Sémantique distributionnelle & approche harrissienne

Serge Fleury & Benoît Habert

http ://www.cavi.univ-paris3.fr/ilpga/ilpga/sfleury/ - fleurys [@] noos [.] fr

http ://www.limsi.fr/Individu/habert - habert [@] limsi.fr

Accès distributionnel au sens

Hypothèse : deux mots ont un sens proche s'ils sont employés dans des contextes très voisins

$$\updownarrow$$

Cooccurrences $\leftrightarrow$ dépendances (récursives)

You shall know a word by the company it keeps [Firth 1957]

[Harris 1988] *Caractériser les mots par leur sélection permet de considérer le type et le degré de recouvrement, d'inclusion et de différences entre mots par rapport à leurs ensembles de sélection.*
... dans la plupart des cas, la sélection d'un mot inclut un ou plusieurs domaines cohérents de sélection.

Associations sur l'axe horizontal 1/3

Phrase	
Ce bébé **agrippe** *très bien le doigt*	
quand on le lui met dans la paume de la main	
Contexte syntaxique	
têtes dépendants	bébé$_{SUJ}$ doigt$_{OBJ}$
Fenêtre graphique	
$\pm$ 4 mots	ce bébé très bien le doigt
+ POS	ce$_D$ bébé$_N$ très$_R$ bien$_R$ le$_D$ doigt$_N$
- « outils »	bébé$_N$ doigt$_N$

Associations sur l'axe horizontal 2/3

- Relations de cooccurrences : *agrippe* avec *ce bébé très bien le doigt*
- Relations de dépendance :

bébé dépend de *agrippe* par la relation SUJ, *doigt* par la relation OBJ

Ou encore *agrippe* est l'*opérateur* et *bébé*, *doigt* les *opérandes*

<entier$_1$> + <entier$_2$> + est un opérateur qui s'applique à deux opérandes

<N$_{hu}$$_{main}$> agripper <N$_{objet}$>

Associations sur l'axe horizontal 3/3

empan

court $\longleftrightarrow$ long

collocations, micro-syntaxe $\longleftrightarrow$ thématique

filtrage

graphique	anti-dictionnaire	$\longleftrightarrow$	POS	syntaxe
bruit		$\longleftrightarrow$		précision
abondance		$\longleftrightarrow$		disette

Ordres d'affinité entre mots

[Grefenstette 94]

1. axe horizontal : observation des cooccurrences entre mots $\rightarrow$ relation d'association
2. axe vertical : comparaison des distributions de cooccurrences ($\pm$ grande similarité des associations des mots 2 à 2) $\rightarrow$ relation de substitution
3. rapprochement des mots qui ont des distributions de co-occurrences similaires $\rightarrow$ partitionnement

Menelas

- projet européen de compréhension de comptes rendus d'hospitalisation
- thèmatique : maladies coronariennes
- genres : extrait de manuel / comptes rendus d'hospitalisation / lettres entre médecins
- 84 839 occurrences – 6 191 formes

Menelas : manuel

Par ailleurs, la comparaison du diamètre de l'obstruction est faite avec le diamètre « supposé normal » de l'artère en amont ou en aval de la sténose. Or des études récentes démontrent que ces segments d'artères sont très souvent le site de rétrécissements diffus. Ceci peut résulter d'une sous-estimation marquée du degré de l'obstruction.

