Comment mesurer la couverture d'une ressource terminologique pour un corpus ?

Goritsa Ninova (1), Adeline Nazarenko (2),
Thierry Hamon (2), Sylvie Szulman (2)

LIPN UMR 7030
Université Paris 13 & CNRS
99, av. J.-B. Clément
93430 Villetaneuse
(1) cylvago@yahoo.fr
(2){prénom.nom}@lipn.univ-paris13.fr

Mots-clés : couverture lexicale, terminologie, statistique lexicale
Keywords : lexical coverage, terminology, lexical statistics

Résumé Cet article propose une définition formelle de la notion de couverture lexicale. Celle-ci repose sur un ensemble de quatre métriques qui donnent une vue globale de l'adéquation d'une ressource lexicale à un corpus et permettent ainsi de guider le choix d'une ressource en fonction d'un corpus donné. Les métriques proposées sont testées dans le contexte de l'analyse de corpus spécialisés en génomique : 5 terminologies différentes sont confrontées à 4 corpus. La combinaison des valeurs obtenues permet de discerner différents types de relations entre ressources et corpus.

Abstract This paper proposes a formal definition of the notion of lexical coverage. This definition is based on four metrics that give a global view over a lexical resource to corpus relationship, thus helping the choice of a relevant resource with respect to a given corpus. These metrics have been experimented in the context of specialised corpus analysis in genomics. 5 terminologies have been confronted to 4 different corpora. The combination of resulting figures reflects various types of corpus *vs* . resource relationships.

1 Introduction

On parle couramment de « couverture lexicale » sans définir clairement ce qu'on entend par là. Différents auteurs mettent sous ce terme différentes notions et mesures. Le problème est d'autant plus complexe que les ressources utilisées comportent souvent des expressions polylexicales dont la projection en corpus peut se faire de différentes manières. Le présent article propose de définir un ensemble de métriques pour cerner cette notion de couverture dans le cas général d'une ressource constituée d'une liste de termes mono- et polylexicaux. Ces mesures sont testées

pour différents couples ressource/corpus. Les premiers résultats obtenus sont encourageants. Ils montrent qu'on peut en effet documenter le comportement d'une ressource pour un corpus donné en préalable à tout traitement, et ainsi guider le choix de la ressource.

Après avoir souligné les enjeux de cette problématique et les questions qu'elle soulève (section 2), nous présentons dans la section 3 un ensemble de métriques. Celles-ci sont exploitées dans la perspective du traitement automatique de corpus de génomique. Les résultats de ces expériences sont présentés et discutés dans la section 4 de cet article.

2 Problématique

2.1 Enjeux

Le traitement de corpus spécialisé fait appel à des ressources sémantiques qu'on appelle généralement spécialisées parce qu'elles décrivent un domaine particulier d'activité. Ces ressources peuvent être de différents types selon les traitements envisagés, mais elles doivent comporter une dimension lexicale dès lors qu'elles sont destinées à l'analyse et l'interprétation de données textuelles.

Les ontologies du web sémantique doivent ainsi être ancrées lexicalement (avec des items lexicaux associés aux noeuds de l'ontologie) si elles doivent servir à indexer des textes. Les techniques d'accès au contenu des documents textuels sont diverses (extraction d'information, question-réponse, outils de navigation ou de résumé) mais elles reposent toutes sur une analyse sémantique partielle des documents, qui implique la reconnaissance de certains éléments du discours (entités nommées et termes du domaine, notamment), leur typage sémantique et leur mise en relation (Nazarenko, 2005). De ce fait, ces techniques reposent également sur des lexiques, terminologies ou thesaurus spécialisés pour identifier le vocabulaire de spécialité. Les catégories sémantiques et les relations lexicales sont utilisées (quand elles existent) pour désambiguïser les textes et en guider l'interprétation.

Dès lors que les applications de traitement automatique des langues (TAL), y compris au niveau sémantique, sont de plus en plus guidées par le lexique, la question du choix des ressources à exploiter prend de l'importance. La situation relève souvent à la fois de la pléthore et de la pénurie. D'un côté, il existe de nombreuses ressources terminologiques, surtout dans des domaines comme la biologie ou la médecine où l'effort d'organisation des connaissances est ancien[1]. Mais d'un autre côté, les « bonnes ressources » sont rares : le degré de spécialisation ou le point de vue représenté par la ressource est généralement différent de celui du texte que l'on cherche à analyser. Ce constat a été fait par (Charlet *et al.*, 1996), toujours dans le domaine de la médecine pourtant reconnu pour la richesse de ses bases de connaissances. Dans la pratique, comme on ne peut ni se passer de ressource, ni en reconstruire de nouvelles pour chaque nouvelle application, on fait souvent avec ce qu'on a. Dans certains cas, on peut spécialiser la ressource et l'adapter en fonction du domaine et de la tâche visés (problématique de l'adaptation lexicale ou « lexical tuning » (Basili *et al.*, 1998)) mais cela suppose néanmoins une ressource initiale.

