

Exploitation de dimensions du traitement de corpus en découverte de connaissances linguistiques

Pierre Zweigenbaum

- ^a STIM/DSI, Assistance Publique – Hôpitaux de Paris, France
- ^b ERM 202, INSERM, Paris, France
- ^c CRIM, INaLCO, Paris, France



Plan

- ✓ Les trois ordres de G. Grefenstette
- ✓ Premier ordre : mots thématiquement proches
- ✓ Second ordre : mots sémantiquement proches
- ✓ Second ordre : alignement en corpus comparables
- ✓ Conclusion

Détecter des régularités



Détecter des régularités



Freiburg, mars 2004



Göteborg, mai 2004

Exploration de corpus : trois ordres d'affinité

(Grefenstette, 1994) (Rapp, 2000+)

- 1 Observation des cooccurrences entre mots :
un mot apparaît dans le contexte d'un autre mot
relation syntagmatique : ↪ association

Exploration de corpus : trois ordres d'affinité

(Grefenstette, 1994) (Rapp, 2000+)

- ❶ Observation des cooccurrences entre mots :
un mot apparaît dans le contexte d'un autre mot
relation syntagmatique : ↪ association
- ❷ Comparaison de deux distributions de cooccurrences :
un mot apparaît dans les mêmes contextes qu'un autre mot
relation paradigmaticque : ↪ substituabilité

Exploration de corpus : trois ordres d'affinité

(Grefenstette, 1994) (Rapp, 2000+)

- ① Observation des cooccurrences entre mots :
un mot apparaît dans le contexte d'un autre mot
relation syntagmatique : ➡ association
- ② Comparaison de deux distributions de cooccurrences :
un mot apparaît dans les mêmes contextes qu'un autre mot
relation paradigmaticque : ➡ substituabilité
- ③ Recherche d'une structure sur les distributions de cooccurrences :
appartenance à un paradigme ➡ classification

Exploration de corpus : unités de discours

- ✓ syntagme
- ✓ phrase
- ✓ paragraphe
- ✓ document
- ✓ fenêtre de N mots
- ✓ ...

Acquisition de connaissances à partir de corpus

✓ Connaissances morphologiques

familles de mots construits (*Hathout et al.*) 

Acquisition de connaissances à partir de corpus

- ✓ Connaissances morphologiques

familles de mots construits (*Hathout et al.*) ERSS

- ✓ Connaissances syntaxiques

probabilités de sous-catégorisation (*Bourigault et al.*) ERSS

Acquisition de connaissances à partir de corpus

- ✓ Connaissances morphologiques
familles de mots construits (*Hathout et al.*) ERSS
- ✓ Connaissances syntaxiques
probabilités de sous-catégorisation (*Bourigault et al.*) ERSS
- ✓ Connaissances sémantiques
couples N-V qualia (*Fabre et al.*) ERSS

Acquisition de connaissances à partir de corpus

- ✓ Connaissances morphologiques
familles de mots construits (*Hathout et al.*) ERSS
- ✓ Connaissances syntaxiques
probabilités de sous-catégorisation (*Bourigault et al.*) ERSS
- ✓ Connaissances sémantiques
couples N-V qualia (*Fabre et al.*) ERSS
- ✓ ... ERSS

Quels traitements pour quelles connaissances ?

Le jeu de la découverte :
secouer son corpus
pour faire émerger
les relations linguistiques
qui le sous-tendent

Quels traitements pour quelles connaissances ?

Le jeu de la découverte :
secouer son corpus
pour faire émerger
les relations linguistiques
qui le sous-tendent

- ✓ Type de connaissance
- ✓ Ordre
- ✓ Unité de discours
- ✓ ...

Quels traitements pour quelles connaissances ?

Le jeu de la découverte :
secouer son corpus
pour faire émerger
les relations linguistiques
qui le sous-tendent

- ✓ Type de connaissance
- ✓ Ordre
- ✓ Unité de discours
- ✓ ...

➡ Quelques exemples

- ✓ Les trois ordres de G. Grefenstette
- ✍ Premier ordre : mots thématiquement proches
- ✓ Second ordre : mots sémantiquement proches
- ✓ Second ordre : alignement en corpus comparables
- ✓ Conclusion

Premier ordre : mots thématiquement proches

- ✓ Associations de mots
- ✓ Premier ordre : mots qui cooccurrent
(plus souvent qu'au hasard)
 - ✓ (distance courte) : collocations : expressions à plusieurs mots, sous-catégorisation
infarctus du myocarde, indexé sur, résoudre un problème
 - ✓ (distance plus grande) : relations thématiques
hôpital, médecin, chirurgie,
hospitalisation, chirurgien, chirurgical

Une (longue) phrase

Le caractère multifactoriel de la maladie asthmatique (prédisposition génétique, facteurs d'environnement - allergènes et polluants -, rôle des infections notamment virales) rend compte du polymorphisme de l'affection et explique le fait qu'aucune définition de l'asthme n'apparaît pleinement satisfaisante dans la mesure où elle n'inclut pas tous les aspects d'une affection très polymorphe dans ses modes de déclenchement, son profil évolutif ou sa sévérité.

