3.A. Explorations sémantiques

- Partir de l'idée que l'on ne comprend **RIEN** à un texte ancien
- Mettre en œuvre des procédures statistiques permettant de structurer les contextes de diverses manières, et en jouant sur les différences (entre catégories, surtout entre périodes)

3.B. Les distributions lexicales

- La statistique « classique » considère presque uniquement des distributions ayant une valeur centrale et une dispersion autour de cette valeur (distributions binomiales, de Laplace-Gauss, etc)
- Les distributions lexicales sont de forme hyperbolique : il n'y a pas de valeur centrale, le calcul de la dispersion (s'il est possible) est complexe
- Les distributions lexicales sont de type « parétien », elles sont difficiles à manipuler, mais il faut bien les connaître pour éviter des erreurs fatales...

3.C. Examen des distributions lexicales

- Langue et texte (ou corpus) : la langue est non-finie, non mesurable, le corpus est fini, mesurable
- Un corpus peut être considéré comme un échantillon de langue, mais il a des propriétés statistiques différentes de celle de la langue, et il est toujours biaisé par rapport à la population d'origine.
- Les paramètres numériques d'une distribution empirique (corpus) varient selon la taille, même si la langue sous-jacente est la même ! => les comparaisons sont très délicates, le seul moyen simple d'éviter les erreur est de comparer uniquement des ensembles de même effectif.
- Des procédures empiriques existent (paquet R), mais elles restent relativement mal maîtrisées

3.D. Contexte et structure

- Le mot isolé n'a pas de signification ! Les articles de dictionnaire comportent tous des définitions et des exemples = ensemble de contextes
- Le sens d'un mot est constitué par la structure de ses relations avec d'autres mots
- L'objectif de la sémantique numérique est donc de mettre en évidence cette structure
- Les procédures numériques sont la formalisation des méthodes des lexicographes ! = « étude des contextes »
- L'outil ESSENTIEL : la distance entre les mots = les cooccurrences
- NB : la simple distance néglige toute syntaxe ; empiriquement : ça marche !

3.E. Cooccurrences !!

- Facile de compter les mots autour d'un pivot
- => n'importe quoi !! on trouve une masse de mots sans rapport particulier avec le pivot
- Il faut impérativement **FILTRER les mots pertinents**
- Tout repose sur *le carré magique*
  fréquence de la cooccurrence
  fréquence totale du pivot
  fréquence totale du cooccurrent
  effectif total du corpus
- Des dizaines de manière de calculer un coefficient ont été proposées : une demi-douzaine courants (Dice, hyperbolique...), il faut faire des essais nombreux et examiner les résultats ; encore des progrès à faire
- Fonction intégrée à TXM et cqpweb, mais scripts ad hoc plus efficaces

3.F. Cooccurrences 2a

- Plusieurs types de comptage et d'exploitation, on peut en distingue au moins 4
- 1. la base : **cooccurrences simples** = une liste, éventuellement triée (selon le coefficient, ou selon tel effectif)
- 2. **cooccurrences structurées** : cooccurrences entre cooccurrents d'un même pivot => définit plus ou moins des types de contexte
  ** deux possibilités :
  A : cooccurrences dans les contextes
  B : cooccurrences dans le corpus
- Les résultats sont sous forme de tableau : visualisation structurée par analyse factorielle (ACP)
- Les 2 modes de calcul aboutissent à 2 tableaux, comparer (comment ?)

3.G Cooccurrences 2b

- 3. **cooccurrences prolongées** : rechercher l'ensemble des cooccurrents des cooccurrents d'un pivot
  on obtient une masse importante difficile à maîtriser
- 4. **cooccurrences généralisées** : calculer tous les cooccurrents de tous les lemmes, et récupérer, pour un lemme donné, les lemmes ayant le « profil » le plus voisin
  on peut limiter l'opération à un ensemble de lemmes prédéterminés, de manière à visualiser la structure de leurs relations
- Les cooccurrences généralisées (alias « sémantique distributionnelle ») sont coûteuses en espace disponible et en temps de calcul, mais elles sont ***très efficaces*** (paquet R disponible)

3.H. Évolutions

- Implique que le corpus utilisé comporte des indications chronologiques, ou que l'on compare des corpus ordonnés dans le temps
- Comparaisons toujours périlleuses du fait de la nature des distributions
- L'analyse chronologique est la voie royale de la sémantique historique et de la sémantique en général : on ne voit jamais mieux les structures que lorsqu'elles bougent
- Presque rien de disponible en matière de procédures : le développement des logiciels d'analyse de texte est très peu orienté vers la sémantique et pas du tout sur les aspects chronologiques

3.I. Courbes chronologiques

- Avantages nombreux : procédures éprouvées, nombreux paquets, résultats très suggestifs
- Une précaution absolue : ne comparer que des effectifs égaux, jamais de fréquences relatives calculées sur des effectifs différents (catastrophes garanties dans ce cas)
- Attention aux effets d'échelle (effectuer une anamorphose si nécessaire : à programmer)
- Nécessité, pour une série, de construire de nombreux graphiques, en modifiant la taille des tranches, et les méthodes de calcul des tendances.
- Un outil très intéressant (à programmer !) : l'évolution chronologique du coefficient de cooccurrence entre deux lemmes

3.J. Comparer deux listes

- Nombreuses variantes
- Le plus simple : deux listes parallèles triées
- Possibilité de feedback : calculer pour deux pivots les coefficients d'un ensemble de lemmes (éventuellement le total de deux listes simples) et trier selon les écarts
- Dans ce dernier cas, possibilité de graphique

3.K. Comparer plusieurs listes

- Rien de prévu pour cooccurrences simples !!
- Diverses possibilités : soit une analyse sur un tableau de fréquence (écarts à l'indépendance), soit construction d'un tableau de coefficients.
- Pour cooccurrences généralisées, AFC sur coefficients de distance au plus proche voisin (marche relativement bien)
- À développer….

3.L. Questions en suspens

- Cooccurrences structurées : on observe que, selon les lemmes, se forment des « nuages » de manière plus ou moins nette ; il y a donc des lemmes sans contextes différenciés, d'autres au contraire qui apparaissent dans des contextes différents (« sens » différents ?) => calculer un coefficient ? Repérer des sous-ensembles de cooccurrents par procédure ?
- Zones du lexique : mots fréquents, mots « médians » et mots rares n'ont pas les mêmes structures sémantiques. Adapter le calcul des coefficients de filtrage ? (facile pour le coefficient de Dice)
- Citations récurrentes : dans certains corpus, réapparition permanente de certaines citations ou de certaines formules => impact sur les cooccurrents => faut-il considérer qu'il s'agit d'un trouble à éliminer, ou d'une structure importante ?
- Liste non exhaustive….
- **La sémantique historique en est à la phase de ses débuts !**

Je vous remercie de votre attention

QUESTIONS ?