


Séminaire « Intégrité et Partage de la Science : les données de la recherche », Grenoble, 12 décembre 2018.



# Statistique lexicale et données textuelles:

De la collecte à la visualisation,  
enjeux des données et des outils.

Emmanuel MARTY, GRESEC, Université Grenoble Alpes

# Plan de la présentation



- **La statistique lexicale:** méthodes, outils et principales fonctionnalités
- **Enjeux des données et des outils dans une recherche** sur la modération des commentaires dans les sites d'actualité en France
  - Collecte
  - Analyse et visualisation
  - Diffusion



# **La statistique lexicale: méthodes et principales fonctionnalités**

# **Pourquoi la statistique lexicale?**

**Analyse de discours** (enquêtes, analyse sur archives)

→ Pourquoi, dans l'univers des mots possibles, certains ont-ils été choisis, d'autres écartés ?

Lien avec la linguistique, les sciences de l'information et de la communication, les sciences politiques, mais usage croissant dans les sciences et techniques.

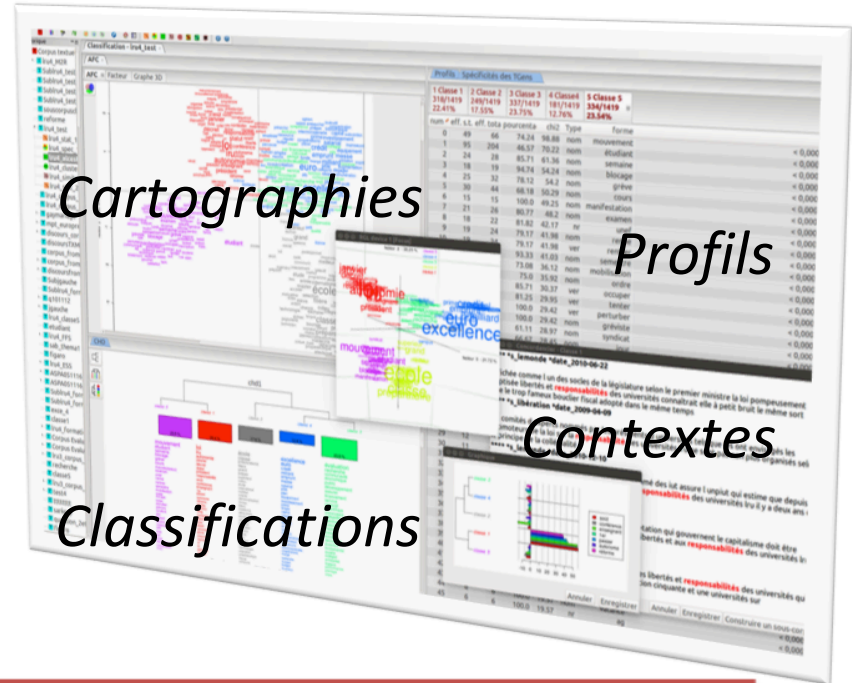
→ **Définir les mots récurrents**, leurs fonctions, leurs relations, leurs utilisations pour **reconstruire du sens**.

# Quelques logiciels de statistique lexicale

---

- **Alceste** ➤ M. Reinert (<http://www.image-zafar.com>)
- **Lexico 3** ➤ A. Salem (<http://lexico3.no-ip.org>)
- **Sphinx Lexica** ➤ Y. Baulac (<http://www.lesphinx-developpement.fr>)
- **Hyperbase** ➤ E. Brunet (<http://ancilla.unice.fr/>)
- **TXM** ➤ S. Heiden (<http://textometrie.ens-lyon.fr/>)
- **IRAMuTeQ** ➤ P. Ratinaud (<http://repere.no-ip.org/Members/logiciel/iramuteq>)

# Analyser des corpus de textes



Cartographies

Profils

Contextes

Classifications

Iramuteq (<http://www.iramuteq.org>) est développé par Pierre

Ratinaud au sein du *Lerass, Univ. Toulouse*

# Deux opérations de base

**Segmentation:** découpage de suites de caractères bornées par deux caractères délimiteurs (=occurrence ou word-tokens). Deux suites identiques constituent deux occurrences d'une même forme graphique (word-type).