Une (longue) phrase

Le caractère multifactoriel de la maladie asthmatique (prédisposition génétique, facteurs d'environnement - allergènes et polluants -, rôle des infections notamment virales) rend compte du polymorphisme de l'affection et explique le fait qu'aucune définition de l'asthme n'apparaît pleinement satisfaisante dans la mesure où elle n'inclut pas tous les aspects d'une affection très polymorphe dans ses modes de déclenchement, son profil évolutif ou sa sévérité.

Phrases plus courtes

A côté des problèmes inhérents à l'identification de l'asthme, les conditions d'une prise en charge correcte du patient asthmatique tiennent à plusieurs facteurs :...

De nos jours, l'asthme est une maladie relativement bien connue des asthmatiques, ce qui n'a pas été toujours le cas.

Un asthmatique peut avoir un asthme d'origine allergique ET intrinsèque.

Proximité thématique

Phrases plus courtes

A côté des problèmes inhérents à l'identification de l'asthme, les conditions d'une prise en charge correcte du patient asthmatique tiennent à plusieurs facteurs :...

De nos jours, l'asthme est une maladie relativement bien connue des asthmatiques, ce qui n'a pas été toujours le cas.

Un asthmatique peut avoir un asthme d'origine allergique ET intrinsèque.

D'une phrase à l'autre

5 à 10 % des patients atteints d'un asthme corticodépendant aux Etats-Unis (3) correspondraient à une ABPA ;
28 % des asthmatiques dont les tests cutanés sont positifs envers *Aspergillus fumigatus*, dans une autre étude américaine, présentent tous les critères d'une ABPA (3,11,28).

D'une phrase à l'autre

5 à 10 % des patients atteints d'un asthme corticodépendant aux Etats-Unis (3) correspondraient à une ABPA ;
28 % des asthmatiques dont les tests cutanés sont positifs envers *Aspergillus fumigatus*, dans une autre étude américaine, présentent tous les critères d'une ABPA (3,11,28).

D'une phrase à l'autre

On sait que l'environnement joue un rôle significatif dans le développement de l'asthme chez les enfants.

Une étude menée en Grande Bretagne (*) met en évidence qu'un enfant de moins de 2 ans exposé à la fumée de cigarette de sa maman présentera assez systématiquement des symptômes de type asthmatiques.

D'une phrase à l'autre

On sait que l'environnement joue un rôle significatif dans le développement de l'asthme chez les enfants.

Une étude menée en Grande Bretagne (*) met en évidence qu'un enfant de moins de 2 ans exposé à la fumée de cigarette de sa maman présentera assez systématiquement des symptômes de type asthmatiques.

Proximité thématique

D'une phrase à l'autre

L'asthme affecte près de 15 millions d'américains dont 5 millions d'enfants.

En milieu rural, 7% des enfants sont asthmatiques, le double le sont en ville.

Proximité thématique

D'une phrase à l'autre

L'asthme affecte près de 15 millions d'américains dont 5 millions d'enfants.

En milieu rural, 7% des enfants sont asthmatiques, le double le sont en ville.

D'une phrase à l'autre

Notamment, elle irrite les muqueuses de la trachée et des poumons et favorise le déclenchement de crises d'asthme. Si vous êtes asthmatique, il est donc essentiel de vérifier la qualité de l'air dans votre région pour anticiper.

D'une phrase à l'autre

Notamment, elle irrite les muqueuses de la trachée et des poumons et favorise le déclenchement de crises d'asthme.
Si vous êtes asthmatique, il est donc essentiel de vérifier la qualité de l'air dans votre région pour anticiper.

Proximité thématique

Cinq phrases

Certains climats et certaines zones géographiques peuvent être profitables aux **asthmatiques** du fait d'un air plus pur, d'un meilleur ensoleillement.

Les séjours climatiques :

Il y a en France de nombreux établissements :
en montagne, en mer ou en plaine.

La qualité de l'air y est meilleure et il y a une diminution du contact avec des substances **allergisantes** ou **allergènes** (**acariens**, **pollens**).

Beaucoup de patients sont satisfaits des **cures thermales**, c'est un lieu de détente et **d'oxygénation**.

Par exemple LA BOURBOULE (au niveau **ORL** et **asthme**) ; AVENE, LA ROCHE POSAY (peau).

Proximité thématique

Cinq phrases

Certains climats et certaines zones géographiques peuvent être profitables aux asthmatiques du fait d'un air plus pur, d'un meilleur ensoleillement.

Les séjours climatiques :

Il y a en France de nombreux établissements :
en montagne, en mer ou en plaine.

La qualité de l'air y est meilleure et il y a une diminution du contact avec des substances allergisantes ou allergènes (acariens, pollens).

