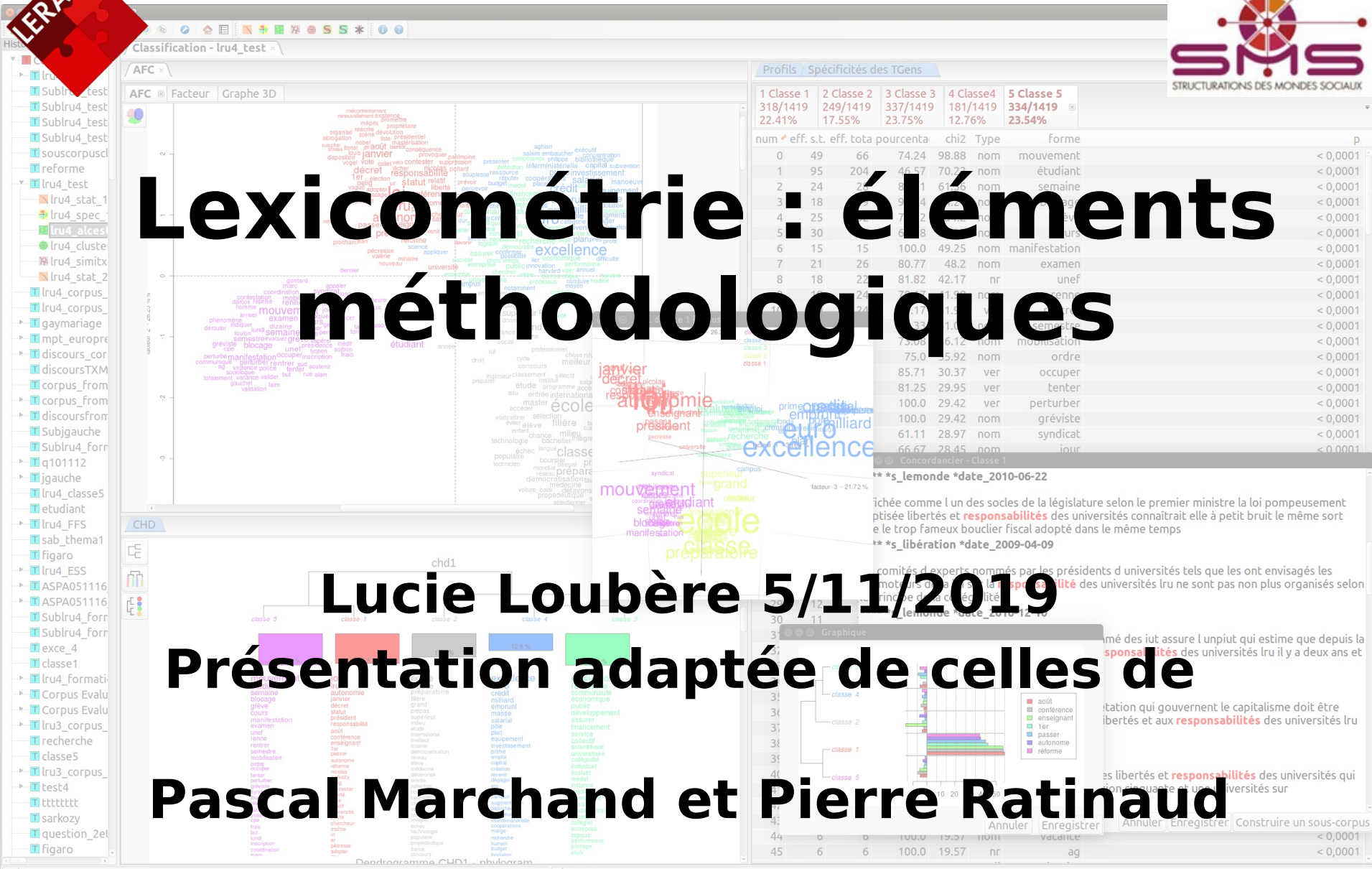


Lexicométrie : éléments méthodologiques



Lucie Loubère 5/11/2019

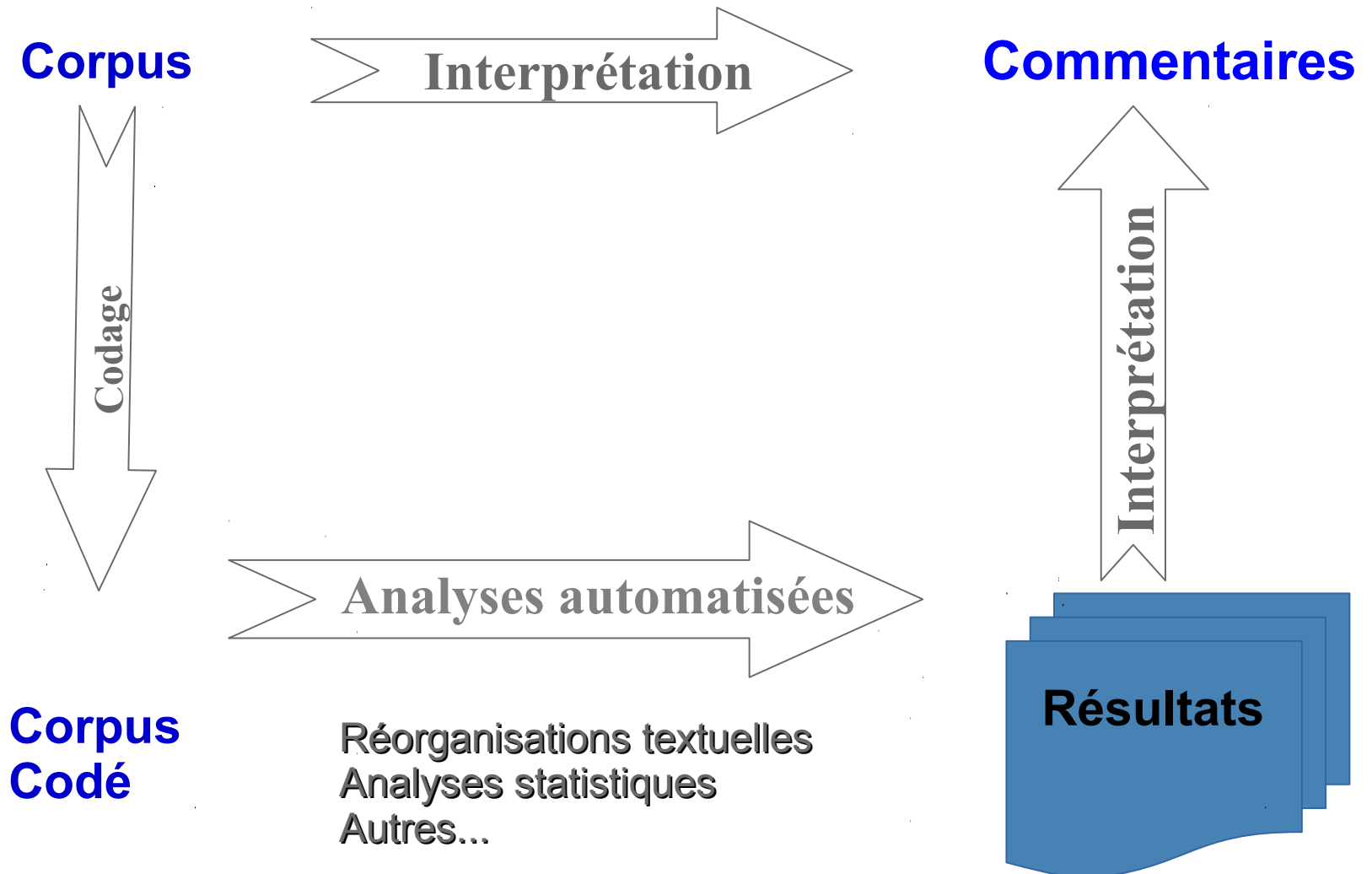
Présentation adaptée de celles de

Pascal Marchand et Pierre Ratinaud

Quelques précautions d'usage...

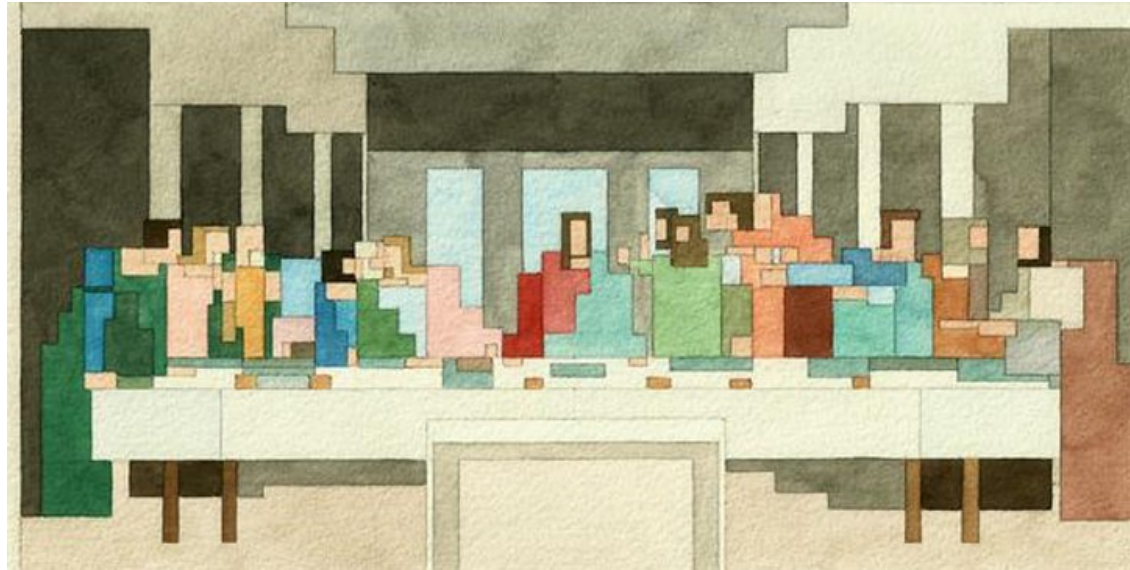
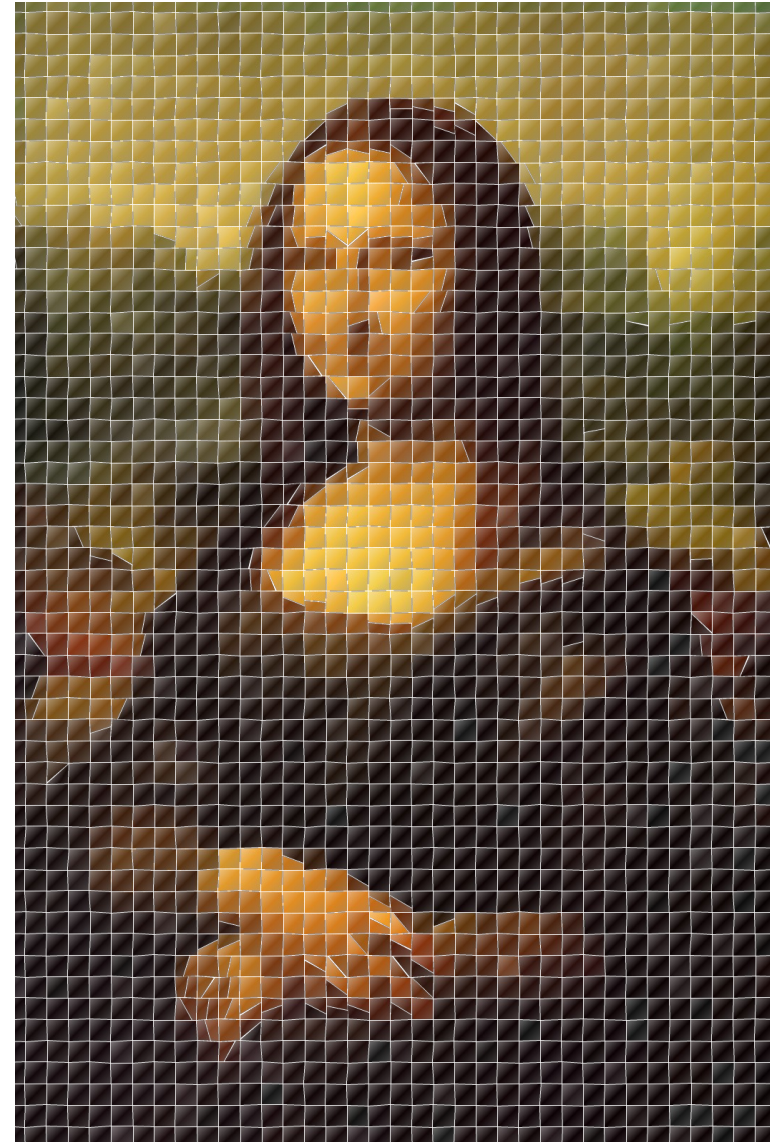
- ✓ La statistique n'est pas la seule approche possible ...
- ✓ *Aux chiffres, on leur fait dire n'importe quoi !*
- ✓ La statistique ne peut pas tout faire: il faut prévoir ce qu'on lui demandera :
HYPOTHÈSES
- ✓ Ce n'est pas la statistique qui garantit la qualité d'une recherche, mais le protocole.