Une question se pose alors : parmi l'ensemble des ressources qui paraissent recouvrir en partie et *a priori* le domaine du corpus à traiter, laquelle ou lesquelles choisir et sur quels critères ? Cette question est d'autant plus importante qu'on doit souvent limiter le nombre de ces ressources pour réduire l'inévitable travail de préparation des données et pour éviter les problèmes d'incohérence. Il est en général trop coûteux d'exploiter en parallèle différentes ressources pour les tester en les comparant au regard de l'application visée. Les experts du domaine ne sont pas toujours d'un grand secours non plus. Même s'ils sont capables de décrire

---

[1] Voir par exemple, UMLS (Unified Medical Language System, http://www.nlm.nih.gov/research/umls/).

le sous-domaine couvert par une ressource et le point de vue qui y est représenté, ils sont d'ordinaire peu à même de mesurer son adéquation proprement lexicale.

Ce problème du choix des ressources est souvent résolu de manière très empirique, ce qui ne permet pas de capitaliser d'une expérience à l'autre. Il est donc important de se doter de critères formels permettant de décrire le comportement d'une ressource par rapport à un corpus donné et d'en guider le choix. C'est l'objet de ce travail : nous proposons un premier ensemble de métriques pour apprécier la couverture d'un corpus par une ressource terminologique.

2.2 Difficultés

Pour des dictionnaires traditionnels, on exprime l'adéquation à un corpus en termes de couverture et on l'apprécie à partir du nombre d'occurrences de mots du corpus qui se rattachent à des entrées du dictionnaire. La couverture est plus difficile à définir pour des ressources terminologiques.

La première difficulté tient à la diversité des ressources terminologiques qui rend problématique leur comparaison. La *nature* de l'information diffère d'une ressource à l'autre. Au-delà des listes de termes, les termes eux-mêmes peuvent être typés et les types peuvent être organisés en hiérarchie (thesaurus). Dans les ressources les plus riches, les termes sont de surcroît liés entre eux par des liens sémantiques. Les ressources ont par ailleurs des *degrés de spécialisation* divers. Il est difficile de comparer un lexique de 10 000 unités qui comporterait de nombreuses unités également présentes dans des dictionnaires généralistes et un lexique de 500 unités dont très peu figurent dans des dictionnaires classiques. Les ressources s'opposent enfin par leur *degré de lexicalisation* : certaines se contentent de lister des étiquettes de concepts ; d'autres considèrent ces étiquettes dans leur dimension lexicale et linguistique. Ces dernières rendent compte des différentes formes sous lesquelles ces unités sémantiques peuvent se réaliser en corpus, jusqu'à associer des règles de désambiguïsation contextuelles aux unités polysémiques (Nédellec, Nazarenko, 2005).

La deuxième difficulté tient au fait qu'on cherche à confronter deux objets qui ne sont pas de même nature. La ressource et le corpus s'opposent comme la langue s'oppose au discours : il faut comparer un ensemble d'éléments de lexique (la ressource) avec un ensemble d'occurrences (le corpus). Par voie de conséquence, il faut aussi comparer des unités potentiellement polylexicales avec des occurrences observées en corpus, nécessairement monolexicales. Comme il s'agit d'apprécier *a priori* l'adéquation des ressources aux corpus, nous ne présupposons en effet aucune étape de reconnaissance terminologique préalable.

Dans ce premier travail, nous focalisons l'étude sur les ressources terminologiques considérées comme des listes de termes, sans exploiter les éventuelles informations qu'elles contiennent concernant leurs règles de variation, leur typage sémantique ou les relations sémantiques qu'ils entretiennent. Les premiers éléments étant posés, il est évidemment nécessaire de poursuivre, par exemple, en prenant en compte la désambiguïsation des termes polysémiques, les liens de variations entre termes et la structure sémantique. Ces points ne sont pas abordés ici.

