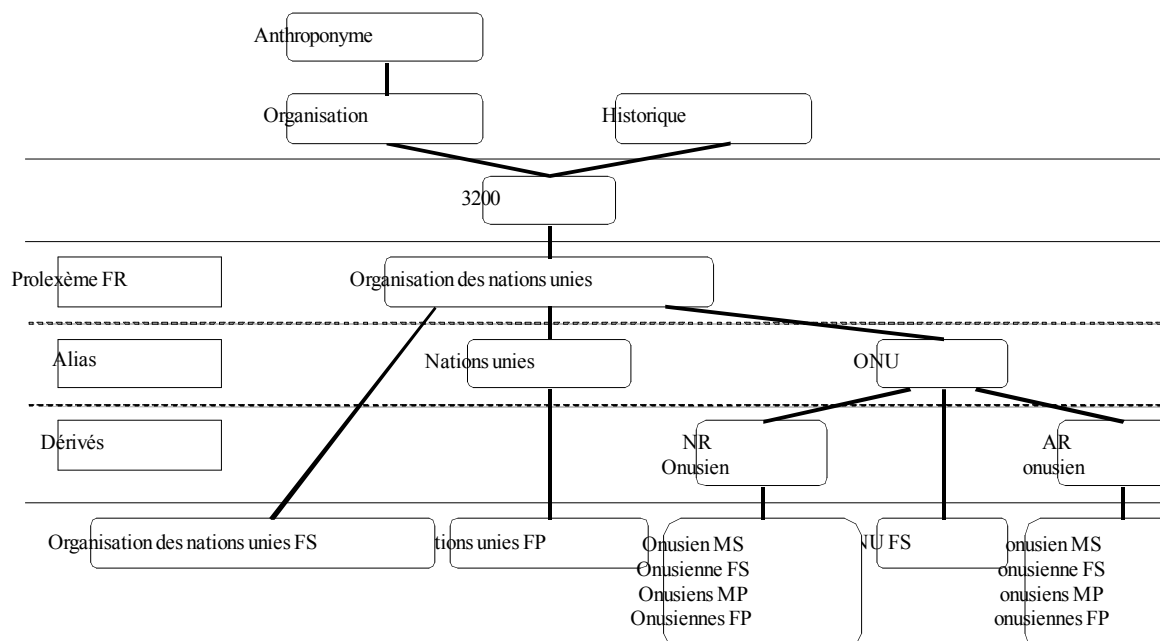


Figure 2 : Exemple avec le nom propre *Organisation des nations unies*

### 3 Des relations sémantiques entre prolexèmes

Le cœur de l'ontologie reste le niveau conceptuel, notamment à travers la notion de *nom propre conceptuel*, représenté par un numéro d'identité unique, le pivot. Un nom propre conceptuel ne correspond pas au référent linguistique, mais à un certain point de vue sur ce référent (diachronique, pragmatique, sociolinguistique, stylistique, thématique...). Le niveau conceptuel est un niveau interlangue. Ainsi, chaque nom propre représentant le même concept possèdera le même pivot. Donc un nom propre est la traduction d'un autre nom propre dans une autre langue s'ils partagent le même pivot.

Nous avons intégré dans le niveau conceptuel les relations sémantiques entre noms propres qui sont communes aux langues traitées (synonymie diachronique, synonymie synchronique, méronymie et prédication). Un exemple en français est donné Figure 3, celui des prolexèmes *Suisse* et *Confédération Helvétique*.

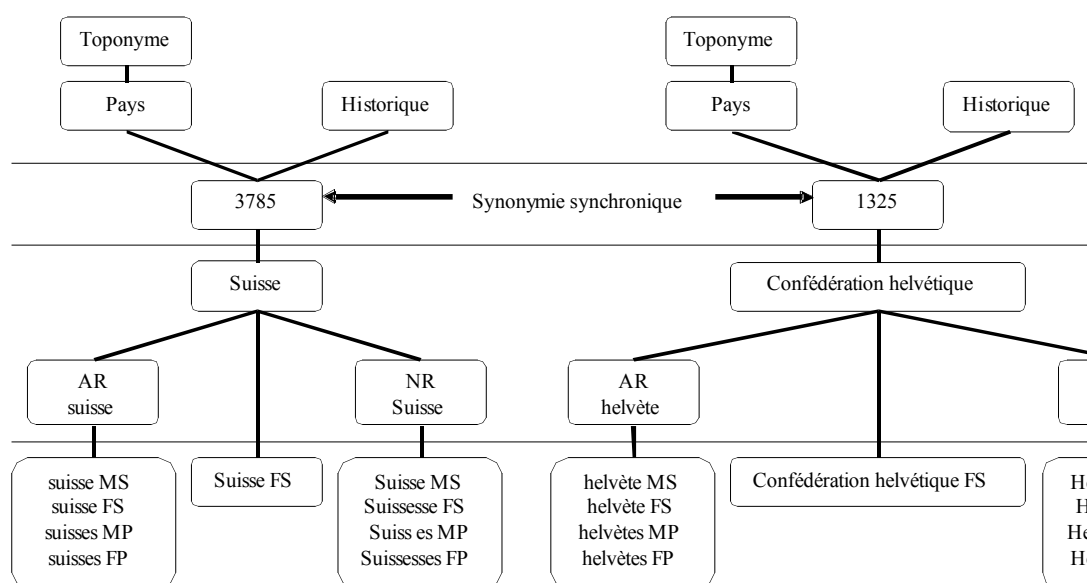


Figure 3 : Exemple en français avec les prolexèmes *Suisse* et *Confédération Helvétique*

