

Le repérage automatique des gloses de nomination seconde

Augusta MELA
Université Montpellier III

Introduction

Les textes numérisés constituent des ressources à partir desquelles on peut chercher des informations lexicales, syntaxiques ou sémantiques. Les méthodes de travail en linguistique s'en trouvent modifiées, en amont, au stade de la recherche des « données », en aval, au stade de l'évaluation et de la validation d'hypothèses.

Or, la façon standard de rechercher les données dans ces ressources consiste à ramener les listes de contextes (les concordances) d'un mot ou d'une suite de mots et le dépouillement de ces listes peut s'avérer fastidieux, les résultats, nombreux, n'étant pas toujours pertinents. Il est donc préférable de sélectionner les lignes qui ont la plus forte probabilité d'appréhender l'information recherchée. Si, par exemple, on s'intéresse au sens d'un mot, spécialisé ou ancien, on cherchera à sélectionner les contextes où ce mot est expliqué voire défini. Les gloses représentent de tels contextes et leurs marques linguistiques permettent d'envisager à la fois leur repérage automatique et la systématisation de leur exploitation puisqu'il existe une correspondance entre ces marques et les relations sémantiques mises en jeu¹.

S'appuyant sur leur étude linguistique, l'exploitation systématisée des gloses entre dans le cadre de l'acquisition automatique de relations lexicales à partir de corpus : cette activité primaire de l'ingénierie linguistique consiste à créer et mettre à jour des ressources lexicales et terminologiques, spécialisées ou générales, qui sont ensuite mises à profit dans de multiples tâches automatisées dont la classification des documents ou la recherche d'information. Ainsi, en recherche d'information, la connaissance de l'existence d'une relation de synonymie entre deux termes permet d'accéder à des extensions du terme premier. Cet accès aux extensions utilise les relations lexicales classiques (synonymie, méronymie, hyperonymie, etc.) entre noms mais également les relations intercatégorielles, de nom à verbe².

Est-il possible de localiser les gloses automatiquement ? Je réponds ici en prenant le cas des gloses introduites par les mots *dit(e)(s)*, signes des plus ambigus et d'un environnement informatique dit pour linguistes : la base textuelle Frantext³ et son interpréteur de langage de requête Stella. Après quelques précisions concernant l'objet du repérage, je décris sa mise en œuvre et ses résultats. Je conclus en généralisant aux autres gloses de nomination seconde.

Champ linguistique couvert

Outre que *dit* peut être la forme conjuguée du verbe *dire* :

(1) Les sports, *dit* Montherlant, sont ceux que les font les mœurs. (Jeux et sports, 1967, p.1291)⁴

les formes *dit(e)(s)* apparaissent sous la forme participe passé, soit en fonction épithète, dans une modalisation autonymique (2), soit en tant que connecteur introduisant une glose de nomination seconde (3) :

(2) Leur étude est l'objet de la théorie *dite* générale. (Histoire générale des sciences, 1964, T.3, Vol.2, p.72)

¹ Relation d'équivalence avec *c'est-à-dire, ou* ; spécification du sens avec *au sens* ; nomination avec *dit, baptisé, nommé* ; hyponymie avec *en particulier, comme, tel* ; hyperonymie avec *et/ou autre(s)*, etc.

² « Par exemple, le lien téléique (du grec « qui indique la finalité ») entre le N *joueur* et le V *mesurer* permet d'accéder à des extensions intercatégorielles du type *joueur de carburant-mesurer du carburant* » (Claveau et al., 2001,731).

³ La base Frantext est accessible par abonnement à l'adresse : <<http://www.atilf.inalf.fr/frantext>>.

⁴ Sauf mention contraire, les exemples cités dans cet article sont extraits de la base textuelle Frantext.

(3) Trenmor entendit murmurer autour de lui que c'était à coup sûr Pulchérie, dite la Zinzolina. (Sand G., Lélia, 1839, p.451)

Dans ce travail, nous cherchons à repérer les gloses de nomination seconde.

Repérer versus extraire

Les gloses correspondent au schéma général « X marqueur Y », X désignant le support de glose et Y l'apport (Steuckardt, ici-même). Dans le cas des gloses de nomination seconde, on peut considérer que X et Y sont des groupes nominaux (GN), le marqueur étant *dit(e)(s)*, *appelé(e)(s)*, *nommé(e)(s)*, ou *baptisé(e)(s)*.

