DES GLOSES DE MOT AUX TYPES DE TEXTES : UN BILAN DIFFÉRENCIÉ

Augusta MELA et Mathieu ROCHE
LIRMM, Université Montpellier III et Université Montpellier II

SOMMAIRE

Résumé : *Qu'il s'agisse d'évoquer une langue étrangère ou spécialisée (1), de procéder à une explication didactique (2), ou de s'assurer que l'interlocuteur attribue la signification adéquate au mot employé, le langage courant fournit de nombreux exemples de commentaires en situation parenthétique qui traduisent, expliquent le sens des mots en discours :*

(1) Si l'on admet que l'état d'un électron n'est pas entièrement décrit par sa position et sa vitesse de translation dans l'espace, mais qu'il est animé en outre d'un pivotement sur lui-même, ou « spin » *mouvement essentiellement quantifié : son moment cinétique est d'un demi quantum et crée un moment magnétique égal à un magnéton de Bohr. (Hist. gen. sciences, 1964, t.3, vol.2)*

(2) Ce sont la dépigmentation, c'est-à-dire l'absence quasi totale des éléments colorés dermiques qui s'opposent normalement à l'action nocive des rayons ultra-violets d'origine solaire et l'anophtalmie, ou réduction de l'appareil oculaire allant le plus souvent jusqu'à sa disparition complète*. (Geze, La spéléologie scientifique, 1965)*

On appelle gloses *ces commentaires sur le mot, qui nous mènent au sens par des « chemins buissonniers » (Steuckardt et Niklas-Salminen 2005).*

Les gloses se manifestent dans les textes par des marques typographiques comme les guillemets et des éléments lexicaux comme appelé, c'est-à-dire, ou*. Ces marqueurs sont polysémiques mais en tirant parti des particularités syntaxico-sémantiques des constructions segment glosé-marqueur-glose, leur repérage automatique est réalisable.*

De plus, ces marqueurs explicitent la nature du lien sémantique entre le segment glosé et sa glose : équivalence avec c'est-à-dire*,* ou *; spécification du sens avec* au sens *; nomination avec* dit*,* alias*,* baptisé *; hyponymie avec* en particulier, comme, tel, par exemple *; hyperonymie avec* et/ou autre(s)*.*

De premières études ont été réalisées sur la base Frantext (Mela 2004, 2005). Nous élargissons ici nos investigations à d'autres types de textes, d'autres marqueurs et d'autres liens sémantiques. Des comparaisons avec la langue anglaise sont menées.

Nous répondons aux questions suivantes :

1) dans quelle mesure peut-on repérer automatiquement les gloses ?

2) dans quelle mesure tel marqueur de glose déclare tel lien sémantique entre le segment glosé et sa glose ?

3) la fréquence de gloses dépend du type de textes (poésie versus *ouvrage didactique) et pour un même lien sémantique, le choix du marqueur dépend du niveau de langue et du type de texte (ouvrage didactique* versus *roman) : dans quelles mesures ?*

en dressant un bilan, par types de textes, par marqueurs et par liens, de ces mesures. Une démonstration du logiciel de repérage sera proposée.

1. Introduction

Le travail que nous présentons a pour origine un questionnement de collègues linguistes : dans le cadre d'une recherche sur « le mot et sa glose », elles se demandaient si le repérage automatique des gloses de mot était possible. Les gloses de mot sont ces commentaires, souvent introduits par des termes de liaison tels que *c'est-à-dire*, *ou*, *autrement dit, ce qu'on appelle*, etc. qui définissent des mots en discours, comme la séquence en italique de l'énoncé suivant :

(1)« 10 % de ces embauches vont porter sur un métier qui monte : le « testing », *c'est-à-dire la maîtrise des méthodologies rigoureuses de test des logiciels* », indique Dominique Duflo, le DRH. (*L'Expansion*, Avril 2006, p. 136)

Un repérage automatique permettrait de ramener plus efficacement les gloses des corpus et de quantifier les phénomènes observés. De plus, dans la perspective d'un travail lexicographique, on espère, étant donné un mot spécifique, pouvoir repérer ses gloses et accéder ainsi à son sens. En effet, si on considère les énoncés suivants :

(2)Quant aux espèces endobiontes des substrats meubles, *appelées* fouisseuses, elles sont légion, ce qui est normal car il est évidemment plus facile de fouir un sable ou une vase que de forer une roche, même tendre. (Peres J-M, *La vie dans l'océan*,1966, p. 72)

(3)Les fouisseuses *sont* des espèces endobiontes des substrats meubles. (exemple construit)

on constate que l'apposition de la glose *appelées fouisseuses* contient, au même titre que la copule *sont* dans l'énoncé définitoire (3), une définition dont *fouisseuses* est le défini (definiendum) et *espèces endobiontes des substrats meubles*, le définissant (definiens).

Ainsi, les gloses de mot sont utiles, au même titre que les définitions, dans l'aide à l'acquisition terminologique, et ce à double titre :

- elles pointent le vocabulaire spécifique et les nouvelles unités qui demandent à être expliquées ;
- elles signent/signalent un texte définitoire comme en (4), voire elles définissent elles-mêmes le mot (1,2) :

(4)Chaînon manquant entre l'apparition de la photographie et le cinéma des Lumière en 1895 , le flipbook (de flip over, feuilleter), aussi *appelé* folioscope, est ce livre animé, dont l'assemblage d'images dans le défilement donne l'impression du mouvement. (*Libération*, Stéphanie Binet, 19 mai 2006)

Nous pensons que leur repérage automatique est plus aisé que celui d'autres énoncés définitoires parce que les pivots du repérage, à savoir les marqueurs *appelé, c'est-à-dire,* etc. sont plus filtrants que la copule *être*, par exemple ; et parce que leur configuration, en apposition au mot glosé, est moins sujette à la variation que ne peuvent l'être les configurations où la définition est en prédication principale.

Nous proposons ici de faire le point sur ces questions de traitement automatique, en mettant les attentes en regard des outillages nécessaires.

