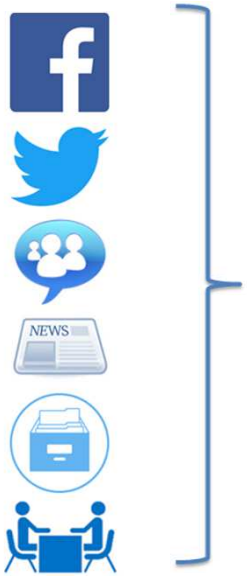
Journées d'étude @Media : cycle de l'information et Digital Methods.
Grenoble 30-31 mai 2018

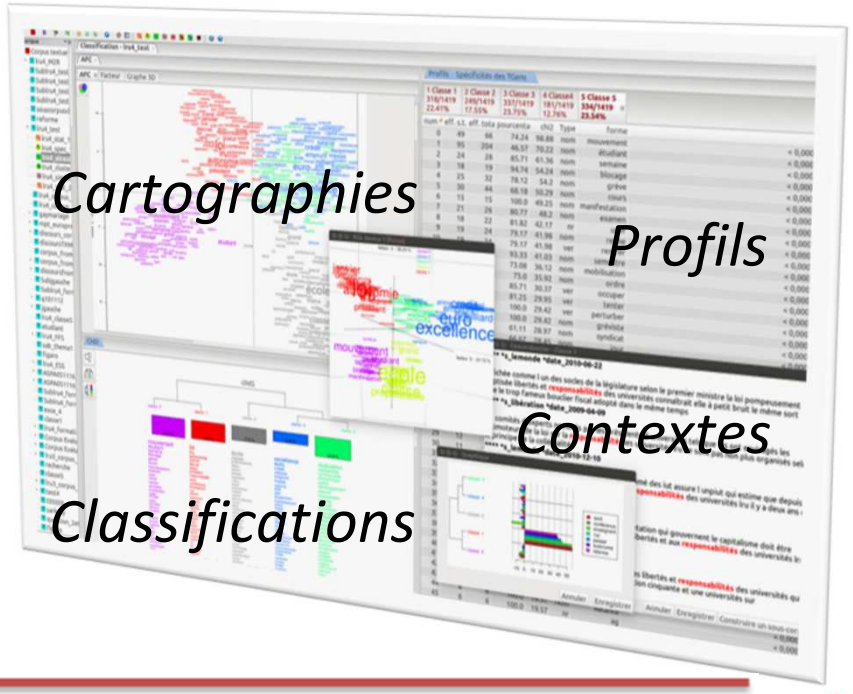
L'analyse lexicométrique des discours en ligne :

méthodes et exemples d'application aux médias et espaces d'expression sur le web

Pascal MARCHAND, LERASS, Université Toulouse 3
Emmanuel MARTY, GRESEC, Université Grenoble Alpes

Analyser des corpus de textes





Iramuteq (http://www.iramuteq.org) est développé par Pierre Ratinaud au sein du *Lerass* et dans le cadre du *LabEx SMS*.

Plan de la présentation

- **Principes de base de la statistique lexicale**
  Segmentation, partition, tableau lexical

- **Questions de recherche et résultats d'analyses: une typologie**
  Mesurer et caractériser la diversité des expressions
  Comparer plusieurs corpus
  Caractériser un processus

Principes de base de la statistique lexicale

Segmentation et partition: deux opérations constitutives de la statistique lexicale

Quelques définitions:

Les questions que se donne la statistique lexicale sont les suivantes : « quels sont les textes les plus semblables en ce qui concerne le vocabulaire et la fréquence des formes utilisées ? Quelles sont les formes qui caractérisent chaque texte, par leur présence ou leur absence ? » (Lebart & Salem, 1994).