Il ne s'agit pas d'extraire les séquences « X_GN1 marqueur Y_GN2 » :

- d'une part, cela supposerait de disposer d'un étiquetage structurel du texte ;
- d'autre part, on se heurterait au fait que X et Y peuvent être arbitrairement distants⁵ comme dans l'énoncé (4) où *médecine* se rapporte à *sciences* et non à *homme* :

(4) Plus les sciences se rapportent à l'homme, *comme* la médecine par exemple, moins elles peuvent se passer de religion... (J. de Maistre, Les Soirées de Saint-Petersbourg, 1821, T.2, p.217)

- cela supposerait de délimiter les frontières de X et Y. Or cette délimitation exige une compréhension fine du texte : dans l'énoncé (5) ci-dessous, le support de glose est le GN maximal *l'ensemble des instructions pour [...] dans le matériel*, alors que dans (6), il se réduit à *des zones* :

(5) En regard, il y a l'ensemble des instructions pour l'arrangement, le rassemblement, le traitement et l'édition des informations enregistrées dans le matériel, dit Software. (Jolley J-L., Le traitement des informations, 1968, p.37)

(6) Des sujétions particulières à des zones dites « réserves intégrales » peuvent être édictées par décret afin d'assurer, dans un but scientifique, sur une ou plusieurs parties déterminées d'un parc national, une protection plus grande de certains éléments de la faune et de la flore. (Jocard L-M., Tourisme et action de l'état, 1966, p.181)

- enfin, une glose pouvant précéder une deuxième glose ou une explication, il est préférable de garder un contexte large. Dans (7) par exemple, l'extraction des séquences *hypothèse implicite due à *Kepler*, et *règle du triangle distantiométrique*, nous priverait de la définition même de la règle en question donnée par la proposition relative :

(7) Les images sur lesquelles portent les règles de l'optique géométrique sont déduites d'une hypothèse implicite due à Kepler, dite "règle du triangle distantiométrique", qui affirme que l'observateur voit un point lumineux au sommet du cône des rayons qui arrivent à l'œil. (Histoire générale des sciences, 1964, T.3, Vol.2, p.203).

Lorsque l'objectif est de repérer les gloses et de les extraire du corpus dans leur contexte large, en l'occurrence celui de la phrase, ces problèmes ne se posent plus⁶ : ils sont réduits : il suffit de définir les contextes distinctifs minimaux dans lesquels le marqueur introduit une glose.

Définition des « patrons lexico-syntaxiques »

À partir de premières observations, on induit les « patrons lexico-syntaxiques » ou « grammaires locales » de ces contextes distinctifs. La projection de ces patrons sur les corpus permet de ramener de nouveaux résultats ; au vu de ces résultats, les patrons sont raffinés.

⁵ Toutefois, à la différence des gloses en *tel*, les gloses en *dit* semblent ne pas pouvoir être éloignées de la tête du groupe nominal support de glose, quitte à rompre la cohésion de celui-ci : Ainsi, dans un énoncé tel que (14) repris ici : « Il est préparé par un mélange de quatre souches, dont la fameuse souche, dite "Sendai", à laquelle sont attribuées les fréquentes complications pulmonaires des gripes », on peut en effet considérer que le support de glose est la séquence entière : « la fameuse souche à laquelle sont attribuées les fréquentes complications pulmonaires des gripes » et que la glose « dite Sendai » est « anticipée ».

⁶ Du moins, ils relèvent de la compétence humaine de celui ou celle qui exploitera les données recueillies.

À côté des marqueurs *dit/dite/dits/dites* d'autres indices en effet, interviennent, à différents niveaux linguistiques. Ces indices peuvent être de bas niveau, comme les ponctuations : « : », « (« , « , »), qui marquent le décrochage linguistique opéré par la glose. On peut considérer alors que le relateur de glose est double : apposition et marqueur lexical.

Les indices peuvent être de plus haut niveau et abstraits. Ainsi, l'absence d'article devant le substantif introduit par la glose indique qu'il n'a pas de valeur référentielle mais notionnelle ; c'est le cas en (8) pour le terme *couche limite* :

(8) Un fluide naturel, de faible viscosité, se comporte sensiblement comme un fluide parfait à faible distance d'un obstacle autour duquel il s'écoule, tandis que dans une couche très mince entourant l'obstacle, *dite* couche limite, s'exercent des efforts tangentiels importants, avec dissipation d'énergie et formation de tourbillons. (Histoire générale des sciences, 1964, T.3, Vol. 2, p.183)

Dans l'environnement Frantext, à chaque mot correspond une catégorie morpho-syntaxique mais on ne dispose pas d'annotations structurelles. Il n'est pas (encore) possible dans cet environnement de rechercher les séquences correspondant au motif « GN1 dit GN2 »⁷. Il faut donc spécifier des motifs de recherche au niveau des mots, prévoyant tous les cas possibles de frontière droite de GN1 et de frontière gauche de GN2.

