**Corela**
Cognition, représentation, langage

**HS-2 | 2005**
**Le traitement lexicographique des noms propres**

# La glose comme outil de désambiguïsation référentielle des noms propres purs

**Montserrat Rangel Vicente**

**Édition électronique**
URL : http://journals.openedition.org/corela/1212
DOI : 10.4000/corela.1212
ISSN : 1638-573X

**Éditeur**
Cercle linguistique du Centre et de l'Ouest - CerLICO

**Référence électronique**
Montserrat Rangel Vicente, « La glose comme outil de désambiguïsation référentielle des noms propres purs », *Corela* [En ligne], HS-2 | 2005, mis en ligne le 02 décembre 2005, consulté le 02 avril 2021. URL : http://journals.openedition.org/corela/1212 ; DOI : https://doi.org/10.4000/corela.1212

# La glose comme outil de désambiguïsation référentielle des noms propres purs

Montserrat Rangel Vicente

## Introduction

1 Avec les noms propres purs, on a à faire du point de vue linguistique à une catégorie difficile à délimiter en raison de sa dimension interdisciplinaire, de l'hétérogénéité des éléments qui la constituent et de la variation à laquelle ces éléments sont assujettis. En effet, la notion nom propre est de nature interdisciplinaire, et concerne plusieurs domaines de savoir comme la linguistique, la logique, l'anthropologie, la sociologie... chacun lui imposant un certain nombre de critères définitionnels. Elle est également hétérogène, en allant, dans le cadre de la gestion de l'information, des anthroponymes aux noms de marque et de produit (en passant par les noms de découvertes humaines non matérielles). Elle est finalement contrainte à des variations de trois sortes :

- des formes propriales différentes avec le même sens[1] et le même référent, comme *JFK* et *John F. Kennedy* ;
- des formes propriales différentes avec des sens différents mais le même référent, comme *Leningrad* et *Saint Petersbourg* ;
- un même nom propre avec différents sens et différents référents, comme *Yves Saint Laurent*, qui peut être une personne, une marque ou un parfum.

2 Mon étude sera consacrée à la désambiguïsation référentielle - par le biais du contexte - des noms propres assujettis à ce dernier type de variation. Comme le constatent Daille

et Morin (2000 : 607), les noms propres nécessitent d'un cotexte large et parfois difficile à formaliser pour déterminer la nature de leurs référents :

> «/.../ dans la plupart des cas, une analyse locale suffit comme pour le Comte de Paris où l'apparition de Comte induit une catégorisation humaine, mais d'autres cas requièrent une analyse sémantique de la phrase /.../ ».

3 Tout nom propre est potentiellement polyréférentiel parce qu'il ne possède pas de sens lexical qui détermine la nature du référent auquel il serait susceptible d'être associé. Cela a comme conséquence que n'importe qui peut donner, par le biais d'un acte de baptême extralinguistique, n'importe quel nom propre à n'importe quelle sorte de référent. La désambiguïsation référentielle du nom propre semble donc devoir passer par une analyse de son contexte d'apparition. Le nom commun catégorisant qui accompagne le nom propre peut servir à ces fins parce qu'il contient des informations qui nous permettent de déterminer la nature du référent de ce dernier. Et c'est dans le but d'utiliser ces informations référentielles inscrites dans le nom commun catégorisant le nom propre que la structure que j'ai nommée glose peut être un outil de désambiguïsation référentielle efficace.

4 J'appelle *glose* l'ensemble des structures qui apparaissent dans le cotexte direct du nom propre et qui du point de vue morpho-syntaxique prennent la forme de syntagmes nominaux apposés et de propositions subordonnées. Dans le cadre de mes études concernant la production de sens du nom propre, j'ai défini deux types de glose du nom propre : les *gloses catégorisantes* et les *gloses caractérisantes*. Les premières sont celles qui contiennent une forme catégorisante ou un pronom sémantiquement marqué[2] qui nous permet d'inclure le référent du nom propre glosé dans une classe référentielle. C'est le cas, par exemple, de

5 */.../ Bonaparte, général en chef et membre de l'institut national*[3].

6 qui nous permet d'inclure le référent de Bonaparte dans deux classes référentielles humaines : celle de *général* et celle de *membre de l'institut national*.

7 Les *gloses caractérisantes* sont celles qui collaborent à la construction du référent discursif sans pour autant l'inscrire directement dans une catégorie référentielle. Prenons l'exemple suivant,

8 */.../ Bonaparte, qui fondait son pouvoir sur l'avilissement des corps aussi bien que sur celui des individus, /.../*[4].

9 Le but de ce genre de gloses est d'expliciter des éléments discursifs qui permettent de construire une certaine représentation du référent du nom propre sans pour autant nous donner des informations explicites sur la nature de ce dernier. Autrement dit, même si de façon implicite des éléments de cette sorte de structure nous permettent de cerner quel type de référent est nommé par le nom propre, aucun élément de ces gloses n'est susceptible d'être formalisé pour permettre l'inclusion du référent du nom propre dans une catégorie. C'est pourquoi, dans le cadre de mon étude, je me limite à la formalisation des gloses catégorisantes et à l'élaboration d'une grammaire locale de ce genre de structures syntaxiques susceptibles de déterminer le nom propre pur pour le désambiguïser référentiellement grâce aux éléments catégorisants qui les caractérisent. Nous verrons dans le point concernant la définition de ces derniers (qui peuvent être des formes catégorisantes ou des pronoms sémantiquement marqués) comment et pourquoi les gloses catégorisantes peuvent participer à la désambiguïsation des noms propres polyréférentiels. Mais avant cela, je vais présenter

le cadre théorique dans lequel j'élaborerai ma grammaire de la glose et les notions qui se trouvent au centre de cette réflexion.

# 1 Approche et choix terminologique

10 Mon étude se situe dans le croisement de deux branches de la linguistique : la sémantique nominale d'orientation praxématique[5] de l'Université Montpellier III et le traitement automatique du langage tel qu'il se fait au sein de l'équipe de recherche CLiC[6] de l'Université de Barcelone. De par mon orientation praxématique, j'analyserai dans le cadre de ce travail les outils mis en œuvre dans le discours par le locuteur en vue de désambiguïser référentiellement le nom propre. En raison de mon intérêt pour le traitement automatique du langage, je me servirai des outils informatiques lors de l'analyse de mon corpus pour automatiser ces dispositifs. En raison de cette double orientation, je vais devoir donner une définition consensuelle des notions qui se recoupent (sans pour autant être équivalentes) dans les deux domaines. C'est le cas de la notion sémantico linguistique de *nom propre*, qui s'oppose à celle *d'entité nommée* issue du TAL. C'est également celui de la désignation *forme catégorisante*, qui renvoie au même phénomène dans les deux disciplines mais dont la spécificité de la définition diffère de l'une à l'autre.

