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

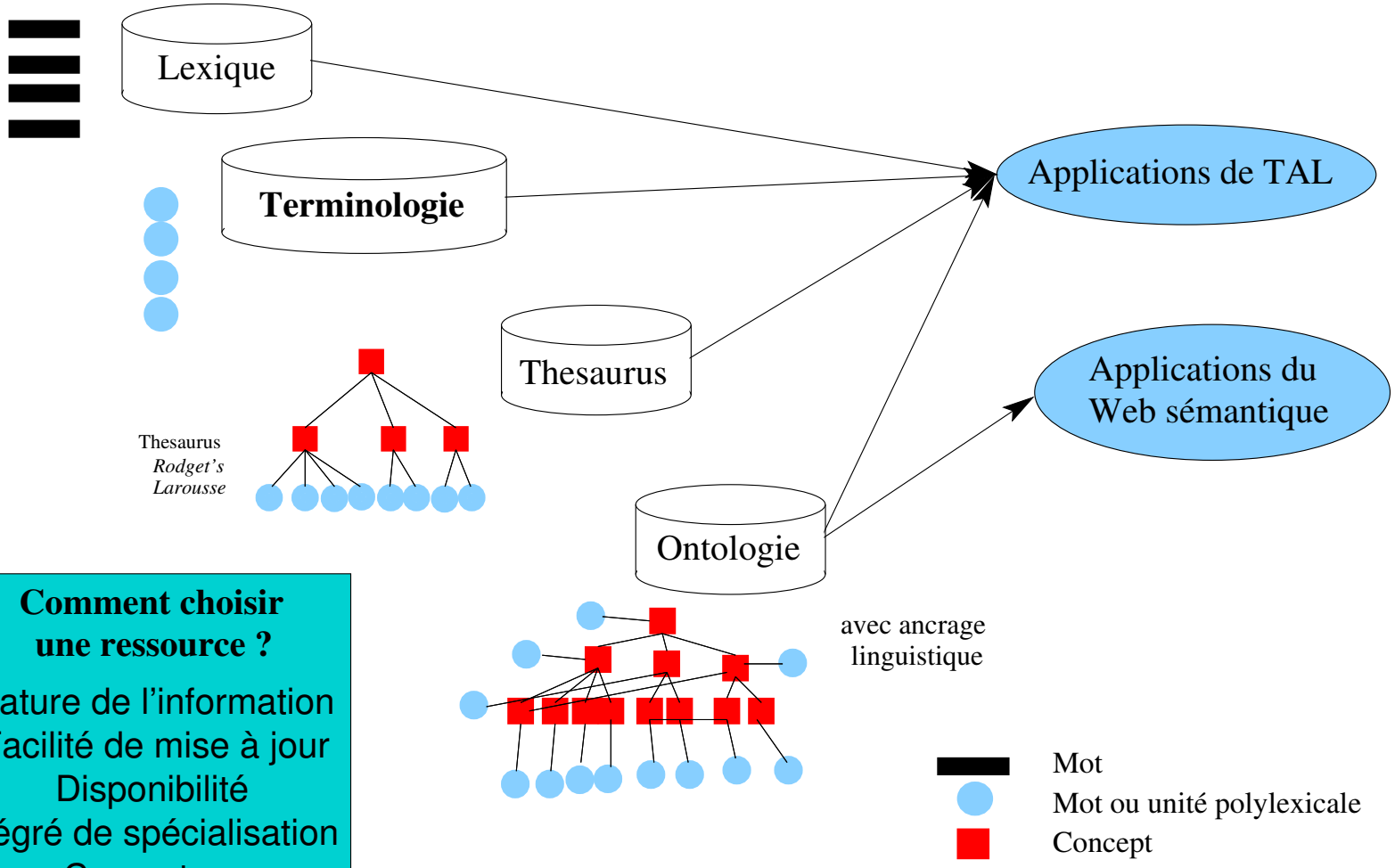
Goritsa Ninova, Adeline Nazarenko, Thierry Hamon, Sylvie Szulman

LIPN – UMR 7030
CNRS – Université Paris-Nord
99 av. J.-B. Clément, F-93430 Villetaneuse
prenom.nom@**lipn.univ-paris13.fr**

Plan

1. Contexte et enjeux
2. Proposition de métriques
3. Expériences
4. Conclusion

Le défi de la réutilisation des ressources



Comment choisir une ressource ?