Ces spécifications, dites « grammaire » des gloses en *dit*, sont codées⁸ et rangées dans un fichier, nommé *Gloses_en_dit*, recopié ci-dessous. Dans cette grammaire, la catégorie frontière droite de GN1 peut être un substantif ou un adjectif. Le substantif peut être réalisé par les catégories S, Np (nom propre) ou R (mot inconnu de l'étiqueteur de Frantext⁹) ; la fonction adjectivale peut être réalisée par un Adjectif (A), un participe présent (Apr, Pr) ou passé (Aps). nom2 représente la frontière gauche de la nomination seconde : un substantif comme *couche* (8), un Np comme *Federal Reserve Bank* (9), ou un mot inconnu.

(9) C'est à ces défauts que s'efforça de remédier la loi de 1913 qui créait une banque *dite* Federal Reserve Bank, répartie entre 12 districts [...] (Lesourd-Gerard, Histoire de l'économie 19^e et 20^e s., T.1, 1968, p.51)

Afin d'écarter les séquences du type *dit le monsieur* ou *dites le mot*, nom2 ne reconnaît pas les substantifs déterminés suivant *dit* ou *dites*. En revanche, les formes *dits/dite* étant moins ambiguës (elles ne sont pas des formes conjuguées de *dire*), elles sont moins contrôlées et peuvent introduire une nomination seconde déterminée (nom2det). Dans ce cas, l'adjectif est souvent frontière gauche de GN2, soit en tant qu'adjectif postposé au substantif (10), soit employé en fonction substantivale (11) :

(10) Après sa mort /1748/, sa femme Pierrette *dite* la veuve Perrin adopte la nouvelle technique du feu de moufle. (Fontaine G., La céramique française, 1965, p.62)

(11) Tholomyès avait Fantine, *dite* la blonde à cause de ses beaux cheveux couleur de soleil. (Hugo V., Les misérables, 1862, T.1, p.154)

C'est la raison pour laquelle, la frontière gauche de nom2det peut être de catégorie A, Apr ou Aps.

⁷ Actuellement, la technique de l'annotation structurelle est encore au stade de prototype. Cf. (Véronis, 2000) pour un bilan des techniques d'annotation automatique de corpus.

⁸ Un tutoriel est disponible sur le serveur, à partir du menu, via les liens *À quoi servent les listes/grammaires*. Par ailleurs, (Mela, 2004) présente la méthode de mise au point des patrons et leur codage en détail.

⁹ Ce cas est fréquent dans le contexte des gloses.

```
objet :
&e(g=S R Np A APr Pr APs) &?&rponct (&robject1| &robject2| &robject3)

objet1:
dit &?&radv &rnom2

objet2:
&e(c=dites g=APs A) &?&radv &rnom2

objet3:
&e(c=(dite|dits) g=APs A) &?&radv (&rnom2|&rnom2det)

adv:
aussi|encore|également

nom2:
&?" &e(g=S R Np c!=(&msaint|Jésus))

nom2det:
&e(g=D) &?" &e(g=S R Np A APr APs)

ponct:
,|/|\(|...|-
```

Figure 1. Fichier *Gloses_en_dit*

Le signe « | » représente l’alternative ; « &? » précède un élément optionnel.

La recherche de ces motifs se fait via le formulaire de Frantext. La recherche de « &robject,Gloses_en_dit » fait appel à la règle objet du fichier *Gloses_en_dit*, qui elle-même fait appel aux règles objet1, objet2 et objet3, etc.

Sur le corpus des Traités et Essais de 1965 à 1980, soit 45 textes d’un total de 2 739 892 mots, cette recherche ramène 78 résultats présentés dans leur contexte sous la forme d’un fichier Web :



Figure 2 : Résultats de la recherche des séquences décrites par la grammaire des gloses en *dit*.

Analyse des résultats

Comme en recherche documentaire, on évalue les résultats en calculant le bruit (proportion de résultats non corrects parmi les résultats), le silence (proportion de résultats non fournis parmi les résultats attendus), la précision (proportion de résultats corrects parmi les résultats) et le rappel (proportion de résultats donnés parmi les résultats attendus).

Le bruit

Dans 10 cas, *dit* apparaît en tant que verbe conjugué. Il s'agit de :

- 7 séquences « dit_Vbe_conjugué Np » telles que « *dit Montherland* » dans l'énoncé(1) ;
 - 3 séquences de la forme « dit_Vbe_conjugué Objet_non_déterminé » telles que « *dit mythe* » dans (12) ou « *qui dit x dit y* » dans les maximes (13) :
- (12) Claude Lévi-Strauss *dit mythe* là où nous disons métaphysique ou foi. (Salleron L., Comment informer honnêtement, 1965, p.15)
- (13) À cet égard, *qui dit langage dit société*. (Gurvitch G., Traité de sociologie, T.2, 1968, p.276)

La forme « dit Np » est nécessaire si l'on veut ramener les gloses telles que *dite Federal Reserve Bank* (9). Pour éviter les 10 cas de bruit précédents, il faudrait pouvoir distinguer *dit* Verbe conjugué de *dit* participe passé. En principe, l'étiqueteur de Frantext les différencie : l'étiquette V correspond aux verbes conjugués alors que l'étiquette Ps est attribuée aux participes passés, mais, dans la pratique, de nombreux participes sont étiquetés V¹⁰. L'étiquetage prévu n'est pas suffisamment fiable, c'est la raison pour laquelle la grammaire des gloses en *dit* n'en tient pas compte pour la forme masculin singulier *dit*.