Menelas : compte rendu

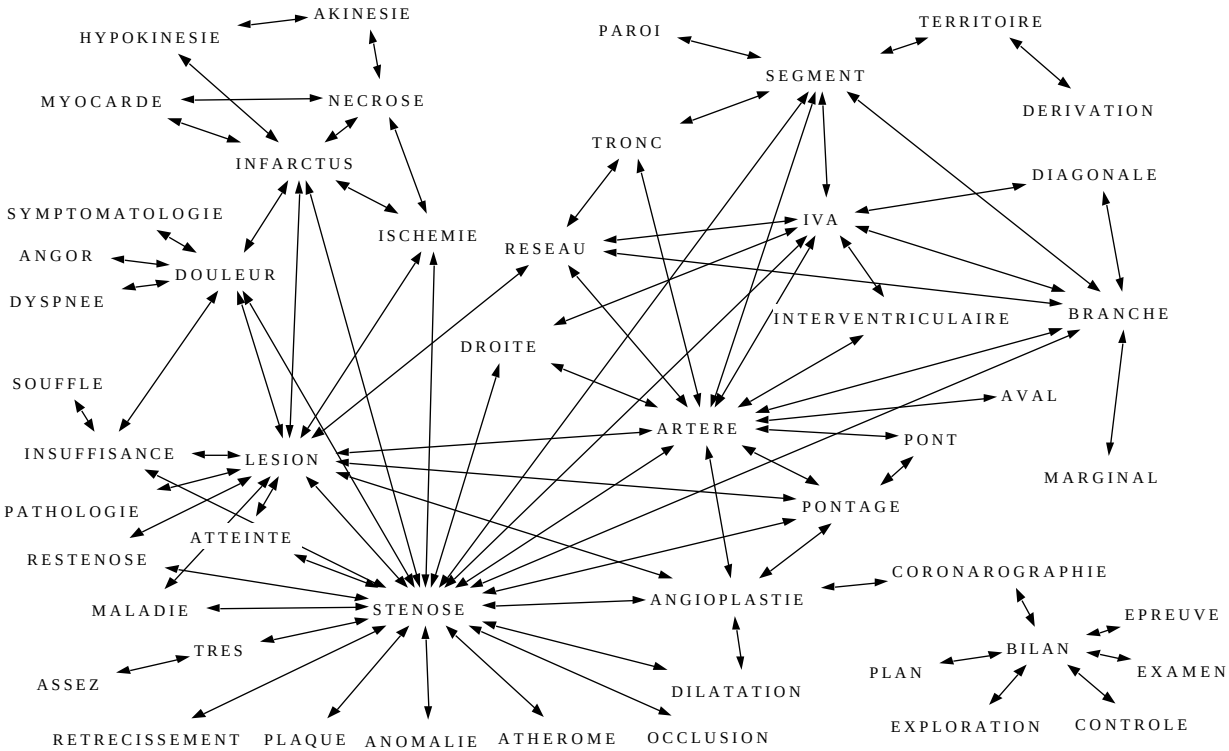
Patient âgé de 70 ans, diabétique, qui a présenté il y a un an une douleur thoracique nocture probablement en rapport avec un infarctus antéro-septal. Il est toujours symptomatique sous la forme d'un angor d'effort qu'il a de lui-même négligé, avec semble-t-il plusieurs épisodes de préchordialgies de repos. La coronarographie met en évidence des lésions bitronculaires. L'occlusion de l'IVA est responsable d'une hypokinésie antérieure. Une sténose serrée, diagonale et circonflexe est responsable de l'angor d'effort.

Menelas : lettre à un confrère

Veux-tu trouver ci-joint les documents concernant Madame MOPABAR11. Je pense qu'il faut d'abord réaliser un pontage aortocoronarien avant de faire le geste sur la carotide. Bien sûr il s'agit d'une malade âgée, avec un risque opératoire et péri-opératoire non nul, mais les lésions et la sténose du tronc gauche sont très menaçantes et compte tenu des douleurs spontanées qu'elle a présenté récemment à la CSR, je crois qu'il faut sans tarder revasculariser son myocarde malgré les risques importants. En te remerciant de ta confiance, et en restant à ton entière disposition.