2.3 État de l'art

La question de la sélection des ontologies pour une application prend de l'importance avec l'augmentation du nombre des ontologies disponibles et la standardisation de leurs formats. Cette préoccupation est au coeur de la problématique du web sémantique. (Buitelaar *et al.*, 2004) montre que la création d'une bibliothèque d'ontologies (OntoSelect) suppose de définir des critères permettant de sélectionner une ontologie particulière. Trois critères sont proposés : les degrés de structuration et de connectivité sont des mesures proprement ontologiques, mais le critère de couverture est établi relativement à une collection de documents. Ce dernier critère est

cependant défini de manière assez fruste[2] : il ne prend qu'imparfaitement en compte la dimension proprement linguistique des « étiquettes de concepts ».

Brewster *et al*. (2004) vont plus loin. Ils proposent d'évaluer les ontologies relativement à un corpus donné. La notion de couverture qu'ils proposent est plus riche que la précédente. Elle repose sur le nombre de termes en corpus qui correspondent à des concepts de l'ontologie, une fois effectués un calcul de variation pour reconnaître des formes de termes non canoniques et une expansion sémantique pour autoriser une adéquation à différents niveaux de généralité. À partir de là, une ontologie est évaluée en fonction du nombre de concepts qui trouvent leur contrepartie en corpus. Ce deuxième travail prend davantage en compte la nature linguistique des réalisations lexicales des concepts en corpus (notion de variation, quasisynonymie entre un hyperonyme et son hyponyme) mais il est centré sur l'évaluation et la cohérence interne d'une ontologie alors que notre objectif est plutôt de guider le choix d'une ressource pour un corpus donné, ce qui confère un autre rôle à la notion de couverture et impose de la définir plus précisément.

Sur le plan lexical, la question de la couverture n'a guère été étudiée[3]. De manière intuitive, on tend à préférer des ressources de grande taille (en nombre d'entrées), avec l'idée qu'elles sont soit plus complètes sur un domaine restreint soit plus génériques et moins liées à un domaine particulier. (Nirenburg *et al*., 1996) critique ce présupposé en soulignant que la taille de la ressource donne une vue très partielle de sa couverture. Dans ce travail, les auteurs cherchent cependant à apprécier la qualité intrinsèque de la ressource alors que nous défendons l'idée qu'une ressource n'a pas de valeur propre et qu'elle n'a de valeur que par les utilisations qui peuvent en être faites. Au total, la question du choix de la ressource étant donné un corpus a moins retenu l'attention que la question ultérieure : une fois cette ressource choisie, comment l'adapter à ce corpus (Basili *et al*., 1998) ?

La statistique lexicale a souligné depuis ses débuts (Muller, 1977 ; Manning, Schütze, 1999) qu'il existe une relation fonctionnelle entre une ressource (un vocabulaire) et un corpus mais elle n'a pas abordé le problème des unités polylexicales que contiennent les terminologies.

3 Proposition de métriques

Afin d'apprécier l'adéquation d'une ressource à un corpus, nous proposons différentes mesures. Il s'agit de caractériser la couverture de la ressource ainsi que son degré de spécialisation.

3.1 Remarques terminologiques

Nous posons les définitions suivantes :

- Le *texte T* du corpus est un ensemble ordonné de mots[4]. Les mots sont définis par leur forme graphique et repérés par leur position dans le texte.

---

[2] «Coverage is measured by the number of labels for classes and properties that can be matched in the document».

[3] La notion de couverture lexicale n'est pas définie dans les ouvrages de statistique linguistique (Oakes, 1998). Quand la question est abordée (Manning, Schütze, 1999, p. 130), c'est uniquement pour apprécier le nombre de mots inconnus dans un texte.

[4] La notion de « mot » est difficile à définir. Nous considérons ici comme mots les unités résultant d'une segmentation du texte, étant donné un algorithme de segmentation clairement défini. Dans les exemples présentés ici, tous les caractères d'espacement et de ponctuation sont considérés comme séparateurs de mots.

- Le *vocabulaire V* du corpus est l'ensemble des vocables, *i.e.* l'ensemble des mots différents du corpus. Les vocables sont des unités monolexicales.

- Le *lexique L* de la ressource est l'ensemble des lexies ou entrées lexicales de la ressource[5], qu'elles soient composées de un ou plusieurs mots, spécialisées ou non.