6 des 7 énoncés de la forme « dit_Vbe_conjugué Np » pourraient être écartés si l'on disposait du genre et du nombre des mots. Il suffirait, dans la grammaire des gloses en *dit*, d'exprimer la contrainte d'accord en genre et nombre entre la frontière droite de GN1 et le mot *dit*, vérifiée quand *dit* est épithète ou connecteur de glose. Ainsi (!) serait exclu, *sports* ne s'accordant pas en genre et nombre à *dit*. L'étiqueteur de Frantext ne fournit pas un étiquetage aussi fin mais d'autres étiqueteurs comme Cordial¹¹ le proposent.

Dans l'environnement Frantext, lorsque le corpus contient une forte proportion de discours directs, la parade consistera à supprimer la catégorie Np dans nom2 ou à ne l'autoriser que précédée de guillemets. On évitera alors les séquences du type *dit Montherland* tout en maintenant les énoncés tels que (14) :

(14) Il est préparé par un mélange de quatre souches, dont la fameuse souche, *dite "Sendai"*, à laquelle sont attribuées les fréquentes complications pulmonaires des gripes. (Schwartz, Nouveaux remèdes et maladies actuelles, 1965, p.133)

La précision augmentera mais le rappel diminuera aussi puisque les gloses telles que *dite Federal Reserve Bank* (9) ne seront plus ramenées.

Les 11 autres cas de bruit relèvent de la modalisation autonymique. Les séquences « dit_PPassé Adj » telles que « *dite générale* » ne sont pas reconnues par la grammaire mais dans 9 cas, l'étiqueteur les analyse en tant que séquences « dit_PPassé Mot_inconnu ». Ainsi les résultats 3 et 4 de la copie d'écran ci-dessus sont ramenés parce que les adjectifs *surgras* et *alternating* sont catégorisés R (mots inconnus du logiciel).

2 derniers cas de bruit résultent de l'étiquetage erroné du mot *sous-développé* : la catégorie S(substantif) possible dans d'autres contextes est incorrecte ici :

(15) Les pays *dits* sous-développés apparaissent d'abord comme un ensemble de collectivités rurales médiocrement liées entre elles. (Gurvitch G., Traité de sociologie, T.1, 1967, p.335)

Notons que la frontière entre modalisation autonymique et glose de nomination seconde peut être floue, humainement parlant. C'est notamment le cas lorsque le mot qui suit *dit(e)(s)* peut être employé en tant que substantif ou en tant qu'adjectif comme le mot *autotrophes*¹² en (16) :

(16) Les végétaux chlorophylliens, *dits* autotrophes, se nourrissent donc en fait d'aliments véritablement minéraux, ce que les animaux sont incapables de faire. (Peres J-M, La vie dans l'océan, 1966, p.10)

Le mot *autotrophes* ayant été catégorisé R (mot inconnu du logiciel) par Frantext, l'énoncé (16) est ramené.

Au total, toutes causes confondues, sur 78 résultats, 21 énoncés, soit environ 1/4 des résultats, ne sont pas des gloses de nomination seconde. La moitié des cas « bruyants » proviennent de l'ambiguïté de la forme masculin singulier *dit* : verbe et participe passé.

¹⁰ Voir la liste des codes grammaticaux de Frantext en annexe.

¹¹ L'analyseur morpho-syntaxique Cordial 'Universités' (pour Windows) est créé et commercialisé par la société Synapse Développement, Toulouse : <<http://www.synapse.com>>.

¹² Une recherche sur le Web avec le moteur Google montre deux emplois d' *autotrophe* dont : « Autotrophes : Organismes puisant l'énergie nécessaire à leurs synthèses de matières organiques dans des sources minérales. » sur <<http://membres.lycos.fr/mad8/EvolVie/glossair.htm>>.

