Mixer les moyens pour extraire les gloses

Augusta Mela*, Mathieu Roche** et Mohamed el Amine Bekhtaoui**

* Université Montpellier 3, 34199 Montpellier Cedex 5, France
Augusta.Mela@univ-montp3.fr
** LIRMM,CNRS, Université Montpellier 2, 34392 Montpellier Cedex 5, France
Mathieu.Roche@lirmm.fr, a.bekhtaoui@gmail.com

Résumé : Nous proposons d'extraire des connaissances lexicales en exploitant les « gloses » de mot, ces descriptions spontanées de sens, repérables par des marqueurs lexicaux et des configurations morpho-syntaxiques spécifiques. Ainsi dans l'extrait suivant, le mot *testing* est suivi d'une glose en *c'est-à-dire* : « *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*». Cette approche ouvre des perspectives pour l'acquisition lexicale et terminologique, fondamentale pour de nombreuses tâches. Dans cet article, nous comparons deux façons d'extraire les unités en relation de glose : patrons et statistiques d'associations d'unités sur le web, en les évaluant sur des données réelles.

1 Introduction

L'acquisition automatique de connaissances lexicales à partir de textes vise à identifier divers types d'unités lexicales (termes, entités nommées, mots composés, mots nouveaux, mots à sens nouveau) ainsi que leurs propriétés syntaxiques et sémantiques. En contexte multilingue, à partir des corpus bilingues, elle consiste à repérer les traductions de ces unités.

Elle constitue une aide précieuse pour la construction de dictionnaires, thesaurus et terminologies, qu'ils soient de langue générale ou spécialisée. Elle intéresse également la recherche documentaire grâce à l' « expansion de requête » puisqu'elle permet de comparer aux index des documents susceptibles de correspondre à la requête de l'utilisateur, non seulement les mots présents dans la requête mais également leurs synonymes, hyperonymes ou hyponymes, voire leurs traductions dans une autre langue, toutes connaissances obtenues en amont par des procédés d'acquisition lexicale.

Trois objectifs peuvent donc être distingués :

l'*extraction* d'unités, ou comment repérer des unités lexicales spécifiques : termes, entités nommées, mots composés ;

l'*alignement* d'unités, ou comment repérer leurs traductions à partir de corpus bilingues ;

la *structuration* de ces unités, c'est-à-dire les relations de ces unités entre elles.

La section 2 présente les approches généralement utilisées en structuration, la section 3 présente notre approche conceptuelle, la section 4 est réservée au logiciel qui en résulte et à partir duquel sont obtenus les résultats expérimentaux rapportés en section 5. La section 6 dessine quelques perspectives de ce travail.

2 L'acquisition de connaissances lexicales à partir de corpus

Notre travail concerne la structuration lexicale. Pour le situer, nous nous restreindrons aux travaux en structuration lexicale et renvoyons le lecteur à ((Bourigault et Jacquemin (2000) ; (Véronis, (2000)) pour un panorama des travaux effectués en extraction et en alignement. Nous reprenons à notre compte la distinction opérée par (Nazarenko et Hamon, 2002) concernant les différents plans de structuration : les relations de sémantique lexicale telles que la synonymie ou l'hyperonymie relèvent de la microstructuration, alors que les classifications d'unités en thèmes ou domaines relèvent de la macrostructuration : à ce niveau, la nature du lien entre unités n'est pas identifiée.

En macrostructuration, les approches utilisées sont essentiellement fondées sur la sémantique distributionnelle de Harris (Langages n°99). Selon cette théorie, l'ensemble des contextes dans lesquels le mot apparaît permet d'en tirer « le portrait » et d'en déterminer le sens. Pratiquement, étant donné un mot, on recherche par quels mots il est substituable (propriétés paradigmatiques), avec quels mots il se combine (propriétés syntagmatiques), ou plus largement, dans quels voisinages il apparaît (propriétés de co-occurrence). Ces questions ne sont pas nouvelles et les lexicographes ont toujours utilisé les corpus pour y répondre. Ce qui change aujourd'hui, c'est la taille des corpus et les outils informatiques capables de systématiser l'approche distributionnelle.