→ Tableau lexical (formes * textes)

La lexicométrie regroupe “ toute une série de méthodes qui permettent d’opérer des ré-organisations formelles de la séquence textuelle et des analyses statistiques portant sur le vocabulaire à partir d’une segmentation ” (Salem, 1986)

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
8324 la	1188 je	528 doit	323 travail
6211 l	1183 ne	527 aussi	310 entre
5815 et	1127 par	509 ont	306 si
5217 les	1117 ce	494 français	297 écon
4908 le	1074 sur	479 y	290 aujo
4631 à	985 qu	462 j	290 hui
4435 des	908 france	453 etat	288 dont
3832 d	855 s	447 sans	283 socia
3051 est	838 aux	434 ou	282 on
2982 en	838 n	425 comme	280 sero
2799 que	816 nos	422 ces	278 mon
2441 une	810 gouvernement	422 tout	278 répu
2425 nous	803 avec	421 son	266 fait
2273 qui	744 mais	413 avons	265 loi
2142 un	711 elle	410 ses	265 où
2060 pour	697 cette	409 même	264 cont
2024 du	695 vous	406 été	263 leurs
1977 dans	693 politique	400 faire	262 actio
1809 il	667 se	390 ils	256 euro
1410 au	651 être	386 faut	243 effor
1393 notre	647 sont	375 entreprises	241 peut
1368 plus	633 leur	362 emploi	236 natio
1275 pas	603 pays	346 bien	235 aven
1214 a	533 tous	342 sa	235 prés

233	développement
231	économie

de	13006 *pre*	**par**	1147 *pre*	**ou**	449 *con*	**savoir**	312 *ver_sup*		
la	8806 *art_def*	**sur**	1133 *pre*	**ces**	449 *adj_dem*	**république**	307 *nom*		
être	6744 *aux*	**qu**	965 *con*	**falloir**	444 *ver_sup*	**aujourd_hui**	302 *adv_sup*		
l	6474 *art_def*	**faire**	956 *ver_sup*	**vouloir**	440 *ver_sup*	**année**	297 *nom*		
et	6100 *con*	**france**	948 *nr*	**comme**	438 *con*	**on**	294 *pro_per*		
les	5481 *art_def*	**s**	892 *pro_per*	**tout**	433 *pro_ind*	**moyen**	292 *nom*		
le	5134 *art_def*	**nos**	885 *adj_pos*	**son**	432 *adj_pos*	**européen**	290 *adj*		
à	4684 *pre*	**pouvoir**	884 *ver_sup*	**public**	432 *adj*	**dont**	290 *pro_rel*		
des	4569 *art_ind*	**aux**	881 *art_def*	**ses**	431 *adj_pos*	**demander**	277 *ver*		
d	3779 *pre*	**avec**	863 *pre*	**ils**	411 *pro_per*	**économie**	276 *nom*		
avoir	3271 *aux*	**gouvernement**	849 *nom*	**économique**	399 *adj*	**monde**	276 *nom*		
en	3091 *pre*	**politique**	811 *nom*	**premier**	396 *adj*	**contre**	273 *pre*		
que	2818 *pro_rel*	**mais**	779 *con*	**national**	390 *adj*	**société**	270 *nom*		
nous	2587 *pro_per*	**cette**	722 *adj_dem*	**mettre**	383 *ver*	**où**	270 *pro_rel*		
une	2534 *art_ind*	**elle**	709 *pro_per*	**prendre**	379 *ver*	**leurs**	270 *adj_pos*		
qui	2390 *pro_rel*	**français**	702 *adj*	**monsieur**	379 *nom_sup*	**europe**	270 *nr*		
un	2239 *art_ind*	**se**	695 *pro_per*	**bien**	376 *nom_sup*	**mesure**	266 *nom*		
pour	2135 *pre*	**social**	694 *adj*	**travail**	374 *nom*	**agir**	266 *ver*		
ne	2135 *adv_sup*	**vous**	686 *pro_per*	**nouveau**	374 *adj*	**an**	263 *nom_sup*		
du	2111 *art_def*	**leur**	666 *pro_per*	**sa**	357 *adj_pos*	**président**	258 *nom*		
dans	2069 *pre*	**pays**	638 *nom*	**permettre**	355 *ver*	**aller**	257 *ver*		
il	1857 *pro_per*	**aussi**	552 *adv_sup*	**effort**	351 *nom*	**objectif**	251 *nom*		
notre	1487 *adj_pos*	**tous**	546 *pro_ind*	**même**	347 *pro_ind*	**jeune**	245 *adj*		
plus	1417 *adv_sup*	**emploi**	543 *nom*	**action**	328 *nom*	**engager**	245 *ver*		
au	1386 *art_def*	**entreprise**	525 *nom*	**si**	320 *con*	**projet**	244 *nom*		
pas	1360 *adv_sup*	**grand**	517 *adj*	**loi**	318 *nom*	**avenir**	244 *nom*		
je	1271 *pro_per*	**y**	503 *pro_per*	**entre**	318 *pre*	**temps**	243 *nom*		
devoir	1244 *ver_sup*	**j**	493 *pro_per*	**droit**	318 *nom*	**service**	241 *nom*		
c	1177 *pro_dem*	**etat**	476 *nr*	**donner**	313 *ver*	**réforme**	241 *nom*		
ce	1170 *pro_dem*	**sans**	467 *pre*	**dire**	313 *ver_sup*	**assurer**	241 *ver*		