Le silence

Dans cet environnement, on ne peut accéder au corpus de façon linéaire. Toutefois, on peut sonder le silence en utilisant une « grammaire » moins restreinte comme la grammaire des appositions en *dit* ci-dessous, censée ramener les gloses en *dit* ainsi que les modalisations autonymiques :

```
objet : &e(g=S R Np A APr Pr APs) &?&rponct (&robject1| &robject2| &robject3)

objet1:
&e(c=dit) &?&radv &?" &?(de|d'|des|du|de la) &rnombis

objet2:
&e(c=dites g=APs A) &?" &?&radv &?(de|d'|des|du|de la) &rnombis

objet3:
&e(c=(dite|dits) g=APs A) &?" &?&radv &?(de|d'|des|du|de la|à|au|aux) &?&e(g=D
c!=&mun) &rnombis

adv:
aussi|encore|également

nom2:
&?" &e(g=S R Np A APs Aca Apr c!=(concert|ton|voix))

nombis:
&?" &e(g=S R Np A APs Aca APr Pr Ps c!=(&msaint|Jésus|concert|ton|voix))

ponct:
,|/|\(|...|-
```

Figure 3. Fichier *Appositions_en_dit*

Cette grammaire, projetée sur le même corpus, ramène 257 résultats, dont la plupart sont des modalisations autonymiques. Les 58 gloses de nomination repérées par la grammaire des gloses en *dit* y figurent ainsi que 3 gloses de nomination seconde non reconnues par la grammaire des gloses en *dit*. Dans 2 cas, il s'agit d'entités nommées non (re)connues en tant que telles par l'étiqueteur. Au lieu donc d'être étiquetées globalement en tant que Np, elles sont analysées mot par mot et, la frontière gauche de GN2 étant un A(dj) :

[dit V] [basse A]-[courtille S] (16) et « [seconde A] [A Np] »

la grammaire des gloses en *dit* ne les reconnaît pas :

- (17) Au lieu *dit* basse-courtille, Loché de Roissy s'établit en 1771. (Fontaine G., la céramique française, 1965, p.132)
- (18) Ainsi, le baccalauréat de philosophie /A/ et le baccalauréat de sciences économiques /B/ sont issus d'une première année commune dite "seconde A", les sections A et B étant différenciées dès le début de la deuxième année du cycle. (Capelle J., École de demain reste à faire, 1966)

Le troisième cas est ambigu parce que le mot *anaérobies* catégorisé Adjectif ([germes S] [dits APs] [anaérobies A]), peut aussi être employé en tant que substantif :

- (19) À l'intervention, le chirurgien retire un pus très épais bien lié ou grumeleux, d'odeur souvent fétide, contenant parfois des gaz lorsqu'on se trouve en présence de germes *dits* anaérobies. (Encyclopédie médicale Quillet, 1965, p.345)

Si enfin, on relâche les contraintes sur la catégorie de *dite/dits/dites* (c'est-à-dire qu'on efface les spécifications « g=APs A » dans la grammaire), on obtient 8 résultats supplémentaires qui sont tous des modalisations autonymiques.

Nous ne pouvons pas dire pour autant que le silence est réduit sur ce corpus à ces trois cas. En effet, le schéma « X marqueur Y » est abstrait. Pour l'utiliser en repérage automatique, il a fallu préciser les catégories possibles de X et de Y, mais il faudrait également préciser le type des insertions éventuelles entre X et Y. Ainsi, la grammaire des appositions en *dit* ne ramène pas un énoncé tel que (20) parce qu'un élément (*en outre*) est inséré entre *certaines terrains* et le marqueur *dits* :

(20) Certains terrains, en outre, dits "phyloxérants", c'est-à-dire plus propices que d'autres à la multiplication de la bête, mirent le cep à la fois en présence d'ennemis en nombre écrasant et d'éléments fonciers indigestes. (Pesquidoux J de, Le livre de raison, T.1, 1925, p.77)

Pour remédier à cette cause de silence, il reste à prévoir toutes les insertions possibles. Ce travail reste à faire.

Enfin, le logiciel Stella limitant le repérage au contexte de la phrase, il n'est pas facile de repérer les gloses inter-phrastiques, laissées aussi sous silence, ou les gloses faussement inter phrastiques, comme *dite "das reich"* dans (21)¹³ qui suivent un point d'abréviation :

(21) La 17e panzer, aux prises avec les nôtres entre Bordeaux et Poitiers, perd dix jours avant que ses colonnes aient réussi à se frayer la route. La 2e panzer s.S. *Dite "das reich"*, partie de Montauban le 6 juin et qui ne peut utiliser les voies ferrées –toutes hors d'usage–voit ses éléments arrêtés dans le Tarn, le Lot, la Corrèze, la haute Vienne ; (Gaulle Ch. de, Mémoires de guerre : L'unité, 1956)

Les données sur le silence et le rappel sont donc imprécises .

Récapitulatif

Notons enfin que, dans le corpus examiné, la ponctuation n'est ni majoritaire (24 gloses précédées d'une ponctuation sur 58 gloses de nomination seconde) ni significative : elle ne permet pas de départager les modalisations autonymiques des gloses de nomination seconde.