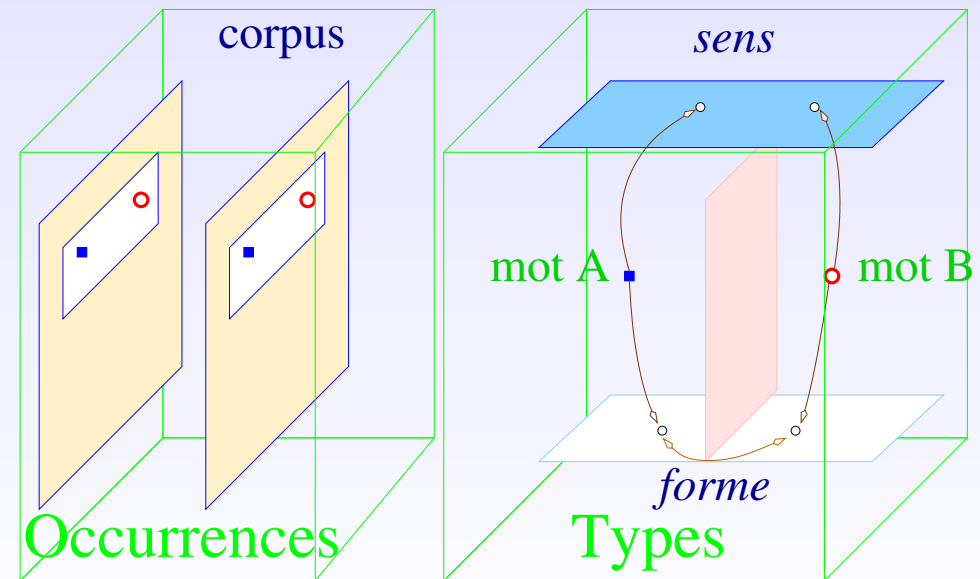
Beaucoup de patients sont satisfaits des cures thermales, c'est un lieu de détente et d'oxygénation.

Par exemple LA BOURBOULE (au niveau ORL et asthme) ; AVENE, LA ROCHE POSAY (peau).

Premier ordre : mots morphologiquement reliés

- ✓ Objectif : repérer des familles morphologiques
- ✓ Mots de forme proche
- ✓ Qui sont reliés thématiquement
- ✓ Unité : fenêtres de mots

(Zweigenbaum & Grabar, 2003)



Corpus de travail

- ✓ Corpus construit à partir du web à travers le catalogue CISMef des sites médicaux francophones
[http ://www.chu-rouen.fr/cismef/](http://www.chu-rouen.fr/cismef/)
- ✓ Étiqueté et lemmatisé : TreeTagger (*Schmid, NEMLAP 1994*)
+ FLEMM (*Namer, TAL 2000*)
fournit essentiellement des dérivations
- ➡ 4 627 documents
- ➡ 5 204 901 mots
- ➡ 2 041 627 mots non grammaticaux

Unité : fenêtre graphique

- ✓ Suppression des mots “outils”
- ✓ Fenêtre glissante, de M mots à gauche et à droite du “mot pivot”

Cooccurrences :

- ✓ Collecte les cooccurrents du mot pivot
- ✓ qui commencent par les mêmes N premières lettres
($N = 4$)

{asthme, asthmatique}

Sélection heuristique des dérivés

- ✍ Pas de dérivation régressive
longueur dérivé \geq longueur base $- 1$
articulation / articulaire, sacrum / sacré
- ✍ Éviter les composés (morphèmes longs)
longueur dérivé \leq longueur base $+ 5$
bronche / bronchopneumonique
- ✍ Fréquence de la règle : le même opérateur morphologique (“règle”) est employé “souvent”
Ex. : la substitution -e / -aire s’applique 72 fois dans les couples trouvés

Exemples de dérivations repérées

Sur 26 noms d'anatomie commençant par **a**
trouvés dans la nomenclature SNOMED Internationale
(376 examinés en tout)

Nom	Adjectif	# cooc	loglike	ch.i.c.m.	suf1	suf2	f
abdomen	abdominal	101	584.21	abdom	en	inal	2
amygdale	amygdalien	8	100.24	amygdal	e	ien	24
aorte	aortique	170	1314.74	aort	e	ique	131
apophyse	apophysaire++	3	39.66	apophys	e	aire	72
appendice	appendiculaire++	19	225.24	appendic	e	ulaire	5
articulation	articulaire	216	1406.34	articula	tion	ire	13
artériole	artériolaire+	15	99.99	artériol	e	aire	72
aréole	aréolaire+	2	27.55	aréol	e	aire	72
astrocyte	astrocytaire	2	28.60	astrocyt	e	aire	72
axone	axonal+	8	93.21	axon	e	al	42

+ association non spécifiée par SNOMED

++ adjectif absent de SNOMED

Précision, rappel, ajouts % SNOMED

Proportions de couples nom-adjectif

Corpus = 150			
seulement dans SNOMED 72 = 49 %	trouvés par le corpus 76 = 51 % rappel	ajoutés par le corpus 61 = 41 % ajouté	erronés 13 = 91 % de précision
SNOMED = 148			

Ajouts : apophysaire, appendiculaire, cardial, cotyloïdien, cristallinien, diaphysaire, hippocampique, intimal, jambier, lysosomal, macrophagique, mastocytaire, myométrial, métatarsien, néphronique, olécrânien, paramétrial, plasmatique, rhinopharyngé, réticulocytaire, tympanique, éosinophilique

