Dégrouper les sens à partir d'attestations

B. Habert

LIMSI – CNRS & université Paris X – Nanterre

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)

On reconnaît un mot à ses fréquentations (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.

Approches

- Dispositifs pour repérer les zones denses dans un graphe des cooccurrences d'un mot
  - [Véronis 04]
`http://www.up.univ-mrs.fr/~veronis/pdf/2003-taln.pdf`
  - [Ferret 04]
`http://www.lpl.univ-aix.fr/jep-taln04/proceed/actes/taln2004-Fez/Ferret.pdf`

- Instruments
  - Partitionner les phrases employant un mot selon les traits qui les caractérisent
  - Manifester les grandes oppositions qui structurent les phrases employant un mot

Communautés de cooccurrents [V04] 1/4

- Graphe des cooccurrences entres les cooccurrents les plus importants d'un mot
  - Extraction de paragraphes comprenant un mot au sein de pages Web
  - Etiquetage morphosyntaxique (Cordial)
  - Filtrage
    - noms et adjectifs
    - anti-dictionnaires (mots outils, mots généraux du Web, mots de fréquence < 10)
    - plancher (5 cooccurrences)
    - pondération des arêtes (place dans les occurrences de chaque noeud)
    - conservation des associations fortes

Communautés de cooccurrents [V04] 2/4

L’ accalmie devrait permettre de récupérer d’ autres traces de fioul en mer , au moyen de barrages flottants tractés par des bateaux , dans le Pertuis breton , séparant l’ île de la côte du Marais poitevin.

accalmie, autre, trace, fioul, mer, barrage, flottant, bateau, pertuis, breton, île, côte

<accalmie, trace>, <accalmie, fioul>, <accalmie, mer>, <accalmie, barrage>, <accalmie, flottant>, <accalmie, bateau>, <accalmie, breton>, <accalmie, île>, <accalmie, côte>, <trace, fioul><trace, mer> . . . <île, côte>

Communautés de cooccurrents [V04] 3/4

- Repérage des « communautés »
  - coefficient de regroupement d'un noeud : propension plus ou moins forte de ses voisins immédiats à être voisins entre eux ($\Rightarrow$ homogénéité sémantique)
  - recherche du noeud le plus « regroupant » et destruction de ce noeud avec ses voisins : obtention d'une première « communauté », puis on continue de manière récursive

Communautés de cooccurrents [V04] 4/4

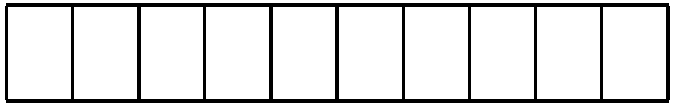
1.1	construction, ouvrage, rivière, projet, retenue, crue
1.2	véhicule, camion, membre, conducteur, policier, groupement
1.3	Algérie, militaire, efficacité, armée, Suisse, poste
1.4	vainqueur, victoire, rencontre, qualification, tir, football

Communautés de cooccurrents [F04] 1/3

- Corpus : 24 mois (1990–94) du journal *Le Monde*
- Fenêtre glissante de 20 mots où ne sont conservés que noms, verbes et adjectifs (lemmatisés), simples ou « en plusieurs mots »
- Restriction aux cooccurrences – attirances – les plus significatives (Information Mutuelle)
- Réseau de 23 000 « mots » et 5,2 millions de cooccurrences
- 2 mesures de similarité/distance entre deux « mots »
  - Information mutuelle immédiate entre les deux mots
  - distance (cosinus) en fonction des cooccurrents partagés ou non (pondération par l'information mutuelle)