## 1.1 Les entités nommantes individualisantes : nom propre vs entité nommée

11 L'approche praxématique dans lequel je m'inscris conteste la conception classique du nom propre[7] représentée entre d'autres par Ullmann (1952) et Grevisse et Goosse (1986) parce qu'elle ne recouvre que les noms propres prototypiques (anthroponymes et toponymes), et parmi eux, seulement ceux qualifiés comme *purs*[8] par Jonasson (1994), en négligeant tout un ensemble d'éléments appartenant à cette catégorie. Elle conteste également la conception actuelle du nom propre, représentée par Kleiber (1981), Gary-Prieur (1994), Jonasson (1994) et Noailly (1999) parce qu'elle restreint la définition du sens au rôle de ce dernier dans la référenciation[9], ce qui relève d'une approche « en langue » du système linguistique qui s'avère peu rentable pour rendre compte du fonctionnement sémantico-discursif du nom propre. En effet, même si cette dernière conception reconnaît le rôle essentiel des informations véhiculées par le nom propre dans le fonctionnement en discours de ce dernier, elle leur nie un statut sémantique[10]. La seule tentative de légitimer le rôle de ces informations dans une étude linguistique est celle de Gary-Prieur (1994) avec sa notion de contenu du nom propre :

> «[Le contenu du nom propre est constitué par] l'ensemble de propriétés du référent initial du nom propre qui interviennent dans l'interprétation de certains énoncés concernant ce nom. /.../ La distinction entre ces deux niveaux de fonctionnement sémantique [le sens et le contenu] vise à expliciter ce qui fait la spécificité du nom propre par rapport au nom commun qui ne fonctionne, lui, qu'à niveau du sens ».

12 Cependant, cette conception résulte d'une description « en langue » orientée essentiellement à la description des fonctionnements prédicatifs ou modifiés[11] du nom propre.

13 La praxématique se fixe comme tâche « d'analyser comment les réglages intersubjectivement stables (ce que l'on considère naïvement comme des significations objectives) restreignent la signifiance, comment aussi, par quelles stratégies, la signifiance ressurgit dans les marges de la signification »[12]. Elle se situe du côté de la production (donc du discours) où « le sens est un produit, la signification un système d'outils (les praxèmes)[13] ; la signifiance est la production même »[14]. Dans ce contexte, le statut de tous les traits sémantiques inscrits en langue sont homogénéisés parce que tous signifient au sujet du rapport du langage au réel que le locuteur établit.

14 Pour la praxématique, le nom propre est un fonctionnement particulier du praxème,

> «c'est une vision délibérée qui extrait l'individu de cette catégorisation [celle opérée par le nom commun] par typisation, pour l'élire en tant qu'être singulier. Le praxème devenu nom propre est alors cause et effet de cette promotion à l'individualité. De manière indissociable il est l'instrument de l'individualisation et la marque de l'individualité ; il est outil de production de sens et sens produit. /.../ La spécificité du nom propre est là. En cela, et en cela seulement, il se distingue fonctionnellement du praxème ordinaire [nom commun] »[15].

15 La particularité du nom propre est donc celle de ne pas opérer, comme le nom commun, une catégorisation descriptive de l'entité nommée mais une *catégorisation individualisante* et que celle-ci soit réglée par un « fait social »[16] (sans intervention d'un sens quelconque) qui stipule qu'une certaine forme linguistique sert à nommer une certaine entité[17] « dans un espace donné de circulation de sens »[18], ce qui met à l'écart des expressions comme le *président Chirac* où seulement *Chirac* suspend sa production de sens au détriment de sa valeur désignationnelle.

16 En effet, la praxématique réserve la notion de nom propre exclusivement aux éléments qui suspendent leur production sémantique[19] pour désigner leur référent et privilégier la désignation individualisante, donc aux noms propres purs[20]. Cependant, même si aucun travail dans ce sens n'a pas encore vu le jour, en définissant le nom propre comme un fonctionnement et non comme un type de catégorie lexicale et en le caractérisant par rapport au type de catégorisation qu'il opère, la praxématique permet de rendre compte aisément du fonctionnement individualisant de certaines entités discursives qui sont constituées d'éléments autres que les noms propres purs, comme *République Française*. Je propose de nommer ce genre de structures *entités nommantes individualisantes* par opposition à la notion de *nom propre* (qui recouvre seulement une partie de ces dernières) et à celle *d'entité nommée* issue du TAL qui, comme nous allons le voir ci-dessous, ne rend pas compte de la réalité qu'elle est censée de recouvrir.

17 *Entité nommée* est la notion utilisée en TAL pour désigner les éléments discursifs monoréférentiels qui coïncident en partie avec les noms propres[21] et qui suivent des patrons syntaxiques déterminés[22]. Du point de vue de leur structure syntaxique, les entités nommées sont *fortes* quand elles sont composées exclusivement de noms propres et *faibles* quand elles sont constituées par un nom propre (ou entité nommée forte) et une forme catégorisante.

18 Même si du point de vue conceptuel j'adhère presque entièrement[23] à la notion d'entité nommée tel que le TAL la définit, je propose de remplacer cette désignation par celle d'entité nommante individualisante. La raison est que qualifier l'entité de *nommée* me semble un choix qui peut induire à l'erreur et convenir plus au résultat de l'actualisation linguistique (la représentation du référent construite par le discours) ou à l'entité du réel visée qu'à la structure linguistique utilisée à ces fins. La solution

désignationnelle qui me semble s'adapter le plus à la réalité que nous voulons désigner est donc celle d'entité *nommante*, à laquelle j'ajoute l'expansion *individualisante* pour expliciter la fonction caractéristique qu'elle opère[24]. Je conserve toutefois pour ma description la différenciation établie en TAL pour les entités nommées entre fortes et faibles parce qu'elle me permet de définir le nom propre (tel que la praxématique le conçoit) comme étant une entité nommante forte[25].

19 L'interprétation de l'entité nommante forte demande la prise en compte d'informations textuelles autres que celles véhiculées exclusivement par la forme commençant par une majuscule[26]. Par ailleurs, le nom propre et la glose catégorisante constituent une unité sémantico-discursive au même titre que le nom propre et la forme catégorisante dans le cas du nom propre mixte, forme prototypique des entités nommantes faibles[27]. C'est pourquoi, même si la glose relève de *l'évidence externe*[28] du nom propre, je vais inclure la structure formée par ces deux éléments (*glose + nom propre*) dans la catégorie des entités nommantes faibles. Ce choix s'explique par le contexte théorique dans lequel je développe ce travail, qui privilégie la construction sémantico-discursive du référent du nom propre issue de l'interaction des éléments linguistiques qui l'accompagnent au détriment de la façon selon laquelle chaque élément opère, de façon isolée, cette construction.