Les vocables du corpus et les lexies étant de natures différentes, on ne peut pas comparer directement le vocabulaire et le lexique. Pour établir cette comparaison, nous considérons la « partie utile » de la ressource et sa « décomposition », ainsi que leurs complémentaires, définis de la manière suivante (fig. 1) :

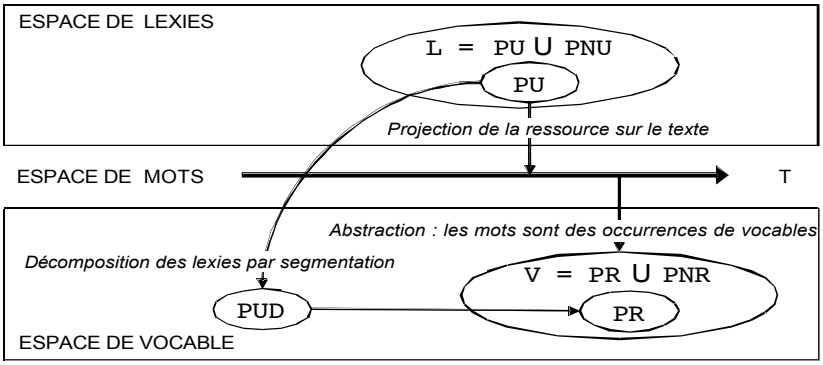

Figure 1 : Construction des ensembles de référence

- La *partie utile de la ressource PU* est l'ensemble des lexies de la ressource qui apparaissent dans le corpus. C'est un sous-ensemble de *L*.

- La *partie utile décomposée PUD* est l'ensemble de tous les vocables des lexies de *PU*. Elle est obtenue par décomposition en vocables élémentaires des lexies de *PU*. En supposant que cette décomposition est faite selon les mêmes règles qui ont permis de segmenter le corpus, cet ensemble de vocables *PUD* correspond aussi à la partie du vocabulaire du corpus qui est reconnue (*PR*) par la ressource. On a donc *PUD=PR*, où *PR* est un sous-ensemble de *V*.

- La *partie inutile de la ressource PNU* est l'ensemble des lexies qui n'ont pas d'occurrence dans le corpus. C'est le complémentaire de *PU* par rapport à *L*.

- La *partie inconnue du vocabulaire du corpus PNR* est l'ensemble des vocables de *V* non reconnus par la ressource. C'est le complémentaire de *PN* par rapport à *V*.

3.2 Mesures

Les métriques que nous proposons pour apprécier l'adéquation d'une ressource terminologique à un corpus sont définies comme des rapports entre les différents ensembles définis ci-dessus. On peut distinguer les mesures qui portent sur les formes et celles qui portent sur les occurrences.

La première mesure permet d'apprécier le degré de spécialité de la ressource par rapport à un corpus. La *contribution* (*Contr*) est la proportion de lexies du lexique qui figurent en corpus. Elle est définie par la formule ci-dessous. Nous désignons par *surplus* (*Surpl*) la proportion de lexies « inutiles ». On retrouve ici la notion d'excès de ressource introduite par (Brewster *et al.*, 2004). La contribution est forte si beaucoup des lexies de la ressource se retrouvent en corpus et donc si le domaine de spécialité de la ressource correspond bien à celui du corpus. À l'inverse,

---

[5] Comme nous l'avons souligné plus haut, nous ne considérons pas à ce stade les autres informations sémantiques apportées par la ressource.

un surplus élevé indique que la ressource est relativement générique et donc potentiellement utile pour des corpus variés. Ces mesures étant indépendantes de la taille de la ressource, on peut comparer les contributions de ressources très différentes.

$Contr = |PU| / |L|$ $Surpl = 1 - Contr = |PNU| / |L|$ où $|X|$ représente le cardinal de X

Une autre mesure permet d'apprécier dans quelle mesure la ressource « couvre » le vocabulaire du corpus. Pour avoir des ensembles comparables, il faut comparer la partie reconnue du vocabulaire et le vocabulaire dans son ensemble. Les deux mesures duales de la *reconnaissance (Rec)* et de l'*ignorance (Ign)* sont définies ci-dessous. La reconnaissance est la proportion des lexies décomposées reconnues en corpus par rapport au nombre total de vocables du corpus. La reconnaissance augmente 1) si on trouve dans le lexique les termes spécialisés employés dans le corpus mais aussi 2) quand la ressource comporte beaucoup de mots de la langue générale comme par exemple les mots grammaticaux. Seule la confrontation des différentes mesures permet de se faire une idée plus précise du comportement d'une ressource. Dans le cas 2, la forte reconnaissance tend à être associée à un surplus important. Une forte reconnaissance combinée à une contribution élevée indique une ressource spécifique bien adaptée au corpus considéré.

$$Re\,c = |PR| / |V| = |PUD| / |V| \qquad Ign = 1 - Re\,c = |PNR| / |V|$$

Parler de « couverture » évoque l'idée d'un corpus tout ou partiellement « couvert » par la ressource. La couverture est donc calculée relativement au corpus plutôt qu'à son vocabulaire. Nous définissons la *couverture* (*Couv*) comme la proportion d'occurrences de mots correspondant à des vocables entrant dans les lexies de la partie utile de la ressource. Dans la formule ci-dessous, $freq_i$ représente le nombre d'occurrences d'une lexie *i* de *PU* non incluses dans une occurrence d'une autre lexie plus large et $longueur_i$ est la longueur de la lexie en nombre de mots. Dans le cas de termes enchâssés (p. ex. *système et système de fichiers*), seule l'occurrence du terme le plus large entre dans la mesure de fréquence. Cette mesure de couverture est indépendante de la taille du corpus, ce qui rend les mesures de couverture d'une ressource comparables même sur des corpus de taille différente.

