Master 2 LSCN (UPEM) – Module “La littérature à l'ère du numérique”
03/10/2019 – UFT LACT (Champs-sur-Marne)

Analyse de corpus assistée par ordinateur : logiciels de textométrie et arbres de mots

Philippe Gambette

LIGM
Université Paris-Est
Marne-la-Vallée

Plan

Panorama des logiciels de textométrie

- Alceste, Hyperbase, Lexico, TXM, Iramuteq
- Exemple d'exploration textométrique
- Analyse textuelle à Paris-Est

Arbres de mots

- Concept et construction des nuages arborés
- Méthodologie et cas d'usage
- Construction des arbres
- Comparaison avec d'autres visualisations
- Implémentations et bibliographie

Recueil et prétraitements de corpus

- Océrisation de textes
- Prétraitements de textes obtenus par OCR
- Extraction de contextes

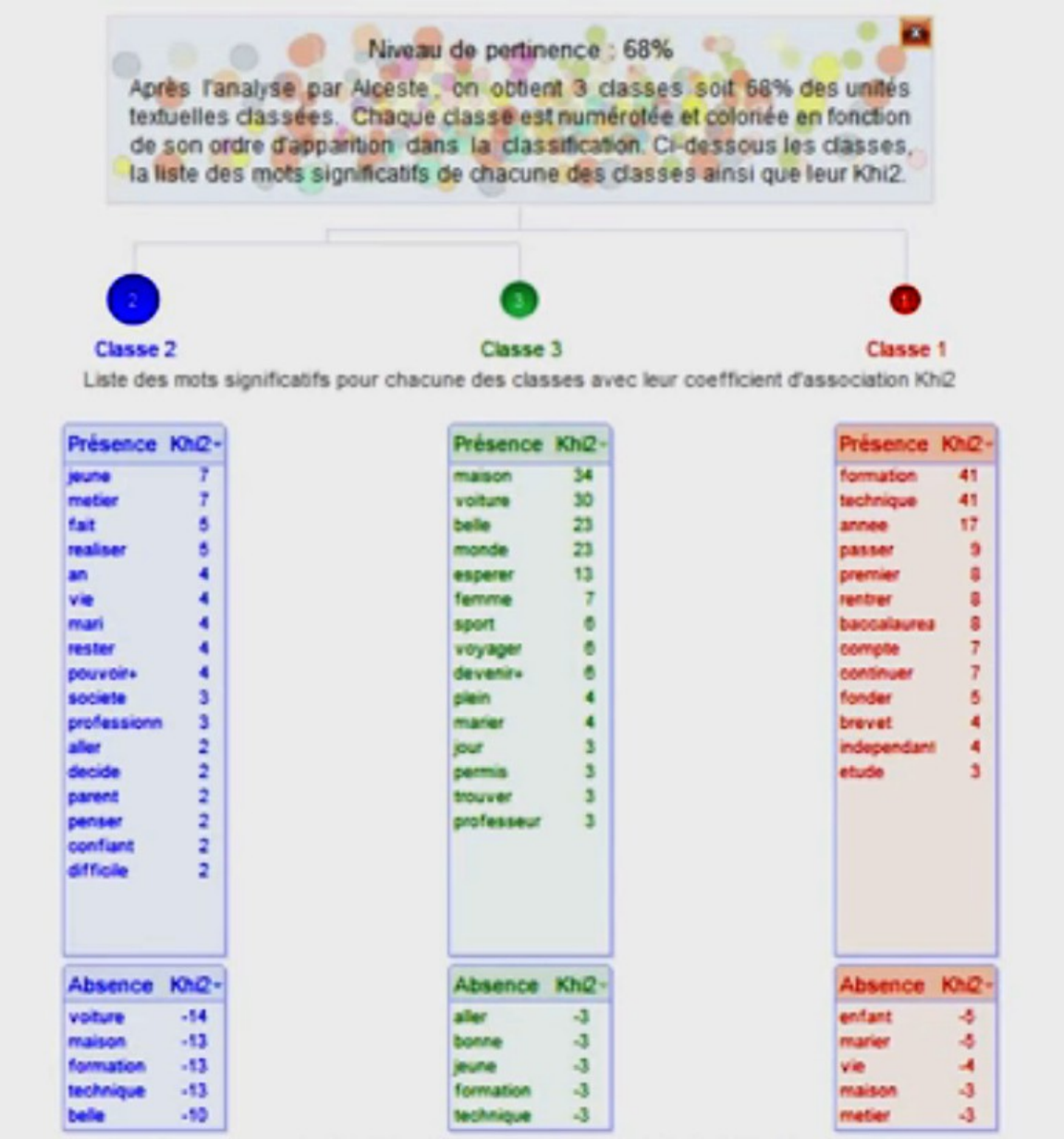
Panorama des logiciels de textométrie

Quelques logiciels de textométrie

Alceste (depuis 1983)

Société IMAGE
http://www.image-zafar.com/Logiciel.html

La méthode Alceste



Répartition des phrases du texte en différentes classes
→ vocabulaire de chaque classe

https://www.youtube.com/watch?v=Hgk6EaBxyeM

Quelques logiciels de textométrie

Hyperbase (depuis 1989)