L'interprétation en ADT



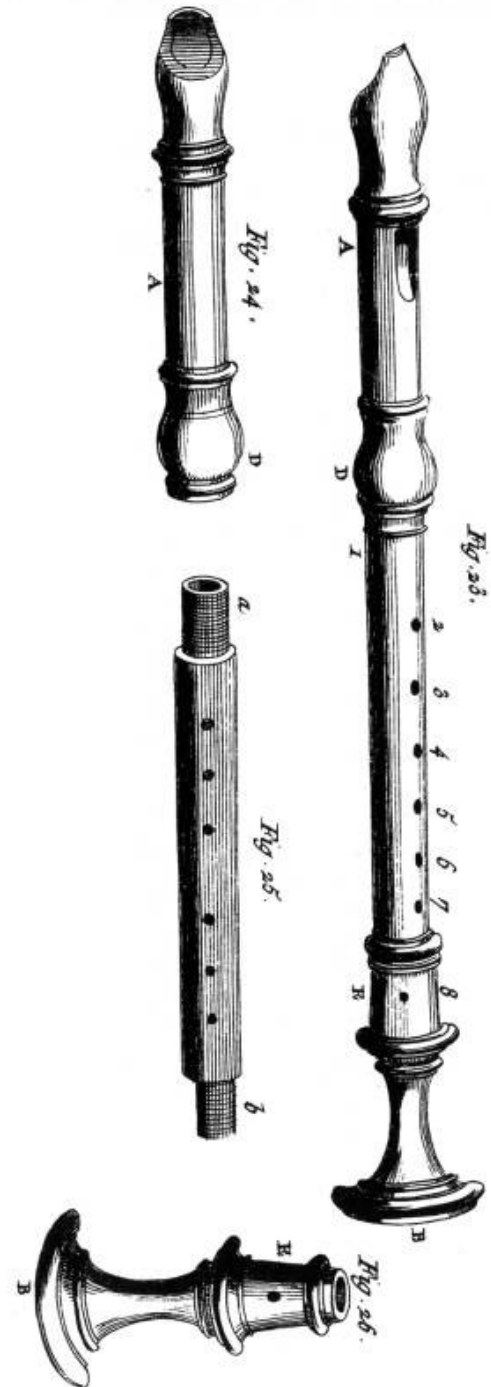
Merci à André Salem

Les résultats... une portion de l'objet de départ



Lebart & Salem (1994)

« Supposons (...) que l'on étudie les histogrammes des longueurs d'ondes correspondant aux couleurs d'un tableau de Rembrandt (pour chacun des pixels d'une reproduction). Il va de soi que l'on utilise une fraction dérisoire de l'information contenue dans l'image d'origine. Il est cependant possible que la forme de l'histogramme (ou d'une fonction plus élaborée des mêmes mesures et données de base) permette de distinguer un Rembrandt d'un Rubens ou d'un Van Dyck » (p.21).



Quelques logiciels de lexicométrie

- **Alceste** ➤ M. Reinert (<http://www.image-zafar.com>)
- **DtmVic** ➤ L. Lebart (<http://lebart.org>)
- **Hyperbase** ➤ E. Brunet
(<http://ancilla.unice.fr/~brunet/pub/hyperbase.html>)
- **Lexico 3** ➤ A. Salem (<http://lexico3.no-ip.org>)
- **SPAD** ➤ Decisia / M. Bécue (<http://www.spad.eu>)
- **Sphinx Lexica** ➤ Y. Baulac (<http://www.lesphinx-developpement.fr>)
- **Taltac** ➤ S. Bolasco (<http://www.taltac.it/it/index.shtml>)
- **TXM** ➤ S. Heiden (<http://textometrie.ens-lyon.fr/>)

- **IRAMuTeQ** ➤ P. Ratinaud (Win, Mac, Linux)
(<http://www.iramuteq.org>)

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, p.135).