Tableau lexical

Parties (variable-s / segments de textes)

Lexique :

- Tokenization
- Reconnaissance
- Lemmatisation
- Statuts statistiques

- Nombre d’occurrences
- Présence / absence

Questions de recherche et résultats d'analyses: une typologie

A. Mesurer et caractériser la diversité des expressions et les controverses en ligne

B. Comparer plusieurs corpus, repérer convergences et divergences de représentations, opinions et/ou idéologies dans différents espaces

C. Caractériser un processus ou l'évolution diachronique d'un discours

A. Mesurer et caractériser la diversité des expressions et les controverses en ligne

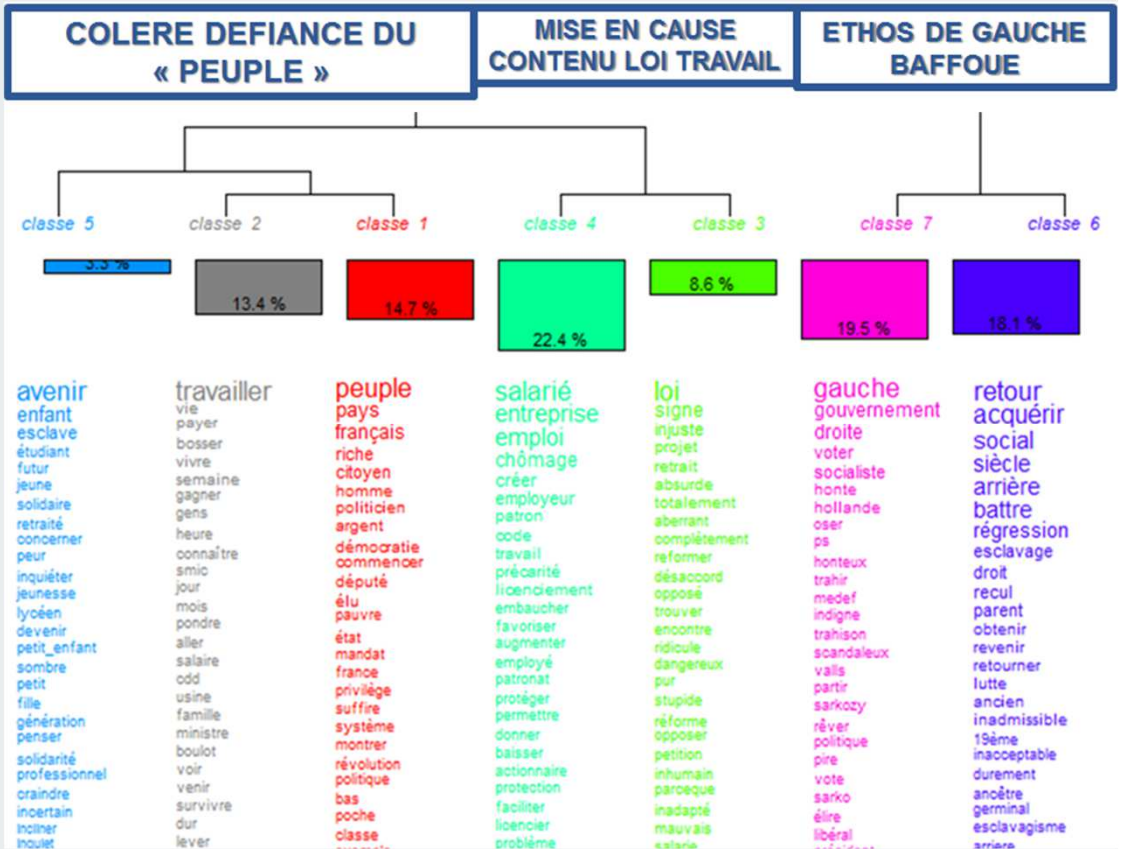