En microstructuration, les approches sont diverses :

- soit elles utilisent la structure même de l'unité (par exemple, le *coussin de sécurité* est un hyperonyme de *coussin de sécurité arrière* et *anticapitaliste* est l'antonyme de *capitaliste*) ;

- soit, suivant l'approche distributionnelle, elles considèrent que les unités en relation apparaissent dans des contextes similaires et elles diffèrent quant à la manière dont elles définissent le « contexte » et calculent la « similarité ». Le contexte peut se limiter aux mots dans l'entourage proche du mot cible ou être étendu au document entier. Représenté dans un espace vectoriel dont la base sont les mots, le mot cible est un vecteur et deux mots sont similaires si leurs représentations vectorielles sont proches (Salton et al., 1975). Cette approche se heurte au fait que deux mots synonymes peuvent ne pas avoir la même distribution (ex : les verbes *appeler* (au téléphone) et *téléphoner* sont proches sémantiquement mais l'un est suivi d'un objet direct et l'autre d'une préposition (*appeler quelqu'un/téléphoner à quelqu'un*) ; d'autre part, leurs contextes similaires peuvent se situer hors de la fenêtre prise en compte pour l'étude. Mieux vaut alors considérer que le contexte pertinent n'est pas l'ensemble des co-occurrents mais l'ensemble des éléments qui sont en relation syntaxique avec le mot cible (par exemple, les objets communs des verbes *appeler* et *téléphoner*, qui peuvent être directs ou indirects, à distance ou pas de ces verbes). De plus, ces calculs de similarité entre mots ne sont significatifs que si les mots ciblés apparaissent souvent dans le corpus ;

- soit elles partent du principe que les relations sémantiques sont exprimées par des marques linguistiques (éléments lexicaux, grammaticaux, paralinguistiques (marques de ponctuation, guillemets). Dans cette approche, des patrons morpho-syntaxiques qui combinent ces marques sont mis au point manuellement et projetés dans les textes pour ramener les items censés être dans la relation sémantique étudiée. Ainsi, dans (Hearst, 1992), le patron « SN, SN et autres SN » décrit les Syntagmes Nominaux (SN désormais) du type *conjonctivites, orgelets et autres infections oculaires* et repère la relation d'hyperonymie entre *infections oculaires* et *conjonctivites* et *orgelets*. Ces patrons peuvent également

s'acquérir de façon semi-automatique : plutôt que de décrire le patron de la relation d'hyperonymie, le système Prométhée (Morin, 1999) part d'une liste préalable de couples de termes qui vérifient la relation, et cherche dans les textes les énoncés qui contiennent ces couples. À partir de ceux-ci, des descriptions génériques sont induites qui, projetées dans les textes, permettent de ramener de nouveaux couples. Une fois validés par l'expert du domaine, les nouveaux couples servent, à leur tour, d'accroche pour repérer d'autres configurations et ainsi de suite jusqu'à stabilisation du nombre de couples et de configurations. Il faut bien sûr disposer de gros corpus ou de textes suffisamment spécialisés pour que les termes en relation soient mis en scène plusieurs fois et puissent signaler de nouvelles manifestations de leur relation.

Quoi qu'il en soit, la qualité des résultats de cette approche dépend de la précision avec laquelle les patrons ont été définis ou de la qualité du corpus qui sert à les induire. C'est pourquoi les patrons sont dépendants du corpus et du domaine considérés.

Enfin, dans l'approche par patrons, la recherche d'un phénomène linguistique est efficace en temps de calcul mais limitée intrinsèquement par le fait que certaines occurrences du phénomène ne sont pas conformes au patron à cause des possibilités d'insertion de la langue.

3 Une approche mixte pour extraire les gloses

Pour nous affranchir de ces limites, et donner des résultats « à tous les coups », nous proposons une approche mixte, reposant sur des patrons mais aussi sur des calculs statistiques d'associations de termes attestées sur le Web. Le logiciel présenté en section 4 est une implémentation de cette solution mixte. Dans ce paragraphe, nous détaillons les bases de notre approche : la section 3.1 précise ce qu'est la « glose » et rappelle les difficultés liées à son extraction automatique. La section 3.2 explique la mise en œuvre de notre approche.

3.1 Le mot et sa glose

Dans le projet pluridisciplinaire PEPS RESENS, linguistes et informaticiens sont associés pour vérifier dans quelle mesure « le mot et sa glose » permettent un accès efficace au sens des mots et en proposer des applications.

Les gloses sont des commentaires en situation parenthétique, souvent introduits par des marqueurs tels que *appelé*, *c'est-à-dire*, *ou* qui signent la relation de sémantique lexicale mise en jeu : équivalence avec *c'est-à-dire, ou* ; spécification du sens avec *au sens* ; nomination avec *dit*, *appelé* ; hyponymie avec *en particulier*, *comme* ; hyperonymie avec *et/ou autre(s)*, etc.