→ Délimiteurs: espace, retour à la ligne, [(« ,.;?!'/\_ \_ »)]

Eventuellement, étape de lemmatisation : rassemblement des formes lexicales liées par une racine commune

12528 de	1195 c	530 sera	341 ai	233 développement
8324 la	1188 je	528 doit	323 travail	231 économie
6211 l	1183 ne	527 aussi	310 entre	229 deux
5815 et	1127 par	509 ont	306 si	227 enfin
5217 les	1117 ce	494 français	297 économique	226 encore
4908 le	1074 sur	479 y	290 aujourd	226 temps
4631 à	985 qu	462 j	290 hui	222 ensemble
4435 des	908 france	453 etat	288 dont	221 vie
3832 d	855 s	447 sans	283 sociale	220 société
3051 est	838 aux	434 ou	282 on	219 depuis
2982 en	838 n	425 comme	280 seront	216 ceux
2799 que	816 nos	422 ces	278 monde	215 donc
2441 une	810 gouvernement	422 tout	278 république	210 toutes
2425 nous	803 avec	421 son	266 fait	209 soit
2273 qui	744 mais	413 avons	265 loi	208 droit
2142 un	711 elle	410 ses	265 où	208 sécurité
2060 pour	697 cette	409 même	264 contre	207 ainsi
2024 du	695 vous	406 été	263 leurs	206 elles
1977 dans	693 politique	400 faire	262 action	206 moyens
1809 il	667 se	390 ils	256 europe	203 cet
1410 au	651 être	386 faut	243 effort	202 autres
1393 notre	647 sont	375 entreprises	241 peut	202 cela
1368 plus	633 leur	362 emploi	236 nationale	199 mesures
1275 pas	603 pays	346 bien	235 avenir	197 jeunes
1214 a	533 tous	342 sa	235 président	195 croissance