- ✓ Les trois ordres de G. Grefenstette
- ✓ Premier ordre : mots thématiquement proches
- ✍ Second ordre : mots sémantiquement proches
- ✓ Second ordre : alignement en corpus comparables
- ✓ Conclusion

Mots sémantiquement proches : second ordre

- ✓ Sens proche \Leftrightarrow usage similaire
- ✓ Premier ordre : Représenter le sens d'un mot par l'ensemble de ses contextes d'usage
- ✓ Vecteur de contextes : vecteur des mots associés
- ✓ Second ordre : les mots qui possèdent des vecteurs de contextes similaires ont des sens proches

(Habert, Nazarenko, Bouaud, Zweigenbaum, 1997–2000)

Préparation des données

Zellig (*Habert et al., 1996*)

- ✓ Corpus Menelas (84 kmots)
- ✓ Syntagmes nominaux obtenus par Lexter ou AlethIPGN
- ✓ Arbres élémentaires (dépendances)
- ✓ Contextes syntaxiques d'occurrence des N et Adj

administration de médicament,
administration de routine,
administration orale

Vecteurs de contexte

Contextes de sténose :

(score d'association = nb de cooccurrences)

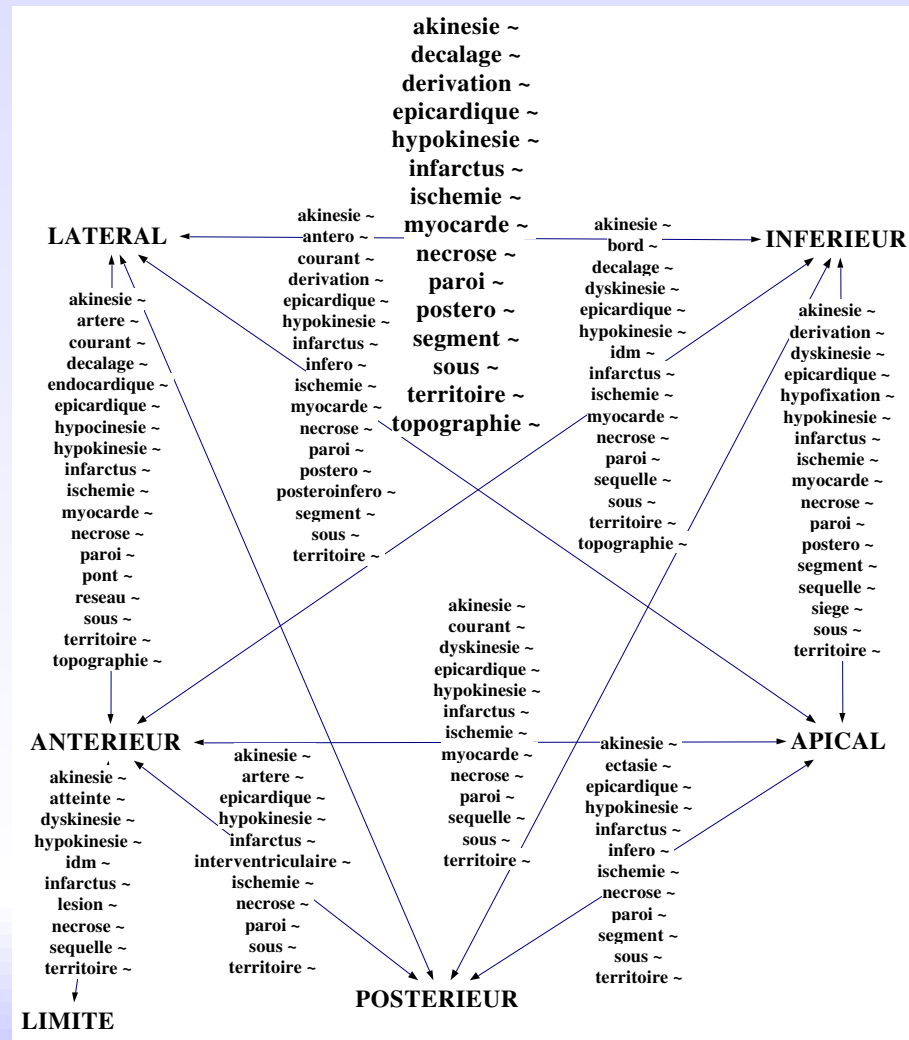
contexte	score	
de artere	10	■
de allure	10	■
de branche	3	■
de carotide	3	■
de debut	3	■
diagonale	3	■
droite	4	■
...		