Gloses ponctuées	24	41%
Gloses non ponctuées	34	59%
Total :	58	

Figure 4. La ponctuation des gloses de nomination seconde

Le tableau suivant récapitule les observations sur les gloses en *dit* :

Causes de bruit :	Nombre d'occurrences		Bruit	Précision	Silence
dit_Vbe_conjugué Np	7	10	21/78 ~ 25%	57/78 ~ 75%	3 cas repérés sur 61 (58+3) attestés
dit_Vbe_conjugué Objet_non_déterminé	3				
dit_PPasseé Mot_inconnu	9				
dit_Ppassé Mot_S_et_A	2				

Figure 5. Analyse des résultats pour la grammaire des gloses en *dit*.

Généralisation : une grammaire des gloses de nomination seconde

La grammaire précédente peut être étendue aux autres marqueurs de gloses de nomination seconde : *appelé(e)(s)*, *nommé(e)(s)*, *baptisé(e)(s)*. Ces marqueurs étant moins ambigus que *dit(e)(s)*, leur « grammaire » s'en trouve simplifiée :

¹³ Cet énoncé s'obtient en recherchant le motif « .dite » et en appliquant un « zoom » sur le premier résultat.

```

objet :
&e(g=S R Np A APr Pr APs) &?&rponct &e(c=(&mappelé|&mnommé|&mbaptisé) g=APs A) &?&radv
&rnom2

adv:
aussi|encore|également |tout d'abord

nom2:
&?" &e(g=S R Np c!=(&msaint|Jésus))|&e(g=D) &?" &e(g=S R Np A APr APs)

ponct:
,|/|\(|...|-

```

Figure 6. Grammaire des nominations secondes autres que les gloses en *dit*.

Projetée sur le même corpus des Traités et Essais de 1965 à 1980, cette grammaire ramène 134 résultats :

The screenshot shows a web browser window with the URL <http://atilf.atilf.fr/Dendien/scripts/categ/browserb.exe?1043;s=1938604605;r=19;>. The search results are displayed on a yellow background. The main heading is "Résultats 1 à 50/134". A search box contains the text: "Les codes 'certains' sont indiqués en vert. Les codes 'incertains' sont indiqués en orange. Cliquez ici pour voir la signification des codes grammaticaux." Below this, there are five search results, each with a "ZOOM" button and a "Retour en haut du document" link. The results are as follows:

- Résultat 1 (Texte sous droits)**: P947/* / ENCYCLOPÉDIE MÉDICALE QUILLET / 1965 page 125 / Les sécrétions de la glande sont évacuées par un canal **excréteur appelé canal** de *Stenon qui aboutit à la cavité buccale où il se termine en regard de la première ou de la deuxième molaire supérieure.
- Résultat 2 (Texte sous droits)**: P947/* / ENCYCLOPÉDIE MÉDICALE QUILLET / 1965 page 163 / Elle a été mise au point en *I 948 par un **chirurgien nommé Swenson**.
- Résultat 3 (Texte sous droits)**: P947/* / ENCYCLOPÉDIE MÉDICALE QUILLET / 1965 page 171 / Elles peuvent également se révéler par une démangeaison **anale appelée prurit** qui peut devenir rapidement insupportable.
- Résultat 4 (Texte sous droits)**: P947/* / ENCYCLOPÉDIE MÉDICALE QUILLET / 1965 page 188 / Certains appartiennent au groupe des champignons et sont responsables de ces **maladies appelées mycoses**, dont l'importance est actuellement en augmentation : les teignes, le muguet, l'actinomycose, etc.

Figure 7. Résultats de la recherche des séquences décrites par la grammaire des gloses de nomination seconde autres que les gloses en *dit*.

La précision est proche de 100% : sur 134 résultats, deux cas seulement ne sont pas des gloses de nomination seconde :

(22) Il est pour la première fois *appelé* faïencier en 1679. (Fontaine G., La céramique française, 1965, p.46)

(23) Ce qui répond le mieux aux vœux des élus locaux, c'est de donner à l'administration territoriale *nommée* la faculté de traiter les affaires et d'essayer de les dénouer à l'échelon même où elles se posent. (Belorgey G., Gouvernement et administration de la France, 1967, p.295)

On peut classer les 4 marqueurs suivant leur productivité sur le corpus examiné, c'est-à-dire le nombre de gloses de nomination seconde ramenées :

Configurations	Marqueurs	Nombre de gloses de nomination seconde ramenées
	appelé(e)(s)	107
	dit(e)(s)	57
Humain nommé :19	nommé(e)(s)	23
Lieu nommé :1		
Objet nommé :3		
Lieu baptisé :1	baptisé(e)(s)	2
Objet baptisé :1		

Figure 8. Comparaison des marqueurs en termes de « productivité » de nominations secondes.