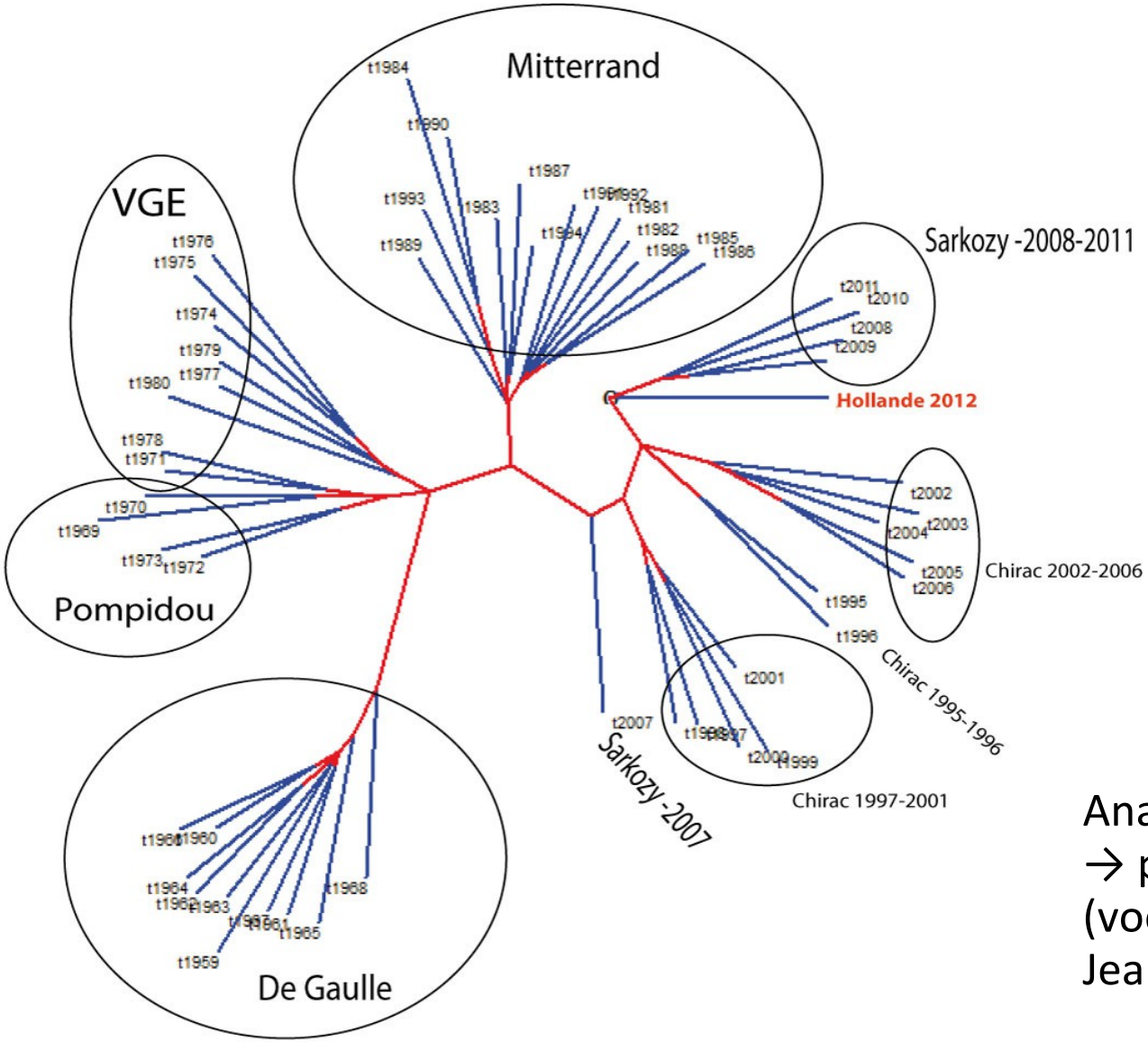
Hyperbase
Web Edition

Université de Nice Sophia-Antipolis
http://logometrie.unice.fr

Alceste (depuis 1983)

Société IMAGE
http://www.image-zafar.com/Logiciel.html

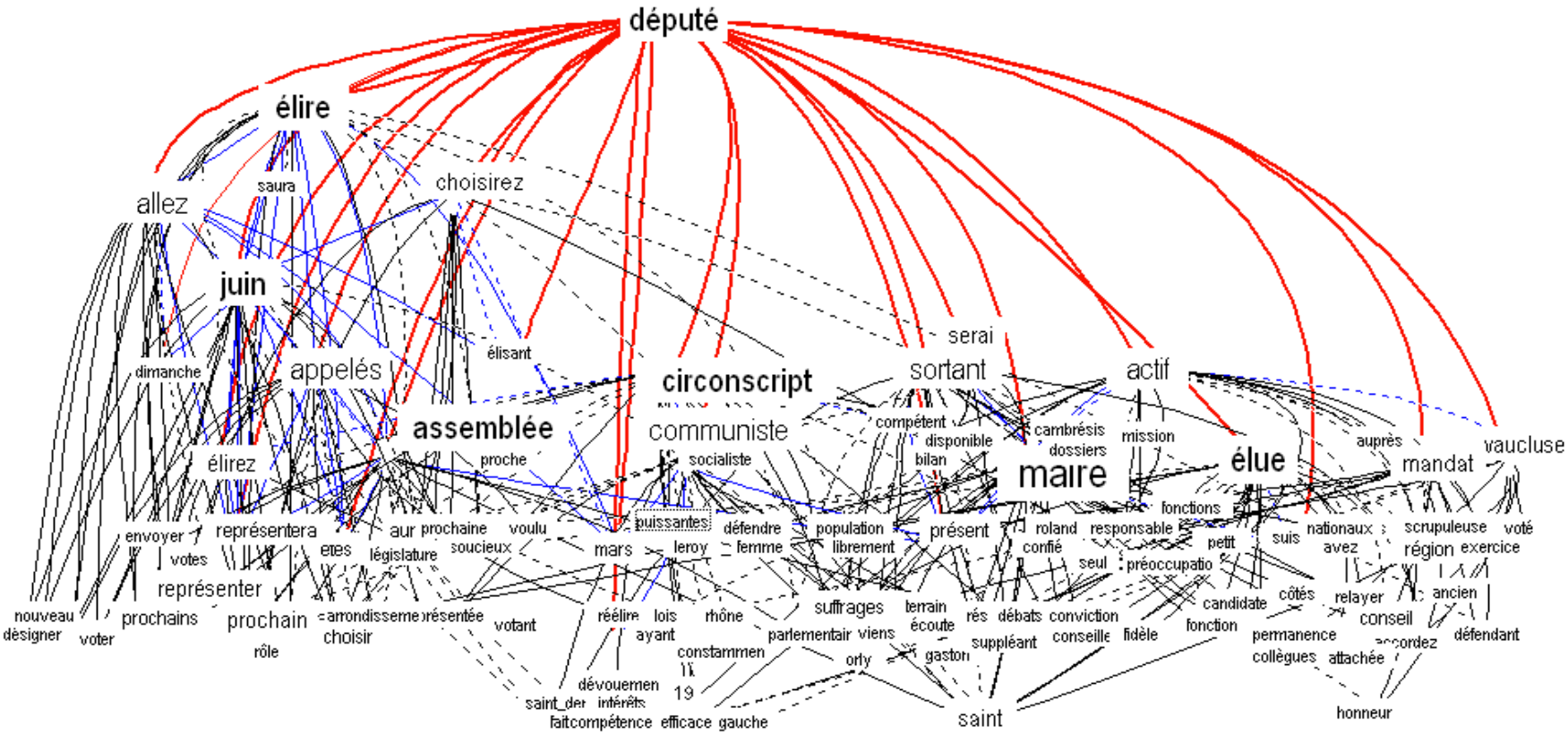
Hyperbase



Analyse arborée
→ proximité des textes
(voeux présidentiels,
Jean-Marc Leblanc)

http://textopol.free.fr/dotclear/index.php?2013/01/05/49-comment-le-discours-de-franois-hollande-se-situe-t-il-par-r