- **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)

Analyse lexicale : 1 - *Tokenization*

Analyse lexicale : 1 - *Tokenization*

- Une suite de caractères bornée par deux caractères délimiteurs est une **occurrence** (*word-tokens*) : *Taille*.
 - espace, retour à la ligne, [(« ,.;?:!'/- _ »)]
- Deux suites identiques constituent deux occurrences d'une même **forme graphique** (*word-type*) : *Index*
- Normes de saisie (Labbé, 1990)
 - <http://halshs.archives-ouvertes.fr/docs/00/43/71/50/PDF/LabbeNormes.pdf>

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

Formes initiales / réduites

Lemmatisation

- On cherche à réduire les déclinaisons d'un mot en les ramenant à leur racine :
 - ✓ les verbes → infinitif
 - ✓ Les noms et adjectifs → masculin singulier
- Exemple :

Le petit chat **est** mort, c'**est** dommage
il **était** sympa le chat

13 occurrences
10 formes

Le petit chat **être** mort, c **être**
dommage il **être** sympa le chat

13 occurrences
9 formes

Formes initiales / réduites

Lemmatisation

➤ Comment juger de la catégorie grammaticale d'un mot ?

➤ Dictionnaire à étiquettes.

pariçots	pariçot	adj	m	p	0.4	1.22	0.14	0.14	
parions	parier	ver			81.96	15.61	0.38	0.2	imp:pre:lp;ind:imp:lp;ind:pre:lp;
paris	pari	nom	m	p	26.61	12.57	12.68	7.97	
paris_brest	paris_brest	nom	m			0.19	0.2	0.19	0.2
parisien	parisien	adj	m	s	2.73	30.61	0.94	12.09	
parisienne	parisien	adj	f	s	2.73	30.61	1.11	10.07	
parisiennes	parisien	adj	f	p	2.73	30.61	0.16	2.3	
parisiens	parisien	nom	m	p	2.73	30.61	1.36	6.40	

➤ **Autres solutions :**

TreeTagger - a language independent part-of-speech tagger

<http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>

Cordial Analyseur

<http://www.synapse-fr.com/>

Lexique 3 (Paris 5)

<http://www.lexique.org/>

Analyse lexicale : 2 – Partition / segmentation

Analyse lexicale : 2 – Partition / segmentation

- Le corpus = l'ensemble des textes (dans un fichier)
- Le texte (uci dans alceste) = un entretien, un chapitre, un livre, discours...
- Le segment de texte (uce dans alceste) = une portion du texte la taille est choisie par le chercheur
- Chacune de ces délimitations est choisie en fonction des éléments à étudier et des hypothèses du chercheur.

Les textes...

Le corpus

**** *art_444 *00_05_cq *libération *quotidien *autres *2004 *moyen
il faudra un vrai courage politique pour que l'art retrouve la place que
l'éducation nationale lui avait accordée. l'art à l'école, voie de démocratie
d'jian jean_michel pour ceux qui sont traversés par le doute quant aux
vertus de l'éducation artistique à l'école, le dernier film de gérard jugnot les
choristes tombe à pic. jamais le cinéma ne rendra un tel hommage à cette
pratique, d'autant que l'histoire est vraie, comme l'est, d'une autre
manière, celle de ces jeunes de banlieues qui, dans l'esquive, le film
d'abdelatif kechiche mettent en scène marivaux dans le jeu de l'amour et
du hasard.

...

**** *art_445 *00_05_cq *libération *quotidien *arts_cul *2004 *moyen
annoncée moribonde, la scène française n'a pas dit son dernier mot. la
preuve au printemps de bourges, qui s'ouvre aujourd'hui. le rap bouge
encore binet stéphanie a la sortie de l'album revoir un printemps en
septembre, les marseillais d'iam portaient sur leurs épaules tous les
espoirs du rap français. après l'explosion des ventes en 1998, la
médiatisation nationale via la radio skyrock, le rap français devient à l'entrée
du millénaire médiocre, uniforme, enfermé dans ses clichés matérialistes
machos racailleux.

04/11/11

ED0255X - Diogo Botelho

2 textes

Les segments de textes ...

**** *art_444 *00_05_cq *libération *quotidien *autres *2004 *moyen

il faudra un vrai courage politique pour que l'art retrouve la place que l'éducation nationale lui avait accordée. l'art à l'école, voie de démocratie
djian jean_michel pour ceux qui sont traversés par le doute quant aux vertus de l'éducation artistique à l'école, le dernier film de gérard jugnot les choristes tombe à pic. jamais le cinéma ne rendra un tel hommage à cette pratique, d'autant que l'histoire est vraie, comme l'est, d'une autre manière, celle de ces jeunes de banlieues qui, dans l'esquive, le film d'abdelatif kechiche mettent en scène marivaux dans le jeu de l'amour et du hasard.

...