Attirance entre mots 1/2

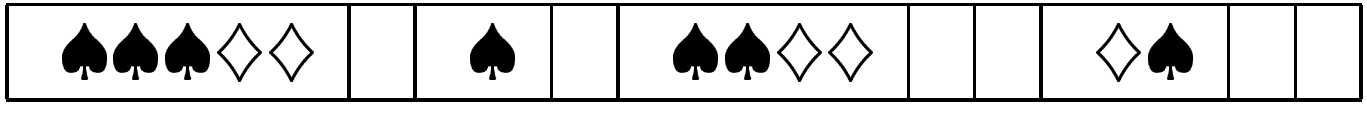
Information mutuelle

un texte 

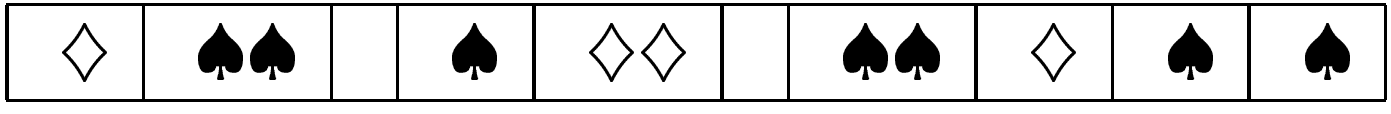
2 mots ♢♢♢♢ ♠♠♠♠♠♠♠♠

Configurations

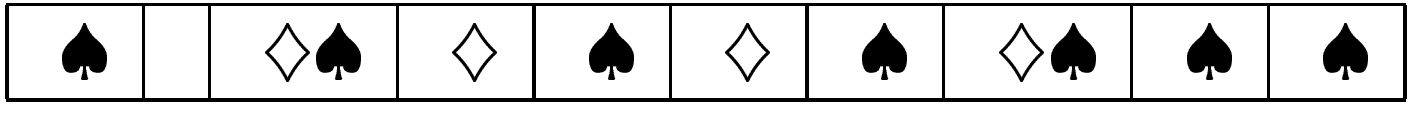
- attirance



- répulsion

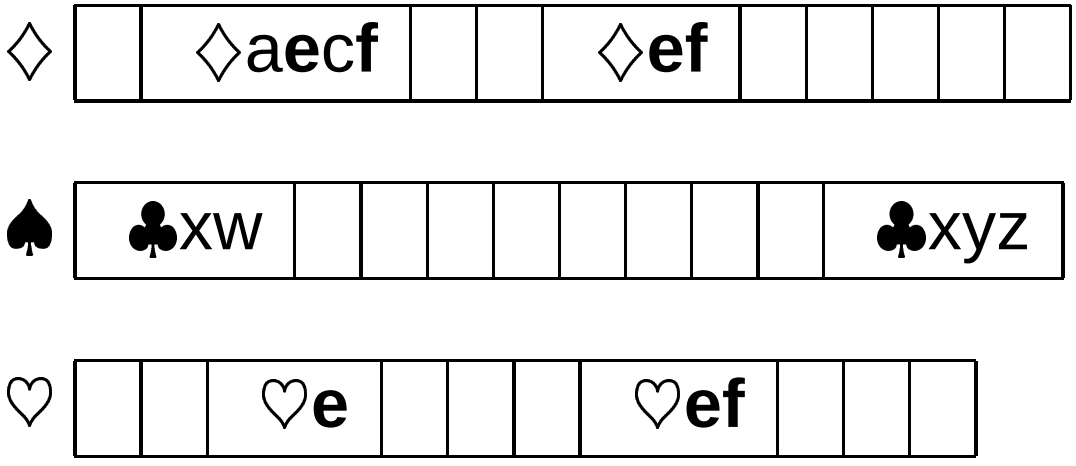


- indépendance



Attirance entre mots 2/2

Partager plus ou moins des voisins



♢ est proche de ♡ (ils partagent **e** et **f**)
♠ est éloigné de ♢ comme de ♡ (aucun partage de voisin)

Communautés de cooccurrents [F04] 2/3

Détection de zones de forte densité

- Recherche d'« embryons » (noyaux)
  - « éclaircissement » du graphe : conservation des k (= 15, ici) plus proches cooccurrents
  - force du lien entre 2 mots =nombre de voisins directs partagés
  - conservation au-delà d'un seuil (second « éclaircissement »)
- Rattachement des mots « orphelins » au noyau dont ils sont le plus proche