Elles sont en apposition au mot glosé, le plus souvent de catégorie nominale.

Leurs typologies syntaxique et sémantique ont été établies par les linguistes (Steuckardt, 2006). Il a également été remarqué que leur fréquence dépend du genre des textes : on en trouve davantage dans des textes didactiques ou de vulgarisation qu'en poésie.

La glose est spontanée. Elle partage cette caractéristique avec la définition, dite naturelle parce qu'elle n'est pas le fruit du travail réfléchi d'un lexicographe. Cependant la glose est parenthétique alors que la définition naturelle est l'objet principal du propos. Ainsi, dans *L'accouchement, également appelé travail, naissance ou parturition, est l'aboutissement de la grossesse, la sortie d'un enfant de l'utérus de sa mère,* la glose en *appelé* du mot *accouchement* : « *également appelé travail, naissance ou parturition* » révèle que *travail,*

naissance ou parturition sont d'autres façons de nommer *l'accouchement* alors que *l'aboutissement de la grossesse, la sortie d'un enfant de l'utérus de sa mère* est la définition naturelle du mot *accouchement*. La définition naturelle a été décrite en vue de son traitement automatique par (Rebeyrolles, 2000). Sa configuration et la copule *être* ne sont pas des marques suffisamment discriminantes mais comme il n'est pas rare qu'elle soit concomitante d'une glose, nous la rechercherons dans les passages didactiques signalés par les gloses.

On peut donc décrire la glose schématiquement par la configuration « *X* marqueur *Y* ».

Nous avons précédemment réalisé le repérage de gloses diverses en adoptant l'approche par patrons : gloses en *ou* (Mela, 2004), gloses en *dit* (Mela, 2005).

Nous souhaitons à présent extraire précisément les entités en relation de glose. (Mela et Roche, 2006) étudie le cas des gloses de nomination en *appelé* et analyse les problèmes que pose l'extraction. Les difficultés sont dues :

- à l'ambiguïté structurelle du langage naturel : quand un SN enchasse d'autres SN, auquel se rapporte la glose ? On peut s'appuyer sur des marques d'accord grammatical, ou typographiques comme dans l'exemple ci-dessous, mais ces marques ne sont pas toujours disponibles ;
  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». (Sud Ouest, Axelle Maquin-Roy, 19 mai 2006)
- aux possibilités d'insertion au sein de la structure de glose : des éléments peuvent s'insérer entre le pivot verbal *appelé* et *X* et *Y* : adverbe, incise ou syntagme coordonné :
  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)
  Avec ce "radio-conducteur", perfectionné en 1890 et *appelé* «cohéreur» par Lodge, la radioélectricité était née.

3.2 Modalités de notre approche

Nous détaillons notre méthode en reprenant le cas des gloses de nomination en *appelé*.

Nous partons du principe ici que *X* et *Y* sont des syntagmes nominaux (SN désormais), ce qui est le cas général. De plus, on prend en compte le fait que *Y* peut être une coordination de SN, comme dans « *L'accouchement, également appelé travail, naissance ou parturition* », donc la variante du schéma abstrait de glose devient : « *X* marqueur Y_i ».

Le patron de glose décrit dans (Mela et Roche, 2006) sert à repérer les gloses ainsi que les SN candidats. Ce sont ensuite des calculs statistiques d'associations sur le Web des couples (SN glosé, SN glosant) qui nous conduiront à choisir les SN glosants les plus pertinents.

Un premier patron détecte Y_1, le premier SN à droite du marqueur. Par exemple : « *Un disque microsillon, appelé disque vinyle* » permet d'extraire *X = Un disque microsillon* et Y_1 = *disque vinyle*.

Un deuxième patron prend en compte l'éventualité d'une coordination en position *Y*, pour extraire d'une séquence telle que de « *Un disque microsillon, appelé disque vinyle ou Maxi* », deux SN *disque vinyle* et *Maxi*.

Le corpus est étiqueté au préalable avec le système de Brill (1994), ce qui nous permet d'utiliser les patrons (Nom-Nom, Nom-Adjectif, etc.) de Daille (1994) pour reconnaître les SN. En cas d'ambiguité structurelle, nous privilégions de façon arbitraire l'extraction des syntagmes nominaux maximaux. Nous pourrions utiliser les statistiques pour choisir le SN le plus pertinent, nous ne l'avons pas réalisé ici. Enfin, d'autres heuristiques ont été ajoutées, par exemple, nous privilégions les unités entre guillemets, ces signes paralinguistiques étant des marques assez discriminantes d'unités glosées ou glosantes.