**** *art_445 *00_05_cq *libération *quotidien *arts_cul *2004 *moyen

annoncée moribonde, la scène française n'a pas dit son dernier mot. la preuve au printemps de bourges, qui s'ouvre aujourd'hui. le rap bouge encore binet stéphanie a la sortie de l'album revoir un printemps en septembre, les marseillais d'iam portaient sur leurs épaules tous les espoirs du rap français. après l'explosion des ventes en 1998, la médiatisation nationale via la radio skyrock, le rap français devient à l'entrée du millénaire médiocre, uniforme, enfermé dans ses clichés matérialistes
mohes rasiloux

4
Segments
de Texte

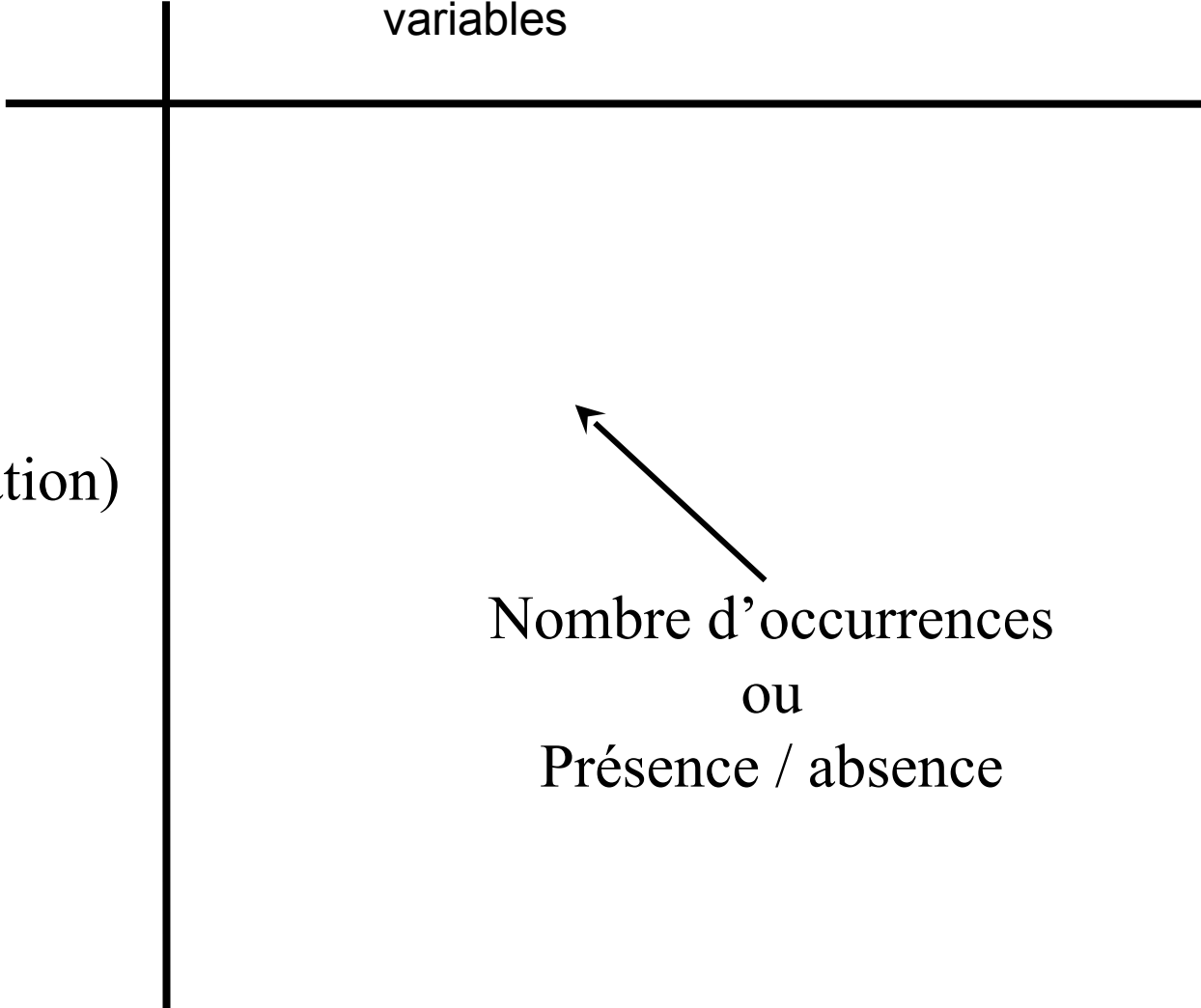


Tableau lexical *formes* * *parties*

Segments
Ou
variables

Lexique:
(tokenization / lemmatisation)

Nombre d'occurrences
ou
Présence / absence



Avec un exemple...

	*source la croix	*source le figaro	*source le monde	*source lhumanité	*source libération
sécurité	15	46	28	43	15
devenir	33	27	35	17	40
établissement	1	7	12	4	3
tirer	13	4	14	6	12
pencher	0	4	3	5	2
brexit	5	7	4	0	3
reporter	3	5	4	0	0
évolution	4	13	7	3	4
chine	1	5	2	0	2
connaissance	1	2	7	5	1
naturel	0	2	3	1	4
attente	4	5	8	3	3
unir	11	9	15	5	11

Analyse lexicale : 3 - Statistiques

La classification lexicale



Corpus
(classe 0)