Analyse des commentaires de pétitions

Bousquet, Marty, Smyrnaios (2017)

-286 836 commentaires du 19 fév au 31 mars ("Je signe parce que"…)
6 597 624 occurrences, 57797 formes lex.

→ **Diversité des expressions du mécontentement:**

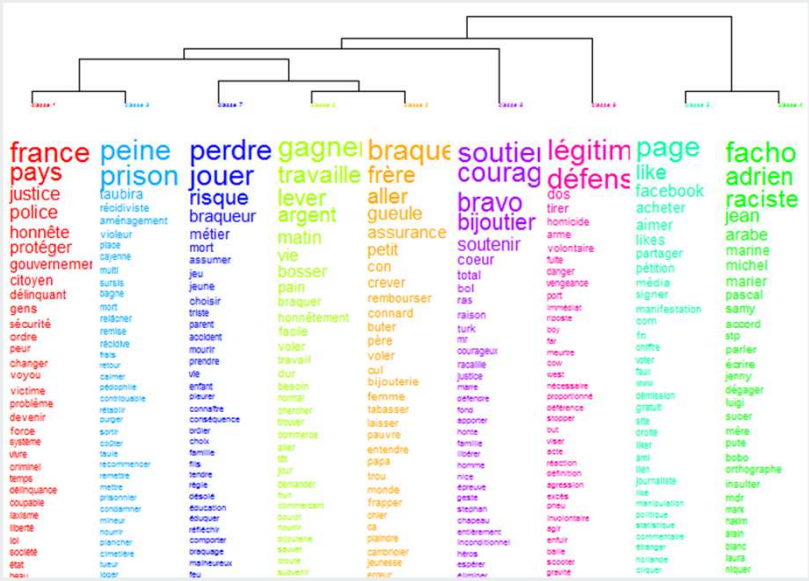
* **pathos:** colère du "peuple" contre les élites
* **logos :** mise en cause argumentée de la loi Travail
* **ethos:** choix politique mettant en danger l'histoire et les valeurs sociales de la France

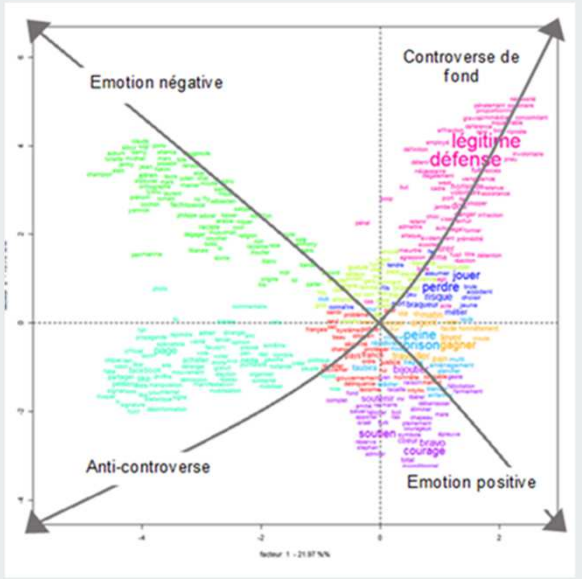


Analyse *Facebook*

Près de 4 millions d'occurrences (167000 textes ; 62000 formes ; 45000 lemmes après correction / homogénéisation)