Graphe de contextes partagés

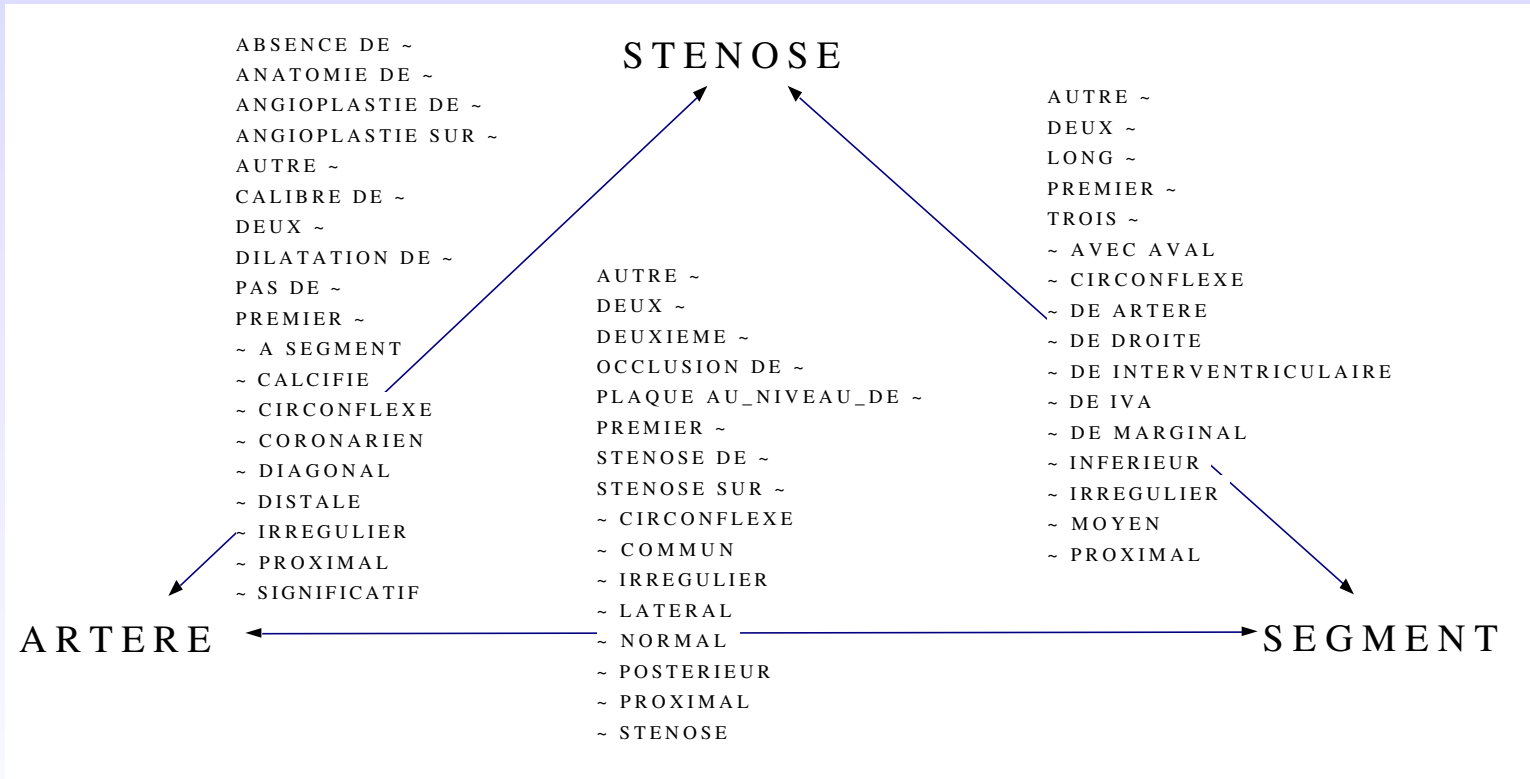
- ✓ Deux mots sont liés par une arête si leur nombre de contextes communs est supérieur à un seuil donné
- ✓ Seuil = 10

Exemple : clique (+)

(Lexter)