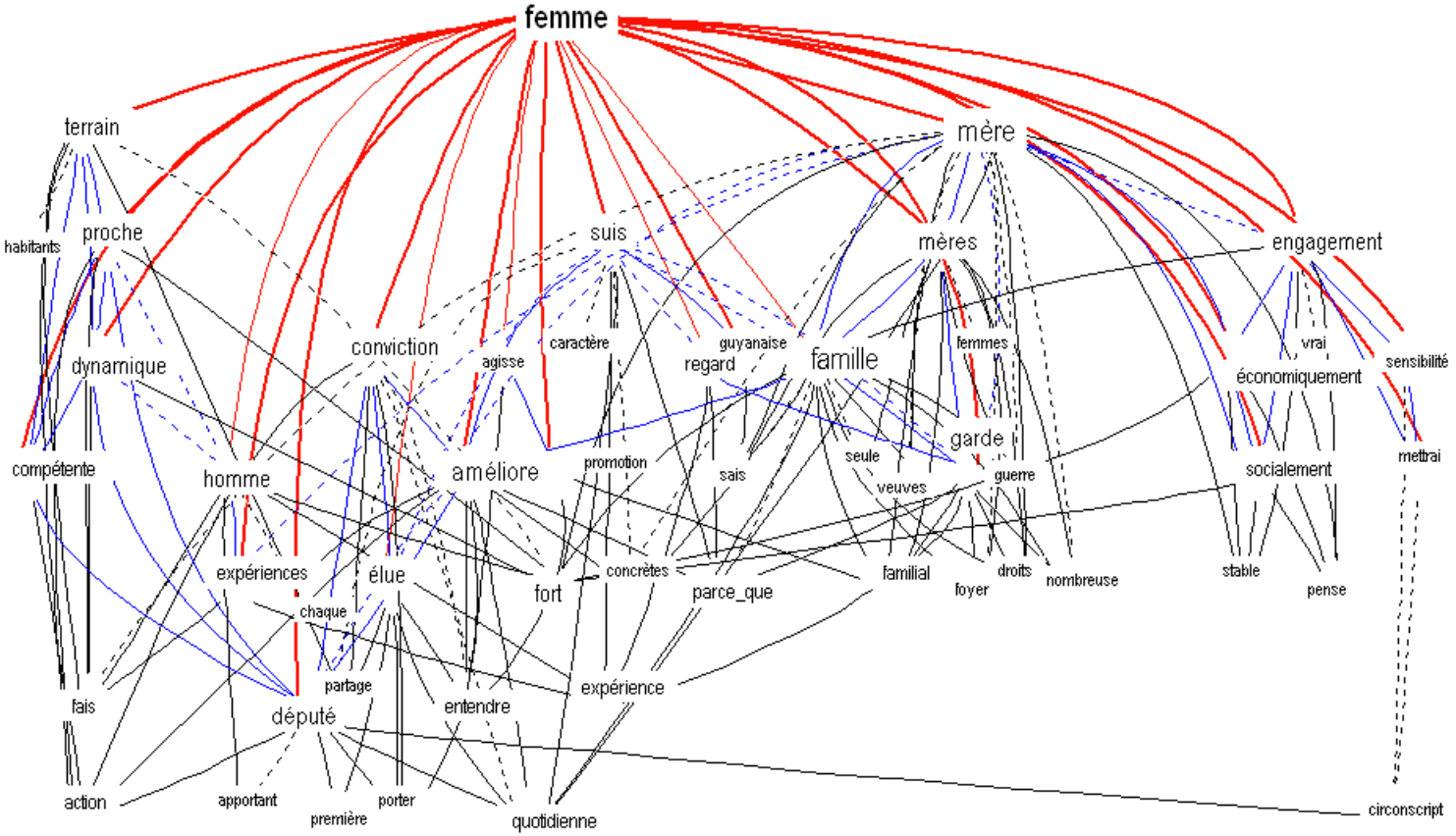
Hyperbase



Réseau de cooccurrents d'un mot
→ graphique des co-occurrents directs et indirects du mot-pôle « député » dans l'ensemble du corpus de Professions de foi

Magali Guaresi (2014) L'approche co-occurrentielle, un bond qualitatif ? L'environnement lexical du lemme « député » dans les Professions de foi des candidates à la députation (1958 – 2002)
https://corela.revues.org/3586?lang=fr

Hyperbase



Réseau de cooccurrents d'un mot
→ graphique des co-occurrents directs et indirects du mot-pôle « femme » dans le corpus des Professions de foi de candidates

Magali Guaresi (2014) L'approche co-occurrentielle, un bond qualitatif ? L'environnement lexical du lemme « député » dans les Professions de foi des candidates à la députation (1958 – 2002)
https://corela.revues.org/3586?lang=fr

Quelques logiciels de textométrie

Hyperbase (depuis 1989)

Université de Nice Sophia-Antipolis
http://logometrie.unice.fr

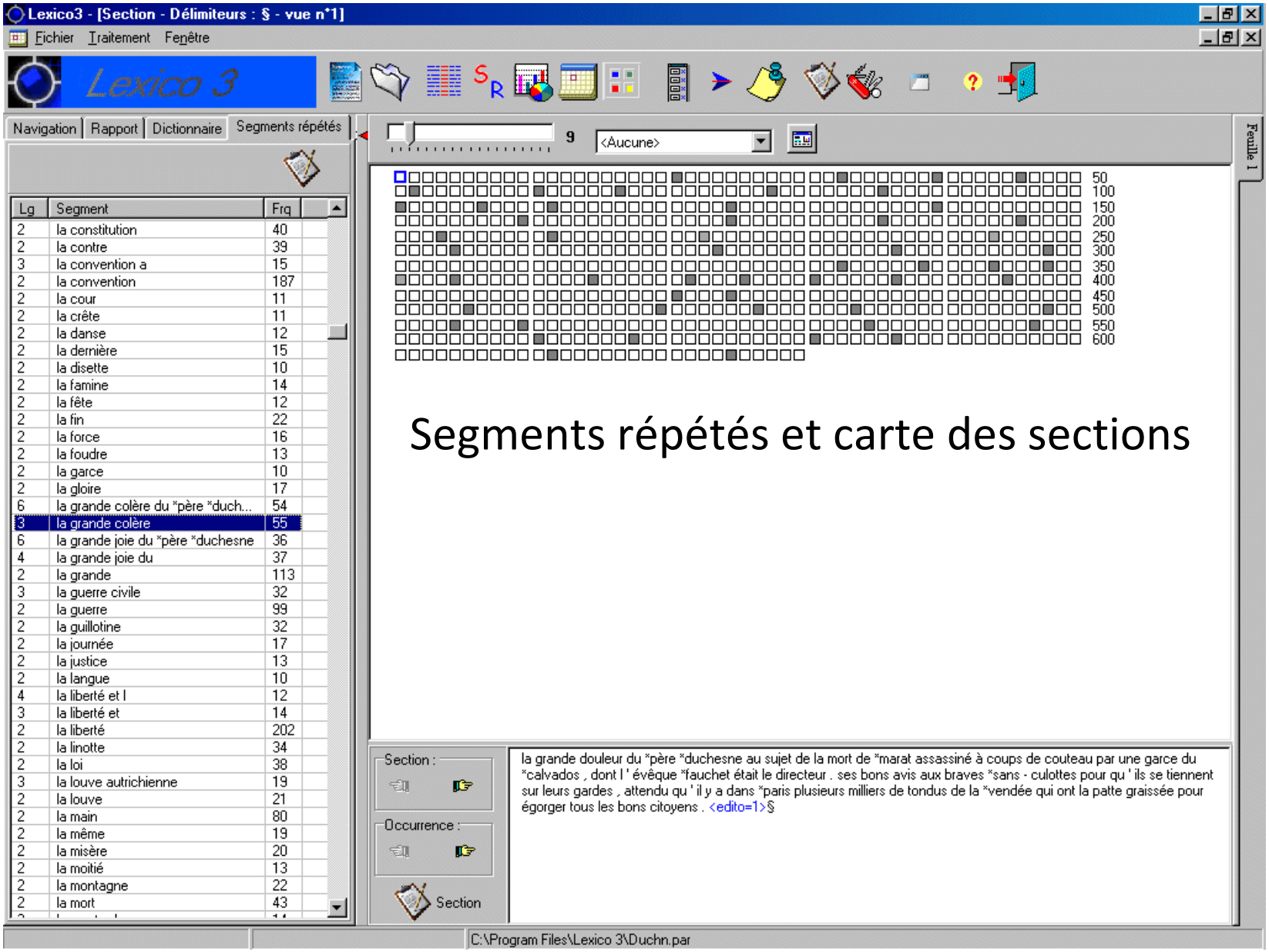
Lexico (depuis 1990)

Université Sorbonne nouvelle
http://lexi-co.com/

Alceste (depuis 1983)

Société IMAGE
http://www.image-zafar.com/Logiciel.html

Lexico



Segments répétés et carte des sections

http://www.tal.univ-paris3.fr/lexico/demoLex3/out6.htm

Quelques logiciels de textométrie

Hyperbase (depuis 1989)

Université de Nice Sophia-Antipolis
http://logometrie.unice.fr

Lexico (depuis 1990)

