Utilisation de corpus de spécialité pour le filtrage de synonymes de la langue générale

Natalia Grabar, Pierre Zweigenbaum

INSERM U729, Paris, F-75006 France
INaLCO, CRIM, Paris, F-75343 France
Assistance Publique - Hôpitaux de Paris, STIM/DSI, Paris, F-75674 France
ngr@biomath.jussieu.fr
pz@biomath.jussieu.fr

TALN 2005, 7 juin 2005

Contexte : variation terminologique

STENOSE DE L'AORTE (WHOART)
Sténose aortique (MedDRA)
Rétrécissement aortique (MeSH)
aorte sténosée
rétrécissement de l'aorte

$\Longrightarrow$ *Neutraliser la variation des termes*

- outils informatiques
- ressources lexicales

Ressources lexicales pour la variation des termes

- *Variantes orthographiques*
  {*anévrisme*, *anévrysme*}, {*rhino-pharyngite*, *rhinopharyngite*}
  (McCray, et al.*, 1994)*
- *Morphologie* : {*kyste*, *kystes*}, {*kyste*, *kystique*}, {*kyste*, *enkysté*}
  - Specialist lexicon d'UMLS *(McCray,* et al.*, 1994)*
  - Projet ACI UMLF *(Zweigenbaum,* et al.*, 2001)*
  - Projet RNTS VUMeF

$\Longrightarrow$ *Synonymie*
{*aorte*, *rétrécissement*}, {*kyste*, *tumeur*}, {*kyste*, *ganglion*}

- Lexiques généraux

Contextes applicatifs

- *Interopérabilité sémantique*
  (Degoulet *et al.*, 1998 ; Bousquet *et al.*, 2000 ; Joubert *et al.*, 2004)
- *Aide au codage*
  (Wolff, 1987 ; Lovis et al.*, 1998 ; Lussier* et al.*, 2001)*
- *Indexation et recherche d'information*
  (Zweigenbaum et al.*, 2001 ; Gaudinat & Boyer, 2002)*
- *Constitution de terminologies*
  (Grabar & Zweigenbaum, 2004)

$\Longrightarrow$ Les performances sont meilleures avec les lexiques du domaine

Adaptation de synonymes de la langue générale aux textes médicaux

- Matériel : synonymes de la langue générale
- Méthodes :
  - adaptation : filtrage à l'aide de corpus
  - évaluation : structuration de terminologies
- Résultats et discussion
- Conclusion et perspectives

Synonymes de la langue générale : Le Robert

- « Rapports analogiques » entre mots : étymologie, définitions, synonymie, antonymie, etc.

$\Rightarrow$ 140 141 séries de « synonymes » :

- *boulimie* : *cynorexie*, *hyperorexie*, *hyperphagie*, *sitiomanie*, *faimcalle*, *appétit*, *avidité*
- *culot* : *fond*, *dépôt*, *résidu*, *benjamin*, *aplomb*, *assurance*, *audace*, *effronterie*, *toupet*, *estomac*

Symétrie des synonymes

- La synonymie relie des lexèmes contextuellement interchangeables
  (Cruse, 1986)
- *boulimie* : *cynorexie*, *hyperorexie*, *hyperphagie*, *sitiomanie*, *faimcalle*, *appétit*, *avidité*
  + *faim*, *fringale*, *frénésie*
- *ventre* : 11 $\Rightarrow$ 19 synonymes
  trouble : 22 $\Rightarrow$ 64 synonymes

$\Rightarrow$ Normalisations vers l'entrée dictionnairique
(*boulimie*, *culot*)

Spécialiser les synonymes

- Contrainte domaniale absente
- *culot* : *fond*, *dépôt*, *résidu*, *benjamin*, *aplomb*, *assurance*, *audace*, *effronterie*, *toupet*, *estomac*
  - *transfusion*

Synonymes de Le Robert : *ensembles hétérogènes*
$\Longrightarrow$ Besoin d'adaptation au domaine médical

Filtrage en corpus : association récurrente

Détection de relations sémantiques entre termes dans les corpus :