## 1.2 Forme catégorisante : une même désignation et deux notions différentes ?

20 Du constat de Daille et Morin cité ci-dessus ensuit que ce n'est que par le co(n)texte que nous pourrons cerner la nature réelle du référent du nom propre. Parce que dans l'analyse des textes nous n'avons presque jamais accès aux conditions de production de l'énoncé (le contexte), nous sommes obligés de centrer notre analyse sur des éléments textuels appartenant au contexte discursif (le cotexte) susceptibles de nous permettre de déterminer la nature du référent du nom propre.

21 Les gloses catégorisantes s'avèrent un outil privilégié pour déterminer la nature du référent du nom propre glosé, et cela grâce aux formes catégorisantes qui les constituent. Pour cela, il suffit de repérer la nature du référent prédiquée par ces formes et de la transposer aux noms propres déterminés par les gloses dans lesquelles elles sont insérées. C'est ce qui font des systèmes tels que MICE (de TALP[29]-CLIC) et PROLEX[30] (de l'Université François-Rabelais de Tours, Laboratoire d'informatique), qui s'appuient sur le cotexte d'apparition des noms propres pour construire des grammaires locales qui permettent d'analyser leur fonctionnement en discours. Pour cela, ils font appel à la notion de *forme catégorisante* qui, telle que le TAL la définit, ne se limite pas à classer sémantiquement l'entité nommante à laquelle elle est associée. En effet, ces mots (qui apparaissent dans le contexte immédiat du nom propre) permettent également de détecter et de délimiter les entités nommantes faibles dont elles font partie, en apportant des informations concernant le référent du nom propre (comme le genre, le nombre et la nature du référent). C'est pourquoi nous allons adopter la définition de la forme catégorisante donnée dans ce contexte au détriment de celle employée jusqu'ici relevant de l'analyse du discours.

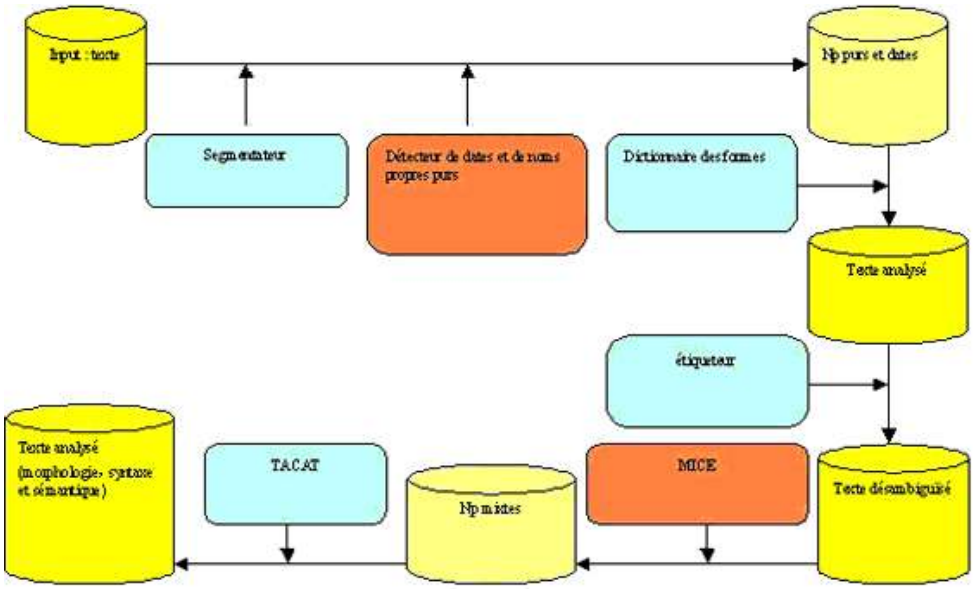
22 Avant d'exposer d'une façon plus concrète comment je pense utiliser cette notion dans le cadre de la désambiguïsation référentielle du nom propre, je vais présenter le système dans lequel ma grammaire locale va être insérée et les éléments qui vont me

permettre de la construire pour parvenir à systématiser les entités nommantes faibles du type *nom propre + glose catégorisante*.

## 2 MICE, le Module d'Identification et Classification des Entités du système TALP-CLIC

23 Je construis ma grammaire de la glose de l'entité nommante forte au sein de l'équipe CLIC qui, parmi ses multiples projets, se consacre au développement de recherches concernant le traitement des entités nommantes et notamment à une typologie spécifique à celles-ci qui permet leur identification et leur étiquetage en vue d'en établir une classification sémantique. Ce système, sensible au contexte comme ceux de Coates-Stephens (1992), de Paik et al. (1996), de Sheretmeyeva (1998) ou encore PROLEX, dispose d'un ensemble de ressources :

- un module informatique d'identification des dates et des entités nommantes fortes par le biais des majuscules (sauf les noms propres qui apparaissent après le point et les sigles) ;
- un dictionnaire de formes qui permet d'identifier l'ensemble des catégories grammaticales auxquelles les éléments analysés peuvent appartenir;
- un désambiguïseur qui identifie la fonction grammaticale que les formes repérées développent dans le cotexte d'apparition et dont le résultat est un document étiqueté morphologiquement ;
- MICE[31], qui est un analyseur superficiel qui permet l'identification et la classification sémantique et typologique des entités nommantes relevées partiellement lors de processus antérieurs en élargissant le cadre d'analyse aux mots qui avec lui constituent l'expression syntagmatique entité nommante faible[32]. Ce parenthiseur est composé de :
  - Une grammaire computationnelle des entités nommantes faibles ;
  - TACAT, élaboré par Atserias et Rodríguez (1998), qui est un analyseur syntaxique superficiel, c'est-à-dire, un chunker.




Système TALP-CLIC

24 MICE est le dispositif du système TALP-CLIC au sein duquel ma grammaire de la glose sera intégrée. Comme je viens de l'exposer, son rôle est celui de délimiter les entités nommantes faibles en syntagmes par le biais de formalismes construits autour des formes catégorisantes qui permettent d'exprimer les caractères syntaxiques et sémantiques des entités nommantes faibles. Je vais vous présenter un peu plus en détail les outils qui composent ce dispositif.

## 2.1 La taxonomie des entités nommantes

25 Les éléments dont MICE dispose pour traiter les entités nommantes faibles sont une typologie des entités nommantes, un dictionnaire des formes catégorisantes qui sont classées en fonction de la taxonomie des entités nommantes et une grammaire computationnelle qui joue le rôle d'analyseur syntaxique.