Nature de l'information
Facilité de mise à jour
Disponibilité
Dégré de spécialisation
Couverture

Des mesures d'adéquation *a priori*

Les expériences étant lourdes

- Préparation des ressources ou de données qui en sont extraites
- Exploitation des ressources dans une application
- Comparaison de différents jeux de résultats

Il faut éclairer le choix des ressources à tester dans les expériences

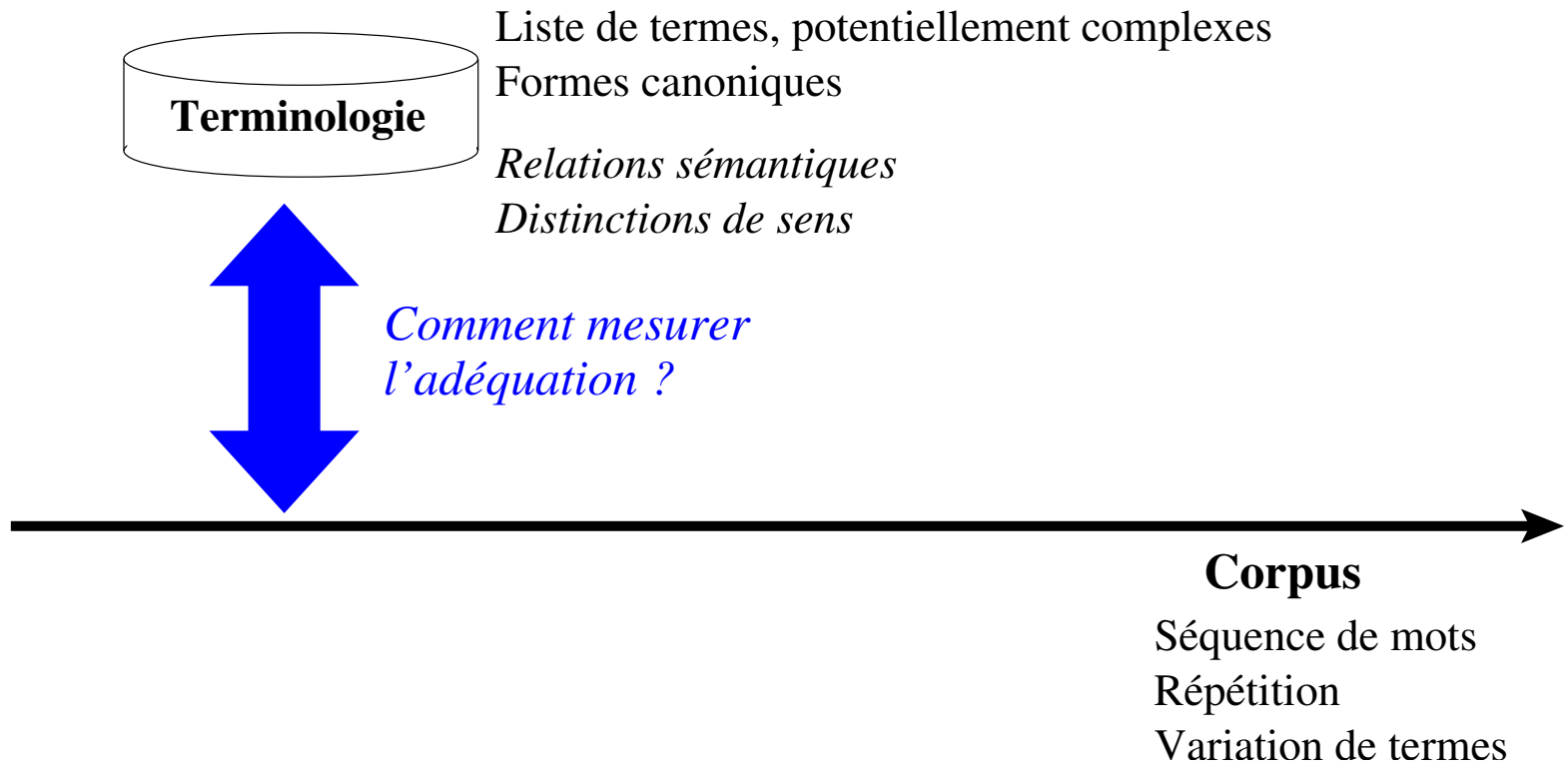
- Les données du problème
  - Le corpus à analyser
  - Un ou plusieurs ressources
- Des critères de choix multiples
  - Degré de spécialisation
  - Nombre de mots du corpus reconnus
  - Potentiel lexical