3.2.1 Extraction des syntagmes candidats par des méthodes de fouille de textes

Une fois la glose en *appelé* détectée dans un corpus, nous recherchons les SN situés entre le marqueur et une frontière droite. Cette frontière droite est soit un verbe conjugué, soit une ponctuation forte. Ainsi, de l'exemple suivant, seront extraits quatre SN, le premier instanciant *X*, le SN glosé : *Le Maxi 45 Tours* et les trois autres sont les SN glosants instanciant les $Y_{i,}$, respectivement : *Maxi 45, Maxi* et *Super 45 tours*.

Le Maxi 45 Tours (aussi appelé Maxi 45 ou même tout simplement Maxi ou encore Super 45 tours) est un format de disque microsillon (ou disque vinyle) très apprécié des disc-jockeys et des collectionneurs.

La section suivante décrit une fonction de rang qui classe les syntagmes Y_i extraits. Cette fonction s'appuie sur des approches de fouille du web pour élaborer un classement de pertinence parmi les SN extraits.

3.2.2 Classement des candidats par des méthodes de fouille du web

Notre méthode est inspirée de (Turney, 2001) qui interroge le Web via le moteur de recherche AltaVista pour identifier des synonymes : étant donné un *mot*, le principe de l'algorithme PMI-IR (Pointwise Mutual Information and Information Retrieval) est de choisir un synonyme de *mot* parmi une liste. Les candidats, notés $choix_i$, correspondent aux questions du TOEFL. Le but est de calculer, pour chaque *mot*, le synonyme $choix_i$ qui donne le meilleur score. Pour calculer le score, l'algorithme PMI-IR utilise différentes mesures fondées sur le nombre de documents dans lesquels les deux termes sont présents (sur une même page ou dans une fenêtre donnée). Ces mesures s'appuient essentiellement sur l'Information Mutuelle (Church and Hanks, 1990). Mais contrairement à la méthode de P. Turney qui utilise l'Information Mutuelle, nous utilisons la mesure Dice (Smadja *et al.*, 1996) dont nos précédents travaux (Roche and Kodratoff, 2009) ont montré le bon comportement dans un contexte de fouille du web.

La mesure de Dice calcule la « dépendance » entre deux SN, la dépendance s'entendant en termes de co-occurrence plus ou moins proche. Appliquée à *X,* le SN glosé et Y_i, le SN candidat glosant, la mesure est définie par la formule (1) :

$$Dice1(X, Y_i) = 2.\ |X \cap Y_i| \ /\ (|X| + |Y_i|) \qquad (1)$$

où $|X \cap Y_i|$ est le nombre de pages web contenant les termes *X* et Y_i l'un à côté de l'autre, $|X|$, le nombre de pages contenant le terme *X* et $|Y_i|$, le nombre de pages contenant Y_i.

Pour calculer $|X \cap Y_i|$, nous utilisons des guillemets dans nos interrogations du Web. Par exemple, si X = *Toulouse* et Y_i = *Ville rose*, les requêtes "Toulouse Ville rose" et "Ville rose Toulouse" sont soumises. La requête "Toulouse Ville rose" ramène aussi bien les occurrences de « Toulouse : Ville rose » que celles de « Toulouse, Ville rose » et celles de « Toulouse (Ville rose) » car généralement les moteurs de recherche ne prennent pas en compte les caractères tels que parenthèses et/ou virgule.

Par la suite, nous « relâchons » les contraintes de proximité entre syntagmes afin de calculer le nombre de pages web où X et Y_i sont présents sur une même page. Dans ce cas, nous appliquons une mesure de Dice appelée Dice2. Le numérateur de cette mesure *(formule (2))*, représente le nombre de fois où X et Y_i apparaissent dans une même page.

$$Dice2(X, Y_i) = 2.\ |X\ AND\ Y_i| \ /\ (|X| + |Y_i|) \qquad (2)$$

Partant de ces deux mesures *Dice1* et *Dice2*, deux types de combinaisons sont proposées : La première, *Diexact,* est la mesure *Dice1* si *Dice1* retourne un résultat, sinon elle vaut *Dice2*. Cette approche privilégie donc *Dice1*. La seconde, *Dibary (formule (3))*, calcule les barycentres de *Dice1 et Dice2* :

$$Dibary_k(X,Y_i) = k\ .\ Dice1(X,Y_i) + (1-k)\ .\ Dice2(X,Y_i) \text{ où } k \in [0,1] \qquad (3)$$

Selon cette approche, lorsque $k > 0.5$ (resp. $k < 0.5$), l'influence de la mesure *Dice1* est plus importante (resp. moins importante) par rapport à *Dice2*. L'utilisateur peut alors affecter les valeurs de k les plus pertinentes au regard des données traitées. Les différentes valeurs de ce paramètre seront discutées dans la section 5 de cet article.