26 La typologie des entités nommantes élaborée par Arévalo et Simón (2000) est construite à partir de taxonomies utilisées pour la classification thématique de dictionnaires et encyclopédies et de la classification de WordNet adaptée au nom propre. La classification des entités nommantes demande d'une typologie complète qui permette de rendre compte de la complexité sémantique de ces structures et qui ne peut pas se voir limitée à la classification proposée par le MUC. Arévalo, M. et Simón, M.-J. (2000), après avoir analysé certaines de ces classifications - comme celles de Sheretmeyeva (1998) ou celle de Paik et al. (1996) -, ont proposé une classification typologique (compatible avec la classification du MUC) susceptible de rendre compte de la structure des entités nommantes apparaissant dans plusieurs types de texte (principalement journalistiques) et facile à adapter aux systèmes d'extraction et de récupération d'information. Elle est structurée avec trois niveaux de spécification dont 6 super-classes (géographie, êtres, organisations, documents et ouvrages, temporels et autres) et 56 sous-classes ou terminaux[33].

Taxonomie des entités nommantes de MICE

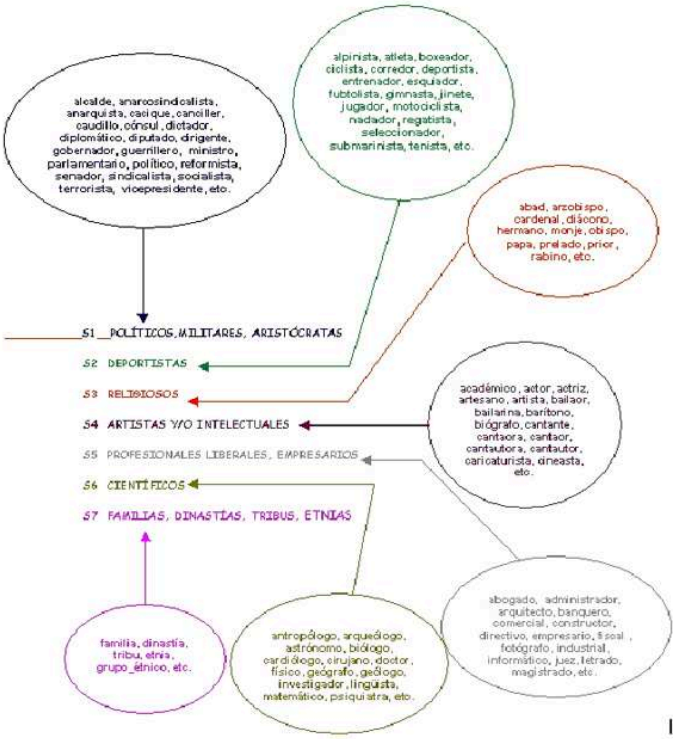
| | | | |
|---|---|---|---|
| NP00G00 | GEOGRAFÍA | | |
| | NCOOGG0 | | GEOGRAFÍA NATURAL |
| | | NP00G10 | UNIVERSO (ASTROS, GALAXIAS...) |
| | | NP00G20 | TIERRA (ACCIDENTES GEOGRÁFICOS, OCÉANOS...) |
| | NP00GP0 | | GEOPOLÍTICA |
| | | NP00G30 | DIVISIONES ADMINISTRATIVAS (PAÍSES, REGIONES, CIUDADES) |
| | | NP00G40 | LUGARES DE ACTIVIDADES HUMANAS (ECONÓMICAS, CULTURALES, ARQUEOLÓGICOS, DE PATRIMONIO HISTÓRICO, ETC.) |
| | NP00GF0 | | LUGARES FICTICIOS, MÍTICOS, MITOLÓGICOS |
| NP00S00 | SERES | | |
| | NP00SP0 | | PERSONAS |
| | | NP00S10 | POLÍTICOS, MILITARES, ARISTÓCRATAS |
| | | NP00S20 | DEPORTISTAS |
| | | NP00S30 | RELIGIOSOS |
| | | NP00S40 | ARTISTAS Y/O INTELECTUALES |
| | | NP00S50 | PROFESIONALES LIBERALES, EMPRESARIOS |
| | | NP00S60 | CIENTÍFICOS |
| | | NP00S70 | FAMILIAS, DINASTÍAS, TRIBUS, ETNIAS |
| | NP00SA0 | | ANIMALES/PLANTAS |
| | NP00SD0 | | SERES DIVINOS, MÍTICOS, MITOLÓGICOS, LEGENDARIOS |
| | NP00SF0 | | SERES DE FICCIÓN |
| NP00O00 | ORGANIZACIONES | | |
| | NP00OE0 | | EMPRESAS |
| | NP00OI0 | | INSTITUCIONES |
| | NP00OO0 | | ORGANIZACIONES |
| | NP00OC0 | | COLECTIVOS, ASOCIACIONES, GRUPOS |

| | | | | |
|---|---|---|---|---|
| NP00D00 | DOCUMENTOS Y OBRAS | | | |
| | NP00DL0 | | LEGALES, POLÍTICOS, HISTÓRICOS | |
| | NP00DA0 | | ARTÍSTICOS, INTELECTUALES | |
| | | NP00D10 | LIBROS, PUBLICACIONES | |
| | | NP00D20 | ARTES PLÁSTICAS | |
| | | NP00D30 | MÚSICA Y DANZA | |
| | | NP00D40 | CINE Y TEATRO | |
| | | NP00D50 | RADIO Y TELEVISIÓN | |
| | NP00DR0 | | RELIGIOSOS | |
| | NP00DV0 | | VIRTUALES (SITIOS WEB) | |
| NP00T00 | TEMPORAL | | | |
| | NP00TF0 | | FECHAS/HORAS | |
| | NP00TP0 | | PERÍODOS, ERAS, SIGLOS | |
| | NP00TE0 | | EVENTOS | |
| | | NP00T10 | NATURALES | |
| | | NP00T20 | SOCIALES | |
| | | NP00T30 | HISTÓRICOS | |
| NP00V00 | VARIOS | | | |
| | NP00VA0 | | ARTEFACTOS, APARATOS | |
| | NP00VD0 | | DENOMINACIONES DE ORIGEN, MARCAS, PRODUCTOS | |
| | NP00VC0 | | CONDECORACIONES, PREMIOS, GALARDONES | |
| | NP00VO0 | | OBJETOS SINGULARES | |
| | NP00VE0 | | ENFERMEDADES, SÍNDROMES | |
| | NP00VN0 | | ANATOMÍA (PARTES DEL CUERPO) | |
| | NP00VQ0 | | ELEMENTOS QUÍMICOS | |
| | NP00VT0 | | TEORÍAS, PRINCIPIOS, LEYES, MÉTODOS | |

| | | |
|---|---|---|
| | NP00VP0 | PLANES, PROGRAMAS, PROCESOS JUDICIALES, ETC. |
| | NP00VM0 | MONEDAS |
| | NP00VI0 | IMPUESTOS |

## 2.2 Le dictionnaire des formes catégorisantes

27 Chaque noyau de la typologie se voit associé les formes catégorisantes qui expriment lexicalement son type sémantique. Chaque forme catégorisante, à son tour, est associée au type sémantique qu'elle représente et à une étiquette - qui suit le standard Eagles[34] - qui la caractérise du point de vue morpho-syntaxique. Outre l'identification et la description syntaxique des noms propres, les formes catégorisantes rendent possible leur classification sémantique et typologique.