Après avoir situé les gloses dans le champ des énoncés définitoires (§ 2), nous illustrons notre démarche en prenant pour fil conducteur l'exemple des indications de dénomination introduites par le marqueur *appelé.* Partant de la description linguistique de cette configuration (§ 3.1), nous enchaînons sur la modélisation et l'implémentation de son repérage (§ 3.2), nous évaluons et analysons les résultats obtenus (§ 3.3).

Nous serons alors à même de répondre à nos collègues linguistes : le repérage des gloses est possible ; la recherche des gloses d'un mot spécifique, comme l'extraction des définitions, est également possible, mais nécessite une analyse syntaxique partielle robuste préalable pour être traitée proprement.

2. Les gloses de mot dans le champ des énoncés définitoires

Menée à l'Université de Provence au cours des quatre dernières années, les études de la glose ont donné lieu à la publication de deux ouvrages collectifs (Steuckardt et Niklas-Salminen 2003) et (Steuckardt et Niklas-Salminen 2005). Selon Agnès Steuckardt, directrice du projet :

> « Pour éclairer le sens d'un mot, l'analyse de corpus privilégie traditionnellement le repérage des associations récurrentes, sans se soucier de la conscience métalinguistique qu'en a ou non le locuteur. À la lumière de l'expérience concrète du travail sur les concordances de mot, il nous a semblé possible de trouver dans les gloses données par les locuteurs aux mots qu'ils emploient un autre accès au sens lexical. »

Une synthèse en ligne (Steuckardt 2006) présente les différentes configurations syntaxiques des gloses de mot et une typologie des marqueurs de glose.

2.1. Forme des gloses

Les gloses de mot sont des configurations définitoires non formelles, parenthétiques. Sur le continuum définitionnel proposé par J. Rebeyrolle (1990, p. 89) qui va des définitions directes aux définitions indirectes (Riegel 1987), les gloses à marqueurs lexicaux (*appelé*, *c'est-à-dire*, *ou*, etc.) se situent au centre, entre les définitions directes (5,6) et les gloses à marqueurs typographiques (7) tels que virgule, parenthèses, crochets, deux points, tiret, etc.

> (5)*Nous appelons* donc définition (D) la mise en relation d'un terme à définir (A) et d'une séquence qui en est la paraphrase, constituée d'un second terme (B) auquel s'adjoint un ensemble de propriétés distinctives (X). (Rebeyrolle, 2000, p. 89)
>
> (6)Avec ironie et non sans excès, *on appelle* « Khmers verts » les groupes d'écologistes qui ont contribué à substituer au corps des ingénieurs et urbanistes qui bitumaient et bétonnaient, un corps d'ingénieurs et d'urbanistes qui découpent avec autant de zèle les chaussées en zones d'exclusion (voitures, piétons, autobus, vélos), multiplient terre-pleins et espaces de verdure... Tous dispositifs qui, paradoxalement, affectent les qualités de partage et de rencontre des espaces publics. (*Le Monde*, Frédéric Edelmann, 27 avril 2006)
>
> (7)Depuis qu'il a été forcé par la Cour suprême des Etats-Unis de reconnaître la victoire de George W. Bush à l'élection présidentielle de 2000, Al Gore se consacre à la présentation d'une conférence illustrée (il l'appelle son « slide show *», sa soirée diapo*) démontrant la réalité du réchauffement global de la planète et l'urgence qu'il y a à le corriger. (*Le Monde*, Thomas Sotinel, 23 mai 2006)

Alors que la définition est la prédication principale des définitions directes (3,5,6), dans les gloses (1,2), la définition est en prédication seconde. Ce sont des définitions comme accidentelles, parenthétiques, insérées « en passant ». Le concept de glose rejoint le concept de *définition non formelle* dite *de substitution*, défini par J. Flowerdew (1992, p. 101) :

> « In contrast to formal and semi-formal definitions, *substitutions* are used most commonly where the definition is not the main focus of the discourse. Instead, they occur "embedded" in the overall discourse structure, inserted, so to speak, "en passant"; their function is not to provide important new information as such, but is a metalinguistic one to explain terms that arise as the lecture progresses. In this way they act as a sort of lubricating device which facilitates comprehension on the part of hearers, as they negotiate their way through the discourse. In the following example, for instance, definitions of the terms "anterior" and "posterior" are embedded in a more general description of a biological specimen:
>
> ... this is the anterior end/ anterior meaning front/ and posterior meaning behind ... »

2.2. Fonctions définitoires des gloses de mot

Si on analyse les énoncés suivants :

> (2)Quant aux espèces endobiontes des substrats meubles, *appelées* fouisseuses, elles sont légion, ce qui est normal car…
>
> (8)L'Arbadetorne, *qui signifie* l'herbe à détourne, a le pouvoir de détourner de son chemin celui qui marche dessus. (*Sud-Ouest*, Karine Robin, 26 mai 2006)
>
> (8)Pour être précis, Manuel Desdín, Cubain de 21 ans, est "atlichnik", *autrement dit* "champion" à l'Institut de physique des basses températures de Kharkov. (*Le Monde*, Jésus Díaz, 26 mai 2006)

on constate que les gloses, comme les définitions directes, peuvent être :

- explicatives : c'est le cas dans les indications de signifié (8,1) et de nouvelle nomination (9), qui amènent le récepteur de l'inconnu vers le connu ;
- didactique : c'est le cas dans les indications de dénomination (2,4) qui amènent le récepteur du connu vers l'inconnu.