Nous avons en fait créé deux relations de synonymie distinctes : la relation de synonymie diachronique qui concerne les noms propres qui ont été renommé pour diverses raisons (historiques, politiques, religieuses...) et la relation de synonymie synchronique qui s'applique à des noms propres ayant approximativement le même référent linguistique, mais selon différents points de vue.

La relation de prédication permet d'établir un lien sémantique entre deux prolexèmes. Cette relation s'inspire au départ de la fonction lexicale appelée Cap, que l'on trouve dans le *Dictionnaire Explicatif et Combinatoire du français contemporain* [Mel'cuk, 1984, 1988, 1992]. Par exemple, *Tours* est le *chef-lieu* de *l'Indre-et-Loire*, *Paris* est la *capitale* de la *France*, *Jacques Chirac* est le *président* de la *République française*, *Ray Norda* est le *patron* de *Novell*, *Mozart* est le *compositeur* de *La flûte enchantée*, *Jacques Chirac* est le *locataire* de *l'Élysée*, *Aaron* est le *frère* de *Moïse*...

Afin de faire un lien entre les noms propres et le lexique général, nous avons défini une relation entre les noms propres conceptuels et l'ontologie proposée par EuroWordnet [Vossen, 1997]. Cette relation associe un pivot à un numéro ILI. Si le nom propre est présent dans la base EuroWordnet, on lui associera son propre numéro ILI. C'est le cas de Paris, qui porte le

numéro 0558236n. Dans le cas contraire, on lui associera le numéro ILI correspondant à son insertion dans la hiérarchie. Par exemple, Jules Verne sera inséré sous le concept *writer*, qui porte le numéro 06438760n.

#### 4 Des liens entre syntaxe et sémantique par grammaire locale

Le prolexème, tout comme la relation de prédication, sont en lien avec une expansion et des grammaires locales [Gross, 1991]. La Figure 4 donne un exemple de grammaire locale attachée à la relation de prédication, par le prédicat "patron". Cette grammaire locale concernera entre autre le lien entre *Ray Norda* et *Novell*.

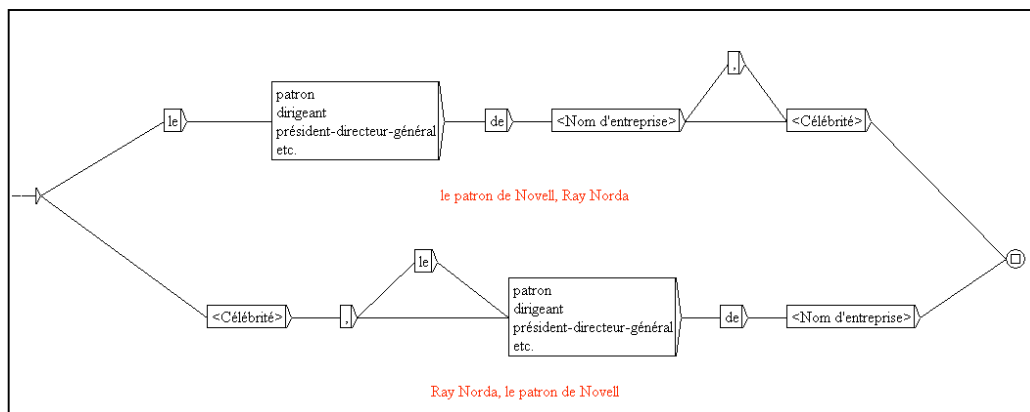


Figure 4 : Un exemple de grammaire locale

#### 5 Conclusion

Le dictionnaire Prolexbase va être utilisé pour différentes applications mono et multilingues : la recherche d'information, l'aide à la traduction et à la rédaction (correction d'orthographe), l'alignement de textes multilingues...

En guise de conclusion, nous donnerons quelques indications sur l'état d'avancement du projet Prolex et sur les outils

#### 6 Références

Gross M. (1991), Un ordinateur peut-il comprendre une langue naturelle ? Analyse automatique et couverture lexicale du français, *Annales des Mines*.

Gruber T. R. (1995), Toward Principles for the Design of Ontologies Used for Knowledge Sharing, *Int. Journal of Human-Computer Studies*, Vol.43, 907-928.

Mel'cuk I. (1984-I, 1988-II, 1992-III), *Dictionnaire explicatif et combinatoire du français contemporain*, Les presses de l'Université de Montréal.

Vossen P. (1997), EuroWordNet: a multilingual database for information retrieval, *DELOS workshop on Cross-language Information Retrieval*, Zurich.