**Association du paradigme des formes catégorisantes avec la taxonomie des entités nommantes dans la classe «Personne »**

28 Le paradigme des entités nommantes est composé d'environ 4000 lèmes et 13.000 formes fléchies. Il est construit à partir de classifications existantes (telles qu'EuroWordNet et celle des encyclopédies), de l'observation empirique de corpus étiquetés à la main et des connaissances du monde des personnes qui l'ont élaboré. Il est structuré hiérarchiquement, ce qui permet de traiter les étiquettes à différents niveaux au moment de faire la grammaire.

29 Structuration hiérarchique du paradigme des formes catégorisantes concernant les entités nommantes «êtres » :

30 *TWS (êtres)*

31 *TWSP (personne)*

32 *TWSPP (métier)*

33 *TWSPPP (homme politique)*

34 *TWSPPM (militaire)*

35 *TWSPPD (sportif)*

36 *TWSPPV (autres)*

37 *TWSPD (titres et dignités)*

38 *TWSPDS (social)*

39 *TWSPDR (religieux)*

40 *TWSA (animaux et plantes)*

41 *TWSF (personnage de fiction)*

## 2.3 L'analyseur syntaxique des entités nommantes

42 La grammaire des entités nommantes de CLIC opère sur le texte déjà analysé et désambiguïsé morphologiquement. Son but est d'analyser aussi bien du point de vue syntaxique que sémantique les éléments composant les entités nommantes faibles du type nom propre descriptif qui, comme nous l'avons vu ci-dessus (Daille et Morin, 2000 : 607), présentent l'avantage de voir précédé le nom propre pur d'une forme classifiante qui élimine l'ambiguïté quant au type de référent visé. Le résultat de l'application de cette partie du système est une entité nommante faible sous forme de syntagme nominal étiqueté morpho-syntaxiquement et sémantiquement :

43 *El (tdms0) presidente (ncns0) González (np000) --> [ El_presidente_González] SN-político (npmss10).*

44 Pour cela, l'outil construit des règles syntactico-sémantiques décrivant la structure des entités nommantes faibles en prenant comme point de départ les formes catégorisantes détectées dans le corpus d'observation (qui sont intégrées dans les règles de manière littérale). Ces règles sont de trois natures différentes : lexicales, regroupantes et syntaxiques.

45 Les **règles lexicales** permettent d'assigner une étiquette sémantique au lexique terminal de la grammaire (mots catégorisants et gentilés).

46 Les **règles regroupantes** permettent de simplifier la notation en regroupant sous un même élément non terminal d'autres éléments non terminaux qui partagent des aspects morpho-syntaxiques.

47 Les **règles syntaxiques** correspondent à des patrons morpho-syntaxiques qui nous permettent de reconnaître les composantes des entités nommantes faibles par le biais de l'analyse des constituants des syntagmes nominaux dans lesquels apparaissent les formes catégorisantes. Autrement dit, ces règles permettent de décrire la structure interne des entités nommantes faibles, comme le montre l'exemple suivant concernant la NE *homme politique* :

48 *np00s10 ==> tw-politico, gentilicio, prep, np, gentilicio.*

49 Cette règle se traduit de la sorte : si on trouve une TW du type *homme politique*, suivi optionnellement d'un gentilé, d'une préposition (*de/del*), d'un *np* (entité nommante forte) et optionnellement d'un *gentilé*, la structure en question fait référence à une personne du type politique. Cette formulation permet de reconnaître des expressions telles que *ministro de Economía polaco, ministro polaco de Economía, diputado [valenciano] de IU et alcalde de Jerez*.

50 Comme je l'ai exposé ci-dessus, la désambiguïsation sémantique opérée par MICE concerne les entités nommantes faibles du type nom propre descriptif. Mais, comment fait-on quand le nom propre n'est pas précédé d'un mot catégorisant qui prédique la classe à laquelle appartient le référent du nom propre ? C'est pour répondre à ce besoin que je vais construire mon outil de désambiguïsation des entités nommantes fortes et qui concerne celles d'entre elles qui présentent la caractéristique d'être accompagnées de gloses catégorisantes.

# 3 Les gloses comme outil de désambiguïsation sémantico-référentielle des entités nommantes fortes

51 Les problèmes qui se dégagent de la polyréférentialité et de la polysémie du nom propre sont surmontables en partie grâce à l'analyse des gloses, qui présentent comme avantage leur forme synthétique, leur proximité immédiate au nom propre et leur régularité formelle facilement systématisable.

52 La formalisation des gloses désambiguïsatrices des entités nommantes fortes[35], qui sera partie intégrante de la grammaire des entités nommantes de MICE, va être construite en premier temps pour l'espagnol, même si une extension au catalan est envisagée. Pour élaborer ma grammaire des gloses, il me faut compter avec le patron de l'ensemble des gloses catégorisantes des entités nommantes fortes. C'est pourquoi je vais construire un corpus de textes journalistiques en ligne (journaux et revues en ligne)[36] borné temporellement (des années 90 à nos jours). Pour cela, je vais extraire l'ensemble de mes occurrences d'un corpus déjà constitué par CLiC de l'agence de presse EFE de soixante millions de mots-texte et des quotidiens espagnols en ligne (El Periódico, El País...).

53 Je m'apprête à traiter deux sortes de gloses catégorisantes qui se différencient en raison de leur nature syntaxique : les propositions de relatif à pronom non ambigu (*donde* et *quien*) et les appositions nominales qui incluent une forme catégorisante qui détermine la classe à laquelle appartient le référent de l'entité nommante forte. La grammaire locale qui concerne ce dernier type de glose est celle qui m'intéresse le plus parce qu'elle me semble la plus rentable. C'est à sa présentation que je consacrerai le dernier point de cet article.

## 3.1 Les relatives à pronom sémantiquement marqué

54 Pour ce qui est des gloses de nature relative en espagnol, je suis obligée de me limiter au traitement de celles introduites par des pronoms relatifs sémantiquement marqués : *(préposition +) quien*, qui reprennent toujours des antécédents humains (des anthroponymes dans le cas des noms propres), et *(préposition +) donde*, dont le référent est toujours un lieu (des toponymes dans le cas des noms propres)[37].