La dernière mesure complète la mesure de couverture. C'est la *densité* (*Dens*), définie par la formule ci-dessous, où f_{PUD} est la fréquence moyenne des lexies de *PU* dans le corpus et f_V est la fréquence moyenne des vocables dans le corpus. C'est une mesure normalisée de la fréquence des lexies utiles en corpus. Pour avoir une mesure indépendante de la taille du corpus, la fréquence moyenne des lexies de *PU* est pondérée par la fréquence moyenne des vocables dans le corpus.

$$Couv = \sum_{i=1}^{PU} freq_i \, x \, longueur_i / |T| \qquad Dens = f_{PUD}/f_V$$

3.3 Exemple

À titre d'exemple, considérons la ressource et le texte suivants :

- L={*système*, *système de fichiers*}
- T=« Il a réparé le système de fichiers »

On a $|L|=2$ et $|T|=7$. Dans ce cas particulier, on a $|V|=|T|=7$. Toutes les unités du lexique se retrouvant en corpus, on a par ailleurs PU=L et PUD=PR={*système, de, fichiers*}.

On obtient donc les mesures suivantes : *Contr*=1, *Rec*=3/7, *Couv*=3/7. Notons que l'occurrence de la lexie *système* qui entre dans l'occurrence plus large de la lexie *système de fichier* n'est pas comptabilisée en tant que telle dans la couverture.

4 Résultats

4.1 Protocole expérimental

Nous avons testé ces métriques dans le cadre de projets de recherche et d'extraction d'information dans le domaine de la génomique. Ce type d'application spécialisée requiert en effet d'exploiter des ressources et le choix de/des ressource(s) à exploiter s'avère souvent délicat. Nous avons considéré différents corpus de génomique et différentes ressources terminologiques *a priori* assez bien adaptées au domaine d'application (Hamon, 2005). À des fins d'évaluation, nous avons complété ces données expérimentales par un autre corpus qui porte sur les plantes carnivores et qui relève d'un domaine un peu différent. Il faudra élargir cette expérimentation en prenant en compte un autre corpus extérieur au champ de la biologie et une ressource dite de « langue générale ».

Nous avons travaillé sur quatre corpus, tous du domaine de la biologie, mais différant les uns des autres par leur style et leurs caractéristiques lexicographiques (tableau 1). Le premier corpus (*Transcript*) est constitué de 2 209 résumés d'articles scientifiques issus de la base Medline[6] à partir de la requête « Bacillus subtilis transcription ». Le second corpus (*Transcript-932* ou *932*) a été construit à partir du premier, en sélectionnant 932 phrases dans lesquelles apparaissent deux noms de gènes. Le troisième corpus (*Drosophile-1199* ou *1199-droso*) (Pillet, 2000) est similaire au second. Il s'agit de 1 199 phrases extraites des résumés de Flybase[7], qui contiennent deux noms de gènes. Le quatrième corpus regroupe différents documents issus du web se rapportant aux plantes carnivores (*Carnivore*).

	Vocabulaire V	Taille T	Fréquence moyenne
Transcript	18 720	405 423	21,66
Transcript-932	3 305	29 848	9,03
Drosophile-1199	3 232	22 691	7,02
Carnivore	27 201	273 605	10,06

Le tableau 1. Caractéristiques lexicographiques des corpus

Pour étudier la couverture des ressources terminologiques, nous avons sélectionné cinq ressources spécialisées publiquement disponibles[8] : 1) les mots clés SwissProt (keywlist) utilisés pour indexer la base de séquençage des protéines, 2) Gene Ontology (GO) qui porte sur les différents types d'organismes vivants, 3) le MeSH qui est dédié à l'indexation de la base de données Medline et rassemble des termes très divers utilisés dans le domaine médical. Nous avons également retenu deux glossaires proposant une grande variété de termes : 4) le glossaire de biochimie et de biologie moléculaire (GlossBioch) qui comporte des termes courants et 5) le glossaire de terminologie de biologie moléculaire (GoMBT).