Exemple : clique

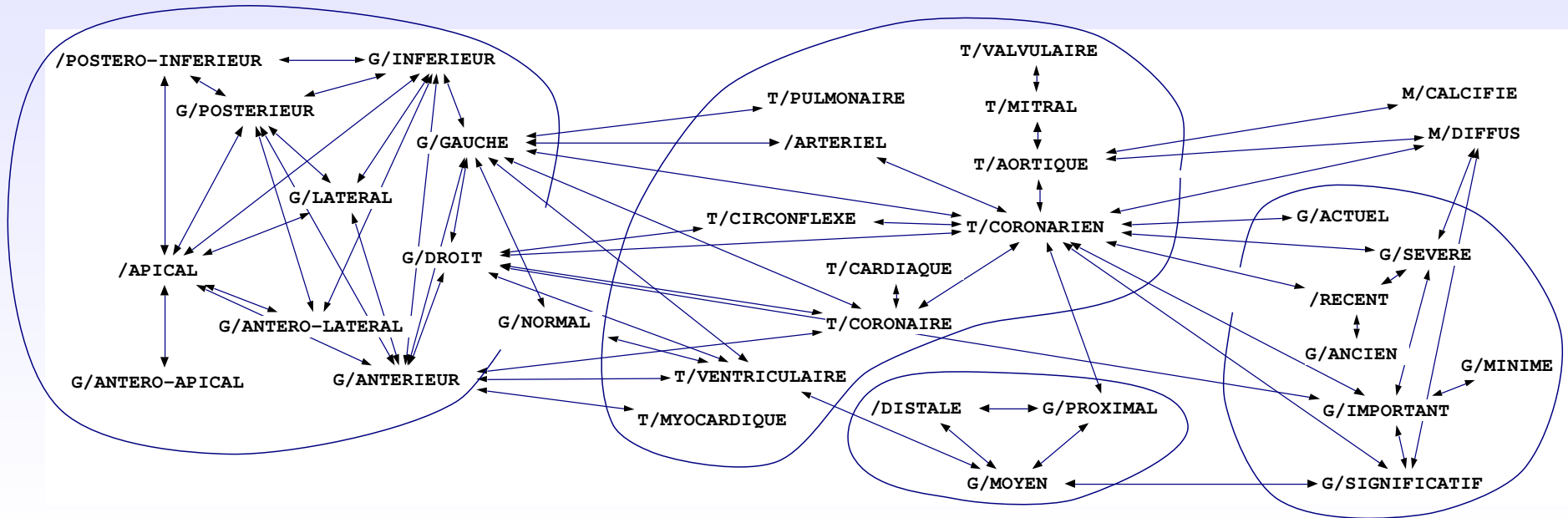


Exemple : projection de classes connues

Axes sémantiques de la nomenclature SNOMED


T = anatomie, G = qualificatifs et termes relationnels,

M = lésions, F = dysfonctions, D = diagnostics...



Discussion

- ✓ Importance de l'interprétation humaine des graphes obtenus
- ✓ Outil d'accès à certaines relations dans un corpus
- ✓ Importance de la préparation des données
 - connaissances linguistiques initiales
sténose d'allure, sténose du début
(Hirschman, 1975)
 - Gestion de la polysémie
au genou de l'artère interventriculaire antérieure
- ✓ Surtout sur un petit corpus ?

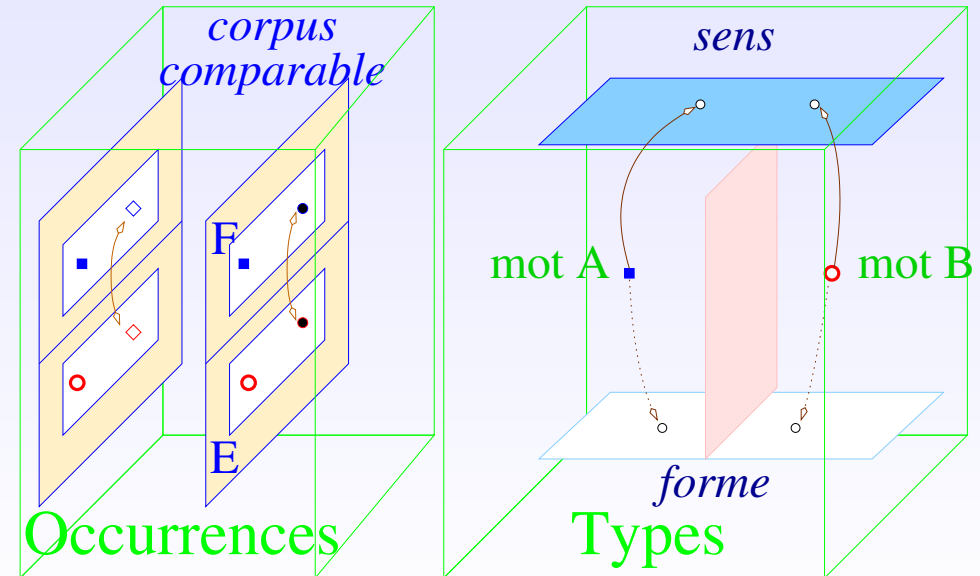
- ✓ Les trois ordres de G. Grefenstette
- ✓ Premier ordre : mots thématiquement proches
- ✓ Second ordre : mots sémantiquement proches
-  Second ordre : alignement en corpus comparables
- ✓ Conclusion

Alignement en corpus comparables : second ordre

- ✓ Second ordre : les mots d'usage similaire partagent leurs mots associés
- ✓ Représenter le sens des mots par des vecteurs de contextes (premier ordre)
- ✓ Les mots qui possèdent des vecteurs de contextes similaires (second ordre) ont des sens proches

Alignement en corpus comparables : traduction






- ✓ Objectif : trouver des équivalents traductionnels pour un mot
- ✓ Unité : fenêtres de mots dans deux corpus monolingues comparables
- ✓ Connaissances : lexique bilingue (partiel)



(Thèse de Yun-Chuang Chiao, juin 2004)

(Chiao & Zweigenbaum, 2002–2004)

Alignement en corpus comparables : corpus

-  Thème : Signes et symptômes (MeSH C23)
-  Français : corpus obtenu à travers le catalogue CISMeF (16 Mmots, puis 54 Mmots)
-  Anglais : corpus obtenu à travers le catalogue CliniWeb (1 Mmots, puis 7 Mmots)
-  Simple segmentation en mots
-  Suppression des mots grammaticaux

Lexique bilingue d'amorçage











- ✓ Collecte des correspondances entre unitermes (termes à un mot) français-anglais dans le Metathesaurus UMLS
- ✓ Dictionnaire médical français avec traductions anglaises
- ✓ Lexique bilingue général (paquetages dictd)