55 Même si mon projet au départ incluait tous les pronoms relatifs, j'ai été obligée de réduire la formalisation à ces deux relatifs parce que le traitement de l'ensemble demanderait de passer par le traitement du sémantisme des verbes, et je ne dispose pas d'outils pour mener à terme une tâche de la sorte. Même si l'apport informatif de ce genre de formalisations permettra de cerner uniquement la nature générale du référent, je pense qu'il vaut la peine de les inclure dans MICE parce qu'elles peuvent désambiguïser des exemples comme le suivant :

56 */.../ Helena, reina de Esparta y esposa de Menelao, ante cuya legendaria belleza sucumbe Paris,* ***quien*** *la rapta y hace su esposa para desgracia de su raza.* [38]

57 Pour mener à terme cette tâche, je devrai rechercher le contexte d'apparition des pronoms relatifs que j'ai présenté comme état sémantiquement transparents. Ensuite, je devrai étiqueter morphologiquement mon corpus (ce que je ferai grâce au dictionnaire des formes et à l'étiqueteur) et formuler des règles qui reprendront l'ensemble de structures relevant de ce genre de gloses grâce à TACAT, l'analyseur syntaxique de MICE. Pour finir, j'associerai à ces gloses des étiquettes correspondant à la nature référentielle de leur antécédent : « lieu » pour *donde* et « personne » pour *quien*. Une phase incontournable de ce travail sera la vérification de la rentabilité de mon outil par le biais de son application à un corpus de test[39], ce qui me permettra de légitimer son inclusion dans le système TALP-CLIC.

## 3.2 Les appositions incluant une forme catégorisante

58 Cette partie de mon travail vise à formaliser des appositions nominales qui déterminant des entités nommantes fortes et qui incluent des formes catégorisantes comme dans les cas suivants :

59 *Paris, el nuevo* ***perfume*** *de Yves Saint Laurent,...*

60 *Paris,* ***capital*** *de Francia, ...*

61 *Paris,* ***príncipe troyano****, ...*

62 Pour élaborer cette partie du projet, je rechercherai l'ensemble des noms propres qu'apparaissent[40] dans mon corpus. Je dois ensuite détecter le patron syntaxique des gloses qui les déterminent. Pour cela, je devrai étiqueter morphologiquement mon corpus à l'aide de l'analyseur de MICE et formuler des règles qui rendront compte de la structure morpho-syntaxique des gloses concernées. La phase suivante consistera à mettre en lien les formalisations auxquelles je serai arrivée par cette première analyse avec la typologie de entités nommantes établie par Arevalo-Simon (2000) qui est associée au répertoire des formes catégorisantes. Pour finir, j'appliquerai ma grammaire locale à une partie de mon corpus étiquetée à la main pour vérifier si elle est assez rentable pour légitimer son inclusion dans le système TALP-CLIC.

63 Je vais analyser plus en détail un sous-ensemble construit à partir d'un nombre limité de noms propres référentiellement ambigus pour vérifier s'il y a une différence entre les gloses désimbiguïsantes et celles dont le seul but est celui de catégoriser référentiellement le référent du nom propre. Pour construire ce sous-corpus, je vais déterminer (grâce à mes connaissances du monde) un ensemble de noms propres ambigus tel qu'il recouvre chaque terminal de la typologie de CLIC (il y en a 56). De cette façon, en plus de recenser les structures morpho-syntaxiques des gloses utilisées pour désambiguïser référentiellement les entités nommantes fortes, je vais pouvoir

augmenter le dictionnaire des mots catégorisants de CLIC. Certains des noms propres déjà sélectionnés pour cette tâche sont *París* (toponyme-administratif, être/mythologique et autres-marque/produit), *Agata Ruíz de la Prada* (être-artiste et autres-marque/produit), *Amazonas* (toponyme-terre et être-mythologie/légendaire), *Oxfam* (organisation-collectif/association et autres-marque/produit), *Guernica* (toponyme-administratif et ouvrages-arts plastiques), *11 de septiembre del 2001* (temporel-date et temporel historique) et *César* (être-politique et autres-prix).

# 4 En guise de conclusion

64 En l'absence de résultats et de diagnostiques d'efficacité (parce que la grammaire des gloses n'a pas encore été construite), je vais vous présenter certaines hypothèses concernant les avantages et les limites de cet outil. Du point de vue syntaxique, cet outil va nous permettre d'établir une typologie formelle des structures susceptibles de compléter les noms propres et de les désambiguïser. Il a en plus comme avantage la relative facilité d'élaboration une fois que les outils (tels que TACAT et le dictionnaire des formes catégorisantes) sont construits et mis à disposition.

65 Ma grammaire des gloses ne se veut être qu'un outil de désambiguïsation référentielle des entités nommantes fortes construit dans le but d'affiner les résultats du système TALP-CLIC. En vue des résultats obtenus par CLIC dans l'analyse des entités nommantes faibles et par les grammaires locales du nom propre d'autres systèmes tels que PROLEX, il semble légitime de croire à l'intérêt de l'inclusion d'une phase de reconnaissance des gloses catégorisantes parmi les premières étapes des programmes de désambiguïsation contextuelle des noms propres en espagnol même si les résultats qui peuvent s'en dégager pourraient se voir surmontés par des analyses syntaxiques plus larges.

66 Le résultat de l'application de ma grammaire locale sera l'insertion du nom propre dans l'entité nommante faible qu'il constitue avec la glose catégorisante qui le détermine. L'entité nommante faible en question est du même type que celle qui résulterait de l'analyse d'un nom propre descriptif, à savoir un syntagme nominal nom propre catégorisé sémantiquement et morpho-syntaxiquement. Cependant, il ne faut pas oublier que, pour ce qui est de la désambiguïsation référentielle du nom propre, nous sommes dans les cadres de l'extraction et de la récupération d'information (qui n'ont en aucun cas l'ambition d'offrir une vision grammaticale de l'aspect analysé)[41]. Il n'en sera pas de même lors de l'utilisation des résultats de ce travail dans la partie linguistique de ma thèse, où le nom propre pur sera différencié du nom propre descriptif (par des critères strictement linguistiques) et la grammaire de la glose sera présentée comme une description syntaxique de certaines structures utilisées par le locuteur pour faciliter la désambiguïsation référentielle des noms propres actualisés dans son discours.

## BIBLIOGRAPHIE

CLIC : http://clic.fil.ub.es/

Praxiling : http://www.univ-montp3.fr/praxiling/

PROLEX : http://tln.li.univ-tours.fr/

Arévalo, M. et Simón, M., 2000, Tratamiento computacional de los nombres propios, doctorat du département de LSI, Universitat Politécnica de Catalunya, Barcelona.

Arévalo, M., Carreras, J., Màrquez, L., Martí, A., Padró, L. et Simón, M., 2002, A Proposal for Wide-Coverage Spanish Named Entities Recognition, Revista de Procesamiento de Lenguaje Natural, 28, Alicante : SEPLN ; pp. 63-80.

