


Techniques quantitatives en HPE: la lexicométrie

*J.
Loulergue*

*RehPer
e*



La lexicométrie, qu'est-ce que c'est?

- Analyse quantitative des données textuelles.
 - Large corpus
 - Décompte d'occurrences et de cooccurrences, extraction des spécificités d'un texte, évolution du vocabulaire à travers le corpus,... Représentations graphiques de ces informations.
 - Exemple: Mayaffre (2014)
- 

Topic-modelling vs Lexicométrie

- Ex: Ambrosino et al. (2018); Malaterre et al. (2019); voir aussi Blei, (2012)
- Cooccurents fréquents
- Etiquetage
- Structure du corpus



Lexicométrie et histoire de la pensée

- Histoire des concepts (Koselleck – *Kritik und Krise*, 1959)
 - Un concept, plusieurs termes.
- Analyse de discours
 - Un terme, plusieurs concepts.



L'analyse de discours

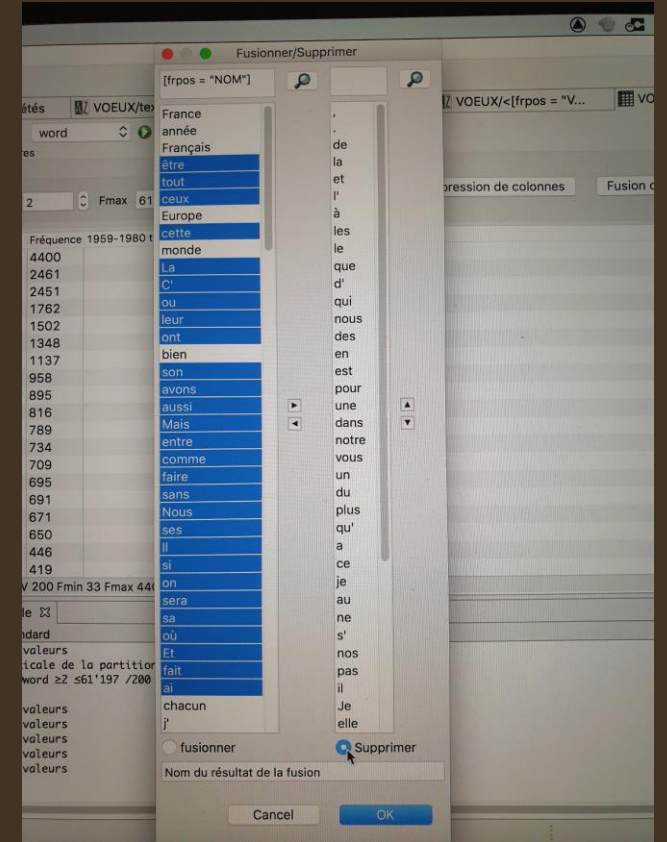
- Zellig S. Harris → Jean Dubois (1967)
- Michel Pêcheux, Régine Robin, Denise Maldidier, Jacques Guilhaumou.
- UMR Analyse de corpus linguistiques
- Revues: Mots, LINX, Cahiers de lexicologie, Semen.
- *Dictionnaire des usages socio-politiques (1992-2006)*, 8 fascicules

Analyse de discours en HPE

- Piguet (1999; 2003): Quesnay et le mot « production »
 - Jacob (2003): évolution du mot « travail » entre le XVIème et le XIXème siècle.
 - Guilhaumou (2003): Sieyès et la langue de l'économie politique
- La lexicométrie est un plus.

Au paradis du quanti?

- Forces et faiblesses des techniques quantitatives?
 - Cherrier & Svorenčik (2018)
 - Les choix en lexicométrie
 - Lemmatisation (TreeTagger): quelles formes comptent?
 - La « stop-list »: quels mots comptent?
 - Interrogation du corpus
- Méthode mixte Cherrier & Svorenčik (2018), Lamoreaux (2015), ou encore Rosenthal (2016).



Optical Character Recognition

- Importance de l'OCR

- ABBY FineReader

- <https://programminghistorian.org/en/lessons/working-with-batches-of-pdf-files>

- OCR et textes historiques

Quelques outils de lexicométrie

- Langage: «R»
- Logiciels
 - Voyant tools (en ligne)
 - Iramuteq
 - Alceste (payant)
 - TXM



UN PEU DE PRATIQUE...



Voyant Tools

- Prenez quelques PDF *lisibles* (OCR)
- Chargez-les sur Voyant Tools

<https://voyant-tools.org/>



TXM (1)

- Préparation du corpus:
 - Documents au format .txt
 - Métadonnées
 - Fichier en .csv
 - Structure: <id>, puis ce que vous voulez
 - Tout mettre dans un même dossier
 - Si corpus trop grand → sur Python: créer le fichier métadonnées automatiquement.
 - Attention au format d'encodage (UTF-8)
 - Pour changer l'encodage d'un texte: Sublime Text (File→save with encoding...)

TXM (2)

- Importation du corpus:

Dans TXM:

- Vérifier que TreeTagger est actif (Fichier → ajouter une extension → WordCloud, TreeTagger software, TreeTagger models.)
- Importation: Fichier → importer... → TXT + CSV
 - Pointer vers votre dossier.
 - Vérifier l'encodage.



TXM (3): Exploration

- Première description du corpus
 - Vérifier la lemmatisation
 - Infos générales
 - Commande lexicque
- Interrogation du corpus
 - Cooccurrences et mise en contexte
 - Requête conditionnelle
 - Partition du corpus
 - Calcul de spécificités