Fournit un alignement (partiel) des mots de contexte

Exemple : vecteur de contexte

- ✓ Vecteur de contexte pour **adénose**, restreint aux mots du lexique d'amorçage
- ✓ Taille de la fenêtre de contexte : ± 3 mots, ± 2 mots
- ✓ Score d'association : cooc, IM, loglike

Vecteur de contexte :

en français	score	converti en anglais
adénome	(11.8) 	adenoma
cellule	(8.9) 	cell
examen	(5.9) 	test
hyperplasie	(14.2) 	hyperplasia
lésion	(8.8) 	lesion
nucléole	(17.4) 	nucleolus
photographie	(13.9) 	photograph
prolifération	(11.9) 	proliferation
prostate	(9.1) 	prostate
prostatique	(11.9) 	prostatic
...		

Exemple : scores de similarité (Fr → En)

- ✓ Mots du corpus anglais qui ont les vecteurs de contexte les plus similaires au vecteur de contexte (converti) français de foie

- ✓ Mesure de similarité : Jaccard, cosinus

français	anglais	similarité	
foie	lung	.270294	█
foie	liver	.231073	█
foie	pain	.174125	█
foie	patient	.162746	█
foie	tumor	.137852	█
foie	disease	.136998	█
foie	primary	.119938	█
foie	treatment	.119257	█
foie	brain	.109586	█
foie	cancer	.105038	█
foie	bone	.104870	█
foie	kidney	.104498	█

Exemple : scores de similarité (En → Fr)

- ✓ Mots du corpus français qui ont les vecteurs de contexte les plus similaires au vecteur de contexte (converti) anglais de **liver**
- ✓ (listes similaires pour les autres mots anglais)

anglais	français	similarité	
liver	foie	.365169	████████
liver	rare	.309686	██████
liver	associée	.292330	██████
liver	alzheimer	.284989	██████
liver	transmissible	.269096	██████
liver	fréquente	.263598	██████
liver	pathologie	.257709	██████
liver	cardiovasculaire	.250468	██████
liver	cardio-vasculaire	.248039	██████
liver	creutzfeldt-jakob	.243688	██████
liver	hépatique	.242475	██████
liver	origine	.240563	██████