Comme nous le fait observer A. Steuckardt (2006) :

> « Visée didactique et visée explicative sont des cheminements inverses d'un même parcours. Du point de vue de l'analyste, l'un comme l'autre ouvrent un accès au sens lexical. »

3. Le traitement informatique

Nous nous intéresserons ici au cas des gloses dites, par raccourci, à « marqueurs » lexicaux. Ces termes de liaison *c'est-à-dire*, *à savoir*, *ou*, a*ppelé*, etc. sont polysémiques mais alliés aux propriétés syntaxico-sémantiques des configurations où ils interviennent en tant qu'introducteurs de glose de mot, ils pourront « marquer » les gloses. Ainsi, lorsqu'il est en configuration de glose comme dans l'énoncé qui suit :

> (9) De remarquables travaux, qui n'avaient pas trouvé d' écho à leur parution, au cours du xixe siècle, ont pris toute leur signification à l' aurore du XXe et ont formé les bases de toute une discipline nouvelle, qui a pris une ampleur et une importance considérables, la génétique, *ou science de l'hérédité*. (Anonyme, *Hist. gen. science*s, t.3 vol.1, 1961, p. 550, Frantext)

le terme *ou* joint un terme en usage (*génétique* ici) à un terme en mention (*science de l'hérédité)* ; le terme en mention s'applique métalinguistiquement au premier ; il est co-possible et son statut métalinguistique est marqué par une absence de détermination (Tamba 1987, pp. 27-28). Projeté sur le corpus de *l'Histoire générale des scien*ces de Frantext[1] (Mela, 2004), le patron « *ou* précédé d'une ponctuation et suivi d'un substantif non déterminé » permet de ramener des gloses en *ou* telles que (10) avec une précision[2] de 97%.

(Mela 2005) traite des gloses en *dit* dans l'environnement Frantext-Stella. Nous détaillons ici le cas des indications de dénomination introduites par *appelé*, sur des corpus diversifiés.

Notons que nous procédons actuellement marqueur par marqueur mais que ces marqueurs pourront agir de concert pour détecter tous types de gloses en une seule passe sur le corpus.

Outre le repérage d'un (ou plusieurs) types de gloses, d'autres fonctionnalités seraient utiles :

- la repérage des gloses d'un mot donné ;
- l'extraction des arguments de la définition, le défini (definiendum) et le définissant (definiens).

Ces trois fonctionnalités ne comportent pas les mêmes difficultés pour des raisons que nous analysons en 5.3. Actuellement seule la première fonctionnalité est implémentée. Nous en évaluons les résultats ; nous spécifions les deux autres fonctionnalités.

Notre article suit les étapes suivantes :

- en section 2, nous définissons le concept de glose ;
- en section 3, nous partons de la description linguistique de la configuration recherchée, pour aboutir au patron, ou motif de recherche ;
- en section 4, le patron est implanté et projeté sur des corpus annotés morpho-syntaxiquement. Nous analysons les résultats.

3.1. Description linguistique de *appelé* dans la glose de dénomination

Le verbe *appeler* appartient à la table 11 de Gross (1975). Les verbes de cette table ont un complément direct substantival et un complément indirect en *à* (*appeler à voter/aux urnes/à ce que Phrase*). Dans la construction indirecte, *appeler* ne dénomme pas. La dénomination peut être en prédication première dans la configuration N_0[3] *appeler* N_1 N_2, ou en prédication seconde dans la configuration N_1, *appelé* N_2. Le participe passé *appelé* peut apparaître dans une autre configuration parenthétique *X, (ce) qu'on a appelé Y*. Mais le plus souvent, il s'agit de simples modalisations autonymiques[4]. Nous avons écarté cette configuration pour l'instant.

Dans la configuration N_1 *appelé* N_2, *appelé* est en position d'épithète détachée (2) ou pas selon qu'il est séparé ou non du mot glosé par une virgule, dont il peut être séparé par un adverbe

[1] La base Frantext est accessible par abonnement à l'adresse :<http://www.frantext.fr/>.

[2] Le terme précision est défini en section 4.

[3] N_0 : sujet formel, N_1 et N_2 sont les compléments du verbe, leur ordre correspondant à leur propriété de présence : obligatoire à facultative (cf. [Gross, 1975, p.13]).

[4] L'*autonymie* peut se définir comme la « propriété linguistique en vertu de laquelle tout mot ou tout élément linguistique peut être employé pour se désigner lui-même ». Par exemple : dans « rose » a 4 lettres, *rose* est autonyme. On parle du signe *rose* et non de ce qu'il dénote, une fleur. La *modalisation autonymique* est un cas particulier de l'autonymie. J. Authier-Revuz [Authier-Revuz, 1995] la définit comme « une opacification, résultant de ou consistant en – selon que l'on parle du résultat ou du processus énonciatif – une référence au monde accomplie en interposant sur le 'trajet' de la nomination la considération de l'objet signe par lequel on réfère. » Par exemple : dans *Le risque existe également de ce qu'on a appelé la «malédiction pétrolière»* (Libération**,** Rueff Judith**,** 17 mai 2006), les guillemets font que l'on s'arrête sur le mot et que le terme entre guillemets n'est plus complètement « transparent », il y a « opacification ».

Dans la base Frantext entière, sur 11occurrences de « ,(*ce) qu'on a appelé* », deux sont des gloses de mot, une est une reformulation de proposition et les autres, des modalisations autonymiques.

comme en (4) : *le flipbook, aussi appelé folioscope*. Il ne se compose pas avec les auxiliaires *être* et *avoir* (*j'ai appelé, est appelé*).
Le mot glosé N_1 est recteur : il impose les contraintes d'accord en genre et nombre. Dans des cas plus rares et à condition que le mot glosé N_1 soit sujet, *appelé* peut lui être antéposé :

> (10)Appelée « bancor », l'unité de compte proposée pour comptabiliser les créances et les dettes était totalement fiduciaire malgré le préfixe (sic) « or ». (1960, *L'univers économique et social*, François Perroux éd., Frantext) (cité dans Steuckardt 2006)

Dans tous les cas, N_2 est attribut du terme recteur N_1.
Enfin, il attend un objet direct (c'est-à-dire sans préposition, ni déterminant contracté).