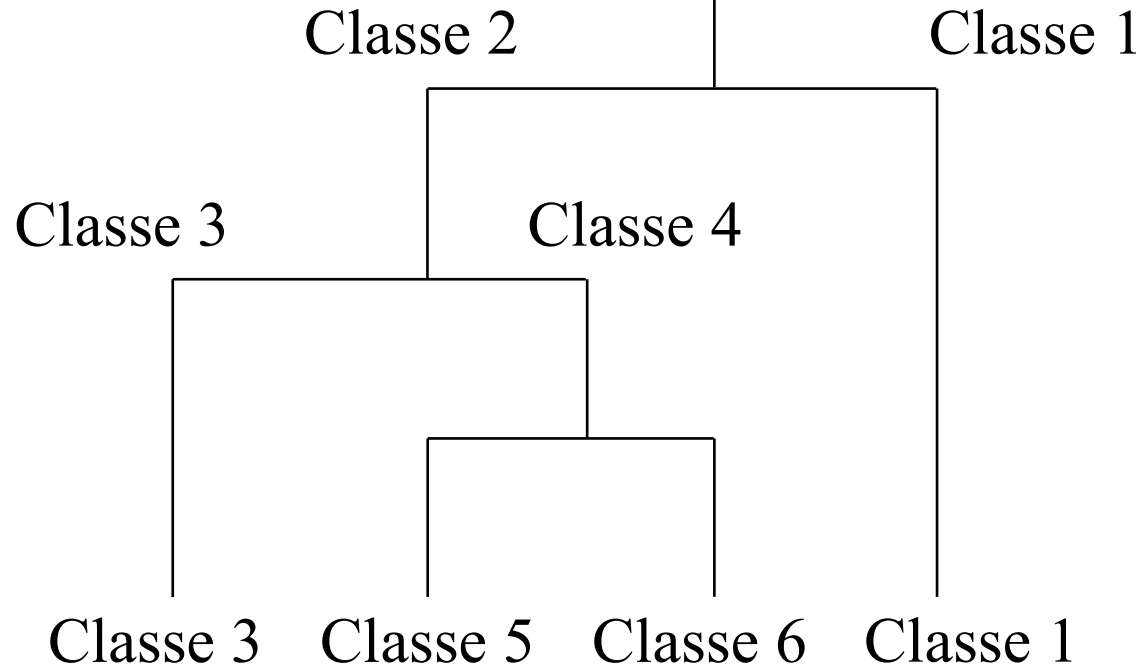


Tableau lexical

Partition (contributions ou paragraphes)



Lexique

- Tokenisation
- Reconnaissance
- Lemmatisation
- statut

Présence / absence

Statistiques :

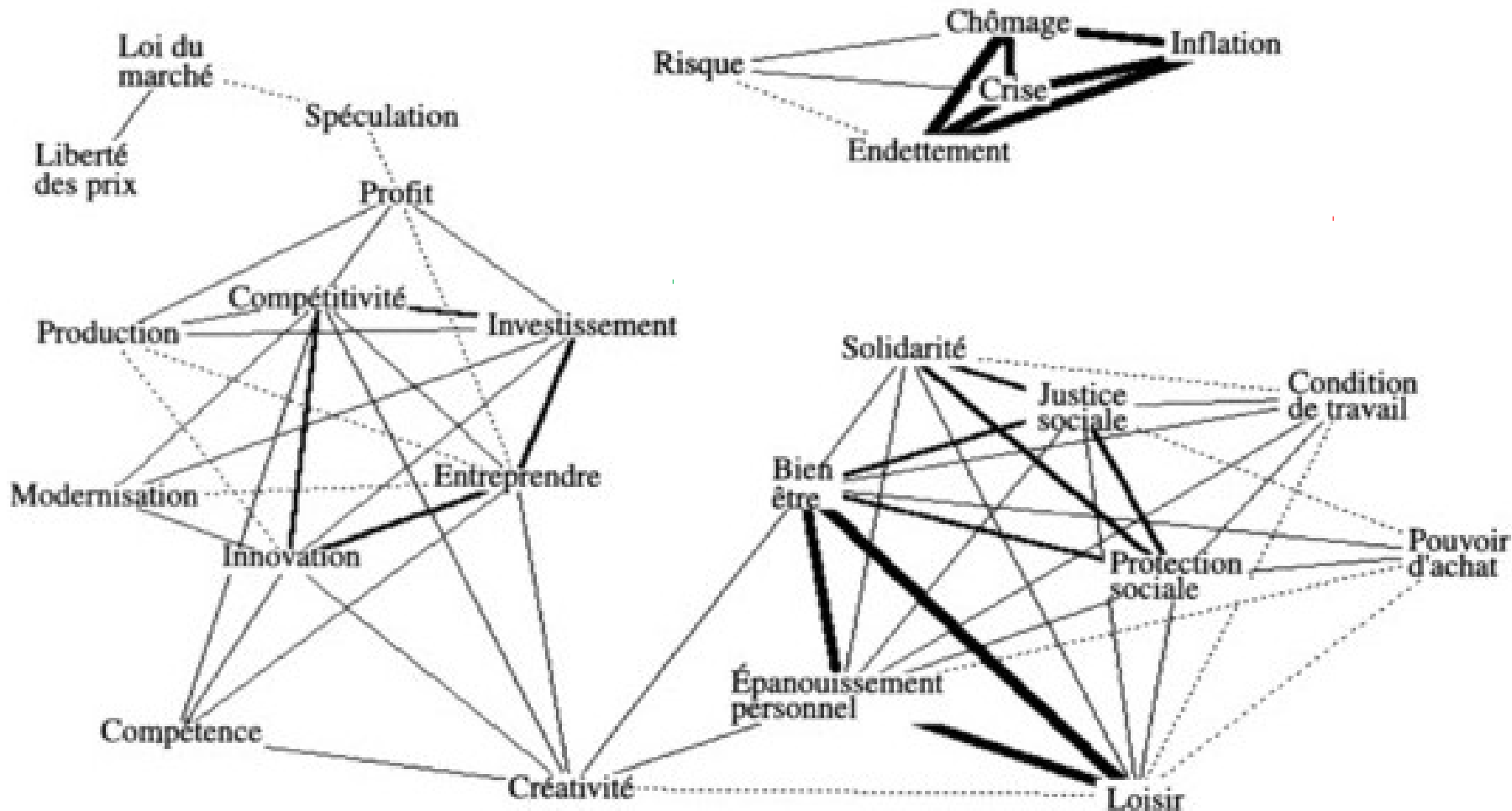
- Factorielles
- Classificatoires
- Arborées
- ...

L'analyse de similitude (ADS)

- Méthode(s) graphique(s) pour l'étude des relations entre les parties d'un ensemble.
- Les matrices :
 - Distances (euclidiennes, maximum, Jaccard, Manhattan...)
 - Similitude (cooccurrences, Jaccard, simple matching, phi...)
- L'ADS est classiquement utilisée pour décrire des représentations sociales, sur la base de questionnaires d'enquête.
 - Flament, 1962 ; Flament, 1981 ; Vergès & Bouriche, 2001.