---

[6] www.ncbi.nlm.nih.gov

[7] Flybase est une base de données structurées et bibliographiques sur la drosophile : http://flybase.bio.indiana.edu/

[8] Ces ressources sont disponibles aux adresses suivantes :
keylist : ftp://ftp.expasy.org/databases/swiss-prot/release/keywlist.txt
GO : http://www.geneontology.org/, version téléchargée en septembre 2002
MeSH : http://www.nlm.nih.gov/mesh/meshhome.html (Medical subject headings, Library of Medicine)
GlossBioch : http://www.portlandpress.com/pcs/books/prod_det.dfm?product=1855780887
GoMBT : http://www.asheducationbook.org/cgi/content/full/2002/1/490

Ressources	MeSH	GO	keywlist	GlossBioch	GoBMT
Taille en nombre de lexies	89 949	16 736	2934	836	263

Tableau 2. Tailles comparées des différentes ressources

4.2 Analyse des résultats

Les calculs des différentes métriques pour les 5 ressources et les 4 corpus ci-dessus sont synthétisés dans les graphiques des figures 2 et 3.

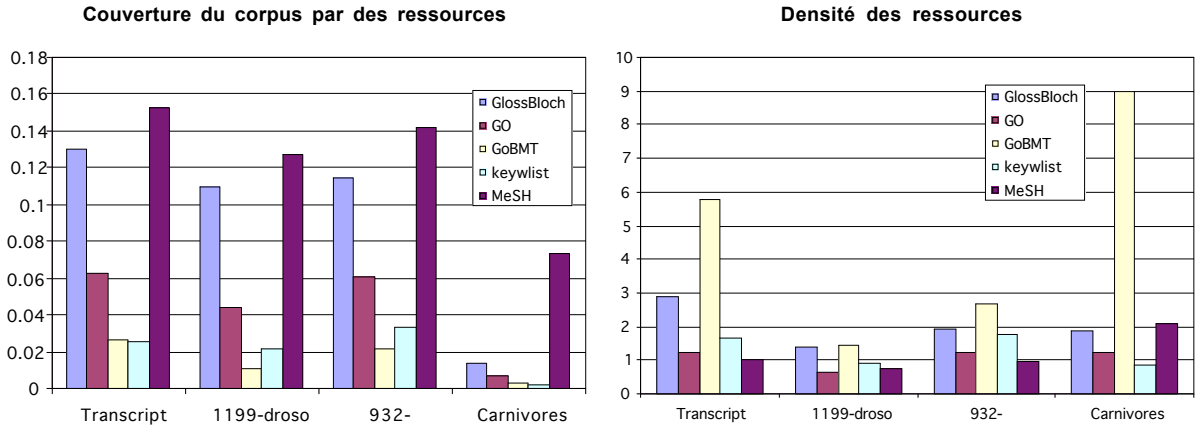

Figure 2. Mesures d'adéquation de différentes ressources à différents corpus : couverture et densité

Les mesures gomment l'effet de taille aussi bien sur les corpus que sur les ressources. Le glossaire GlossBioch a une couverture similaire à celle de MeSH qui comporte pourtant 50 fois plus de termes (fig. 2). Le comportement des ressources est comparable pour le corpus *Transcript* et son sous-corpus *Transcript*-932 (fig. 3). On peut donc envisager de sélectionner une ressource à partir d'un sous-corpus sans chercher à projeter la ressource sur l'intégralité du corpus, ce qui facilite les expérimentations.

La contribution fait exception cependant. Elle est à la fois sensible à la taille de la ressource et à celle du corpus : on remarque qu'elle est moindre pour un petit corpus (GloBioch pour *transcript-932*) et pour les ressources volumineuses (MESH pour *Transcript*). Malgré cette sensibilité aux effets de taille, c'est une mesure intéressante : une forte contribution pour une petite ressource est un bon indicateur de pertinence (*cf.* GoBMT et keywlist pour *Transcript*).

La troisième remarque concerne les deux mesures de reconnaissance et de couverture qui paraissent assez bien corrélées. On note une grande stabilité dans le sens et l'ampleur de leur écart : un corpus est d'autant mieux couvert que son vocabulaire est reconnu. C'est donc l'absence de corrélation qui est significative. Nos expériences montrent par exemple que le glossaire GlossBioch a une couverture nettement supérieure à celle de GO sur *Transcript*, pour une reconnaissance similaire. C'est le signe que GlossBioch reflète mieux la langue de spécialité du corpus *Transcript*, en dépit de sa taille modeste (fig. 3), et la preuve que la taille des ressources n'est pas un critère suffisant. Dans ce cas particulier, les mesures font apparaître un comportement des ressources contraire aux intuitions initiales des biologistes qui recommandaient à tort d'utiliser GO.