Ces mesures de *popularité* utilisant le Web sont particulièrement bien adaptées pour filtrer les SN et leur association retrouvés fréquemment sur le web. Elles sont donc particulièrement bien adaptées pour traiter des données issues d'un domaine général utilisant un vocabulaire souvent présent sur le Web.

Le logiciel qui a été développé est présenté dans la section suivante. Il utilise différents moteurs de recherche (Google, Yahoo et Exalead) pour calculer les mesures de Dice. Les expérimentations à partir d'un échantillon de 22 textes contenant des gloses ayant montré que l'API (Application Programming Interface) *proposée par Yahoo avait un excellent comportement, nous avons choisi ce m*oteur de recherche pour la suite de nos expérimentations.

Extraction des définitions

Outre les couples (X,Y_i) en relation de glose de dénomination, nous extrayons les définitions de mots glosés qui se trouvent dans la même phrase que la glose. Notre méthode s'appuie sur des marqueurs simples tels que les verbes définitoires (par exemple, la copule « est ») pour identifier le début d'une définition et une ponctuation forte comme frontière droite. Dans l'exemple de la section 3.2.1, et rapporté ci-dessous :

Le Maxi 45 Tours (aussi appelé Maxi 45 ou même tout simplement Maxi ou encore Super 45 tours) est un format de disque microsillon (ou disque vinyle) très apprécié des disc-jockeys et des collectionneurs.

La définition du mot glosé *le maxi 45 Tours* est *le Maxi 45 Tours est un format de disque microsillon (ou disque vinyle) très apprécié des disc-jockeys et des collectionneurs.* Des règles plus fines de détection de frontière droite de définition seront proposées dans de futurs travaux.

4 Logiciel et interface graphique

Les patrons de SN sont pris en compte dans un « fichier paramètre » de notre logiciel. Une fois les SN extraits, les calculs décrits dans cet article et implantés en Java sont effectués. La figure 1 montre les résultats obtenus après exécution des différentes phases :

- partie de haut/gauche : visualisation du corpus,
- partie de haut/droite : visualisation du corpus étiqueté,
- partie de bas/gauche : visualisation des syntagmes extraits,
- partie de bas/droite : visualisation des syntagmes Y_i classés par les mesures statistiques (avec les valeurs associées)