3.2. De la description linguistique à l'implémentation

Notre repérage de la glose de dénomination en *appelé* s'appuie sur un étiquetage morpho-syntaxique des mots du corpus. Puisque nous ne disposons pas de structuration syntagmatique du texte, le motif de filtrage doit donc coller à la linéarité du texte.
Le schéma abstrait *X(,)appelé Y* n'est pas opérationnel pour plusieurs raisons :

- on a vu qu'il ne couvrait pas les cas d'antéposition de *appelé* (11) ;
- par ailleurs, il suppose que X et Y sont des groupes substantivaux ;
- des éléments peuvent s'insérer entre le pivot verbal *appelé* et X et Y : adverbe, incise avant (4) ou après (13) *appelé,* coordination (14) :

> (4)Chaînon manquant entre l'apparition de la photographie et le cinéma des Lumière en 1895 , le flipbook (de flip over, feuilleter), aussi *appelé* folioscope, est ce livre animé, dont l'assemblage d' images dans le défilement donne l'impression du mouvement. (*Libération*, Stéphanie Binet, 19 mai 2006)
>
> (13) L' idée est la suivante : à l'espace des hypothèses est associé un autre espace *appelé*, pour des raisons de similarité avec les mécanismes de l'évolution naturelle, espace "génotypique". (Corpus Scientifique)
>
> (11) Avec ce "radio-conducteur", perfectionné en 1890 et *appelé* «cohéreur» par Lodge, la radioélectricité était née. (<http://www.elec.unice.fr/pages/phototheque/photos.html>)

La présence de la virgule devant *appelé* ne nous a pas parue significative. Pour toutes ces raisons, nous nous en sommes tenus au simple patron :

```
(P)  appelé(e)(s)/Participe_passé suivi d'un mot autre que (à/au/aux)
```

Nos corpus sont étiquetés avec l'étiqueteur WinBrill[1] (Brill 1994). Aux accords en genre et nombre près, WinBrill distingue 5 sortes de participes passés :

EPAR : (sg|pl)	Verbe « être », non conjugué, participe passé
APAR : (sg|pl)	Verbe « avoir », non conjugué, participe passé
VPAR : (sg|pl)	autre Verbe, non conjugué, participe passé après « avoir »
ADJ1PAR : (sg|pl)	Participe passé après « être », adjectival ou verbal
ADJ2PAR : (sg|pl)	Participe passé adjectival, singulier (non après auxiliaire)

Pour *appelé*, trois étiquettes sont donc disponibles : VPAR:(sg|pl), ADJ1PAR:(sg|pl) et ADJ2PAR:(sg|pl). Dans la configuration qui nous intéresse, *appelé* n'est pas composé. Il sera donc, sauf erreur d'étiquetage, étiqueté ADJ2PAR.
Traduit en langage des expressions régulières Perl, notre patron (P) devient :

```
(P)_version Perl    appelée?s? \/ADJ2PAR:(sg|pl) (?!(à|au|aux)
```

4. Expérimentations

4.1. Corpus utilisés

Nous avons utilisé trois corpus de domaines différents :

- *Corpus Scientifique :* le livre « Apprentissage Artificiel[2]» d'Antoine Cornuéjols et Laurent Miclet (éditions Eyrolles) sur lequel nous nous appuyons dans nos expérimentations, est un corpus utilisé et prétraité dans le cadre de DEFT'06[3], défi francophone de fouille de textes ;

[1] On peut en savoir plus sur WinBrill et le télécharger gratuitement à partir du site de l'ATILF (http://www.atilf.fr).
[2] http://www.editions-eyrolles.com/Livre/9782212110203.
[3] DEfi Fouille de Textes : <http://www.lri.fr/ia/fdt/DEFT06/>.

- *Corpus Journalistique :* 71 articles de presse du 20 mai 2006 relevés sur Europresse[1] ;
- *Corpus Littéraire :* un sous-corpus de Frantext réduit à l'année 1950, tous genres sauf théâtre et poésie (14 œuvres) ;

à partir desquels, trois traitements sont effectués :

- Un premier filtre morphologique sélectionne les phrases contenant des occurrences de « appelé (e) (s) » ;
- L'étiqueteur de Brill est appliqué sur ces phrases préalablement mises au format ;
- Le patron morpho-syntaxique (P) est appliqué et départage deux listes de phrases. Les phrases reconnues par le filtre, sont appelées Positifs : elles sont censées contenir des gloses de dénomination. Les phrases non retenues, les Négatifs, sont censées ne pas contenir de gloses de dénomination ;
- Une évaluation peut alors être effectuée sur chacune des deux listes :
    - Elle consiste à vérifier si les Positifs sont réellement des gloses de dénomination ou pas. Lorsque ce sont des gloses de dénomination, ce sont des Vrais Positifs, notés *VP*, et nous les classons en 3 sous-groupes : définitions, synonymies, hyperonymies. Lorsque ce ne sont pas des dénominations, ce sont des Faux Positifs, notés *FP*. Nous les classons également en sous-groupes, suivant qu'il s'agit de dénominations propres par NPs et autres désignateurs rigides ou que les sources de mauvais aiguillage soient autres : étiquetage erroné, présence d'un complément introduit par une préposition autre que *à*, etc.
    - Nous examinons également les Négatifs, pour vérifier que tous le sont vraiment et sinon analyser les raisons du « silence ». Les Faux Négatifs, notés *FN*, sont alors les gloses de dénomination passées sous silence. Les autres sont les Vrais Négatifs, notés *VN*.

4.2. Mesures d'évaluation

La mesure d'évaluation appelée *précision* doit être calculée. Une telle mesure fréquemment appliquée dans le domaine de l'apprentissage artificiel (Cornuéjols et Miclet 2002) est donnée par la formule ci-dessous :

$$\text{précision} = VP/(VP+FP)$$

Une précision de 100% signifierait que toutes les phrases ramenées par notre patron sont des gloses de dénomination. Si elle ne vaut pas 100%, c'est qu'il y a du « bruit ».

Pour vérifier que notre patron ramène toutes les gloses de dénomination du corpus, nous utilisons une autre mesure d'évaluation appelée *rappel* (Cornuéjols et Miclet 2002). Pour ce faire, une évaluation « manuelle » consiste à examiner cette fois les Négatifs et à vérifier que tous le sont vraiment et sinon analyser les raisons du « silence ». Les Faux Négatifs seraient les gloses de dénomination passées sous silence.