Sur ce corpus, le marqueur *appelé(e)(s)* est le marqueur de nomination qui offre la meilleure précision et le plus grand nombre de gloses de nomination seconde.

Projetée sur la base Frantext catégorisée, la grammaire des gloses de nomination seconde autres que les gloses en *dit* ramène 2415 résultats. L'examen des premiers résultats confirme leur bon taux de précision.

Conclusion

Est-il possible de repérer automatiquement les gloses? Il convient de différencier la réponse car la qualité des résultats dépend de facteurs linguistiques et techniques. À l'intérieur du même type de gloses de nomination seconde, *appelé* et *nommé* sont des marqueurs plus précis que *dit* ; *dits* et *dite* sont plus précis que *dit* et *dites* pour des raisons que nous avons analysées. Le contexte « *dit* Substantif_non_déterminé » est plus distinctif que « *dit* Déterminant Substantif ».

L'efficacité du repérage dépend également des outils informatiques utilisés. Le choix de l'environnement Frantext présente des avantages certains pour le linguiste : il permet de disposer rapidement à la fois de textes numérisés, d'un outil d'étiquetage de ces textes et d'un outil de recherche de séquences textuelles. Mais cet environnement est limité. Outre l'information structurelle, des étiquettes fines font défaut : nous avons vu comment le genre et le nombre des entités permettraient d'éliminer certains cas de bruit. Les outils utilisés sont exécutés par le serveur de Frantext¹⁴, à distance, avec des conséquences sur le temps d'exécution. Ces outils sont « fermés » : le linguiste n'a pas la possibilité d'enrichir les bases de connaissances lexicales (par exemple ajouter un mot et sa catégorie) et grammaticales sur lesquelles la procédure d'étiquetage se base. Ainsi, un mot inconnu restera inconnu. Enfin, on peut avoir à sonder des textes d'un autre type que ceux de la base Frantext.

¹⁴ Et non « en local », sur notre propre machine.

Toutes ces limites peuvent être dépassées. L'étiqueteur Cordial permet d'annoter nos propres textes et utilise un jeu d'étiquettes plus fines. L'environnement Intex¹⁵ permet d'annoter des informations structurelles (GN, p.ex.) suivant des grammaires que nous pouvons définir suivant nos besoins. Les « grammaires » présentées dans cet article sont transposables dans ces autres environnements, et donc améliorables.

Bibliographie

- Anscombe, J.-Cl. (resp.) (1995) *Théorie des topoi*. Éditions Kimé. Paris
- Auger A (1997) : Repérage des énoncés d'intérêt définitoire dans les bases de données textuelles, Thèse de doctorat, Université de Neuchâtel
<http://www.unige.ch/cyberdocuments/unine/theses2000/AugerA/these_body.html>
- Authier-Revuz, J. (1994) : L'énonciateur glosateur de ses mots : explicitation et interprétation. *Langue française*, 103, 91-102.
- Authier-Revuz, J. (1995) : *Ces mots qui ne vont pas de soi*. Paris, Larousse.
- Authier-Revuz, J. (1982) : La mise en scène de la communication dans les discours de vulgarisation scientifique. *Langue française*, 53, 34-47.
- Claveau, V. et al. (2001) : Acquérir les éléments du lexique génératif : quels résultats et à quels coûts ? *Traitement automatique des langues*, 42, n°3, 729-753.
- Bosredon, B., Tamba, I., Petit, G. (resp.) (2001) Linguistique de la dénomination. *Cahiers de praxématique* 36.
- Bouillon, P. et Viegas, E. (éd.) (2001) : Lexiques sémantiques. *Traitement automatique des langues*, 42, n°3, Hermès.
- Bourigault, D. et Slodzian, M. (1999) : Pour une terminologie textuelle. *Terminologies nouvelles*, 19, 29-32. <<http://www.cfwb.be/franca/termin/charger/rint19.pdf>>
- Desgraupes, B. (2001) : *Introduction aux expressions régulières*. Paris, Vuibert Informatique.
- Fuchs, C. (1982) : La paraphrase entre langue et discours. *Langue française*, 53, 22-33.
- Habert, H., Fabre, C., Issac, F. (1998) : *De l'écrit au numérique*. Paris, InterEditions.
- Habert, B. (2002) : Outiller les linguistes/ outiller la linguistique : par où, par qui commencer ? Intervention à la table ronde TAL et enseignement, TALN'02.
<<http://www.limsi.fr/Individu/habert/Cours/index.html>>
- Hearst M. (1998) : Automated discovery of Wordnet relations. In C. Fellbaum, ed. *WordNet : an electronic lexical database*, Language , Speech and Communication, chapitre 5, pp. 131-151. Cambridge, Massachussets, MIT Press
<<http://www.sims.berkeley.edu/~hearst/publications.shtml>>
- Julia, C. (2001) : *Fixer le sens ? La sémantique spontanée des gloses de spécification du sens*. Paris, Presses de la Sorbonne Nouvelle.
- Malaisé, V., Zweigenbaum, P. et Bachimont, B. (2004) : Repérage et exploitation d'énoncés définitoires en corpus pour la construction d'ontologie. *TALN 2004*.
<www.lpl.univ-aix.fr/jep-taln04/proceed/actes/taln2004-Fez/Malaise-Zweigenbaum-et-al.pdf>
- 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.) <<http://www.univ-montp3.fr/~amela/Publications.html>>