P. Vergès, 2001, « L'analyse des représentations sociales par questionnaires », *Revue française de sociologie*, 42 (3), 537-561

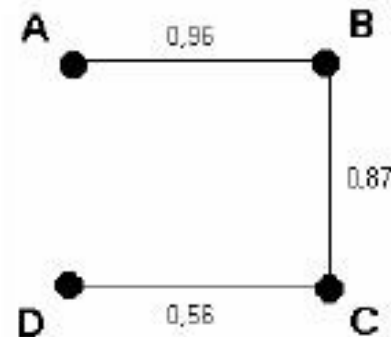
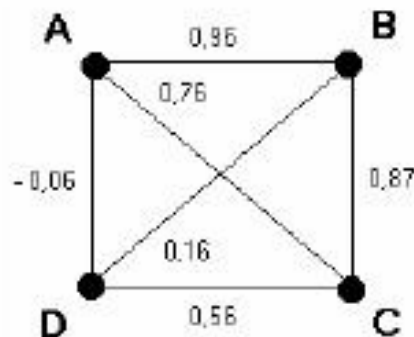
FIGURE XII. – France, graphe des relations données
(entre 32 termes décrivant l'univers socio-économique) par au moins 30 % des étudiants



L'analyse de similitude (ADS)

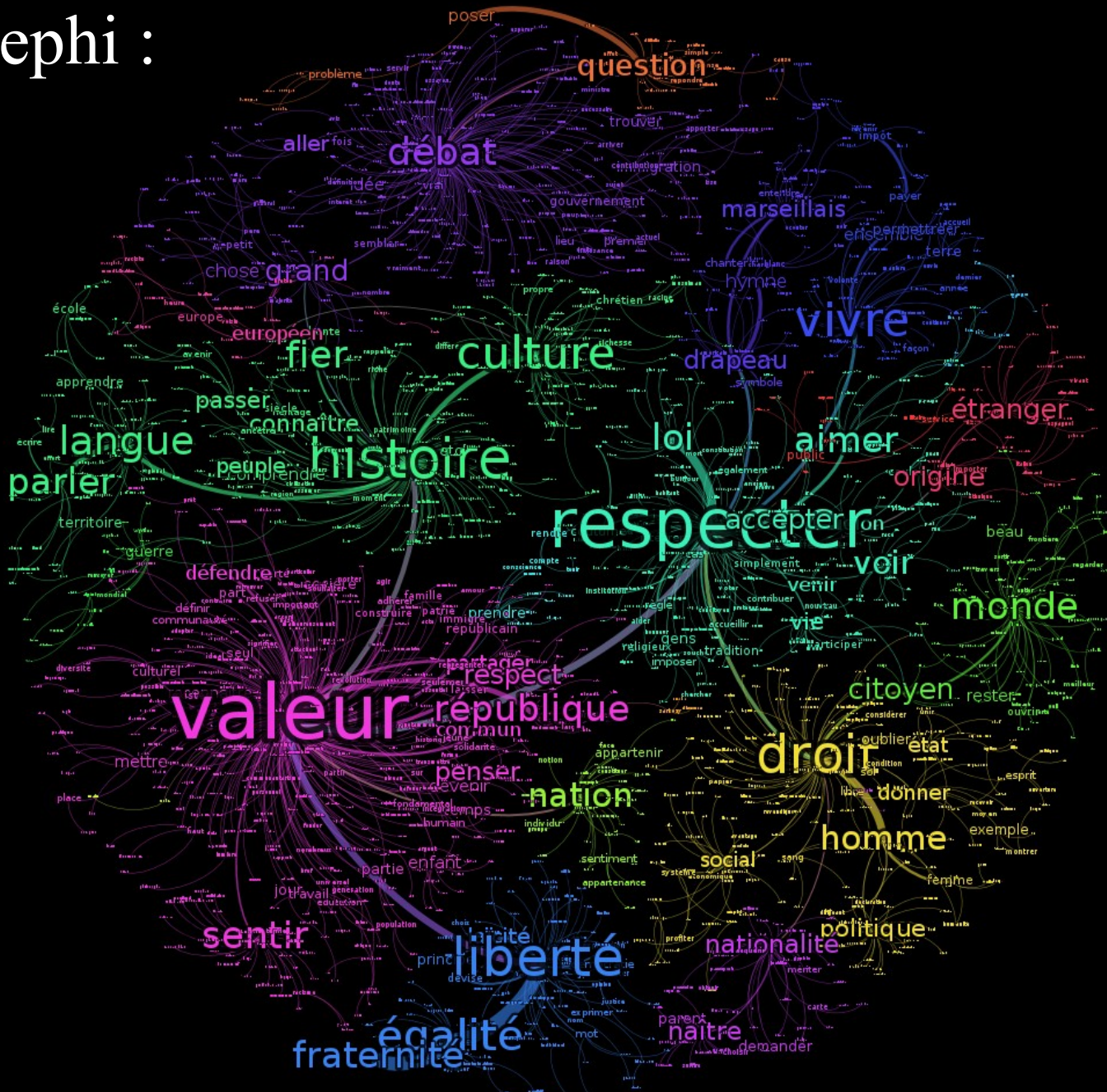
- étudier la proximité et les relations entre les éléments d'un ensemble, généralement sous forme d'*arbres maximum* :
 - le nombre de liens entre deux items évoluant « comme le carré du nombre de sommets » (Flament & Rouquette, 2003 : 88), l'ADS cherche à réduire le nombre de ces liens pour aboutir à « un graphe connexe et sans cycle » (Degenne & Vergès, 1973 : 473).
 - L'« arbre maximum ») est créé par les arêtes les plus fortes du graphique. C'est l'arbre le plus simple que l'on peut obtenir, mais c'est aussi le plus lourd (en termes d'information).
 - On considère toutes les « cliques » possibles (ex. ABCA, BCDB) et on élimine les liens les plus faibles (ex. entre A et C et entre B et D).