Le dernier point porte sur la densité. Elle permet d'apprécier la fréquence des lexies en corpus. De manière surprenante, la plus forte densité s'observe pour une petite ressource très spécialisée (glossaire GoBMT, fig. 2) et pour le corpus le plus différent thématiquement (*Carnivore*). Seule l'analyse détaillée des lexies de la partie utile du glossaire permet de comprendre ce résultat contre-intuitif. Moins de 10% des lexies figurent dans le corpus mais ces lexies ont de fortes fréquences. On trouve notamment *can* (676 occ.), *fish* (121 occ.) *tel* (8 occ), tous les trois décrits dans la ressource comme des noms de gènes. Ce sont des mots ambigus reconnus à tort comme

noms de gènes dans le corpus *Carnivore*. Une forte densité peut ainsi aussi bien refléter une bonne adéquation de la ressource en termes de spécialisation que des phénomènes d'ambiguïté. Une simple mesure de fréquence pondérée n'apparaît donc pas suffisamment éclairante. Il faudrait sans doute considérer le profil lexical des lexies de la partie utile de la ressource par rapport à l'ensemble des vocables du corpus pour pouvoir prédire la nature sémantique de la couverture. Ce profil devrait permettre d'apprécier la dispersion des fréquences et donc de mieux repérer des correspondances artificielles entre certains termes spécialisés et des occurrences de mots courants.

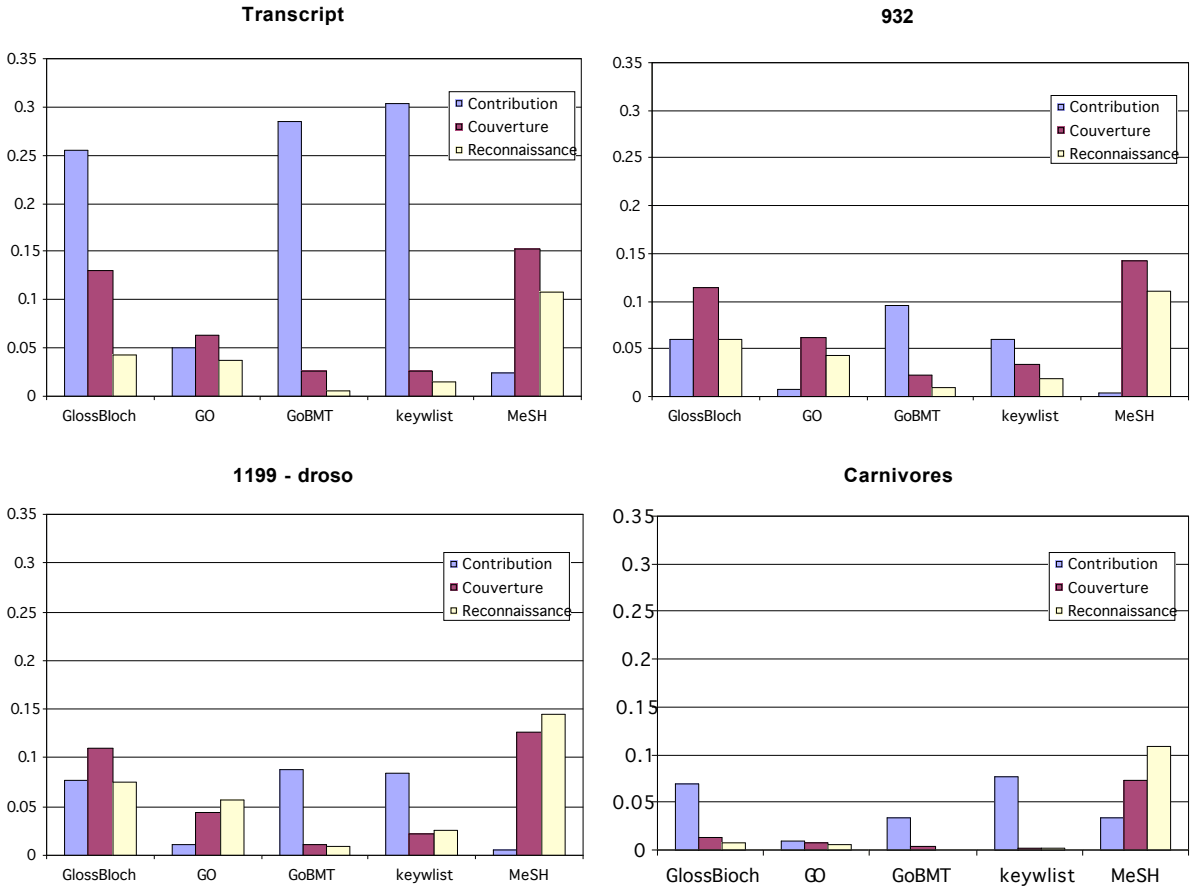

Figure 3. Mesures de contribution, couverture et reconnaissance de 5 ressources sur 4 corpus : *Transcript, Transcript-932 (932), Drosophile (1199-droso)* et *Carnivore*