La couverture, une notion mal définie

- Couverture = Nombre de mots de la ressource présents dans le corpus
- Couverture = Nombre de termes de la ressource figurant dans le corpus

$\Longrightarrow$4 mesures complémentaires

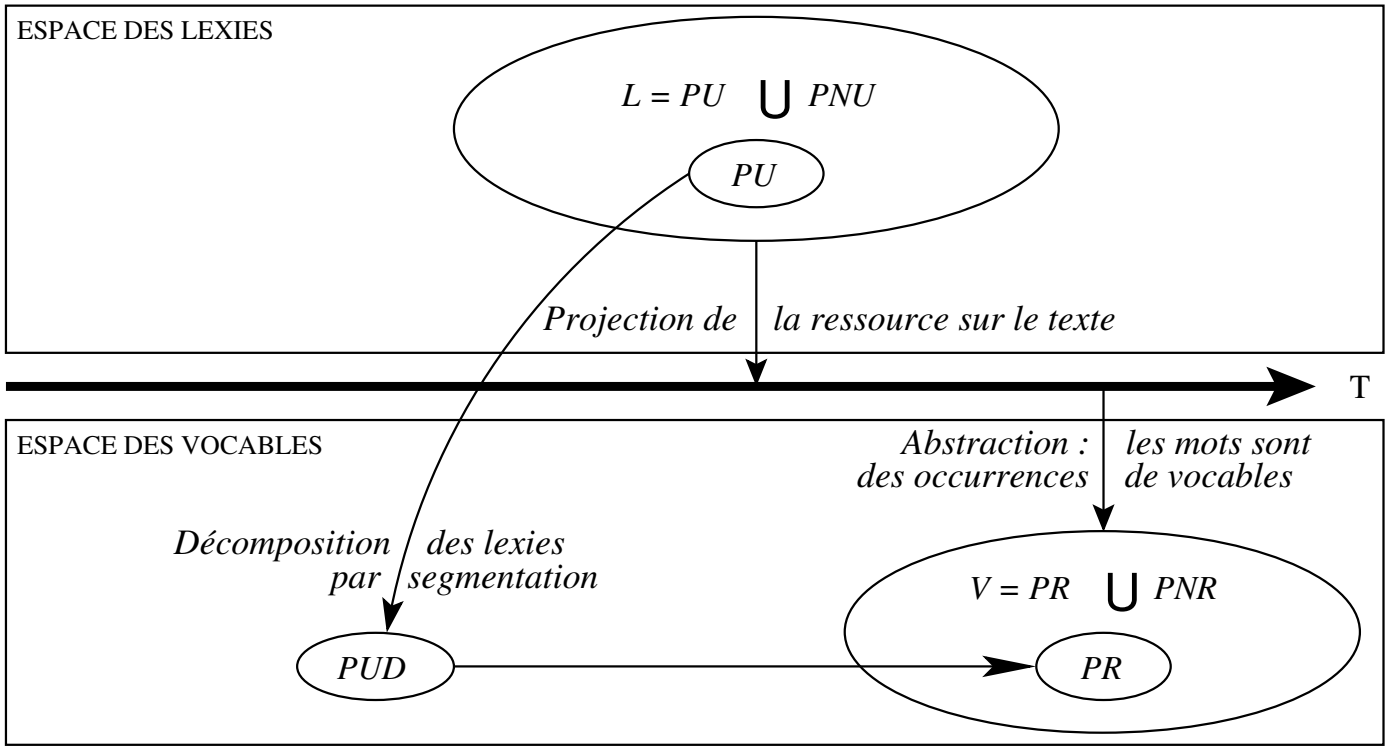
Objectif



Double difficulté

- Mettre en correspondance des mots et des termes
- Mettre en correspondance les lexies et les occurrences des vocables

Définition des ensembles



T : ensemble ordonné de mots

L : entrées lexicales de la ressource

PR : vocables du corpus entrant dans les lexies de PU

PNU : lexies qui n'ont pas d'occurrence dans le corpus

V : vocables apparaissant dans le corpus

PU : lexies de la ressource apparaissant dans le corpus

PUD : vocables des lexies de PU

Exemple

– Ressource

« *Données* », « *Base de données* », « *Logiciel de base de données* »,

« *Système de fichiers* »

– Texte

« *Il a installé un système de gestion de base de données pour gérer toutes ses données.* »

– Ensembles de travail

$|T|$ = 16

$V = \{$il, a installé, un, système, de, gestion, base, données, pour, gérer, toutes, ses$\}$

$L = \{$données, base de données, logiciel de base de données, système de fichiers$\}$

$PU = \{$données, base de données$\}$

$PUD = PR = \{$données, base, de$\}$