12528 de	1195 c	530 sera	341 ai	233 développement							
8324 la	1188 je	528 doit	323 travail	231 économie							
6211 l	1183 ne	527 aussi	310 entre								
5815 et	1127 par	509 ont	306 si	13006pre	de	1147pre	ou	449con	savoir	312ver_sup	
5217 les	1117 ce	494 français	297 écon	8806art_def	la	1133pre	ces	449adj_dem	république	307nom	
4908 le	1074 sur	479 y	290 aujour	6744aux	être	965con	falloir	444ver_sup	aujourd'hui	302adv_sup	
4631 à	985 qu	462 j	290 lui	6474art_def	l	956ver_sup	vouloir	440ver_sup	année	297nom	
4435 des	908 france	453 etat	288 dont	6100con	et	948nr	comme	438con	on	294pro_per	
3832 d	855 s	447 sans	283 socia	5481art_def	les	892pro_per	tout	433pro_ind	moyen	292nom	
3051 est	838 aux	434 ou	282 on	5134art_def	le	885adj_pos	son	432adj_pos	européen	290adj	
2982 en	838 n	425 comme	280 sero	4684pre	à	884ver_sup	pouvoir	432adj	dont	290pro_rel	
2799 que	816 nos	422 ces	278 mon	4569art_ind	des	881art_def	aux	431adj_pos	demander	277ver	
2441 une	810 gouvernement	422 tout	278 répu	3779pre	d	863pre	avec	411pro_per	économie	276nom	
2425 nous	803 avec	421 son	266 fait	3271aux	avoir	849nom	gouvernement	399adj	monde	276nom	
2273 qui	744 mais	413 avons	265 loi	3091pre	en	811nom	politique	396adj	contre	273pre	
2142 un	711 elle	410 ses	265 où	2818pro_rel	que	779con	mais	390adj	société	270nom	
2060 pour	697 cette	409 même	264 cont	2587pro_per	nous	722adj_dem	cette	383ver	où	270pro_rel	
2024 du	695 vous	406 été	263 leurs	2534art_ind	une	709pro_per	elle	379ver	leurs	270adj_pos	
1977 dans	693 politique	400 faire	262 actio	2390pro_rel	qui	702adj	français	379nom_sup	europe	270nr	
1809 il	667 se	390 ils	256 euro	2239art_ind	un	695pro_per	se	376nom_sup	mesure	266nom	
1410 au	651 être	386 faut	243 effor	2135pre	pour	694adj	social	374nom	agir	266ver	
1393 notre	647 sont	375 entreprises	241 peut	2135adv_sup	ne	686pro_per	vous	374adj	an	263nom_sup	
1368 plus	633 leur	362 emploi	236 natio	2111art_def	du	666pro_per	leur	357adj_pos	président	258nom	
1275 pas	603 pays	346 bien	235 aven	2069pre	dans	638nom	pays	355ver	aller	257ver	
1214 a	533 tous	342 sa	235 prés	1857pro_per	il	552adv_sup	aussi	351nom	objectif	251nom	
				1487adj_pos	notre	546pro_ind	tous	347pro_ind	jeune	245adj	
				1417adv_sup	plus	543nom	emploi	328nom	engager	245ver	
				1386art_def	au	525nom	entreprise	320con	projet	244nom	
				1360adv_sup	pas	517adj	grand	318nom	avenir	244nom	
				1271pro_per	je	503pro_per	y	318pre	temps	243nom	
				1244ver_sup	devoir	493pro_per	j	318nom	service	241nom	
				1177pro_dem	c	476nr	etat	313ver	réforme	241nom	
				1170pro_dem	ce	467pre	sans	313ver_sup	assurer	241ver	

# Deux opérations de base



**Partition:** La statistique mesure des différences

- Comparer les parties d'un tout (modalités de variables) pour tester des hypothèses
- Pour des entretiens: comparaison du lexique H/F, Age, CSP
- Pour un corpus médiatique: comparer le discours des plusieurs médias, l'évolution dans le temps, etc.

# Tableau lexical



**Parties:** variable-s et modalités

**Formes lexicales:**

Mots issus de la  
segmentation et  
lemmatisation



- Nombre d'occurrences

# Tableau lexical



	LHumanité	LaCroix	LeFigaro	LeMonde	LePoint	Libération	NouvelObs
yougoslavie	1	4	3	2	0	1	0
considérable	13	10	21	4	3	8	1
controversé	5	6	14	5	3	3	0
sgen	4	2	0	5	0	1	0
garraud	7	3	0	1	0	0	0
originalité	2	9	7	2	2	2	0
chine	10	10	24	5	3	8	1
naturel	18	25	34	12	10	7	1
controverse	5	15	16	8	0	4	0
mener	81	68	80	53	17	46	2
projection	9	10	2	4	0	1	0
sensibilisati	5	4	2	1	0	0	0
constitutif	2	7	5	2	0	1	0
lnder	0	14	0	0	0	0	0
commentaire	15	25	16	16	3	4	0
radicalisati	4	10	14	9	0	0	0
rythme	8	21	5	5	5	1	0
bienvenu	3	2	1	1	1	2	0
cependant	37	86	66	29	12	18	0
tisser	4	8	2	1	0	3	1
souci	17	51	20	19	6	6	1
prétendu	7	3	4	5	0	0	0
défiler	6	10	18	14	5	12	0
peur	41	85	85	44	12	38	5





# **Collecte, analyse et visualisation des données**

La modération des commentaires sur les sites  
d'actualité français: le cas des attentats de Paris  
le 13 novembre 2015 (Smyrnaiois & Marty 2017)

# Questions de recherche et protocole



Quel est le cadre socio-économique du travail de modération, quel est l'**impact des conditions de travail sur ce qui est rendu visible** aux internautes?