Arévalo, M., Civit, M. et Martí, A., 2004, MICE : A module for Named Entities Recognition and Classification, International Journal of Corpus Linguistics, 9 : 1, Amsterdam/Philadelphia : John Benjamins Publishing Company, pp. 53-68.

Atserias, J. et Rodríguez, H., 1998, TACAT : Tagged corpus Analyser Tool, Repport LSI-RT-2-98, Universitat Politéctica de Catalunya, Barcelona.

Chinchor, N., 1987, Muc-7 Named Entity Task Definition, consultable sur le site http://www.itl.nist.gov/iad/894.02/related_projects/muc/proceedings/ne_task.html

Coates-Stephens S., 1992, The Analysis and Acquisition of Proper Names for Robust Text Understanding, thèse doctorale, Department of Computer Science, City University, London.

Daille et Morin, 2000, Reconnaissance automatique des noms propres de la langue écrite : les récentes réalisations, Traitement Automatique des Langues (TAL), 41(3), pp. 601-622.

Gary-Prieur, M.N., 1994, Grammaire du nom propre, Paris : Presses Universitaires de France.

Grevisse, M. et Goosse, A., 1986, Le bon usage, Belgique : Duculot, Gembloux.

Jonasson, K., 1994, Le nom propre. Constructions et interprétations, Paris : Duculot.

Kleiber, G., 1981, Problèmes de référence : descriptions définies et noms propres, Paris : Klincksieck.

Kripke, S., 1972/1982, La logique des noms propres (Naming and Necessity), Paris : Minuit.

Lafont, R. et Gardès-Madray, F., 1976 [1989], Introduction à l'analyse textuelle : index des termes linguistiques, Montpellier : Editions de l'Université Paul Valéry.

Maurel, D., 2004, Les mots inconnus sont-ils des noms propres ?, Septièmes journées internationales d'Analyse statistique des Données Textuelles (JADT 2004), Louvain-la Neuve, Belgique, 10-12 mars.

Noailly, M. 1999, «La querelle des noms propres », Modèles linguistiques 20-1, pp. 107-112.

Paik W., Liddy E., Yu E., McKenna M., 1996, Categorizing and Standarizing Proper Names for Efficient Information Retrieval, Corpus Processing for Lexical Acquisition.

Rangel Vicente, M., 2003, «Nom propre et dialogisme : la construction de la représentation de Napoléon Bonaparte en France et en Espagne », in Cassanas, Demange, A., Laurent, B. et Lecler, A.,(éds.) Dialogisme et nomination, Montpellier : Presses de l'Université Paul Valéry, pp. 199-214.

Rangel Vicente, M., 2004, «Le nom propre en discours : statut et fonctionnement des informations référentielles », in Dufour, F., Dutilleul-Guerroudj, E. et Laurent, B., (ed.), La nomination, quelles problématiques, quelles orientations, quelles applications ?, Montpellier : Presses de l'Université Paul Valéry, pp. 129-140.

Sheremetyeva Sv., Cowie J., Niremburg S., Zajac R., 1998, A Multilingual Onomasticon As a Multipurpose NLP Resource, Proceedings of ELRA Conference, Granada.

Siblot, P., (éd.) 1987, «Signifiance du nom propre », Cahiers de praxématique, 8 (Le nom propre), pp. 97-116

Ullmann, S., 1952, Précis de sémantique française, Bern : A. Francke.

## NOTES

**1.** J'expliquerai ma conception sémantique du nom propre dans le point 1.1.
**2.** En espagnol, c'est le cas des pronoms relatifs *quien* et *donde*, qui renvoient à des antécédents dont le référent ne peut être que de nature humaine pour le premier et locative pour le deuxième.
**3.** De Staël, 1817, *Considérations sur la révolution française* : 2.
**4.** De Staël (1817 :14).
**5.** http://www.univ-montp3.fr/praxiling/
**6.** *Centre de Llenguatge i Computació* de l'Université de Barcelone (http://clic.fil.ub.es/).
**7.** » Les noms propres n'ont pas de sens et, par conséquent, la notion de signification ne s'applique pas à eux. La fonction d'un nom propre est *l'identification pure* : distinguer et individualiser une personne ou une chose à l'aide d'une étiquette spéciale ». Ullmann (1952 : 3).
« [Le nom propre est un signe qui établit] l'association durable d'un nom X et d'un objet x » Kleiber (1981 : 279). Il ajoute « /.../ le lien entre un nom propre et l'objet qu'il désigne n'est pas un lien qui relève de la sémantique linguistique ». Kleiber (1981 : 381)
**8.** Les noms propres purs sont des formes mono-lexicales utilisées en général pour désigner des personnes et des lieux et qui sont spécialisées dans le rôle de nom propre.
**9.** Même si je ne conteste pas la légitimité de l'approche référentielle du sens, je ne restreins pas pour autant le fonctionnement de ce dernier à son rôle dans le pointage du réel.
**10.** « Le fait que l'emploi du nom propre ne puisse pas se faire dans une situation de vacuité informationnelle absolue ne veut pas dire ipso facto que cet apport nécessaire d'information soit de l'ordre sémantique ». Noailly, M. (1999 : 108).
**11.** Kleiber (1981 : 332) défini le nom propre modifié comme celui qui « se présente accompagné de déterminants qui lui font perdre le caractère *unique* ou *singulier* fréquemment assimilé à la marque spécifique qui l'oppose aux noms communs ».
**12.** Lafont et Madray (1976 [1989] : 102).
**13.** Le praxème est l'appréhension dynamique du lexème.
**14.** Lafont et Madray (1976 [1989] : 102).
**15.** Siblot, P. (1987 : 91).
**16.** Buyssens (1973 :27) : « /.../ ce qui caractérise le nom propre, c'est que son emploi est réglé par un fait social ».
**17.** A condition de que chaque lien soit établit par une convention d'ordre social entre les interlocuteurs dans un contexte donné, un même nom propre peut servir à désigner plusieurs référents, ce qui explique le phénomène de la polyréférentialité que j'aborde dans ce travail.

**18.** Siblot, P. (1987 : 93). Cette notion donne une nature essentiellement contextuelle au lien de désignation rigide défini par Kripke (1972/1982) et permet de rendre compte du phénomène qui m'occupe, à savoir la polyréférentialité du nom propre.

**19.** Cela ne veut pas dire que le praxème en fonctionnement nom propre ne produise pas de sens. Pour avoir une illustration de la production de sens du nom propre, voir Rangel Vicente (2003).

**20.** C'est pour cette raison que dorénavant la designation *nom propre* concernera exclusivement les noms propres purs.

**21.** La notion d'entité nommée concerne les noms propres mais aussi les dates et les mesures.

**22.** « On the level of entity extraction, Named Entities (NE) were defined as proper names and quantities of interest. Person, organization, and location names were marked as well as dates, times, percentages, and monetary amounts ». Chinchor (1997).