- Les synonymes apparaissent à proximité dans les textes
  $\Longrightarrow$ Calcul d'associations entre mots (log likelihood ratio)

Filtrage en corpus : contextes spécifiques

- Marqueurs de coordination
  Examen cardio-vasculaire : `pas` *de dyspnée,* `pas` *d'orthopnée*
  Examen cardiaque : bruits ... sans souffle `ni` *bruit surajoutés*
- Patrons lexico-syntaxiques de synonymie
  L'œdème `est défini comme` *un gonflement palpable ...*
  Le rhinopharynx `appelé` *cavum est situé sous la base ...*

Évaluation en usage : structuration de terminologies

Détection de relation hiérarchique entre deux termes par « inclusion lexicale »

acides gras / acides gras indispensables

- Matériau : 19 638 termes du thesaurus MeSH « à plat »
- Normalisation des variantes de termes :
  - Caractères (casse, accents)
  - Ordre de mots
  - Ressources morphologiques
  - Ressources synonymiques
- Comparaison avec la structure hiérarchique originale du MeSH
  complétude (rappel), exactitude (précision)

Résultats et discussion

- Filtrage des synonymes
- Évaluation du filtrage :
  - Évolution quantitative des relations induites
  - Évaluation par rapport à la structure originale
  - Analyse détaillée

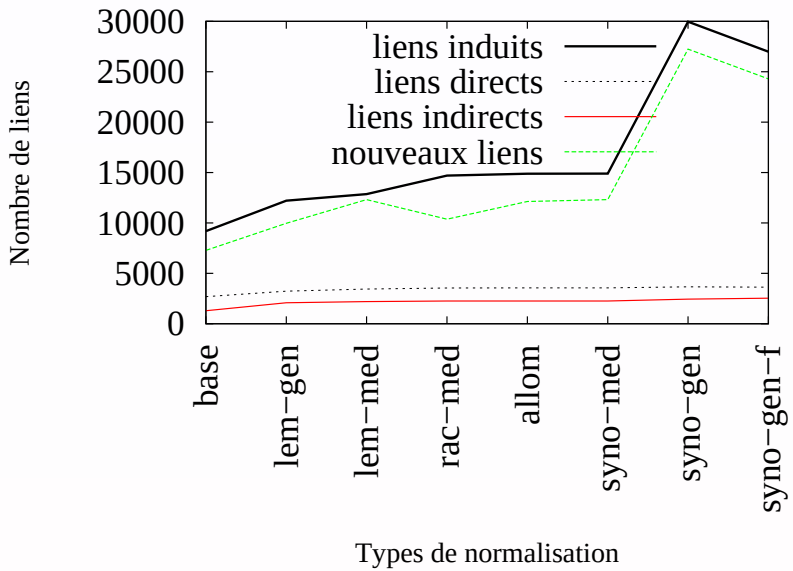
Filtrage des synonymes : quantités

- 140 141 séries initiales
- Calcul d'associations entre les mots : 15 589 couples
- Marqueurs de coordination : 1 736 couples
- Patrons de synonymie : 46 couples

$\Longrightarrow$ 16 154 couples (réduction de presque 90 %)