Ex : Honnete, honnête, honnéte, honneté, hônete, honete, honête, honnêté, jonnête, hônnête, hônnète, honnëte, honnêt, honnètte, honnaite, honnete, honète, honneté, hônnete, honête, honnêté, honnêt, honnétte, onette, etc…





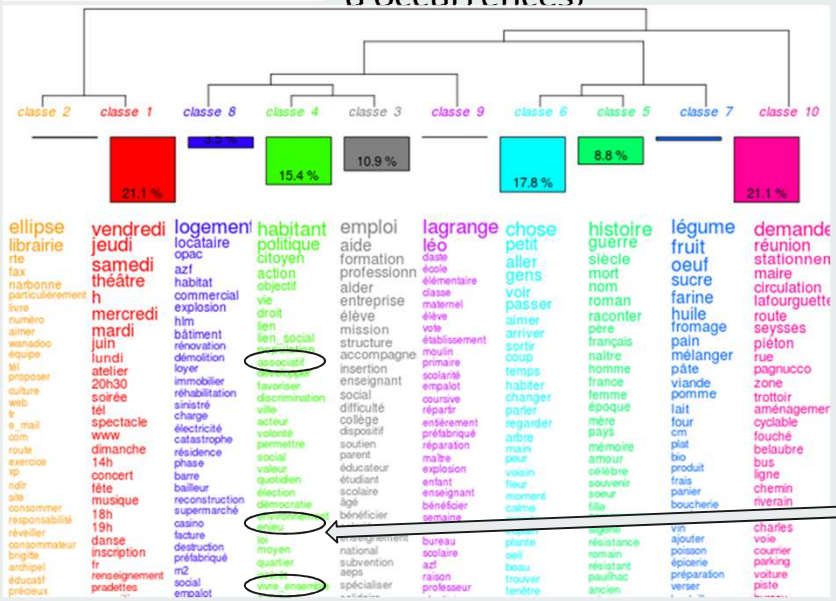
Alexia DUCOS, 2018

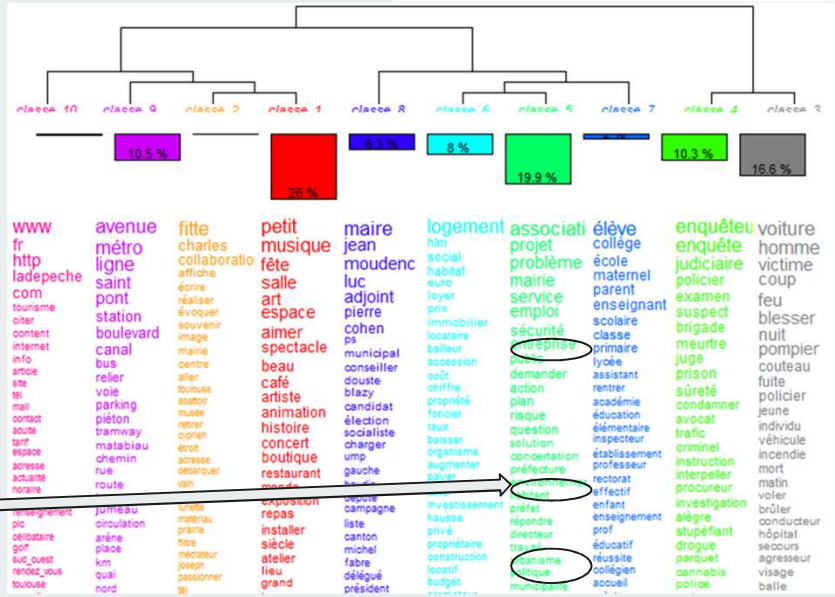
B. Comparer plusieurs corpus, repérer convergences et divergences