Université Sorbonne nouvelle
http://lexi-co.com/

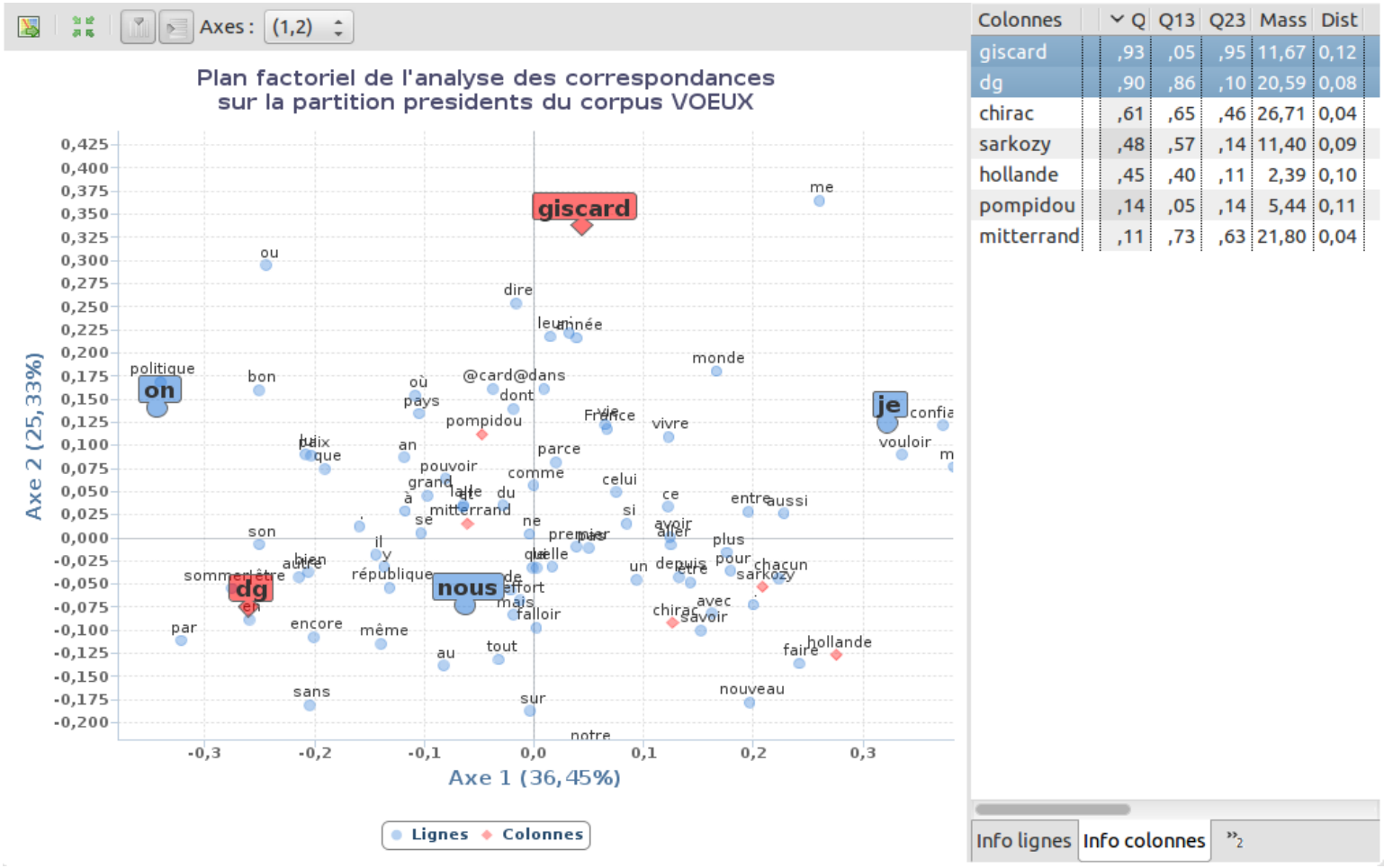
Alceste (depuis 1983)

Société IMAGE
http://www.image-zafar.com/Logiciel.html

TXM (depuis 2008)

ENS Lyon
http://textometrie.ens-lyon.fr/

TXM



Colonnes	Q	Q13	Q23	Mass	Dist
giscard	,93	,05	,95	11,67	0,12
dg	,90	,86	,10	20,59	0,08
chirac	,61	,65	,46	26,71	0,04
sarkozy	,48	,57	,14	11,40	0,09
hollande	,45	,40	,11	2,39	0,10
pompidou	,14	,05	,14	5,44	0,11
mitterrand	,11	,73	,63	21,80	0,04

Info lignes | Info colonnes

Analyse factorielle des correspondances

→ Affichage des points lignes (mots) et des points colonnes (discours) pour les discours de voeux présidentiels

http://txm.sourceforge.net/doc/manual/manual39.xhtml

Quelques logiciels de textométrie

Hyperbase (depuis 1989)

Université de Nice Sophia-Antipolis
http://logometrie.unice.fr

Lexico (depuis 1990)

Université Sorbonne nouvelle
http://lexi-co.com/

Alceste (depuis 1983)

Société IMAGE
http://www.image-zafar.com/Logiciel.html

TXM (depuis 2008)

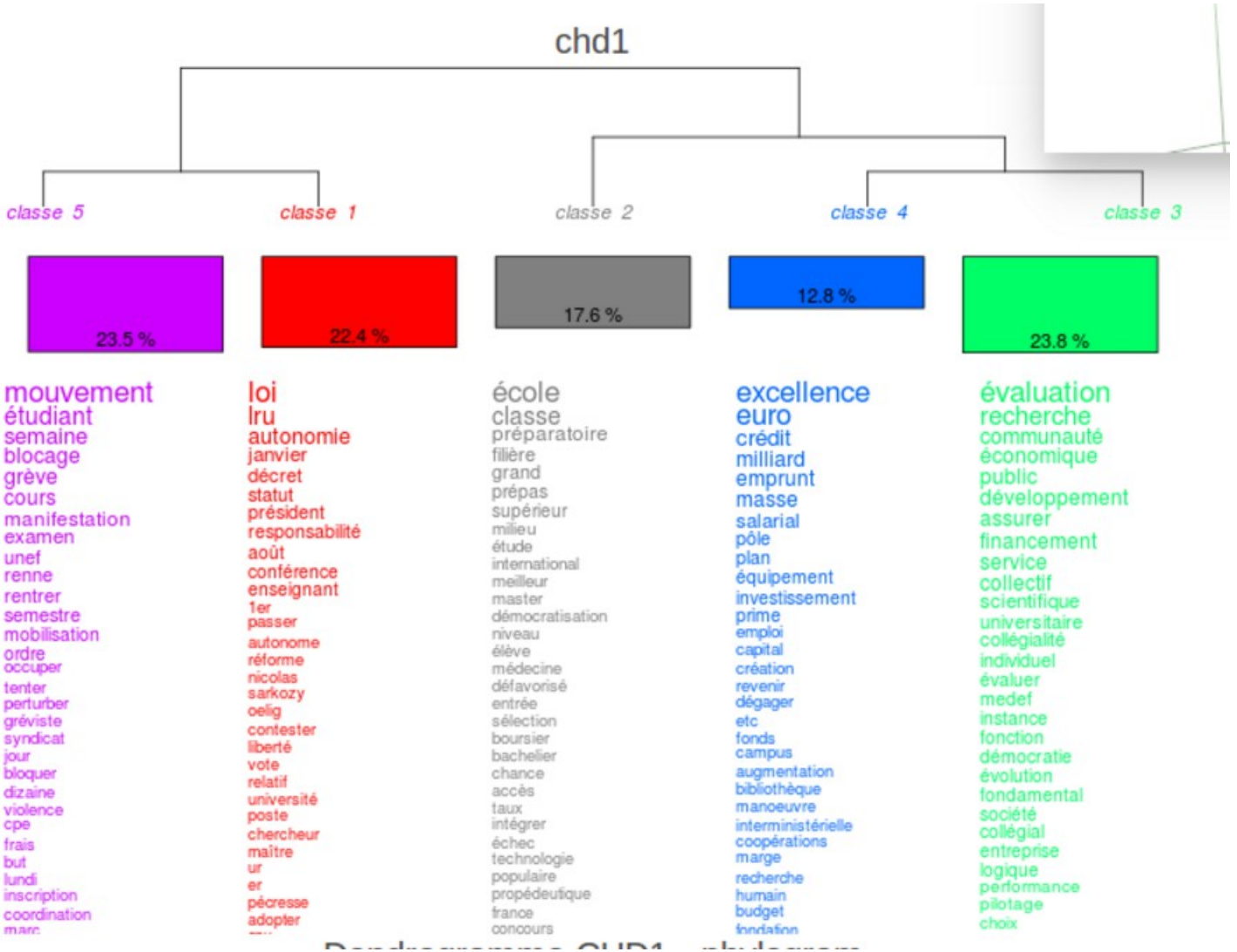
ENS Lyon
http://textometrie.ens-lyon.fr/