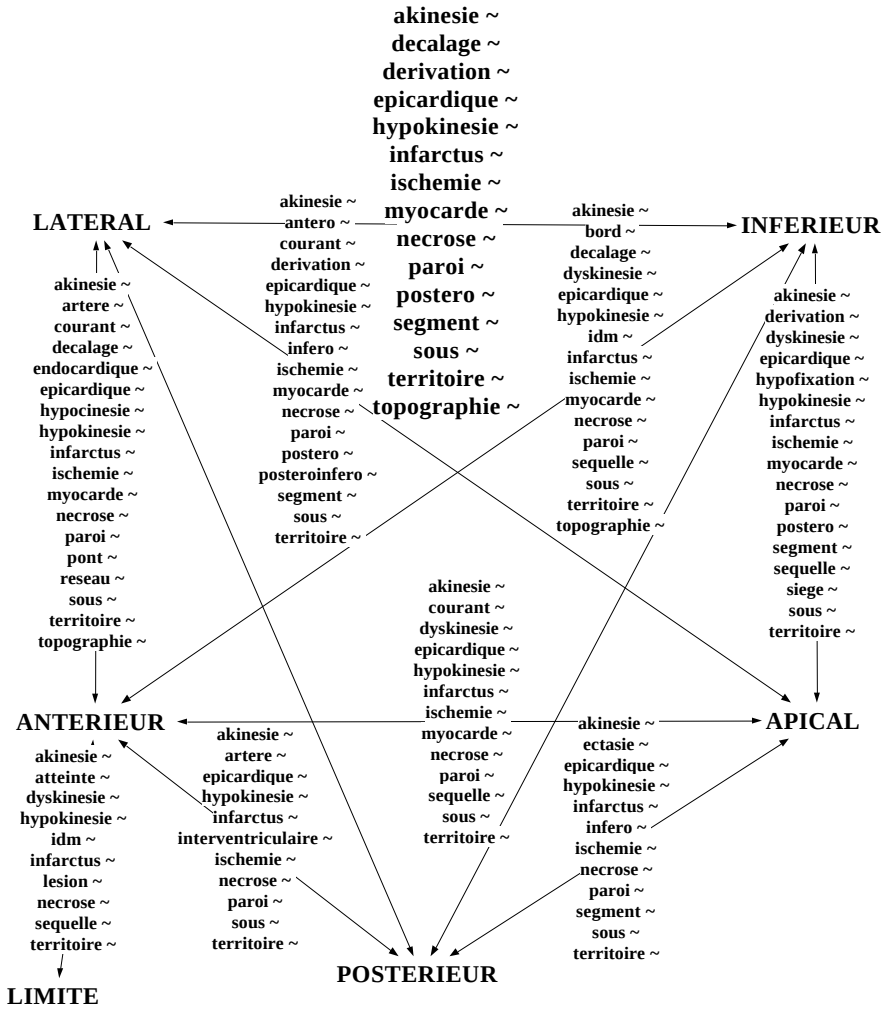
Des regroupements à des classes

Menelas + extracteur de SN AlethIPGN
arête entre 2 mots $\equiv$ nombre de contextes partagés $\geq$ seuil (ici 10)



Partages de contextes et « classes »

Menelas + Lexter



couvrir des grammaires sémantiques : Ha

- Pas de classification *a priori* des sens
- Les relations de dépendance entre opérateur et (ensemble d') opérandes sont par contre objectivables (relations à distinguer des relations de cooccurrence)

Langue générale probabilité spécifique et stable pour un mot d'apparaître comme opérateur ou comme opérande avec un autre (explique les réductions : *Xavier boit* $\emptyset \equiv$ *Xavier boit [régulièrement] de l'alcool*)

Sous-langages (domaines d'activité spécialisés) : sélections booléennes (en terme de possible/impossible)

Méthode d’analyse des sous-langages

- constitution de corpus spécialisés (immunologie, comptes rendus d’hospitalisation)
- normalisation syntaxique (en phrases élémentaires) pour faciliter la mise en évidence des régularités distribution-nelles (passage de l’actif au passif, restitution des verbes des nominalisations. . .)
- obtention de classes d’opérateurs et d’opérandes (« amorçage » éventuel par recours à des spécialistes ou à des lexiques spécialisés)
- des distributions des classes dégagées sont induits des patrons informationnels (grammaires sémantiques)

Statut des grammaires « à la Harris »

- les patrons informationnels d'un sous-langage donné forment un « langage pivot ».
- analyse par A. Daladier d'un corpus comparable à celui de Harris *et al*. : partage des patrons informationnels
- les changements de vraisemblance dans les sélections reflètent les évolutions conceptuelles. Ex. : *L'anticorps*$_A$ *apparaît*$_V$ *dans les tissus*$_T$ $\longrightarrow$ *L'anticorps*$_A$ *apparaît*$_V$ *dans les cellules*$_C$.
  - corpus constitué comme diachronique (1936-1966)
  - le flou nécessaire : points d'hésitation et d'ajustement

Pauvreté linguistique fait loi !

- Harris *et al*. : dépendances syntaxiques de phrases élémentaires, après normalisation manuelle
- possible aujourd'hui
  - traitement automatique de données très volumineuses (entre 100 et 1 000 millions de mots) ;
  - transformations limitées (part d'interprétation à prendre en compte)
  - sous-ensemble des dépendances possibles (parsage partiel) mais en grand nombre
- influence sur les résultats : dépendances syntaxiques fines et formes les plus fréquentes / simples cooccurrences et formes les moins fréquentes (variation du nombre de traits disponibles)

Compromis finesse / nombre de traits

[Grefenstette 96]

- 400 000 mots de l'encyclopédie Grolier contenant un hyponyme d'*institution* dans WordNet
- 2 types de contextes
  - dépendances syntaxiques (Sextant)
  - fenêtre $\pm$ 10 mots : lemmes des N, V, A
- distance : Jaccard pondéré
- pierre de touche : succès quand un mot et son plus proche voisin relèvent de la même catégorie dans Roget
  - contextes syntaxiques : meilleurs pour les 600 mots les plus fréquents
  - contextes graphiques : meilleurs pour les mots moins fréquents
- Ecrémage des contextes syntaxiques $\leftrightarrow$ diminution des points de comparaison

Sous-langages et langue aujourd'hui

- les déferlements de données textuelles sous forme électronique rend plus cruciale la prise en compte de l'hétérogénéité de ces données (le Web n'est pas un corpus)
- peut-on vraiment associer des probabilités aux sélections en langue ?
- au sein d'un sous-langage, tous les types d'énoncés ne se prêtent pas à la découverte de grammaires sémantiques