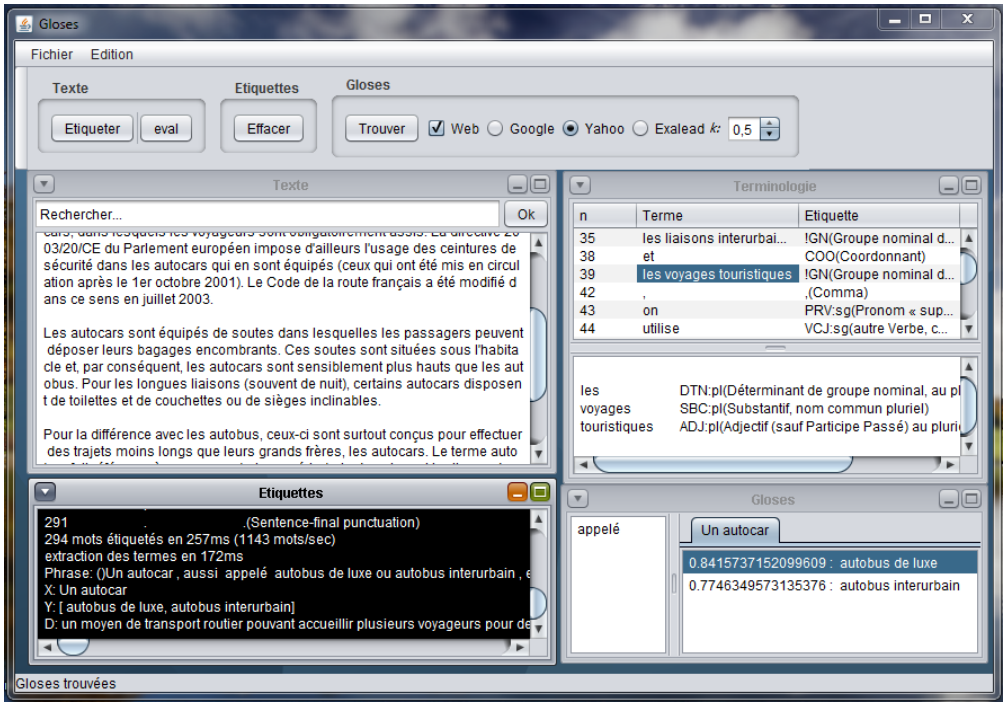

FIG. 1 – *Capture d'écran du logiciel de recherche de gloses.*

5 Expérimentations

5.1 Protocole expérimental

Le corpus que nous utilisons est constitué de 84 phrases contenant des gloses en « *appelé* » et nous permet d'extraire 219 SN candidats Y_i. Des informations concernant la qualité des SN extraits sont données plus bas. Les couples (X, Y_i) jugés pertinents ont la note 1 et les couples très pertinents, la note 2. Lorsqu'un SN n'est extrait que partiellement, le couple (X, Y_i) correspondant est évalué à 1. Les SN non pertinents sont évalués à 0. L'ensemble des résultats est détaillé dans la section suivante.

Les mesures de précision et rappel utilisées pour estimer la qualité de nos différentes approches sur la base des évaluations faites manuellement sont les suivantes :

Précision = nombre de SN extraits pertinents / nombre de SN extraits

Rappel = nombre de SN extraits pertinents / nombre de SN pertinents

Une précision de 100% signifie que tous les syntagmes extraits sont pertinents. Un rappel de 100% signifie que tous les syntagmes pertinents ont été extraits. Par ailleurs, une combinaison de ces deux critères d'évaluation est classiquement appliquée selon la formule de la F-mesure rappelée ci-dessous :

$$F\text{-}mesure(\beta) = (\beta^2+1).(Précision \,.\, Rappel) \,/\, (\beta^2 \cdot Précision + Rappel)$$

Le paramètre β permet de régler les influences de la précision et du rappel. Il est très souvent fixé à 1 (ce qui est le cas dans nos expérimentations). Dans ce cas, le même poids est attribué à ces deux mesures d'évaluation.

La section suivante décrit les résultats obtenus.

5.2 Résultats

5.2.1 Comparaison des méthodes

L'évaluation de la qualité des SN instanciant l'élément glosé *X* est la suivante : 6 SN *X* sont évalués à 1 (pertinent), 68 à 2 (très pertinent). Comparons à présent les méthodes d'extraction des SN glosants, notés Y_i. Le tableau 1 donne les résultats obtenus en terme de pertinence (évaluation à 0, 1 ou 2) à partir des calculs décrits en section 3.

Méthodes / Évaluation	**Nb de Y_i extraits évalués à 1**	**Nb de Y_i extraits évalués à 2**	**Nb de Y_i extraits évalués à 0**
Patron de glose	3 (3,75 %)	74 (**92,5 %**)	3 (3,75 %)
Patron et coordination des Y_i	4 (3,67 %)	101 (**92,66 %**)	4 (3,67 %)
extraction étendue	16 (7,30 %)	150 (**68,49 %**)	53 (24,21 %)

TAB. 1 – *Evaluation des méthodes d'extraction de gloses.*

Les résultats montrent que l'évaluation à 2 (très pertinent) est majoritaire. L'extraction étendue des Y_i produit un quart de syntagmes non pertinents ou partiellement pertinents. Cela correspond aux faits que les SN suivant *appelé* ne sont pas forcément en relation de glose avec *X* ou que le SN pertinent n'est pas toujours maximal. Ces situations sont illustrées par les extraits de corpus suivants où les SN non pertinents ou maximalisés à tort ont été soulignés :

Il faut préciser que le pouvoir n'a fait que louvoyer durant toute l'année en accordant le dialogue à des faux représentants appelés par dérision, "délégués taïwan".

Un cône volcanique, également appelé cône de cendres ou encore cône de scories en fonction des matériaux qui le composent, est la structure d'origine volcanique en forme de montagne ou de colline formée par l'empilement de téphras au cours d'une seule éruption (volcan monogénique) ou de plusieurs éruptions (stratovolcan).