Iramuteq (depuis 2009)

IRaMuTeQ

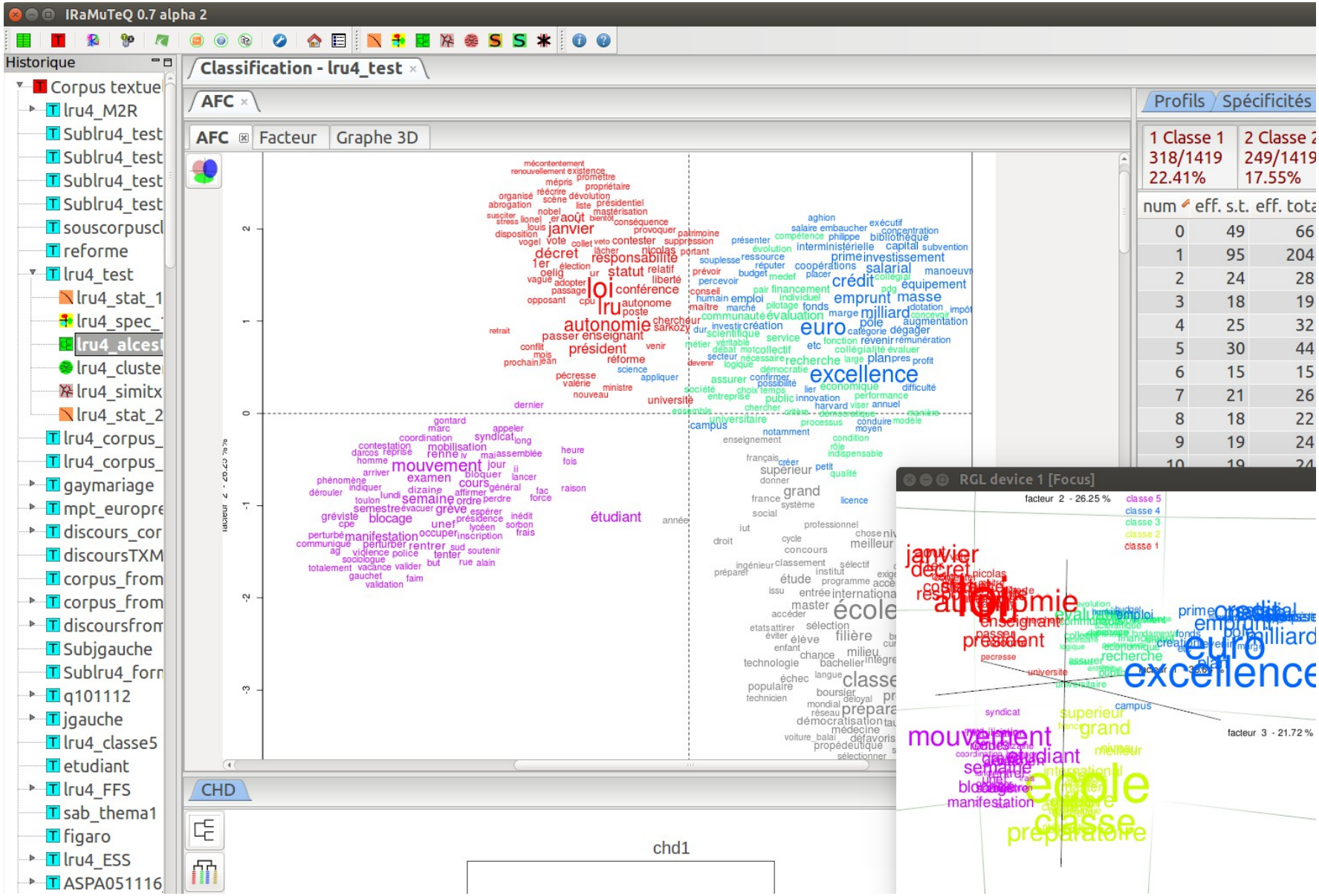
Université de Toulouse
http://www.iramuteq.org/

Iramuteq



http://www.iramuteq.org/captures-decrans/interface/interface-classification/image_view_fullscreen

Iramuteq



http://www.iramuteq.org/captures-decrans/interface/interface-classification/image_view_fullscreen

Quelques logiciels de textométrie

Hyperbase (depuis 1989)

Hyperbase Web Edition

Université de Nice Sophia-Antipolis
http://logometrie.unice.fr

Lexico (depuis 1990)

LEXICO3 version 3.6

Université Sorbonne nouvelle
http://lexi-co.com/

Alceste (depuis 1983)

Société IMAGE
http://www.image-zafar.com/Logiciel.html

TXM (depuis 2008)

ENS Lyon
http://textometrie.ens-lyon.fr/

Iramuteq (depuis 2009)

IRaMuTeQ

Université de Toulouse
http://www.iramuteq.org/

Caractéristiques des logiciels de textométrie

Approches **exploratoires**

→ explorer, générer des hypothèses : **visualisations**

→ évaluer la pertinence d'une hypothèse :

- indicateurs **statistiques**

- **retour au texte**

Exemple d'exploration : articles scientifiques

Etape 1) Récupération du corpus (Scopus) et formatage (Lexico 3)