Communautés de cooccurrents [F04] 3/3

1.1	manifestant, forces_de_l'ordre, préfecture, agriculteur, protester, incendier, calme, pierre
1.2	conducteur, routier, véhicule, poids_lourds, camion, permis, trafic bloquer, voiture, autoroute
1.3	fleuve, lac, rivière, bassin, mètre_cube, crue, amont, pollution, affluent, saumon, poisson
1.4	blessé, casque_bleu, soldat, milicien, tir, milice, convoi, évacuer, croate, milicien, combattant
2.1	eau, mètre, lac, pluie, rivière, bassin, fleuve, site, poisson, affluent, montagne, crue, vallée
2.2	conducteur, trafic, routier, route, camion, chauffeur, voiture, chauffeur_routier, poids_lourds
2.3	casque_bleu, soldat, tir, convoi, milicien, blindé, milice, aéroport, blessé, incident, croate

Partitionner des phrases

- Corpus : LM10 (*Le Monde* 1991–2000)
- Instrument utilisé : Cope [Jardino 2004] partitionne un ensemble d'individus en classes en fonction des traits qu'ils présentent
- Individus : les 5 318 phrases comportant le lemme *barrage*
- Traits : approche dite « sac de mots » $\Rightarrow$ les « mots », les lemmes, les lemmes « pleins », les lemmes pleins sans noms propres
- Résultat : k classes (ici $k = 4$) et leurs traits significatifs

Phrases en *Barrage* dans LM10

- Etats
  - « nu » *L' accalmie devrait permettre de récupérer d' autres traces de fioul en mer , au moyen de* ***barrages*** *flottants tractés par des bateaux , dans le Pertuis breton , séparant l' île de la côte de le Marais poitevin.*
  - lemmatisé *le accalmie permettre de récupérer de autre trace de fioul en mer , au moyen de* ***barrage*** *flottant tracter par de le bateau , dans le pertuis breton , séparer le île de le côte de le Marais Poitevin.*
  - réduits aux lemmes « pleins » *accalmie permettre récupérer autre trace fioul mer* ***barrage*** *flottant tracter bateau pertuis breton séparer île côte Marais Poitevin*
  - idem sans noms propres *accalmie permettre récupérer autre trace fioul mer* ***barrage*** *flottant tracter bateau pertuis breton séparer île côte*

Classes de phrases : texte nu 1/3

1. tentait, RN, ont-été-levés, Abdelkader, a-été-tué, tué, Bouziane, Cisjordanie, CNIR, CRS, raffineries, Jérusalem, Jouques, tués, Hébron
2. faire-barrage, match, tour, candidat, Coupe, RPR, matches, socialistes, Davis, député, deuxièmes, élection, FN, meilleur, disputer, Le-Pen, pts, circonscription, quarts, vainqueur, fait-barrage, Ecosse, Fare, Lens, phase, Samoa, Euro, à-l'extrême, me, Mondial, office, Ukraine, Claude, disputeront, joueurs, Lionel, mercredi-20-octobre, participeront, poule
3. Gabcikovo, Danube, Euphrate, débit, Garonne, gravures, Gorges, Slovaquie, Charlas, milliard, Syrie, réservoir, interdépartementale, Birecik, million, différend, étiage, Lozère, Nagymaros, rupestres
4. Pacifique, Inga, Duras, Marguerite

Classes de phrases : texte nu 2/3

- Thèmes/acceptions
  - opposition politique : faire barrage à
  - obstacle physique à la circulation : barrages routiers ou policiers
  - ouvrage d'art : barrage hydraulique
  - opposition sportive : match de barrage
  - livre : barrage contre le Pacifique

Classes de phrases : texte nu 3/3

1. tentait, RN, ont-été-levés, Abdelkader, a-été-tué, tué, Bouziane, Cisjordanie, CNIR, CRS, raffineries, Jérusalem, Jouques, tués, Hébron
2. faire-barrage, match, tour, candidat, Coupe, RPR, matches, socialistes, Davis, député, deuxièmes, élection, FN, meilleur, disputer, Le-Pen, pts, circonscription, quarts, vainqueur, fait-barrage, Ecosse, Fare, Lens, phase, Samoa, Euro, à-l'extrême, me, Mondial, office, Ukraine, Claude, disputeront, joueurs, Lionel, mercredi-20-octobre, participeront, poule
3. Gabcikovo, Danube, Euphrate, débit, Garonne, gravures, Gorges, Slovaquie, Charlas, milliard, Syrie, réservoir, interdépartementale, Birecik, million, différend, étiage, Lozère, Nagymaros, rupestres
4. Pacifique, Inga, Duras, Marguerite