Le rappel est donné par la formule ci-dessous :

$$\text{rappel} = VP/(VP+FN)$$

Un rappel de 100% signifierait que toutes les gloses de dénomination ont été extraites du corpus.

Notons que la somme des positifs et négatifs (*nb =VP+FP+VN+FN*) recouvre l'ensemble du corpus à filtrer.

En règle générale, il est important de déterminer un compromis entre le rappel et la précision. Pour cela, nous pouvons utiliser une mesure prenant en compte ces deux critères d'évaluation en calculant le *Fscore* (Rijsbergen 1979) :

$$\text{Fscore} = (\beta^2+1)*\text{précision}*\text{rappel}/(\beta^2*\text{précision}+\text{rappel})$$

Le paramètre β de la formule ci-dessus permet de régler les influences respectives de la précision et du Rappel. Il est très souvent fixé à 1 pour accorder le même poids à ces deux mesures d'évaluation.

4.3. Évaluation de (P) sur les trois corpus

4.3.1. Préliminaires : où s'arrêtent les gloses de dénomination ?

La délimitation précise des gloses recherchées est guidée par la tâche à réaliser. On cherche des gloses pour trouver du sens. Si c'est du sens lexical, nous nous intéresserons davantage à la

[1] Europresse, <http://www.bpe.europresse.com/>, correspond à 19 titres de la presse nationale et régionale, dont *Le Monde*, *Les Echos*, *Libération*, *Sud-Ouest*, etc.

fonction *catégorisante* qu'à la fonction *individualisante*[1] des noms, donc nous privilégierons les dénominations communes alors que les désignations rigides, par noms propres telles que *un psychologue américain appelé Gesell*, ou autres noms de lieu, d'enseigne (15), de produit etc. seront de moindre intérêt. En revanche, si l'on cherchait à constituer un dictionnaire de noms propres, elles deviendraient intéressantes.

> (12) Le comité de salut public arrête que la maison nationale ci-devant *appelée les menus-plaisirs*, située rue Bergère, servira désormais pour l' institut national de musique établi par les décrets de la convention nationale. (Anonyme, Enseign. mus. 1. enseign. off., 1950 p. 8, Frantext)

De plus, le sens trouvé doit être le plus riche possible. La définition, la plus complète possible. Telle qu'elle est définie en (5), la définition correspond au schéma A = B+C, où A est le terme défini, B une classe (l'hyperonyme) et C la caractéristique, la « détermination distinctive de B[2] ». Ce schéma est illustré par l'énoncé suivant :

> (13) On appelle carré un losange dont les côtés sont égaux et les angles sont droits.
>
> où A = *carré*, B = *losange* et C = *dont les côtés sont égaux et les angles sont droits*.

Face à l'énoncé suivant :

> (14) Il utilise pour ce faire un mécanisme *appelé* chunking qui produit des "chunks" (littéralement "gros morceaux") ou macro-opérateurs. (Corpus Scientifique)

si on prend le schéma de glose *X marqueur Y* au pied de la lettre, X = *un mécanisme* et Y = *chunking*, le schéma correspond[3] à une définition pauvre : *le chunking est un mécanisme.* On perd l'information donnée par la détermination distinctive C.

Si au contraire, on considère *X marqueur Y* comme un schéma abstrait, qui ne colle pas nécessairement à la linéarité du texte, alors, appliqué à l'énoncé (17), la correspondance entre schéma de glose et schéma de définition est totale : *B+C*$\leftrightarrow$*X = un mécanisme qui produit des chunks* … et A$\leftrightarrow$Y = *chunking*. On rend compte de la définition complète contenue dans (17).

Autrement dit, on doit admettre que dans sa réalisation textuelle linéaire, le groupe substantival X du schéma *X marqueur Y* peut être discontinu, sa tête restant à gauche du marqueur et recteur de celui-ci, et ses compléments (ppassé, pprésent, relative) étant placés à droite de la glose.

Pour ces raisons, nous avons considéré que les cas tels que (17) étaient des définitions prototypiques (étiquetées *D* dans notre évaluation[4]).

Notons qu'en abstrayant le schéma de glose comme nous venons de l'argumenter, nous continuons au niveau de l'analyse sémantique, la démarche que nous avons dû entamer lors de l'analyse syntaxique (§ 3.2). Mais ce faisant, nous faisons un écart : telle qu'elle est définie actuellement dans le projet aixois, la glose n'englobe pas les cas d'hyper/hyponymie. Ainsi, pour A. Steuckardt[5], la séquence *le textile connu sous le nom de ramie* de l'énoncé suivant :

> (15) Le textile *connu sous le nom de ramie* provient de l'écorce des tiges de l'ortie de Chine, *appelée encore ramie* (Blanquet, Technol. mét. habill., 1948, p. 51, in TLFI, s.v. *RAMIE*).

n'est pas une glose parce que *ramie* n'est pas une dénomination de *textile*, les deux termes *ramie* et *textile* étant simplement dans un rapport hyperonyme-hyponyme. Pour nous, cette séquence est intéressante pour l'acquisition lexicographique, parce qu'elle contient la définition *la ramie est un textile qui provient de l'écorce des tiges de l'ortie de Chine (appelée encore ramie)*.

4.3.2. Évaluation du patron (P)

Le patron (P) que nous évaluons résulte de la phase « manuelle » d'analyse linguistique. Étant établi uniquement sur des critères linguistiques généraux, il est lâche et son rappel est maximal (100%, cf. tableau ci-dessous). Sauf dans le cas du corpus Littéraire où il chute de 25% à cause d'un seul cas de silence, dû à un mauvais étiquetage, mais le nombre de gloses de dénomination est trop faible sur ce corpus pour que le résultat soit significatif.

Des allers et retours entre analyse linguistique et résultats sur corpus permettront d'améliorer la précision du repérage et les scores obtenus, tout en gardant un rappel acceptable.

---

[1] Ces termes sont de P. Siblot et cités dans (Kleiber 1996, p. 584).