<annee=2015> <type=article> **<doc=a1>** background: lateral, or horizontal, gene transfers are a type of asymmetric evolutionary events where genetic material is transferred from one species to another. in this paper we consider lgt networks, a general model of phylogenetic networks with lateral gene transfers which consist, roughly, of a principal rooted tree with its leaves labelled on a set of taxa, and a set of extra secondary arcs between nodes in this tree representing lateral gene transfers. an lgt network gives rise in a natural way to a principal phylogenetic subtree and a set of secondary phylogenetic subtrees, which, roughly, represent, respectively, the main line of evolution of most genes and the secondary lines of evolution through lateral gene transfers. results: we introduce a set of simple conditions on an lgt network that guarantee that its principal and secondary phylogenetic subtrees are pairwise different and that these subtrees determine, up to isomorphism, the lgt network. we then give an algorithm that, given a set of pairwise different phylogenetic trees t0, t1, . . . , tk on the same set of taxa, outputs, when it exists, the lgt network that satisfies these conditions and such that its principal phylogenetic tree is t0 and its secondary phylogenetic trees are t1, . . . , tk.
<annee=2015> <type=article> **<doc=a2>** this article presents an innovative approach to phylogenies based on the reduction of multistate characters to binary-state characters. we show that the reduction to binary characters' approach can be applied to both character- and distance-based phylogenies and provides a unifying framework to explain simply and intuitively the similarities and differences between distance- and character-based phylogenies. building on these results, this article gives a possible explanation on why phylogenetic trees obtained from a distance matrix or a set of characters are often quite reasonable despite lateral transfers of genetic material between taxa. in the presence of lateral transfers, outer planar networks furnish a better description of evolution than phylogenetic trees. we present a polynomial-time reconstruction algorithm for perfect outer planar networks with a fixed number of states, characters, and lateral transfers.
<annee=2015> <type=article> **<doc=a3>** background: many problems in comparative biology are, or are thought to be, best expressed as phylogenetic "networks" as opposed to trees. in trees, vertices may have only a single parent (ancestor), while networks allow for multiple parent vertices. there are two main interpretive

https://drive.google.com/file/d/0B28vcDIRmKwobjZlT0JOZzE1MnM/view?usp=sharing

Exemple d'exploration : articles scientifiques

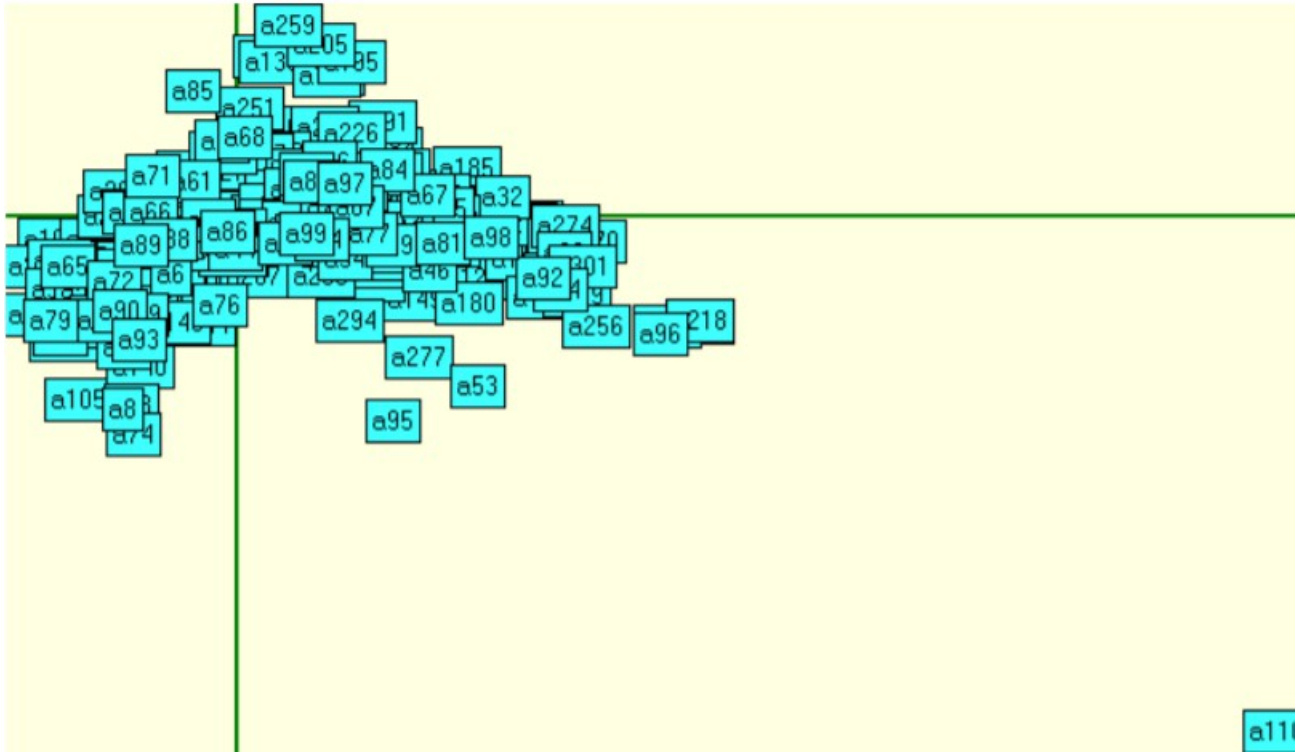
Etape 1) Récupération du corpus (Scopus) et formatage (Lexico 3)

<annee=2015> <type=article> **<doc=a1>** background: lateral, or horizontal, gene transfers are a type of asymmetric evolutionary events where genetic material is transferred from one species to another. in this paper we consider lgt networks, a general model of phylogenetic networks with lateral gene transfers which consist, roughly, of a principal rooted tree with its leaves labelled on a set of taxa, and a set of extra secondary arcs between nodes in this tree representing lateral gene transfers. an lgt network gives rise in a natural way to a principal phylogenetic subtree and a set of secondary phylogenetic subtrees, which, roughly, represent, respectively, the main line of evolution of most genes and the secondary lines of evolution through lateral gene transfers. results: we introduce a set of simple conditions on an lgt network that guarantee that its principal and secondary phylogenetic subtrees are pairwise different and that these subtrees determine, up to isomorphism, the lgt network. we then give an algorithm that, given a set of pairwise different phylogenetic trees t0, t1, . . . , tk on the same set of taxa, outputs, when it exists, the lgt network that satisfies these conditions and such that its principal phylogenetic tree is t0 and its secondary phylogenetic trees are t1, . . . , tk.
<annee=2015> <type=article> **<doc=a2>** this article presents an innovative approach to phylogenies based on the reduction of multistate characters to binary-state characters. we show that the reduction to binary characters' approach can be applied to both character- and distance-based phylogenies and provides a unifying framework to explain simply and intuitively the similarities and differences between distance- and character-based phylogenies. building on these results, this article gives a possible explanation on why phylogenetic trees obtained from a distance matrix or a set of characters are often quite reasonable despite lateral transfers of genetic material between taxa. in the presence of lateral transfers, outer planar networks furnish a better description of evolution than phylogenetic trees. we present a polynomial-time reconstruction algorithm for perfect outer planar networks with a fixed number of states, characters, and lateral transfers.
<annee=2015> <type=article> **<doc=a3>** background: many problems in comparative biology are, or are thought to be, best expressed as phylogenetic "networks" as opposed to trees. in trees, vertices may have only a single parent (ancestor), while networks allow for multiple parent vertices. there are two main interpretive

https://drive.google.com/file/d/0B28vcDIRmKwobjZlT0JOZzE1MnM/view?usp=sharing

Exemple d'exploration : articles scientifiques

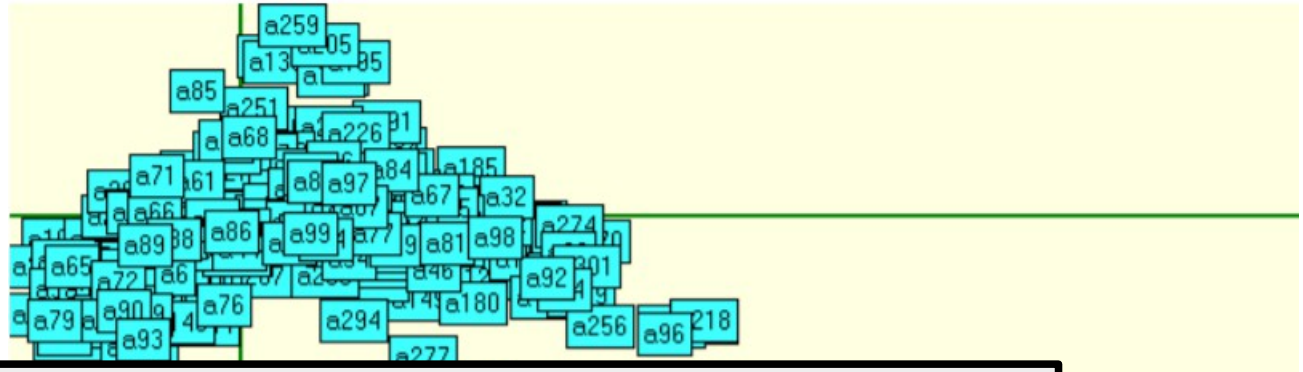
Etape 2) Analyse factorielle des correspondances