5 Conclusion et perspectives

Pour permettre de caractériser avec une certaine fiabilité et une certaine reproductibilité le comportement d'une ressource lexicale pour un corpus donné, nous avons défini et testé un ensemble de métriques qui donne une idée de la « couverture », notion vague mais très couramment utilisée qui prend de l'importance avec l'augmentation du nombre de ressources disponibles. Ces métriques ne peuvent prétendre suppléer une analyse précise de l'apport d'une ressource : elles visent à éclairer le choix des ressources et des traitements à mettre en œuvre. Les expériences que nous avons menées montrent l'intérêt de ce type de métriques mais nous avons également souligné les limites des mesures proposées. Il faudrait définir une mesure de densité plus riche que nous ne l'avons fait et, pour compléter l'image globale de couverture que nous cherchons à construire, tenir compte de la répartition des occurrences des lexies de la ressource. La notion de couverture lexicale telle qu'elle est définie ici doit par ailleurs être étendue pour prendre en compte les variantes de lexies, leurs types sémantiques et même leurs relations sémantiques.

Références

BUITELAAR P., EIGNER T., DECLERCK T. (2004), OntoSelect: A Dynamic Ontology Library with Support for Ontology Selection, In *Proc. of the Demo Session at the Int. Semantic Web Conf.*, Hiroshima, Japan, Nov. 2004.

BREWSTER, C., Alani, H., DASMAHAPATRA, S. and WILKS, Y. (2004), Data Driven Ontology Evaluation. In *Proc. Of the Int. Conf. on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal.

BASILI R., PAZIENZA M.T., STVENSON M., VELARDI P., VINDIGNI M., WILKS Y. (1998), An Empirical Approach to lexical Tuning, In *Proc. of the Workshop on Adaptating Lexical and Corpus Ressources to Sublanguages and Applications (First Int. Conf. on Language Resources and Evaluation LREC 1998)*, P. VELARDI (ed.), May, Grenada.

CHARLET J., BACHIMONT B., BOUAUD J., ZWEIGENBAUM P. (1996), Ontologie et réutilisabilité : expérience et discussion, in *Acquisition et Ingénierie des Connaissance*, N. Aussenac , P. Laublet and C. Reynaud (éd.), pp. 69-87, Cépaduès-Editions, Toulouse.

HAMON H. (2005), Indexing specialized documents : are terminological resources sufficient ?, in *Actes des 6èmes journées Terminologie et Intelligence Artificielle (TIA 2005*), pp. 71-82, Rouen.

HOVY E. (2001), Comparing sets of semantic relations in ontologies. In *Semantics of Relationships*, R. GREEN, C.A. BEAN and S.H. MYAENG (eds.), chapter 6, Kluwer , Dordrecht, NL.

PILLET V. (2000), *Méthodologie d'extraction automatique d'information à partir de la littérature en science en vue d'alimenter un nouveau système d'information. Application à la génétique moléculaire pour l'extraction de données sur les interactions*. Thèse doctorat, Aix-Marseille III.

MANNING C., SCHÜTZE H. (1999). *Foundations of Statistical Natural Language Processing*, The MIT Press.

MULLER C. (1977), *Principes et méthodes de statistique lexicale*, Hachette Université, Paris.

NAZARENKO A. (2005). Sur quelle sémantique reposent les méthodes automatiquesd'accès au contenu textuel ? *Sémantique et corpus*, A. CONDAMINES (coord.), ch. 6, pp. 211-244, Hermès/Lavoisier.

NEDELLEC C., NAZARENKO A. (2005), *Ontology and Information Extraction : a necessary symbiosis*, in Ontology Learning and Population, P. Buitelaar, P. Cimiano, B. Magnini (eds), IOS (to appear).

NIRENBURG S., MAHESH K. and BEALE S. (1996), Measuring semantic coverage, in *Proc. of the 16th Conf. on Computational Linguistics (COLING'96)*, Copenhagen Denmark, ACL, pp. 83-88.