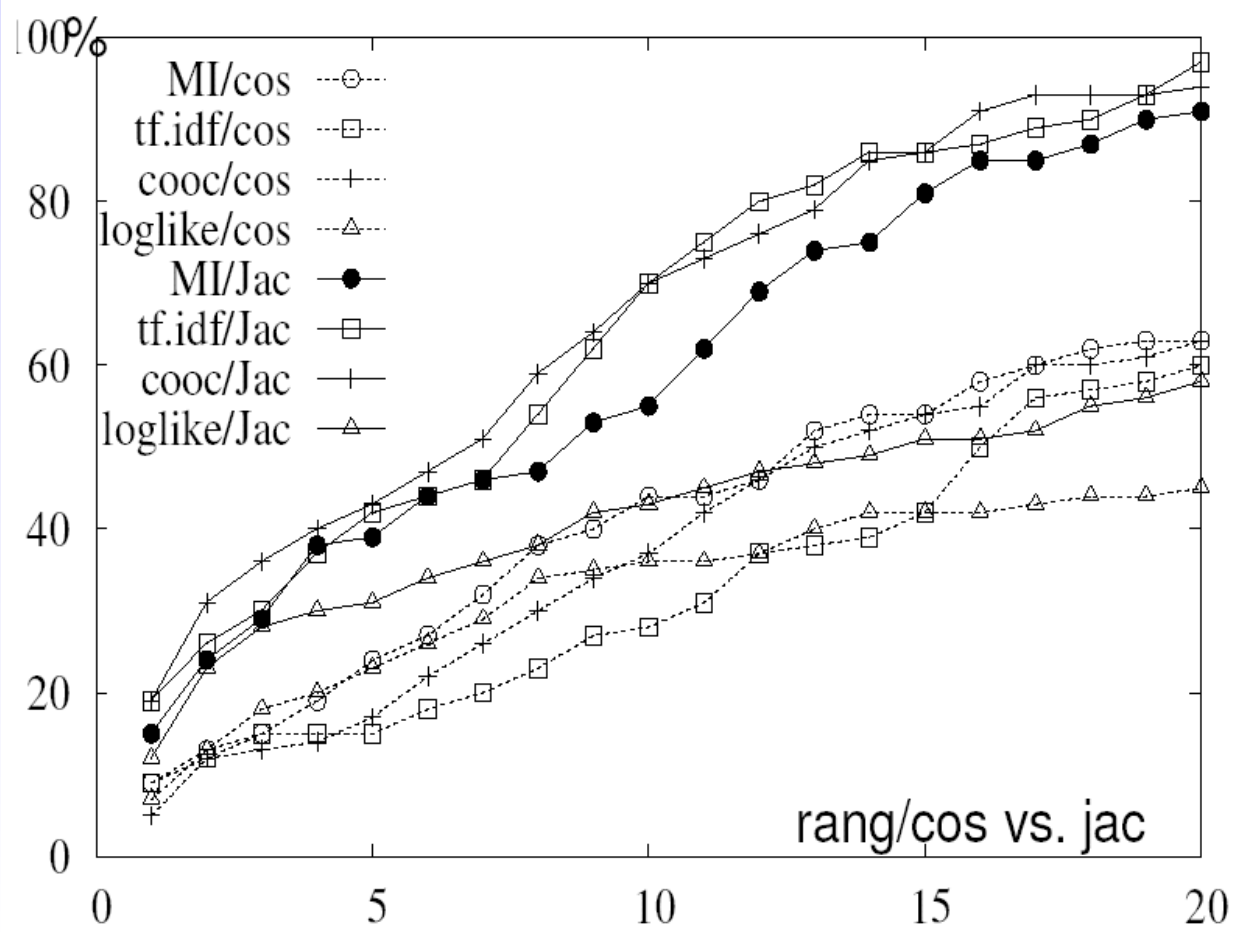
Combinaison des rangs

Moyenne harmonique des rangs initiaux pour les correspondants de foie

candidats	rang _{FrEn}	rang _{EnFr}	MH	nouveau rang
lung	1	4	1.60	2
liver	2	1	1.33	1
pain	3	31	5.48	4

Proportion de traductions correctes dans les meilleurs rangs


✓ Mots fréquents hors lexique (meilleure situation)



Discussion

Dans quel usage effectif ce type de performance est-il utile ?

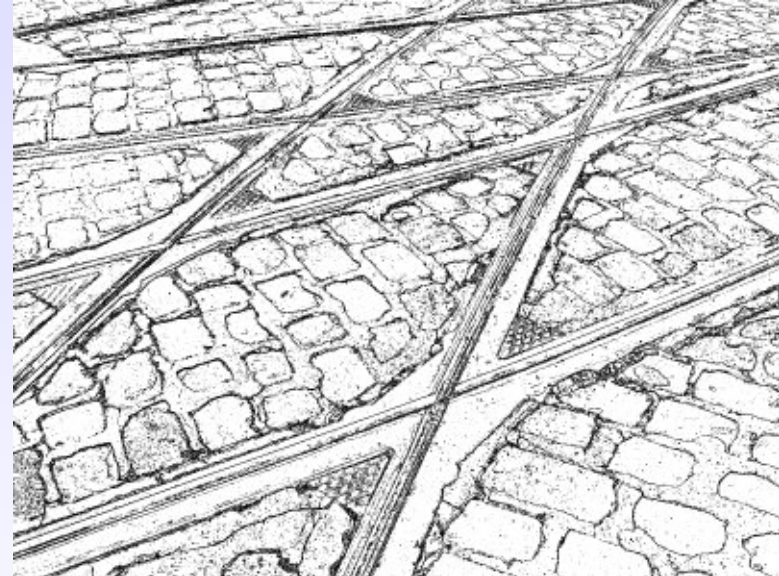
- ✓ Lexicographie, terminologie ?
- ✓ Recherche d'information translangue ?

- 
- ✓ Les trois ordres de G. Grefenstette
 - ✓ Premier ordre : mots thématiquement proches
 - ✓ Second ordre : mots sémantiquement proches
 - ✓ Second ordre : alignement en corpus comparables

 Conclusion

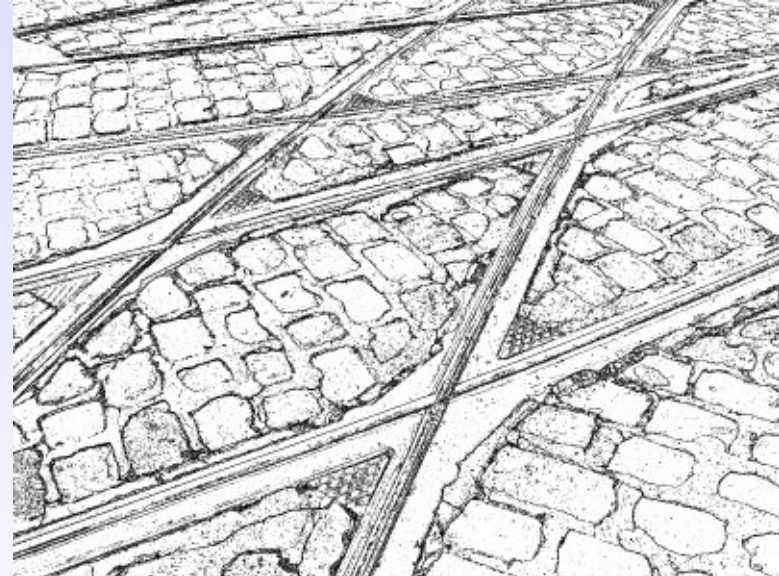
Conclusion

Quel rôle pour les méthodes automatiques ?



Conclusion

Quel rôle pour les méthodes automatiques ?



- ✓ Débroussaillage : faciliter l'examen de masses de textes
- ✓ Accès à l'évident vs accès aux pépites
- ✓ Accès au fréquent vs accès au rare