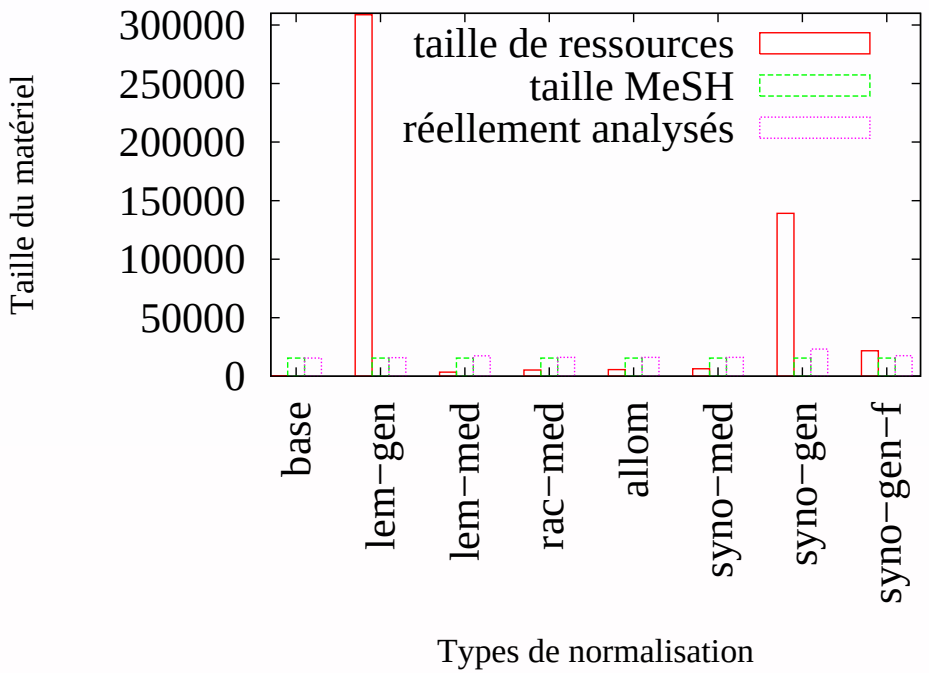
Évaluation : usage de ces synonymes comme ressource additionnelle pour la normalisation de variantes de termes

Évolution quantitative des relations induites



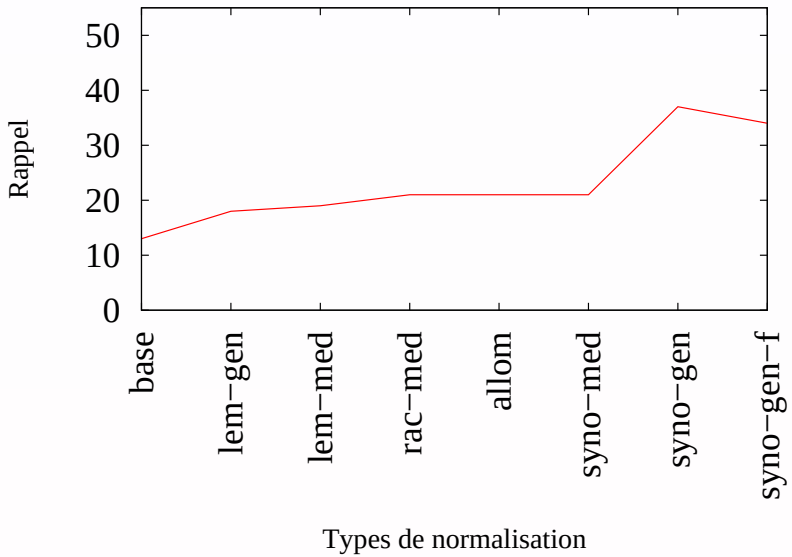
- Étapes comparées : *syno-gen* (29 969), *syno-gen-f* (26 986)
- Filtrages ⇒ *Diminution du volume de relations induites* (-2 983)

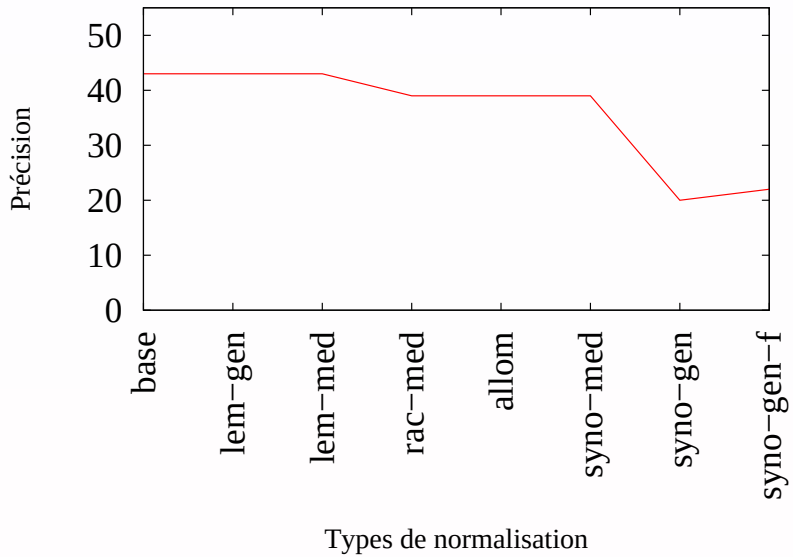
Utilisation effective des ressources lexicales