Classes de phrases : lemmes

1. Rn, finale, Davis, vainqueur, blesser, gendarmerie, Abdelkader, automobiliste, blinder, Irlande, championnat, Bouziane, Liban, Samoa, Cisjordanie, Cnir, crs, affronter, émeute, 96, Jouques, mercredi-20-octobre, Rennes, Ukraine, 17-décembre-1997, à-l'aube, Hébron, poule, slovène,
2. milliard, réservoir, bassin, Assouan, mètre-cube, hydraulique, retenue, affluent, Gabcikovo, Trois-gorges, Epala, coût, centrale, produire, Danube, capacité, dollar, gorge, irrigation, Euphrate, Cher, débit, Garonne, gravure, Petit-saut, Cnr, Charlas, Slovaquie, Naussac, artificiel, étudier, engloutir, mise-en-eau, saumon, inonder, interdépartemental, Rochebut, Serre, décennie, delta, étiage, potable, Villerest, archéologique, Birecik, slovaque, différend, en-1994, Lozère, Montluçon, Nagymaros, rupestre, thermique, turbine
3. candidat, Fn, circonscription
4. Duras, Marguerite

Classes de phrases : lemmes « pleins »

1. match, Rn, finale, Fntr, Davis, vainqueur, Belgique, blesser, intercepter, Abdelkader, automobiliste, blinder, Irlande, Angleterre, championnat, escargot, Alger, Bouziane, centre-ville, Ecosse, Samoa, Cisjordanie, Cnir, crs, enclave, Lens, affronter, heurt, irlandais, Uck, 96, Jérusalem, Jouques, mercredi-20-octobre, Rennes, Ukraine, 17-décembre-1997, Hébron, poule, slovène
2. Loire, candidat, milliard, Chambonchard, Rpr, Gabcikovo, Trois-gorges, Epala, centrale, produire, Danube, Serre-de-la-fare, Euphrate, Cher, débit, Haute-loire, Fn, royer, Garonne, gravure, Petit-saut, Cnr, financement, Charlas, circonscription, Slovaquie, Naussac, Tours, Udf, étudier, hongrois, utilité, Bratislava, interdépartemental, législatif, Rochebut, sécheresse, Serre, étiage, Villerest, archéologique, Birecik, Dumez, Lalonde, slovaque, traité, Brice, différend, impact, Lozère, Montluçon, Nagymaros, rupestre, thermique
3.
4. Duras, Marguerite

lasses de phrases : *id.* sans noms propres

1. match, policier, matin, filtrant, soldat, véhicule, serbe, disputer, palestinien, gendarme, finale, vainqueur, blesser, division, casque, gendarmerie, intercepter, automobiliste, bosniaque, circuler, blinder, conducteur, raffinerie, harki, week-end, championnat, escargot, gazer, patrouille, voyageur, char, centre-ville, crs, enclave, euro, affronter, balle, couvre-feu, émeute, irlandais, 96, mercredi-20-octobre, 17-décembre-1997, poule, slovène
2. candidat, électeur, circonscription
3. mètre-cube, irrigation, amont, poisson, nappe, engloutir, saumon, bas, inonder, potable, archéologique, rupestre, thermique, turbine
4.

Perspectives

- Objectif commun : dégrouper les sens
- Techniques
  - zones denses dans le graphe des cooccurrences
  - distance entre individus-phrases dans l'espace des *n* traits
- Données de test
  - similaires : contextes du mot *barrage*
  - opposition Web / presse « généraliste » (tranches différentes)
- Nature des résultats : acceptions / oppositions thématiques
- Généralisation/stabilisation ?
- Rapport « indistinct » avec une approche distributionnelle du sens