$PNU = \{$logiciel de base de données, système de fichiers$\}$

Proposition de métriques

- Contribution du lexique
- Reconnaissance du vocabulaire
- Couverture du corpus
- Densité

Contribution du lexique

Mesures

– « Proportion de lexies figurant en corpus »

$$Contribution = \frac{|PU|}{|L|}$$

– « Proportion de lexies inutiles »

$$Surplus = 1 - Contr = \frac{|PNU|}{|L|}$$

où $|X|$ est le cardinal de X

Contribution du lexique

Exemple

$T = \{$Il a installé un système de gestion de base de données pour gérer toutes ses données.$\}$

$L = \{$données, base de données, logiciel de base de données, système de fichiers$\}$

$PU = \{$données, base de données$\}$

$$Contribution = \frac{|PU|}{|L|} = 2/4$$

$$Surplus = \frac{|PNU|}{|L|} = 2/4$$

$\Longrightarrow$ Indication : degré de spécialité et potentiel de la ressource

Reconnaissance du vocabulaire

Mesures

« proportion des lexies décomposées reconnues en corpus par rapport au nombre total de vocables du corpus »

$$Reconnaissance = \frac{|PR|}{|V|} = \frac{|PUD|}{|V|}$$

$$Ignorance = 1 - Reconnaissance = \frac{|PNR|}{|V|}$$

Reconnaissance du vocabulaire

Exemple

$T = \{$Il a installé un système de gestion de base de données pour gérer toutes ses données.$\}$

$V = \{$il, a installé, un, système, de, gestion, base, données, pour, gérer, toutes, ses$\}$