- Ressources de la langue générale (*lem-gen*, *syno-gen*)
  Utilisation faible $\Longrightarrow$ *Pertinence limitée*

Évaluation par rapport à la structure originale (1/2)





- Avec injection de connaissances : rappel ⇑, précision ⇓
- *syno-gen*, *syno-gen-f* : rappel ⇓ (- 3 %), précision ⇑ (+ 2,5 %)

Évaluation par rapport à la structure originale (2/2)

Évolution de la précision :

- Connaissances supplémentaires $\Longrightarrow$
  - Augmentation de la « distance sémantique » entre termes
  - « Risque » d'induction de relations incorrectes

Rappel faible :

- Une seule approche de structuration : inclusion lexicale
  - Combinaison d'approches différentes

  $\Longrightarrow$ Structuration plus complète

  (Kavanagh, 1995 ; Grabar & Jeannin, 2002)
- Induction de relations autres que hiérarchiques

Relations non induites suite aux filtrages

		Relations correctes :
69	2,3 %	Relations directes du MeSH
106	3,5 %	Relations indirectes du MeSH
		Relations potentiellement incorrectes :
2 808	94,1 %	Relations non-MeSH

$\Longrightarrow$ Prépondérance de relations potentiellement incorrectes

Relations directes du MeSH

Liens de synonymie non confirmés dans les corpus, par exemple :

- *carcinome* : *adénocarcinome*, *épithélioma*
  adénocarcinome / épithélioma squirrheux
- *céramique* : *porcelaine*
  céramiques / porcelaine dentaire
- *attraction* : *gravitation*, *pesanteur*
  gravitation / modification pesanteur

Relations indirectes du MeSH

Liens de synonymie non confirmés dans les corpus, par exemple :

- *narcose* : *anesthésie*, *hypnose*
  anesthésie / hypnose dentisterie
- *personnalité* : *soi*
  personnalité / concept soi
- *thérapeutique* : *traitement*
  thérapeutique / traitement par art

Relations non-MeSH, potentiellement incorrectes

Liens de synonymie non confirmés dans les corpus, par exemple :

- *grosseur* : *abcès*, *obésité*, *volume*, etc.
  abcès / volume sanguin
  obésité / volume sanguin
- *anéantissement* : *absorption*, *sidération*, etc.
  absorption / sidération myocarde
- *combustible* : *acétylène*, *infusion*, etc.
  acétylène / infusion goudron

$\Rightarrow$ *carcinome* : *adénocarcinome*, *épithélioma*
adénocarcinome / épithélioma mixte
(*adénocarcinome / épithélioma squirrheux*)

Conclusion

- *Normalisation de variantes de termes*
  $\Longrightarrow$ Utilisation de ressources lexicales appropriées
- *Degré de spécialisation de ressources lexicales*
  $\Longrightarrow$ Incidence sur les traitements automatiques
- *Ressources lexicales médicales non disponibles*
  $\Longrightarrow$ Utilisation de ressources de la langue générale
  $\Longrightarrow$ Adaptation, filtrage

- Adaptation de synonymes de la langue générale aux textes médicaux
- Évaluation à travers la structuration de termes

Perspectives

- *Recherche et utilisation de synonymes médicaux*
  $\Longrightarrow$ Pertinence plus élevée
- *D'autres approches pour le filtrage de synonymes généraux*
- *Évaluation du filtrage des synonymes par d'autres applications*