**23.** Je conteste, dans le cadre de ce travail, l'inclusion des dates et des mesures dans la catégorie des entités nommantes individualisantes parce que leur statut reste à définir dans le cadre praxématique. J'espère pouvoir m'y consacrer lors de futures recherches.

**24.** En raison de l'utilisation récurrente de la notion d'entité *nommante individualisante* dans le cadre de ce travail, j'utiliserai à partir de maintenant la forme raccourcie *entité nommante* pour y référer.

**25.** Etant donné que la notion d'entité nommante forte est synonyme de celle de nom propre (tel que la praxématique conçoit ce dernier), je vais les utiliser indistinctement dorénavant dans cet article.

**26.** Comme l'expose Maurel (2004), chaque mot commençant par majuscule ne constitue pas à lui tout seul un nom propre en raison de l'existence des entités nommantes faibles de type nom propre mixte et des noms propres composés de plusieurs mots commençant par majuscule (comme *Banco Real*) qui ne constituent une entité nommante que quand ils apparaissent ensemble.

**27.** Même si du point de vue syntaxique la structure *nom propre + glose* et le nom propre mixte sont à différencier, je ne pense pas qu'il en soit de même dans le cas du traitement de l'information parce qu'elles ont un fonctionnement identique, à savoir celui d'inscrire le référent individualisé par le nom propre dans une catégorie de référents de laquelle l'entité nommée est détachée. Pour plus d'information sur cette conception du fonctionnement du nom propre, voir Rangel Vicente (2004).

**28.** McDonald (1993) définit *l'évidence externe* comme le contexte dans lequel apparaît le nom propre et qui donne des pistes sur sa classification sémantique. Il l'oppose à *l'évidence interne*, qui est constituée par les séquences de mots comprises dans le nom propre, comme par exemple les majuscules et d'autres marques comme Inc., &, Ltd., etc.

**29.** *Centre de Tecnologies i Aplicacions del Llenguatge i la Parla* de l'Universitat Politècnica de Catalunya.

**30.** http://tln.li.univ-tours.fr

**31.** Module d'Identification et Classification des Entités.

**32.** Les aspects essentiels de ce système sont présentés dans Arévalo et al. (2002 et 2004).

**33.** Si CLIC a tenu à sous-diviser les super-classes principales, c'est avec le but d'intégrer la plus grande quantité possible d'information sémantique qui puisse être utile à d'autres applications (par exemple, pour la désambiguïsation ou la résolution des coréférences).

**34.** Projet européen qui essaie de créer un standard d'annotations morpho-syntaxiques de corpus pour toutes les langues européennes. Pour les noms, elles sont composées de 7 chiffres. La première pour la catégorie grammatical, la deuxième pour le type de substantif (commun ou propre), la troisième pour le genre, la quatrième pour le nombre, la cinquième et la sixième pour le type sémantique (ou classification typologique) des entités nommantes, et la septième pour le degré du nom.

**35.** J'envisage également d'analyser les gloses catégorisantes des entités nommantes faibles (du type nom propre descriptif) pour faire une étude comparative de cet ensemble avec celui des gloses des entités nommantes fortes.
**36.** J'ai choisi le genre discursif journalistique pour deux raisons. Premièrement, comme le signale Jean-Yves Antoine, la fréquence d'utilisation des gloses est susceptible de varier en fonction du genre discursif dans lequel elles sont utilisées, et les textes journalistiques font souvent appel à cette structure syntaxique par sa forme synthétique et par sa proximité au nom propre pour apporter des informations sur le référent de ce dernier. Deuxièmement, pour des raisons pratiques, parce que les outils de traitement de l'information s'appliquent très souvent à des corpus de ce genre.
**37.** Pour le catalan, il s'agit de *qui* pour les référents humains et de *on* pour les référents locatifs.
**38.** « Hélène, reine de Sparte et épouse de Ménélas, face à la beauté de laquelle succombe **Paris, qui** la kidnappe et en fait son épouse pour le malheur de sa race » (http://www.culturaclasica.com/nuntii2004/mayo/troya2.htm).
**39.** Goldstandard est le corpus de test utilisé par CLIC. Etiqueté manuellement, il est tiré du corpus de l'agence de presse EFE et il est constitué de 800 000 mots-texte.
**40.** Lancer ma recherche à partir des mots catégorisants se révèle une tâche lourde et pas forcément fructueuse par plusieurs raisons. Premièrement, il y en a trop (environ 13.000 formes).Deuxièmement, il se peut que le dictionnaire de CLIC ne recense pas toutes les formes catégorisantes susceptibles de catégoriser les Np ambigus. Pour finir, certaines de ces formes (comme par exemple *presidente* et *director*) sont polysémiques.
**41.** La partie morpho-syntaxique de l'analyse se fait dans les autres phases du système TALP-CLIC, dont l'approche est plus « en langue ». Il ne faut pas oublier pour autant que leur résultat sert de point de départ à ce module plus sémantique (pour ce qui est des noms propres) qui, à son tour, sera analysé par la partie de TACAT externe à MICE.

## RÉSUMÉS

Tout nom propre est potentiellement polyréférentiel par le fait d'être associé à son référent par des liens extralinguistiques et non par un sens lexical qui caractériserait le type de référents auxquels il est susceptible d'être appliqué. Cela a comme conséquence que le texte écrit offre comme seul moyen de désambiguïsation référentielle du nom propre le cotexte. Mon étude se situant à la charnière du traitement automatique du langage et de la linguistique praxématique, je m'apprête à élaborer la grammaire automatisée d'une structure que je propose d'appeler glose et qui s'avère, dans le cadre de la sémantique nominale, un outil privilégié de désambiguïsation référentielle du nom propre. Le corpus utilisé sera composé essentiellement de textes journalistiques en ligne en langue espagnole.

All proper nouns are potentially multireferential given that they are associated to their referent by extralinguistic links and not in a lexical sence, which would characterise the type of referent to which it is susceptible to be applied to. Consequently, in the writing text there is only the textual context as a means of removing all referential ambiguity from the proper noun. My research hinges on automatic language processing and praxematic linguistics. As such, I will attempt to develop computational grammar from a structure that I will call glose and which, within the framework of nominal semantics, turns out to be a privileged tool for clarifying the

proper noun's referential ambiguity. The corpus used for this study will essentially be composed of on line Spanish journalistic texts.

## INDEX

**Mots-clés :** nom propre (pur), entité nommante (individualisante), forme catégorisante, glose (catégorisante), polyréférentialité

## AUTEUR

### MONTSERRAT RANGEL VICENTE

Université de Barcelone (Espagne) - Université Montpellier III (2ème année de doctorat en Sciences du Langage)