Arbre de similitude



Arbre maximum

Avec Gephi :



Les spécificités lexicales

- Si l'on considère une *forme* lexicale particulière dans un corpus, les occurrences de cette *forme* peuvent se distribuer:
 - de façon équilibrée dans toutes les *parties* (hasard)
 - ou certaines *parties* peuvent révéler une fréquence de cette *forme* plus élevée que d'autres (écart au hasard).
- A ce calcul, qui fait intervenir la comparaison d'une distribution observée à une distribution équilibrée (ou « théorique »), est associée une probabilité (« Modèle hypergéométrique », Lafon, 1984).

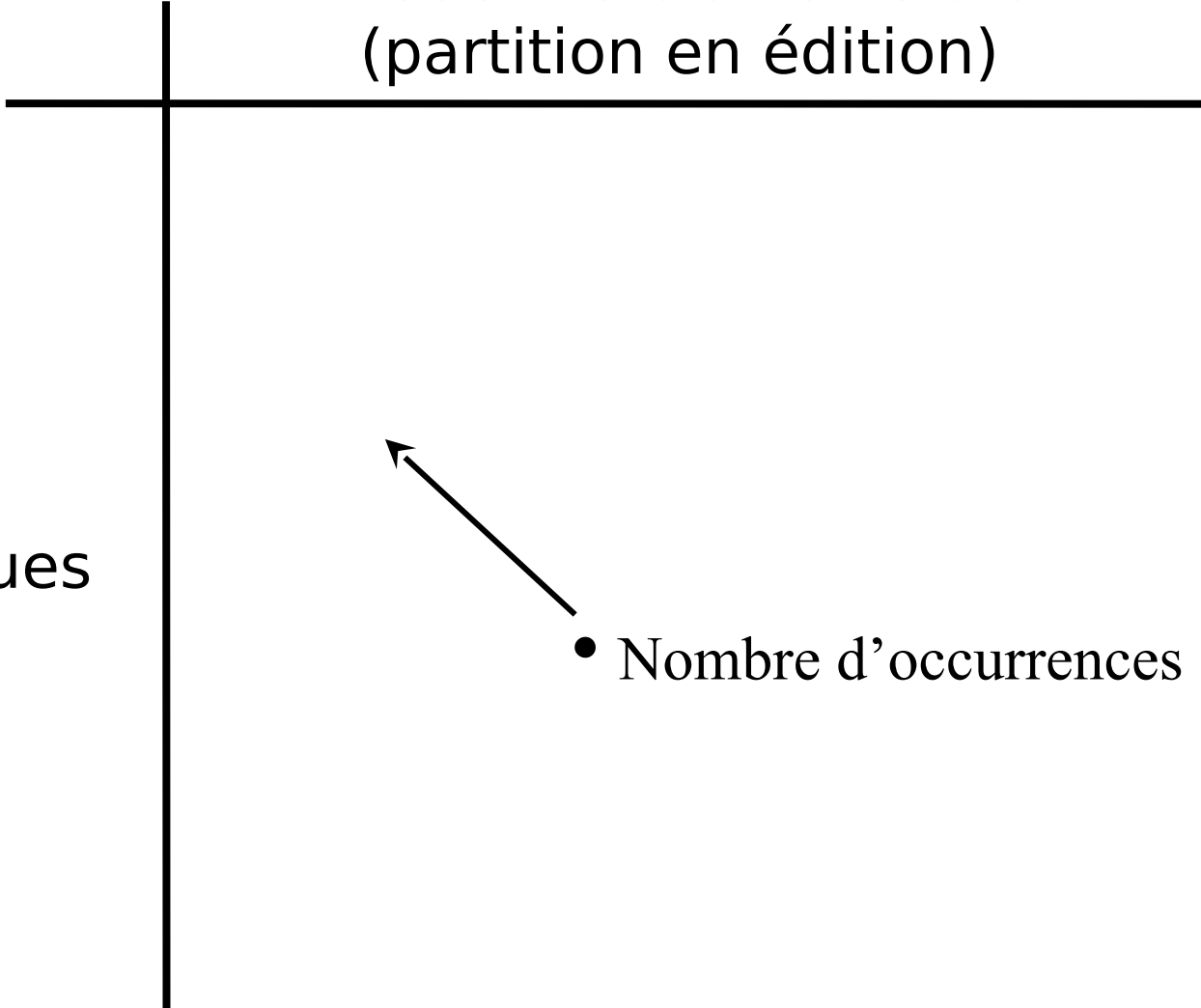
Tableau lexical

Modalité de variable
(partition en édition)

Lexique :

- ✓ Tokenization
- ✓ Reconnaissance
- ✓ Lemmatisation
- ✓ Statuts statistiques

• Nombre d'occurrences



Avec un exemple...

	*source la croix	*source le figaro	*source le monde	*source lhumanité	*source libération
sécurité	15	46	28	43	15
devenir	33	27	35	17	40
établissement	1	7	12	4	3
tirer	13	4	14	6	12
pencher	0	4	3	5	2
brexit	5	7	4	0	3
reporter	3	5	4	0	0
évolution	4	13	7	3	4
chine	1	5	2	0	2
connaissance	1	2	7	5	1
naturel	0	2	3	1	4
attente	4	5	8	3	3
unir	11	9	15	5	11