Différents médias

Natacha Souillard, 2018

Médias associatifs (10 journaux et bulletins de quartier : 2,6 millions occurrences) / PQR (4000 articles : 2,4 millions d'occurrences)





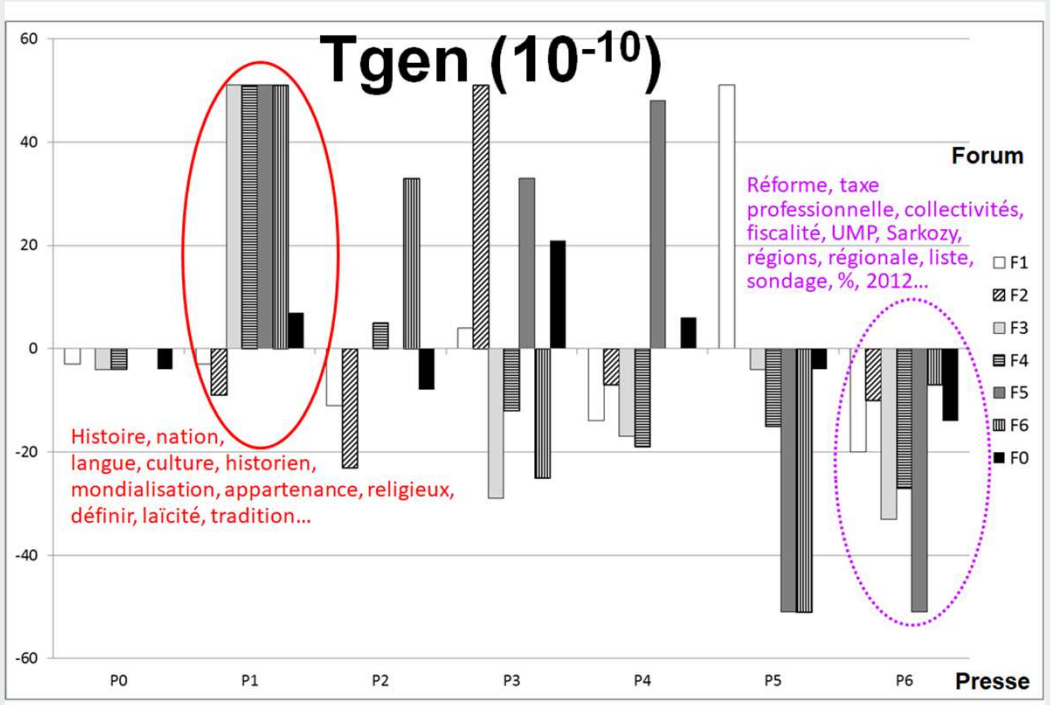
Le débat sur l'identité nationale

Forum : 18240 contributions

- 6 classes -> *Tgens*

Factiva : 1436 articles de presse

- 6 classes -> projection



(d'après Marty, Marchand & Ratinaud, *BMS*, 2012)