Néanmoins, globalement, notre méthode permet d'extraire un nombre bien plus important de SN pertinents que le seul patron. En considérant les évaluations 1 et 2 comme pertinentes, la précision, le rappel et la F-mesure sont reportés dans le tableau 2.

Méthodes / Mesures d'évaluation	**Précision**	**Rappel**	**F-mesure**
Patron de glose	**0.96**	0.46	0.62
Patron et coordination des Y_i	**0.96**	0.63	0.76
Extraction étendue	0.75	**1**	**0.86**

TAB. 2 – *Précision, Rappel, F-mesure suivant les méthodes d'extraction de gloses*

Ce tableau montre que la meilleure F-mesure est obtenue par l'extraction étendue des syntagmes (cf. section 3.2). Ceci est dû au rappel élevé malgré une précision assez faible comparativement aux approches par patrons. A contrario, ces dernières donnent une excellente précision mais un rappel assez faible. Ceci signifie que peu de SN glosants sont extraits mais ceux-ci sont d'excellente qualité. Remarquons que l'utilisation de règles de coordination se révèle assez efficace (excellente précision et rappel honorable).

Enfin, nous avons évalué le nombre de *définitions* pertinentes obtenues. Nous obtenons 40 définitions dont 2 sont évaluées à 1 (pertinentes) et 38 à 2 (très pertinentes). Ceci confirme l'intérêt de notre approche.

5.2.2 Évaluation de *Diexact* et *Dibary*

Évaluons à présent la qualité des classements de SN obtenus par fouille du Web.

Prenons la phrase suivante :

Le Notre Père (aussi appelé par son nom latin, Pater Noster ou, par déformation phonétique, "patenôtre").

L'approche par patron extrait uniquement le SN *patenôtre* alors que notre méthode d'extraction étendue donne quatre SN candidats : *son nom latin*, *Pater Noster*, *déformation phonétique*, *patenôtre*. Les deux mesures *Diexact* et *Dibary* classent différemment ces SN, comme le montre le tableau ci-dessous :

Dibary			*Diexact*
k = 0 ...	k = 0.5	… k = 1	
...	*Pater Noster* *patenôtre* *son nom latin* *déformation phonétique*	...	*Pater Noster* *son nom latin* *déformation phonétique* *patenôtre*

TAB. 3 – *Exemples de classements selon les mesures.*

L'expert a considéré les SN *Pater Noster* et *patenôtre* pertinents puisqu'en relation de glose avec l'élément glosé *X Le Notre Père,* et les SN *son nom latin* et *déformation phonétique* non pertinents.

Pour comparer de manière globale nos algorithmes, nous allons calculer la somme des rangs des SN glosants considérés pertinents par l'expert. La méthode qui donne les meilleurs résultats obtient la somme la plus faible. Cette méthode qui permet d'évaluer les fonctions de rang possède un comportement similaire aux approches fondées sur les courbes ROC (Receiver Operating Characteristics) et le calcul de l'aire sous ces dernières (Roche et Kodratoff, 2006).

Dans notre exemple, la mesure *Dibary* (avec *k=0.5*) peut alors être désignée comme la mesure la plus adaptée. En effet, les sommes des SN glosants pertinents de *Dibary* et *Diexact* sont respectivement 3 (1+2) et 5 (1+4).

La somme des rangs des SN glosants pertinents obtenus à partir de notre corpus d'évaluation composé de 84 phrases, est reportée dans le tableau ci-dessous pour chaque méthode et chaque paramètre *k* de la mesure *Dibary*. Les résultats montrent que la méthode qui globalement obtient le meilleur résultat est *Diexact.* Nous pouvons également remarquer que l'influence de *k* pour *Dibary* est faible. Notons que pour *k=1* avec la méthode *Dibary* seul *Dice1* est pris en compte. Ainsi, les syntagmes non co-occurents sur le web ont un score nul et sont donc ordonnés aléatoirement. Ceci confirme que notre classement prenant en compte les deux mesures *Dice1* et *Dice2* se révèle bien sûr plus intéressant qu'un classement partiellement aléatoire.

Méthode	***Diexact***	***Dibary***					
		k=0	*k=0.2*	*k=0.4*	*k=0.6*	*k=0.8*	*k=1*
Somme	**323**	329	329	329	329	329	364

6 Conclusion et perspectives

Nous venons de présenter une méthode pour extraire des SN en relation de glose en *appelé*. Ces SN sont différents noms de la même entité. Cette méthode mixe patrons et statistiques d'associations d'unités sur le web, en les évaluant sur des données réelles.