$PUD = PR = \{$données, base, de$\}$

$$Reconnaissance = \frac{|PR|}{|V|} = 3/12 = 1/4$$

$\Longrightarrow$ La reconnaissance augmente

si le lexique est « bien » spécialisée par rapport au corpus

si la ressource comporte beaucoup de mots de la langue générale

Couverture du corpus

– Mesure : « proportion d'occurrences de mots correspondant à des vocables entrant dans les lexies de la partie utile de la ressource »

$$Couverture = \frac{\sum_{i=1}^{|PU|} freq_i \times longeur_i}{|T|}$$

– Exemple

$T = \{$Il a installé un système de gestion de base de données pour gérer toutes ses données.$\}$

$V = \{$il, a installé, un, système, de, gestion, base, données, pour, gérer, toutes, ses$\}$

$PU = \{$données, base de données$\}$

$$Couverture = (1 + 3)/16 = 1/4$$

Densité

Mesure

$$densité = f_{PUD} / f_V$$

f_{PUD} : fréquence moyenne des lexies de PU dans le corpus

f_V : fréquence moyenne des vocables dans le corpus

$\Longrightarrow$ mesure normalisée de la fréquence des lexies utiles en corpus

$\Longrightarrow$ mesure indépendante de la taille du corpus

Densité

Exemple

$T = \{$Il a installé un système de gestion de base de données pour gérer toutes ses données.$\}$

$V = \{$il, a installé, un, système, de, gestion, base, données, pour, gérer, toutes, ses$\}$

$PUD = PR = \{$données, base, de$\}$

$f_{PUD} = (2 + 1 + 3)/3 = 2$

$f_V = (1 + 1 + 1 + 1 + 1 + 3 + 1 + 1 + 2 + 1 + 1 + 1 + 1)/13 = 16/13$

$Densité = \frac{f_{PUD}}{f_V} = 2 * 13/16 = 13/8 = 1,625$

Données expérimentales

– 4 corpus

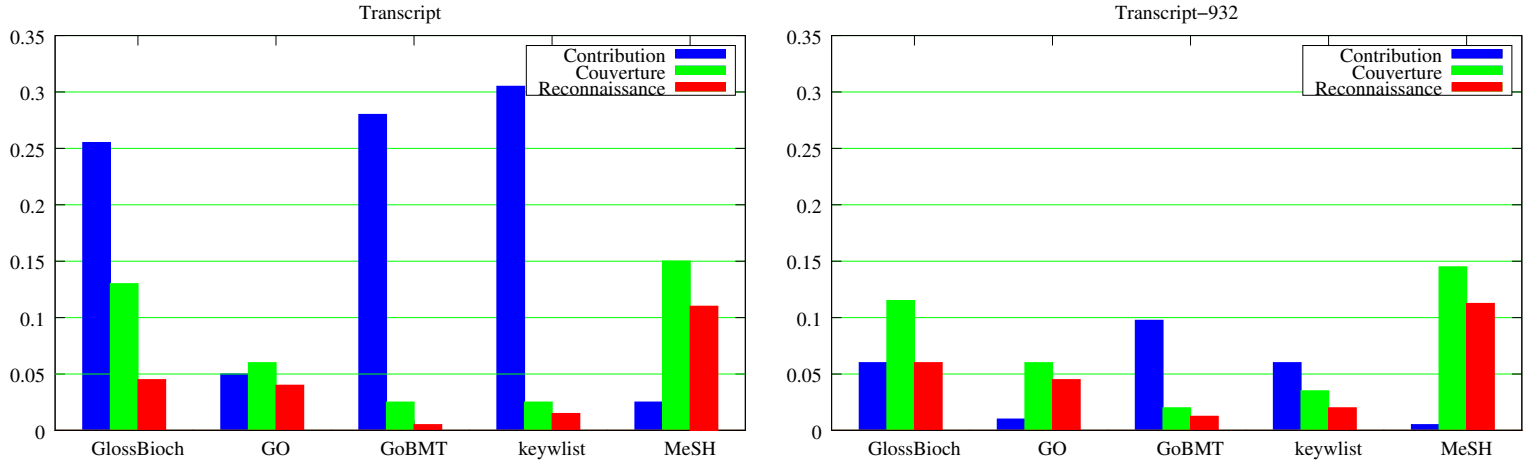
NB : Transcript-932 est un sous-ensemble que Transcript