→ Enquête par entretiens et observations sur place

→ Analyse d'un corpus de commentaires soumis à modération.

## Questions opérationnelles:

- Quels sont les thèmes principaux des commentaires ?
- Y a-t-il une typologie de commentaires acceptés et refusés?
- Y a-t-il des erreurs de modération et/ou des stratégies de contournement?



# **Collecte des données et constitution du corpus**





# Collecte des données et constitution du corpus

Travail de mise en forme et de codage pour traitement par le logiciel

→ Modification de la mise en forme initiale des données: structure, étiquetage, etc.

```
1 **** *s_depmid *dh_13nov_2024 *nomcontrib_love all *typemod_COM_APRIORI *raisonmod_null *resultmod_ACCEPTED
2 très bonne nouvelle!!
3
4 **** *s_depmid *dh_13nov_2024 *nomcontrib_666 *typemod_COM_APRIORI *raisonmod_null *resultmod_ACCEPTED
5 Un déchet humain de moins
6
7 **** *s_depmid *dh_13nov_2024 *nomcontrib_v2010 *typemod_COM_APRIORI *raisonmod_null *resultmod_REFUSED
8 1 de tué, 50 de convertis, ça nous mène où tout ça ?
9
10 **** *s_depmid *dh_13nov_2024 *nomcontrib_lardon3leretour *typemod_COM_APRIORI *raisonmod_null *resultmod_REFUSED
11 Bon débarras !!!! Au tour des autres, maintenant
12
13 **** *s_depmid *dh_13nov_2024 *nomcontrib_JeanMouclade *typemod_COM_APRIORI *raisonmod_null *resultmod_REFUSED
14 "modass, il y a 4 heures Je ne vois pas quel droit il avait : c'était un combattant, il est mort en combattant. ----
15
16 **** *s_depmid *dh_13nov_2024 *nomcontrib_newday *typemod_COM_APRIORI *raisonmod_null *resultmod_REFUSED
17 J'en suis tout retourné, c'est affreux et au suivant.
18
19 **** *s_depmid *dh_13nov_2024 *nomcontrib_Kajuu81 *typemod_COM_APOSTERIORI *raisonmod_null *resultmod_ACCEPTED
20 "les commentaires vont aller bon train ; sur l'ennemi intérieur, les étrangers qui viennent manger nos ressources et nous :
21
22 **** *s_depmid *dh_13nov_2024 *nomcontrib_MOZAMOZA *typemod_COM_APOSTERIORI *raisonmod_null *resultmod_REFUSED
23 J'voudi Il n'est pas question ce soir de dignité. Il est question de haine. Qui de haine envers tous ces assassins :
24
25 **** *s_depmid *dh_13nov_2024 *nomcontrib_lardon3leretour *typemod_COM_APOSTERIORI *raisonmod_null *resultmod_ACCEPTED
26 et malgré ça, la censure trouve encore le moyen de museler les Français !!!!
27
28 **** *s_depmid *dh_13nov_2024 *nomcontrib_monterosso *typemod_COM_APRIORI *raisonmod_null *resultmod_REFUSED
29 il serait souhaitable que ceux qui ont l'intention de jeter de l'huile sur le feu s'abstiennent de commentaires iniques.
30
31 **** *s_depmid *dh_13nov_2024 *nomcontrib_Marcello31 *typemod_COM_APRIORI *raisonmod_null *resultmod_REFUSED
32 Benzéma ???
```

# Collecte des données et constitution du corpus



## ENJEUX DE LA COLLECTE

Corpus « exogène »

→ Le chercheur n'est pas maître des conditions de sa constitution

→ Restrictions dans l'usage:

- \* les titres de presse n'ont pas souhaité être cités nommément

- \* la publication dans une revue de certains commentaires refusés (incitation à la haine, etc.) a fait l'objet d'arbitrages

→ Les chercheurs ne sont pas « propriétaires » des données, on leur a concédé un « droit d'exploitation » (sous conditions), donc délicat de les mettre à disposition pour d'autres.