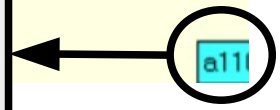
Agarwal, Gambette & Morrison, 2016

Exemple d'exploration : articles scientifiques

Etape 2) Analyse factorielle des correspondances



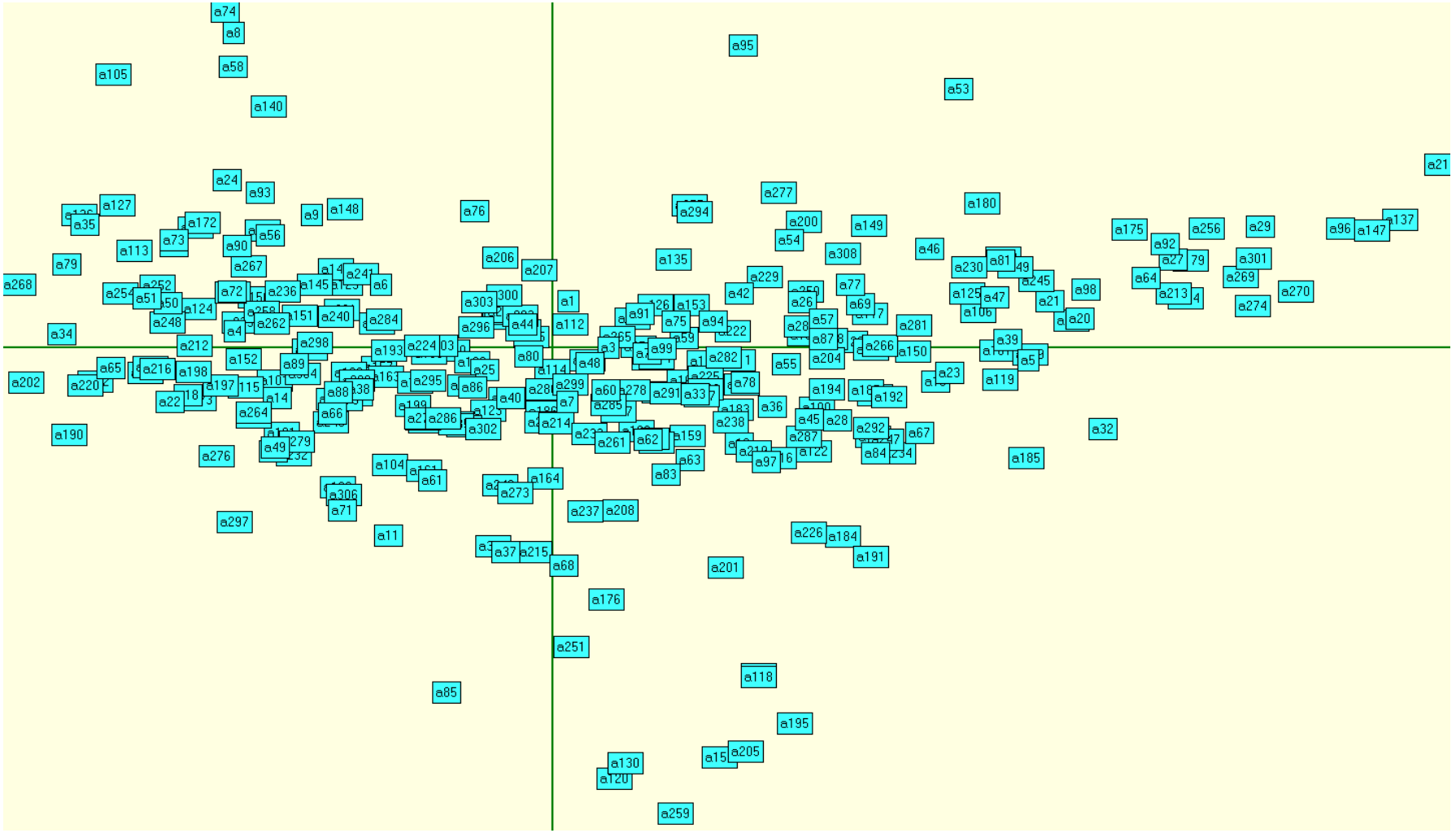
we study the asymptotic behavior of a new type of maximization recurrence, defined as follows. let k be a positive integer and p k(x) a polynomial of degree k satisfying p k(0) = 0. define a 0 = 0 and for n ≥ 1, let a n = max 0≤i<n{a i+n kp k(i/n)}. we prove that lim n→∞a n/n n = sup{pk(x)/1-x k : 0≤x<1}. we also consider two closely related maximization recurrences s n and s' n, defined as s 0 = s' 0 = 0, and for n ≥ 1, s n = max 0≤i<n{s i + i(n-i)(n-i-1)/2} and s' n = max 0≤i<n{s' i + (3 n-i) + 2i(2 n-i) + (n-i)(2 i)}. we prove that lim n→∞ s'n/3(3 n) = 2(√3-1)/3 ≈ 0.488033..., resolving an open problem from bioinformatics about rooted triplets consistency in phylogenetic networks.



Agarwal, Gambette & Morrison, 2016

Exemple d'exploration : articles scientifiques

Etape 2) Analyse factorielle des correspondances (sans a110)



Agarwal, Gambette & Morrison, 2016

Exemple d'exploration : articles scientifiques

Etape 3) Analyse statistique

Termes **sur-représentés** à gauche

Terme	Frq Tot.	Frq Pa...	Spécif
gene	412	342	44
hgt	110	110	34
reconciliation	70	68	19
transfer	134	111	15
genes	88	77	14
and	1320	792	14
methods	208	154	13
species	261	186	12
lineage	32	32	11
model	130	101	11
method	168	122	10
event	34	33	10
phylogeny	52	47	10
accurate	32	31	9
data	227	155	9
horizontal	77	63	9
events	232	160	9
population	26	26	9
genome	49	44	9
likelihood	39	36	9
evolutionary	279	188	9
coalescent	31	30	9
consensus	24	24	8
inference	40	36	8
role	30	29	8
detection	24	24	8
families	27	26	8
sorting	29	28	8
inferring	47	41	8
vertical	23	23	8