	Vocabulaire (V)	Texte (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

– 5 ressources

Ressources	MeSH	GO	keywlist	GlossBioch	GoBMT
Taille en nombre de lexies	89 949	16 736	836	2 934	263

Pertinence d'une analyse sur échantillon

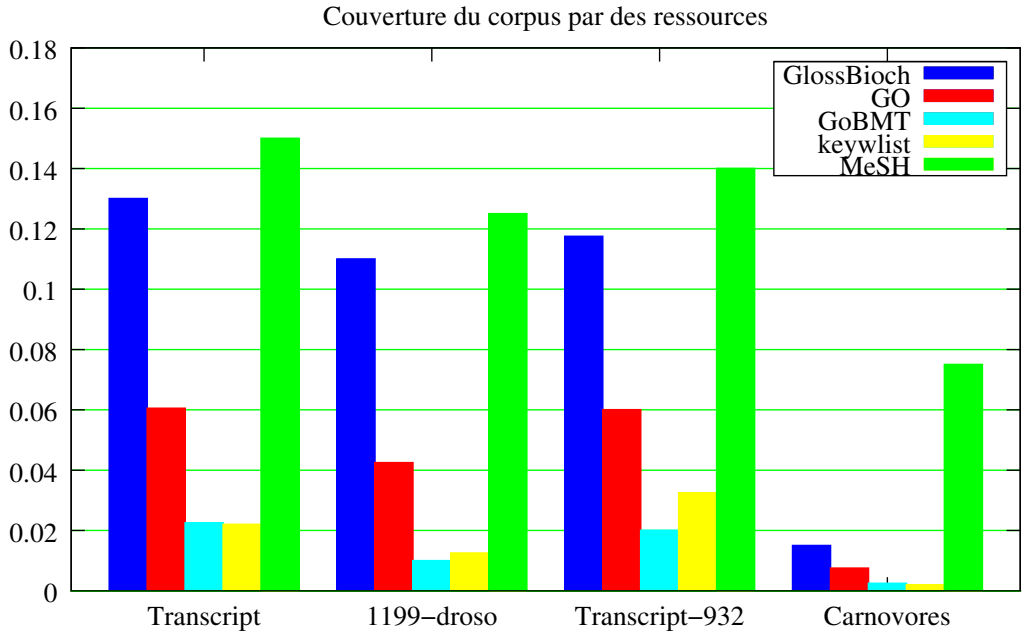


Les mesures gomment l'effet de taille (sauf contribution)

- sur les corpus
  Le comportement des ressources est comparable pour le corpus Transcript et son sous-corpus Transcript-932
- sur les ressources
  Le glossaire GlossBioch a une couverture similaire à celle de MeSH qui comporte pourtant 50 fois plus de termes

$\Longrightarrow$ On peut envisager de sélectionner une ressource à partir d'un sous-corpus