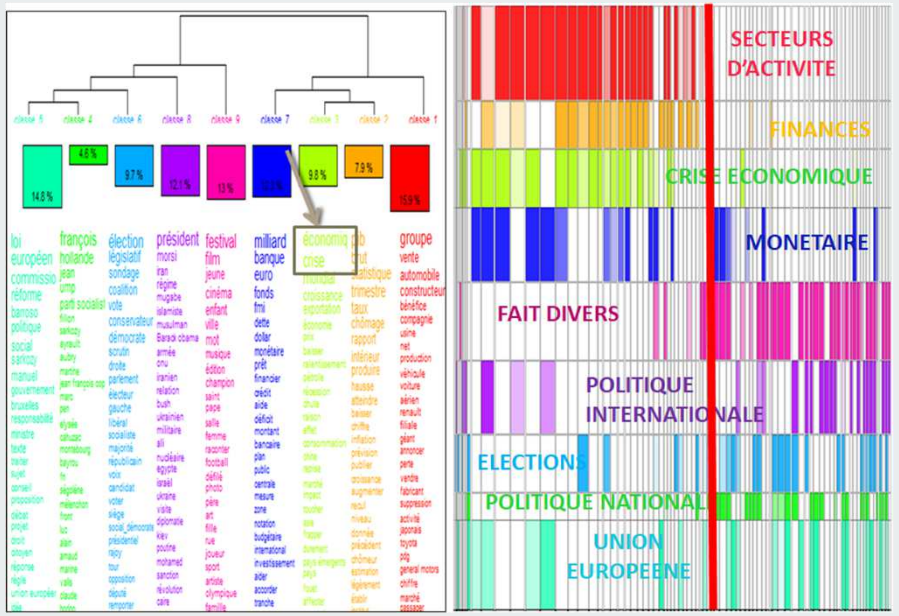
C. Caractériser un processus ou l'évolution diachronique d'un discours

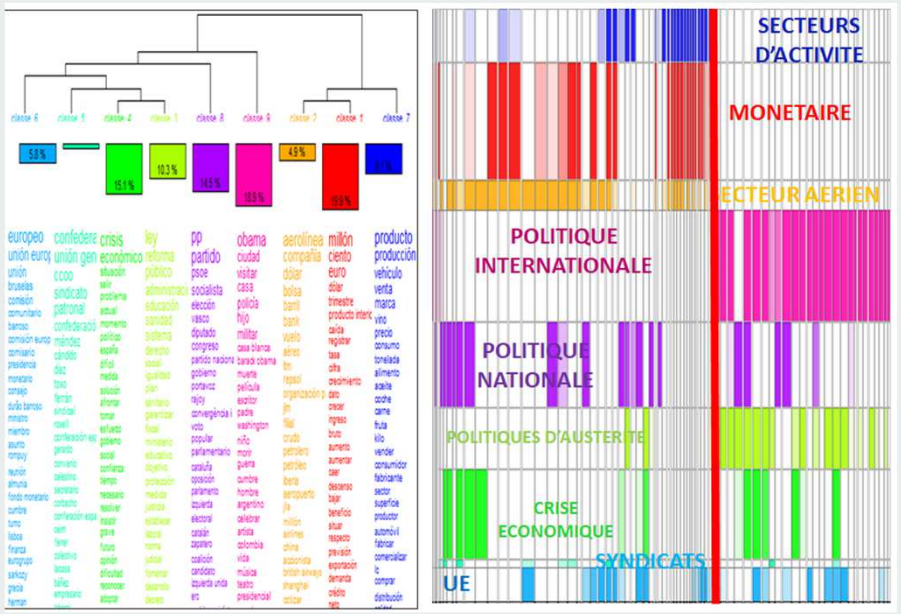
La construction médiatique de la crise

Mariola Moreno, 2018

17 850 textes, 7 962 140 occurrences, 66 579 formes

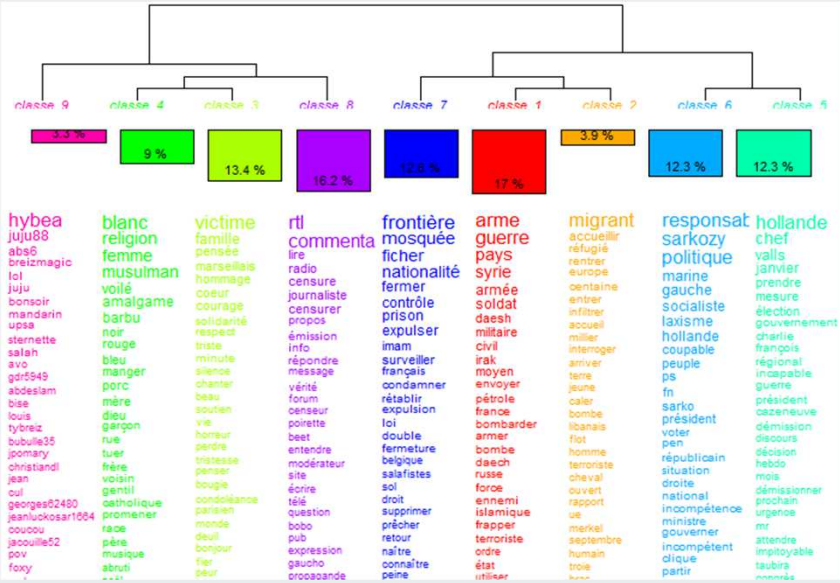
38 147 textes, 14 183 078 occurrences, 100 473 formes