¹⁵ Intex est un environnement de développement linguistique gratuit sous certaines conditions d'utilisation et téléchargeable à l'adresse <<http://grellis.univ-fcomte.fr/intex/overview.html>>.

- Morin, E. (1999) : *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques*. Doctorat en Informatique, Université de Nantes. <<http://www.sciences.univ-nantes.fr/info/perso/permanents/morin/publications.html>>
- Mortureux, M.-F. (éd.) (1982) : La vulgarisation. *Langue française*, 53.
- Muresan, S & Klavans, J. (2002) A method for automatically building and evaluating dictionary resources. *Proceeding of the Language Resources and Evaluation Conference (LREC 2002)* <<http://www1.cs.columbia.edu/~smara/publications/papers.htm>>
- Pearson, J.(1999) : Comment accéder aux éléments définitoires dans les textes spécialisés ?. *Terminologies nouvelles*, 19. <<http://www.rifal.org/>>
- Rebeyrolles, J. (2000) : *Forme et fonction de la définition en discours*. Doctorat en Sciences du langage, Université de Toulouse-le-Mirail, Toulouse II. <<http://www.univtlse2.fr/erss/membres/rebeyrol/>>
- Rey-Debove, J. (1997) : *Le métalangage*. Paris, Armand Colin.
- Riegel, M. et Tamba, I. (éd.) (1987) : La reformulation du sens dans le discours. *Langue française*, 73.
- Steuckardt, A. et Niklas-Salminen, A. (éd.) (2003) : Le Mot et sa Glose. *Langues et langage*, 9, Aix-en-Provence, Publications de l'Université de Provence.
- Tamba-Mecz, I. (1994) : *La sémantique*. Paris, Presses Universitaires de France.
- Véronis, J. (2000) : Annotation automatique de corpus : panorama et état de la technique. In *Ingénierie des langues*, Pierrel, J.-M. (éd.), Paris, Hermès, 111-129. <www.up.univ-mrs.fr/~veronis/pdf/2000hermes4.pdf>

Augusta.Mela@univ-montp3.fr
Université Paul-Valéry Montpellier III
Route de Mende
34199 Montpellier Cedex 5 France

Annexe : Liste des codes grammaticaux de Frantext

Attention : Ces codes sont à utiliser "tels quels" en respectant majuscules et minuscules. Ils peuvent être éventuellement précédés des préfixes "i" ou "c ». Par exemple :

- **&e(g=cCc)** désigne une conjonction de coordination reconnue "avec certitude" par le programme de catégorisation.
- **&e(g=iCc)** désigne une conjonction de coordination reconnue "avec incertitude" par le programme de catégorisation.
- **&e(g=Cc)** désigne une conjonction de coordination reconnue "avec ou sans certitude" par le programme de catégorisation.

A	adjectif (sauf cas Aca, Apr, Aps)
Aca	adjectif cardinal
APr	adjectif/participe présent
APs	adjectif/part. passé
Adv	Adverbe
Avn	Partie d'une négation (par ex. cas de ne / n' , ou pas / point / guère ... si associés à ne ou n')
Cc	Conjonction coordination
Cs	Conjonction subordination
D	Déterminant (sauf cas Dca, Dg)
Dca	car. dét (cardinal ayant le rôle d'un déterminant : <i>deux pigeons s'aimaient</i>)
Dg	amalgamés (au/aux/du/des)
E	exclamatif
Ep	présentatif (voici, voilà, ...)
Ger	gérondif (<i>en</i> lié à un participe présent)
Inf	infinitif
Inj	interjection (ah, oh, ha, ho, ...)
Int	interrogatif
Np	Nom propre
Nu	numeral card.
Ono	onomatopée
P	Pronom (sauf cas Per, X)
Per	Pronom personnel
Pp	Préposition
Pr	Participe présent sauf cas APr, Ger
Ps	Participe passé (sauf cas APs)
S	Substantif
V	Verbe (sauf participes et infinitif)
R	mot inconnu du logiciel
X	mot non traité (que/qu', où, sinon)