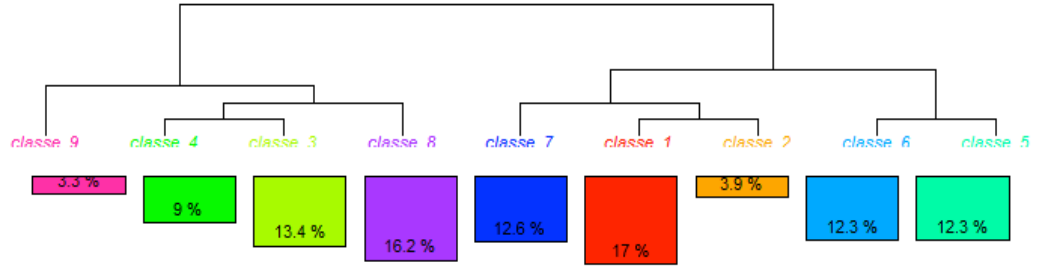
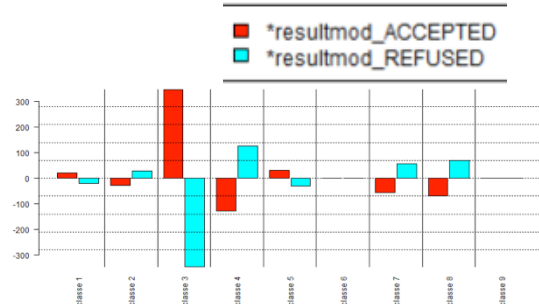


# **Analyse et visualisation des données: l'importance des outils**

# Classification hiérarchique descendante (CHD) et analyse factorielle des correspondances (AFC)

CHD (Reinert 1983) : identifier et représenter le lexique d'un corpus sous forme de « classes lexicales », i.e. clusters de mots significativement employés ensemble (sur la base d'un tableau de contingence)

29017 commentaires,  
1 017 978 occurrences,  
25943 formes lexicales



hybea juju88 abs6 breizmagic lol juju bonsoir mandarin upsa sternette salah avo gdr5949 abdeslam bise louis tybreiz bubulle35 jpmary christiandl jean cul georges62480 jeanlucokosar1064 coucou jacouille52 pov foxy	blanc religion femme musulman voilé amalgame barbu noir rouge bleu manger porc mère dieu garçon rue tuer frère voisin gentil catholique promener race père musique abruti	victime famille pensée marseillais hommage cœur courage solidarité respect triste minute silence chanter beau soutien vie horreur perdre tristesse penser bougie condoléance parisien monde deuil bonjour fier peur	rtl commenta lire radio censure journaliste censurer propos émission info répondre message vérité forum censeur poirette beet entendre modérateur site écrire télé question bobo pub expression gaucho probaande	frontière mosquée ficher nationalité fermer contrôle prison expulser imam surveiller français condamner vérité forum expulsion loi double fermeture belgique salafistes sol droit supprimer prêcher retour naître connaître peine	arme guerre pays syrie armée soldat daesh militaire civil irak moyen envoyer pétrole france bombarder armes bombe daech russe force ennemi islamique frapper terroriste ordre état utiliser	migrant accueillir réfugié rentrer europe centaine entrer infiltrer accueil millier interroger arriver terre jeune caler bombe libanais fiot homme terroriste cheval ouvert rapport ue merkel septembre humain trois bun	responsat sarkozy politique marine gauche socialiste laxisme hollande coupable peuple ps fn sarko président voter pen républicain situation droite national incompétence ministre gouverner incompétent dique partir	hollande chef valls janvier prendre mesure élection gouvernement charlie français régional incapable guerre président cazeneuve démission discours décision hebdo mois démissionner prochain urgence mr attendre impitoyable taubira concrète
---	--	--	---	--	---	--	---	--