Dans nos futurs travaux, nous souhaitons mener une analyse contrastive anglais/français sur corpus afin de donner un nouvel éclairage du phénomène des descriptions spontanées de sens et cerner sa pertinence interlinguistique. Une première prospection sur des corpus alignés français-anglais révèle des correspondances assez régulières entre gloses alignées, certaines gloses marquées lexicalement pouvant être traduites par des gloses sans marqueur lexical et vice-versa. Grâce à l'alignement, le repérage des gloses d'un corpus pourrait servir à pointer les gloses non marquées dans le corpus aligné correspondant et contribuer à améliorer l'acquisition lexicale multilingue de mots nouveaux et de leurs traductions.

Références

Bourigault, D. et C. Jacquemin (2000). Construction de ressources terminologiques, pp. 215-233, dans *Ingéniérie des langues*. Ed. Pierrel, J-M., Hermès Sciences.

Brill, E. (1994). Some Advances in Transformation-Based Part of Speech Tagging. *Proceedings of AAAI,* Vol. 1, p. 722-727.

Church, K.W. et P. Hanks (1990). Word Association Norms, Mutual Information, and Lexicography, *Computational Linguistics*. Vol 16, p22-29

Daille, B. (1994). *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques.* Thèse de doctorat - Université Paris 7

Habert, B., A. Nazarenko., A. Salem (1997). *Les linguistiques de corpus*, Armand Colin, Paris

Hearst, M., (1992). Automatic Acquisition of hyponyms from large text corpora. *Actes de Coling-92*, Nantes, pp. 539-545.

Langages n° 99 (1990). Les grammaires de Harris et leurs questions, Anne Daladier et al.

Mela, A. (2005). Le repérage automatique des gloses de nomination seconde, *Langues et langage*, "Les marqueurs de la glose", A. Steuckardt (resp.), Publications de l'université de Provence.

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.), 25p.

Mela A. et M. Roche (2006). Des gloses de mot aux types de textes : un bilan différencié. Dans *les actes du colloque "Corpus en Lettres et Sciences sociales : des documents numériques à l'interprétation", Albi, juillet 2006. Texto! [en ligne]* vol. XI, n°2

Morin, E. (1999). *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques*. Doctorat en Informatique, Université de Nantes.

Nazarenko, A. et T. Hamon (2002). Structuration de terminologie : quels outils pour quelles pratiques ? *TAL*. Vol. 43, n° 1/2002, pp. 7-18

Rebeyrolles, J. (2000) : *Forme et fonction de la définition en discours*. Doctorat en Sciences du langage, Université de Toulouse-le-Mirail, Toulouse II.

Roche, M. et Y. Kodratoff (2009). Text and Web Mining Approaches in Order to Build Specialized Ontologies. In *Journal of Digital Information (JoDI)*, Vol 10

Roche, M. et Y. Kodratoff (2006). Pruning Terminology Extracted from a Specialized Corpus for CV Ontology Acquisition. In *Proceedings of onToContent'06 workshop (Ontology content and evaluation in Enterprise) - OTM'06*, LNCS, p1107-1116

Salton G., A. Wong, C. S. Yang. (1975). A vector space model for automatic indexing. *ACM*, Vol. 18, No.11, pp. 613-620.

Smadja, F., K. R. McKeown, V. Hatzivassiloglou (1996). Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, Vol 22, num 1, p. 1-38,

Steuckardt, A. (2006). *Du discours au lexique : la glose*, Séminaires de l'ATILF, disponible sur : <http://www.atilf.fr/atilf/seminaires/historique.h tm#Steuckardt_2006-03 >.

Turney, P.D. (2001). Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. *Proceedings of the 12th European Conference on Machine Learning (ECML)*, LNCS, p.491-502

Véronis, J., (2000). Alignement de corpus multilingues, pp. 151-171, dans *Ingéniérie des langues*. Ed. Pierrel, J-M., Hermès Sciences.

Summary

We propose to extract lexical knowledge from text by exploiting the "glosses" of words, spontaneous descriptions of meaning, identifiable by lexical markers and specific morpho-syntactic patterns, as in the following passage where the word "testing" is followed by the marker "*c'est-à-dire*" : *"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"*. This approach offers new research directions for acquiring lexical items and terminology, fundamental for many tasks. In this paper, we compare two ways of extracting the "glose relationships", using local grammars or statistical associations of units on the web, and we evaluate them on real data.