[2] Le terme est de Kleiber et Tamba (Kleiber et Tamba 1990, p. 24).

[3] Nous utilisons les symboles « $\leftrightarrow$ » pour la correspondance entre le schéma de la définition et le schéma de glose ; « = » pour la correspondance entre le schéma de glose et sa réalisation textuelle.

[4] Les corpus et le détail des évaluations sont disponibles sur :
<http://www.lirmm.fr/~mroche/Recherche/glose.html>

[5] Communication personnelle (avril 2006).

Le patron (P) est un point de départ et constitue un dispositif dont on peut faire varier les paramètres (précision et rappel) en fonction de la tâche poursuivie et de la taille du corpus examiné : une analyse linguistique manuelle sur un petit corpus privilégiera un rappel maximal ; alors que l'extraction automatique de définitions à partir d'un gros corpus aura intérêt à resserrer le patron, quitte à laisser sous silence une partie des gloses hors calibre. Nous détaillons en § 5 comment le dispositif peut être adapté.
Le tableau ci-dessous présente le résultat global obtenu pour (P) à partir des trois corpus. Ce tableau montre que les résultats les plus significatifs en terme de Fscore sont obtenus sur le corpus Scientifique. En effet, avec un tel corpus très spécialisé, il est nécessaire de définir des termes. Sur l'ensemble des corpus, le rappel est élevé montrant que le patron (P) ramène toutes les gloses pertinentes (moins une). La faible précision pour les corpus Journalistique et Littéraire est due à une présence importante de dénominations propres, que nous n'avons pas cherché à exclure dans un premier temps.

Corpus	***nb***	***VP***	***FP***	***VN***	***FN***	***Précision***	***Rappel***	***Fscore***
Scientifique	70	25	8	37	0	75,8%	100%	**86,2%**
Journalistique	51	6	10	35	0	37,5%	100%	**54,5%**
Littéraire	58	3	11	44	1	21,4%	75%	**33,5%**

5. Analyse des résultats

Nous analysons les résultats sous l'angle des améliorations du repérage, des types de gloses obtenus, de l'extraction du definiendum Y et du definiens X, et de la recherche des gloses d'un mot donné.

5.1. Améliorations possibles du repérage

Plusieurs pistes amélioreraient sensiblement les scores précédents. Certaines sont à notre portée, d'autres demandent une investigation plus poussée.
Le patron (P) actuel n'exclut que la préposition *à* et les déterminants contractés avec *à*. Nous ne souhaitions pas exclure *a priori* les énoncés où des incises débutant par une préposition s'insèrent entre *appelé* et son complément direct, comme dans l'énoncé suivant :

> (16) On y parvenait par un escalier en bois blanc *appelé*, dans l'argot du bâtiment, échelle de meunier. (Balzac. H de, *Le cousin Pons*, 1847, p. 751, Frantext)

Cette stratégie s'avère coûteuse sur nos corpus puisque nous n'avons ramené aucun de ces cas alors que nous comptons 3 cas de bruit en présence de préposition tels que :

> (17) Les clubs sont fondés à souhaiter que des assurances soient prises pour leurs joueurs *appelés en* sélection et à réclamer une indemnisation. (*L'Équipe*, 19 mai 2006, Corpus Journalistique)

Il semble donc qu'on ait intérêt, quitte à laisser sous silence quelques cas, à étendre l'interdiction à toutes les prépositions.
Les deux propositions suivantes amélioreraient également la précision mais toujours au détriment du rappel, car elles nécessitent de prévoir une position pour le definiendum Y (cf. § 5.3) dans le patron qui deviendrait par exemple :

```
appelé(e)(s) ?ADV ( ?! Prep) Y_Substantival
```

Les modalisations autonymiques (B-MA dans notre évaluation) telles que (21) pourraient être évitées en excluant les adjectifs de la position du definiendum Y :

> (18) J'entre dans le monde *appelé réel* comme on entre dans de la brume. (Green. J, *Journal* T.5, 1950, p. 267, Corpus Littéraire)

Les dénominations rigides, notées *DP* dans notre évaluation, constituent la cause principale (plus de 2/3 des FP) du bruit actuel. On pourrait, à condition qu'ils soient reconnus en amont, exclure les noms propres de la position Y. Une étude contrastive des dénominations communes *versus* rigides (présence ou pas du déterminant, etc.) reste à faire pour voir si un filtrage des dénominations rigides comme *les menus-plaisirs* de (15) est faisable.

5.2. Quels types de glose obtient-on ?

Parmi les Vrais Positifs, outre les définitions complètes (17), on obtient de simples synonymies (étiquetées *S* dans notre évaluation) comme celle qui lie *praticien* et *recors*[1], dans l'énoncé suivant :

> (19) Le praticien, vulgairement *appelé* recors est l'homme de justice par hasard, il est là pour assister l'exécution des jugements, c'est, pour les affaires civiles, un bourreau d'occasion. (Balzac H. de, *Cous. Pons*, 1847, p. 173, Corpus Littérraire)

Les cas d'hyperonymie/hyponymie (étiquetés *H* dans notre évaluation) sont ceux où un hyperonyme de Y précède *appelé* et où la détermination distinctive est nulle :

> (20) Nous ne respirons peut-être pas à votre hauteur, mais nous avons un viscère *appelé* cœur. (Bazin H., *La mort du petit cheval*, 1950, p. 106, XIII, Corpus Littéraire)

Souvent la glose se greffe sur une autre glose ou sur un énoncé définitoire. Dans ce cas, deux niveaux de définition co-opèrent. Ainsi, dans l'énoncé (22), par un jeu de circulation sémantique, le definiens de la prédication principale rejoint le definiens de la prédication seconde, les deux definiens s'amplifiant mutuellement.

5.3. L'extraction du definiens X et du definiendum Y est-elle possible ?

L'extraction du definiens X et du definiendum Y nécessite de prévoir les positions de X et à Y dans le patron. Si ces positions sont contiguës au pivot verbal, le patron sera trop strict. Si on autorise des fenêtres entre *appelé*, X et Y – douze mots pour rendre compte de (13), le patron risque d'être trop lâche.