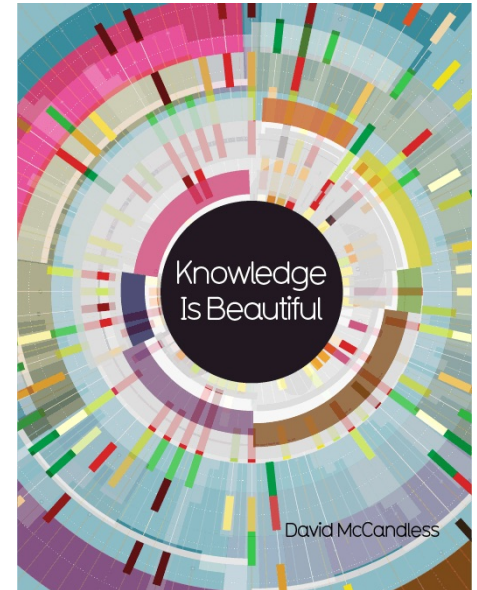
# Analyse et visualisation des données



## ENJEUX DE L'ANALYSE ET DE LA VISUALISATION

### LIES A LA VISUALISATION ELLE-MEME

- Produit des effets heuristiques liés à des dimensions sémiotiques voire esthétiques
- Les modes de visualisation reposent sur des principes et des postulats porteurs de schémas interprétatifs différents (effet de coupure dans la CHD, contiguïté de l'AFC)



# Analyse et visualisation des données

## ENJEUX DE L'ANALYSE ET DE LA VISUALISATION LIES AUX OUTILS ET AUX RESSOURCES ASSOCIEES

- Disposer d'outils logiciels permettant des paramétrages fins et transparents. Logiciel libre (pas juste gratuit): éviter l'effet « boîte noire » des algorithmes propriétaires (Rieder & Rohle, 2010) par l'ouverture du code source (accessible sur Github, par exemple)

Ex: Projet [Digital Methods Initiative](#)

- Disposer d'une documentation technique universitaire sur la nature même des fonctionnalités et calculs mobilisés, présence d'une communauté universitaire d'utilisateurs (liste iramuteq users) permettant l'échange de ressources (corpus, dictionnaires de lemmatisation, bouts de scripts, etc.)







# **Diffusion des recherches et réappropriation des données**

# Diffusion et réappropriation des données

## DES PAS IMPORTANTS...

- une communauté d'usage des logiciels pour se former et échanger (URFIST, mailing list, RSN)
- une communauté scientifique attachée à l'accès ouvert (hors bouquets payants, etc.): Actes des Journées d'Analyse des Données Textuelles (JADT)
- communauté des développeurs universitaires qui travaille à l'**interopérabilité** des outils et corpus: faciliter l'export vers d'autres outils d'un même corpus



# Diffusion et réappropriation des données



...MAIS ENCORE DU CHEMIN A PARCOURIR

- Les corpus (données brutes ou retraitées) devraient pouvoir être accessibles, mais restrictions de droits (y compris sur des bases de données presse type Europresse ou Factiva)
- Nécessité de consigner les protocoles d'analyse étape par étape dans un vrai plan de gestion des données avec:
  - « carnet de bord » des modifications de fichiers,
  - paramètres logiciels appliqués (pour reproductibilité)
  - mise à disposition et diffusion des fichiers de résultats d'analyse: question de l'espace de stockage et des modalités de partage

# **Pour conclure**

**Des conditions et dispositifs de recherche souvent imparfaits du point de vue de l'intégrité et du partage des données:**

**→ penser en amont la gestion des données**

**→ favoriser la dimension collective ou communautaire de la recherche à toutes les étapes (collecte, analyse et traitement, publication, diffusion/évaluation)**