Des contrastes observables



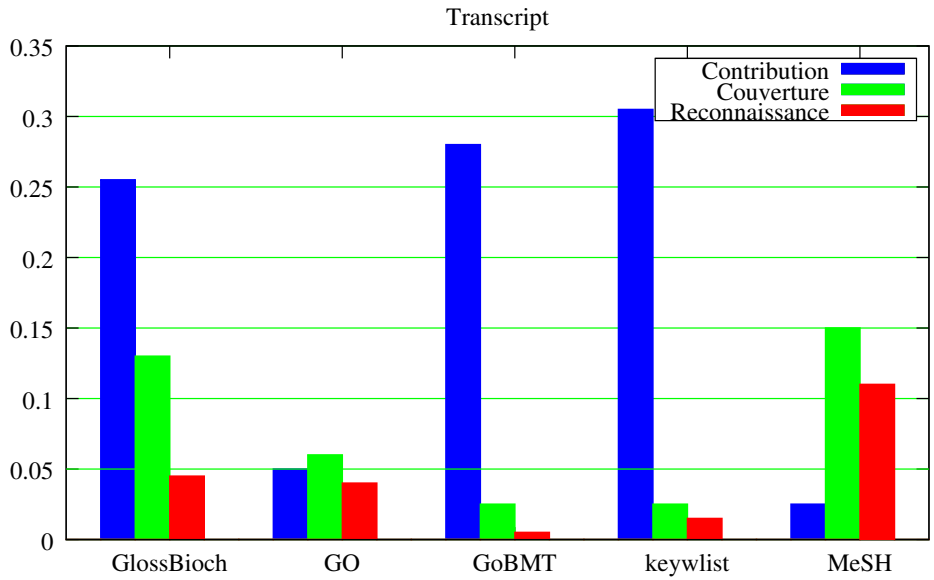
– Couverture des 3 corpus de génomique

  $\Longrightarrow$ GlossBioch et / ou MESH

– Couverture du corpus de botanique (Carnivore)

  $\Longrightarrow$ MESH mais pas GlossBioch

La taille ne fait pas tout !

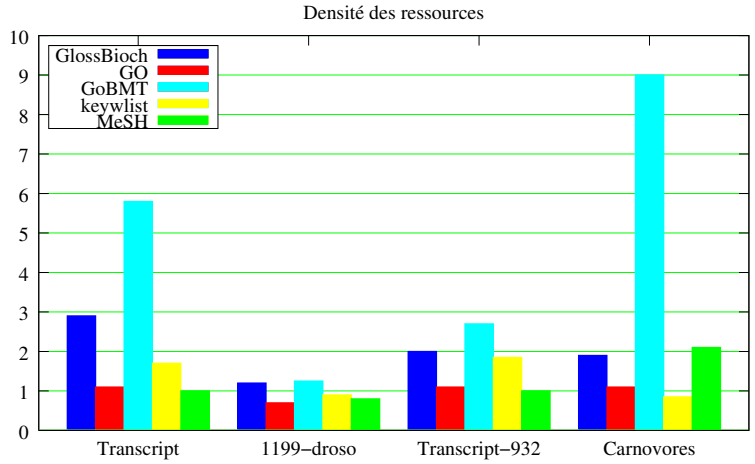


Glossaire GlossBioch $\leftrightarrow$ GO

- couverture de GlossBioch $\gg$ couverture de GO sur Transcript
- reconnaissance de GlossBioch $\simeq$ reconnaissance de GO

  $\Longrightarrow$ le glossaire GlossBioch est plus pertinent que GO malgré sa taille modeste

Une mesure de densité insuffisante (1)



- Remarque : La plus forte densité s'observe
  - pour une petite ressource très spécialisée (glossaire GoBMT)
  - pour le corpus le plus différent thématiquement (Carnivore)
- Analyse
  - Moins de 10% des lexies figurent dans le corpus
  - Ce sont des lexies aux très fortes fréquences
    can (676 occ.) fish (121 occ.) tel (8 occ) : noms de gènes ou mots de sens courants

Une mesure de densité insuffisante (2)

– Une simple mesure de fréquence pondérée ne rend pas compte du degré de spécialisation

– Il faut considérer le profil lexical des lexies de la partie utile de la ressource

Conclusion

Bilan

Des mesures pour apprécier *a priori* l'adéquation des ressources terminologiques à des corpus

→ Economiser du temps d'expérimentation

→ Capitaliser une expertise d'une expérience à l'autre

→ Formaliser les procédures

Une mesure donne un éclairage insuffisant mais la combinaison des 4 mesures fait apparaître des contrastes exploitables

Perspectives : prendre en compte des ressources plus riches

→ Prendre en compte les phénomènes de variation terminologique

→ Prendre en compte les relations sémantiques

→ Prendre en compte les distinctions de sens