Par ailleurs, du fait que X est souvent discontinu (cas de fausses hyper/hyponymies tels que (17)), la question du rattachement de la détermination distinctive se pose.

Pour être traitées proprement, ces deux tâches nécessitent une analyse structurelle des corpus, au moins partielle. Plutôt que de raisonner en termes de fenêtre de mots, on pourrait alors raisonner en termes de présence optionnelle d'un groupe syntagmatique. La question du rattachement de la détermination distinctive de X reviendrait à reconnaître juste après *appelé* les configurations possibles de cette détermination distinctive : proposition relative, proposition construite autour d'un participe passé ou d'un gérondif.

Autre problème, lié cette fois à l'ambiguïté structurelle de la langue : comment reconnaître automatiquement l'empan du terme glosé lorsque des GNs enchassés précèdent *appelé* (24,25) ? On pourrait s'appuyer sur des marques d'accord grammatical, ou typographiques comme les guillemets en (24), mais ces marques ne sont pas toujours disponibles (25) :

> (21) Le jour du drame ces deux employés étaient occupés à des travaux de maintenance sur les pompes du réseau de captage des « lixiviats », autrement *appelés* « jus de décharge ». Ils avaient été retrouvés asphyxiés par les gaz délétères, sur une plate-forme de sécurité, installée dans un puits.(*Sud-Ouest*, Axelle Maquin-Roy,19 mai 2006)
>
> (22) D'autant plus que les intermédiaires spécialisés dans la valorisation de sites (*appelés* "brownfields developers" aux États-Unis) sont plutôt rares en France, avec quelques exceptions, comme Terra Verde Capital... (*La Tribune*, 19 mai 2006)
>
> La question de la recherche des gloses d'un mot spécifique se ramène à la question 5.3. Ainsi, chercher la définition de *espace génotypique* de l'exemple (13) nécessite de pouvoir exprimer dans le patron de recherche que Y = *espace génotypique* ; pour cela il faut que Y (et sa position) soit prévue dans le patron.

5.3.1. Énoncés définitoires et typologie des textes

On sait (Rebeyrolle 2000, Chap. 8) que l'on peut caractériser les textes à partir de leurs propriétés définitoires : la fréquence mais aussi la forme (prédication principale *versus* seconde) des définitions sont des indicateurs de types de corpus. Les textes didactiques sont propices aux définitions en prédication principale. Les définitions en prédication seconde y sont également nombreuses. À l'opposé, les textes poétiques ne contiennent ni définitions ni gloses. Dans Europresse, les définitions formelles en *appeler (On appelle X, X s'appelle*) sont rares. La plupart des définitions en *appeler* sont des gloses. La glose se rencontre aussi dans les textes littéraires de Frantext alors que la définition, posée en prédication première, y est moins représentée.

[1] recors : subst. masc. Terme du droit : Personne qui assistait un huissier dans les opérations d'exécution en qualité de témoin et dont la présence est aujourd'hui facultative. Synon. praticien. (TLFI)

On voit que, dans l'accession au sens, la voie par les gloses est complémentaire de la voie par la définition, tant par les configurations linguistiques utilisées que par le type de textes à exploiter.

6. Conclusion

Moyennant une adaptation à la tâche visée et au corpus traité, le patron linguistique (P) permet de repérer des gloses de dénomination. La recherche des gloses d'un mot spécifique, comme l'extraction des définitions, est également possible, mais nécessite une analyse syntaxique partielle robuste préalable pour être traitée proprement. Ce sera notre prochaine étape.

Les perspectives d'utilisation des énoncés définitoires sont multiples. Ils sont utiles pour l'acquisition lexicographique, terminologique. D'ores et déjà, on peut utiliser le Web comme dictionnaire encore plus efficacement en sélectionnant parmi les concordances d'un mot donné celles qui glosent le mot. La recherche de « webzine, c'est-à-dire » sur Google[1] place en 1ère position le résultat suivant :

> Résultats 1 - 8 sur un total d'environ 10 pour "webzine, c'est-à-dire". (0,23 secondes)
> Expose Libre - Webzine francophone dédié à Magic The Gathering
> C'est un **webzine, c'est à dire** un magazine normal, sauf qu'au lieu d'être imprimé et vendu dans des kiosques à journaux, celui ci est publié sur internet. **...**
> www.exposelibre.org/?numero=0&article=infos - 8k

alors que la recherche de « webzine » demande d'aller chercher la définition sur les sites référés. L'annotation linguistique du Web[2] (Kilgariff 2003), (Kilgarriff et Grefensette 2003) rendra la recherche de gloses de mot d'autant plus efficace.

L'étude des énoncés définitoires sur des corpus alignés (Pearson 2000), (Suarez de la Torre 2004) révèle que souvent les gloses à marqueur lexical sont traduites par des gloses sans marqueur lexical et vice-versa : grâce à l'alignement, le repérage des gloses d'un corpus peut alors servir à pointer, dans l'autre corpus, les définitions en correspondance, même si elles ne sont pas marquées lexicalement.

Des travaux pourraient être menés consistant à étudier les gloses présentes en corpus parallèles multilingues. En effet, le fait de détecter de manière efficace les gloses en français doit permettre l'observation de similitudes linguistiques selon les langues, étude particulièrement utile pour les applications de traduction automatique par exemple. Par ailleurs, une telle étude permettrait de mettre en œuvre des méthodes de construction automatique de dictionnaires multilingues.

Nous remercions Agnès Steuckardt pour sa disponibilité et sa générosité intellectuelle, Antoine Cornuéjols et Laurent Miclet pour nous avoir confié la version électronique de leur ouvrage.

BIBLIOGRAPHIE

AUGER, A. 1997. *Repérage des énoncés d'intérêt définitoire dans les bases de données textuelles*, Thèse de doctorat, Université de Neuchâtel Disponible sur : <http://www.unige.ch/cyberdocuments/unine/theses2000/AugerA/these_body.html>. (Consulé le 28.05.2006)

AUTHIER-REVUZ, J. 1995. *Ces mots qui ne vont pas de soi*, Paris, Larousse.