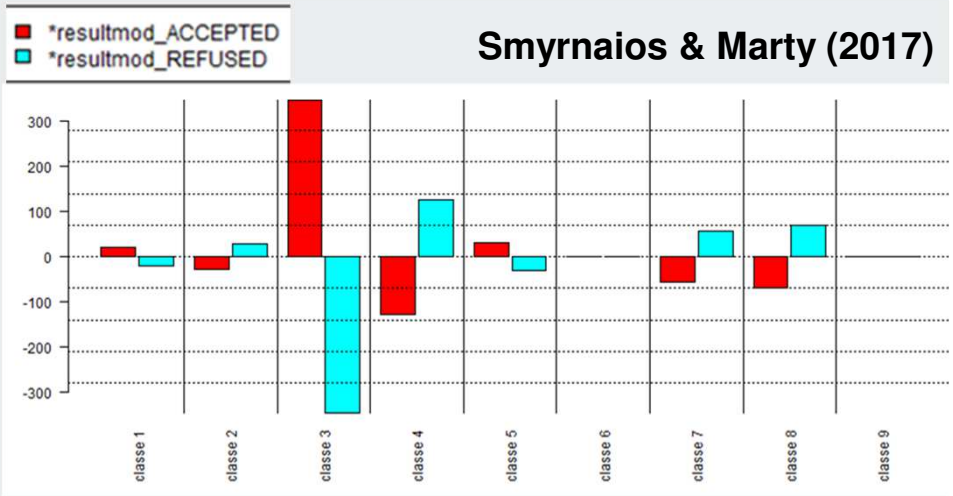
La modération de commentaires d'actualité

Commentaires apposés entre le 13 et le 16 novembre au bas d'articles traitant des attentats du 13 novembre, sur trois sites d'actualité, modérés en sous-traitance par la société Atchik. 29017 commentaires, 1 017 978 occurrences, 25943 formes lexicales



3 attitudes dans le processus de modération:
- refus immédiat (commentaires violents et/ou racistes)
- acceptation évidente (messages compassionnels et/ou descriptifs)
- examen minutieux et arbitrage (commentaires délicats, possiblement insultants ou diffamants)

Smyrnaios & Marty (2017)



Pour conclure:

Des outils statistiques pouvant être mobilisés pour différents contextes et corpus, et pour des questions de recherche différentes:

- La CHD comme première étape d'identification et de caractérisation d'une diversité d'expressions
- L'AFC pour identifier des distances/proximités, des polarisations et/ou des logiques structurantes
- La comparaison de profils de classe, facilitée par l'export de Types Généralisés pour mettre en regard des corpus différents, voire hétérogènes
- La mobilisation de variables indépendantes (ou méta-données) présidant à la partition pour identifier les liens significatifs entre éléments textuels et extra-textuels (locuteur ou source, temporalité, situation, caractérisation exogène du discours, etc.)

Pour conclure:

Statut de la statistique dans l'analyse de discours:

- administration d'une preuve statistique de l'organisation matérielle d'un discours (identité lexicale, spécificités, liens significatifs à certaines variables ou méta-données)

- dimension heuristique des indicateurs pour la reconstruction d'hypothèses interprétatives sur la signification des significativités, au regard des hypothèses et corpus mobilisés