BOUVERET, M. 1998. Approche de la dénomination en langue spécialisée, *Meta, XLIII,3*.

BRILL E. 1994. Some Advances in Transformation-Based Part of Speech Tagging, *Actes de AAAI, Vol. 1*, pp. 722-727. Disponible sur : http://citeseer.ist.psu.edu/brill94some.html

CORNUÉJOLS, A. et MICLET, L. (avec la participation d'Yves KODRATOFF). 2002. *Apprentissage artificiel : Concepts et algorithmes*, Eyrolles.

FLOWERDEW, J. 1992. Salience in the performance of one speech act : the case of definitions, *Discourse Process*, 15 (2), pp. 165-181.

GROSS, M. 1975. *Méthodes en syntaxe*, Paris, Hermann.

KILGARRIFF, A. et GREFENSETTE, G. 2003. Introduction to the Special Issue on the Web as Corpus, *Computational Linguistics*, vol. 29/3, pp. 333-347. <http://mitpress.mit.edu/>

[1] Test effectué le 4 Juin 2006.

[2] Cf.<http://citeseer.ist.psu.edu/568007.html> et WebCorp :<http://www.webcorp.org.uk/>

KILGARRIFF, A. 2003. Linguistic search engine, in Kiril-Simov, (éd.), Shallow Processing of Large Corpora : workshop tenu conjointement à *Corpus Linguistics 2003*, Lancaster, England, pp. 53-58. <http://citeseer.nj.nec.com/568007.html>
KOSKAS, E. et KREMIN, H. (resp.). 1984. La dénomination, *Langue française*, *76*.
KLEIBER, G. 1984. Dénomination et relations dénominatives, *Langages, 76*, La dénomination, Koskas et Kremin (resp.).
KLEIBER, G. et TAMBA, I. 1990. L'hyponymie revisitée : inclusion et hiérarchie, *Langue française*, *98*.
KLEIBER, G. 1996. Noms propres et noms communs :un problème de dénomination, *Meta*, *XLI*, *4*.
MELA, A. 2004. Linguistes et "talistes" peuvent coopérer : repérage et analyse des gloses, *Revue Française de Linguistique Appliquée, IX (1)*, *Linguistique et informatique : nouveaux défis*, B. Habert (resp.). Disponible sur <http://www.univ-montp3.fr/~amela/PUBLICATIONS/>
MELA, A. 2005. *Les gloses de nomination seconde, Les marqueurs de glose*, Aix-en-Provence, Publications de l'Université de Provence. Disponible sur :
<http://www.univ-montp3.fr/~amela/PUBLICATIONS/>
MORTUREUX, M.-F. (resp.). 1990. L'hyponymie et l'hyperonymie, *Langue française*, 98.
PEARSON, J. 1999. Comment accéder aux éléments définitoires dans les textes spécialisés ?, *Terminologies nouvelles*, 19. Disponible sur : <http://www.rifal.org/3_information.html>. (Consulté le 25.05.2006)
PEARSON, J. 2000. Une tentative d'exploitation bi-directionnelle d'un corpus bilingue, *Cahiers de grammaire n°25*, Sémantique et corpus, A. Condamines (resp.). Disponible sur :
<http://www.univ-tlse2.fr/erss/>. (Consulté le 27.05.2006)
REBEYROLLE, J. 2000. *Forme et fonction de la définition en discours*, Thèse de Doctorat en Sciences du langage, Université de Toulouse-le-Mirail, Toulouse II. Disponible sur :
<http://www.univ-tlse2.fr/erss/>. (Consulté le 27.05.2006)
REBEYROLLE, J. et TANGUY, L. 2000. Repérage automatique de structures linguistiques en corpus : Le cas des énoncés définitoires, *Cahiers de grammaire n°25*, Sémantique et corpus, A. Condamines (resp.). Disponible sur :<http://www.univ-tlse2.fr/erss/>. (Consulté le 27.05.2006)
RIEGEL, M., PELLAT, J.-C. et RIOUL, R. 1994. *Grammaire méthodique du français*, Paris, PUF.
RIEGEL, M. et TAMBA, I. (resp.) 1987. La reformulation du sens dans le discours, *Langue française*, 73.
RIEGEL, M. et TAMBA, I. 1987. Présentation, *Langue française*, 73, pp. 3-4.
RIEGEL, M. 1987. Définition directe et indirecte dans le langage ordinaire : les énoncés définitoires copulatifs, *Langue française*, 73, pp. 29-53.
STEUCKARDT, A. et NIKLAS-SALMINEN, A. 2003. *Le mot et sa glose*, Aix-en-Provence, Publications de l'Université de Provence.
STEUCKARDT, A. et NIKLAS-SALMINEN, A. 2005. *Les marqueurs de glose*, Aix-en-Provence, Publications de l'Université de Provence.
STEUCKARDT, A. 2006. Du discours au lexique : la glose, Séminaires de l'ATILF, Disponible sur : <http://www.atilf.fr/atilf/seminaires/historique.htm#Steuckardt_2006-03>. (Consulté le 27.05.2006)
SUAREZ DE LA TORRE, M.M. 2004. *Análisis contrastivo de la variación denominativa en textos especializados : del texto original al texto meta*, Tesis Doctoral, Universitat Pompeu Fabra. Disponible sur : <http://www.tdx.cbuc.es/TESIS_UPF/AVAILABLE/TDX-0217105-130025//tmst1de1.pdf>. (Consulté le 27.05.2006)
TAMBA, I. 1987. « Ou » dans les tours du type : « un bienfaiteur public ou évergète », *Langue française*, 73, pp. 16-28.
TAMBA-MECZ, I. 1994. *La sémantique*, Paris, PUF.
VAN-RISBERGEN, C.J. 1979. *Information Retrieval*, 2nd edition, London, Butterworths.