Termes **sous-représentés** à gauche

Terme	Frq Tot.	Frq Pa...	Spécif
child	24	1	-7
2	47	8	-7
m	37	4	-7
split	79	19	-7
e	53	10	-7
constructing	53	8	-8
galled	40	4	-8
construct	48	7	-8
graph	60	11	-8
binary	75	16	-8
t	51	6	-9
input	83	16	-9
given	137	36	-9
distance	120	30	-9
d	33	1	-9
consistent	54	7	-9
class	34	2	-9
x	47	4	-10
polynomial	62	7	-11
problem	240	72	-11
number	206	56	-12
level	93	14	-13
1	71	7	-13
a	1675	701	-13
is	922	360	-13
leaves	54	2	-14
if	95	12	-15
set	252	68	-15
k	69	4	-16
o	73	4	-17
time	191	39	-18

Agarwal, Gambette & Morrison, 2016

Exemple d'exploration : articles scientifiques

Etape 3) Analyse statistique

Termes **sur-représentés** à gauche

Terme	Frq Tot.	Frq Pa...	Spécif
gene	412	342	44
hgt	110	110	34
reconciliation	70	68	19
transfer	134	111	15
genes	88	77	14
and	1320	792	14
methods	208	154	13
species	261	186	12
lineage	32	32	11
model	130	101	11
method	168	122	10
event	34	33	10
phylogeny	52	47	10
accurate	32	31	9
data	227	155	9
horizontal	77	63	9
events	232	160	9
population	26	26	9
genome	49	44	9
likelihood	39	36	9
evolutionary	279	188	9
coalescent	31	30	9
consensus	24	24	8
inference	40	36	8
role	30	29	8
detection	24	24	8
families	27	26	8
sorting	29	28	8
inferring	47	41	8
vertical	23	23	8

Spécificité
>2 ou <-2 :
statistiquement significatif !

Termes **sous-représentés** à gauche

Terme	Frq Tot.	Frq Pa...	Spécif
child	24	1	-7
2	47	8	-7
m	37	4	-7
split	79	19	-7
e	53	10	-7
constructing	53	8	-8
galled	40	4	-8
construct	48	7	-8
graph	60	11	-8
binary	75	16	-8
t	51	6	-9
input	83	16	-9
given	137	36	-9
distance	120	30	-9
d	33	1	-9
consistent	54	7	-9
class	34	2	-9
x	47	4	-10
polynomial	62	7	-11
problem	240	72	-11
number	206	56	-12
level	93	14	-13
1	71	7	-13
a	1675	701	-13
is	922	360	-13
leaves	54	2	-14
if	95	12	-15
set	252	68	-15
k	69	4	-16
o	73	4	-17
time	191	39	-18

Agarwal, Gambette & Morrison, 2016

Exemple d'exploration : articles scientifiques

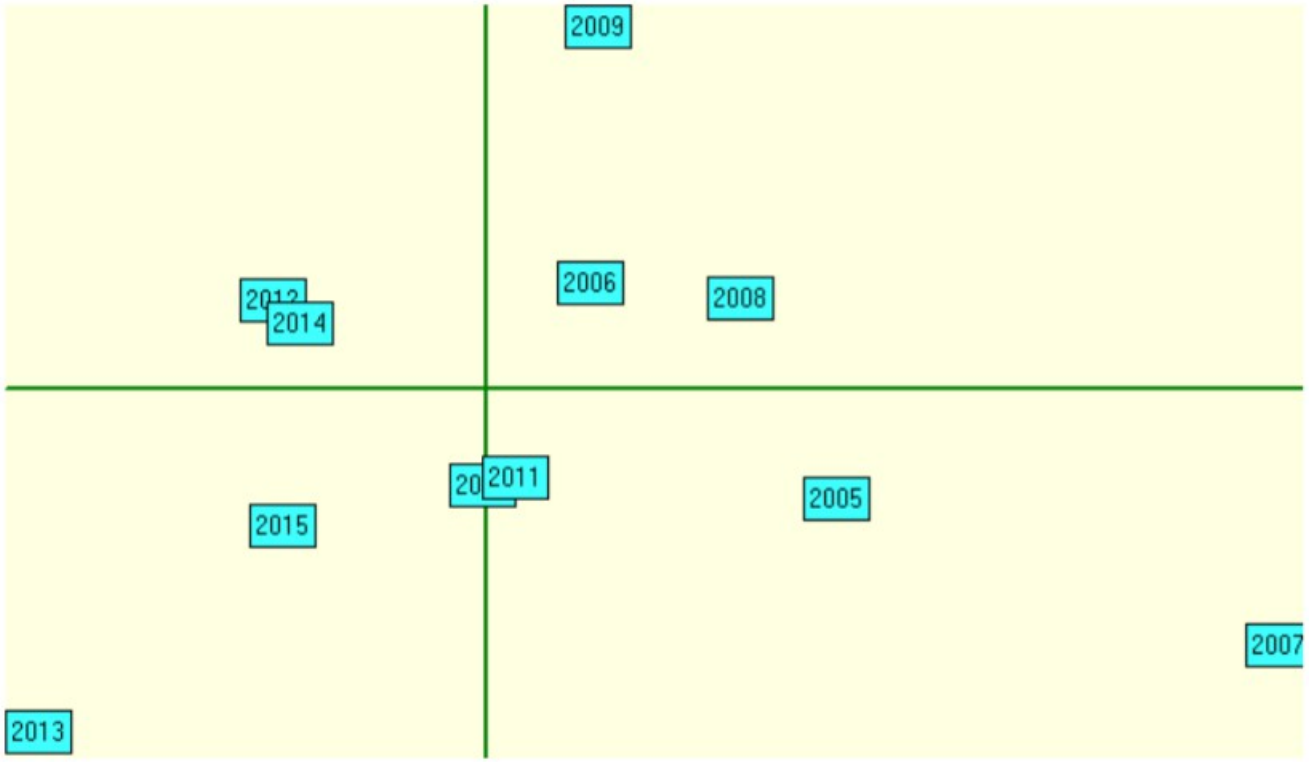
Etape 4) Retour au texte : concordance

n is a normal network , a binary tree - child network , or a level - k network . rec
table networks include normal and tree - child networks , they claim that important e
general networks , and 5 / 4 for tree - child and normal networks . we also show tha
hat the number of leaf - labelled tree - child and normal networks with â„“ leaves ar
ons , or lateral gene transfers . tree - child reticulate networks (tc networks) ar
ary level - 2 networks and binary tree - child networks are also encoded by their tri
etworks that is more general than tree - child networks . background : the advent of
hich pairs of individuals are parent and child . new methods to automate this process
rtices that are not leaves have a tree - child . background : phylogenetic networks a
for normal networks , for binary tree - child networks , and for level - k networks
it possible the generalization to tree - child time consistent (tctc) hybridization
of phylogenetic networks , called tree - child phylogenetic networks , and we provide
l algorithms for reconstructing a tree - child phylogenetic network from its path mul
omputing the distance between two tree - child phylogenetic networks and for aligning
tworks and for aligning a pair of tree - child phylogenetic networks , are provided .
s also a metric on the classes of tree - child phylogenetic networks , semibinary tre
sis and comparison of metrics for tree - child time consistent phylogenetic networks
they are metrics on any class of tree - child time consistent phylogenetic networks
t only to establish properties of tree - child time consistent phylogenetic networks
uction , but also to generate all tree - child time consistent phylogenetic networks
sis and comparison of metrics for tree - child time consistent phylogenetic networks
ain tight bounds on the size of a tree - child time consistent phylogenetic network .
ed them as regular , tree sibling , tree child , or galled trees . we show that , as
netic networks , which generalize tree - child time consistent phylogenetic networks

Agarwal, Gambette & Morrison, 2016

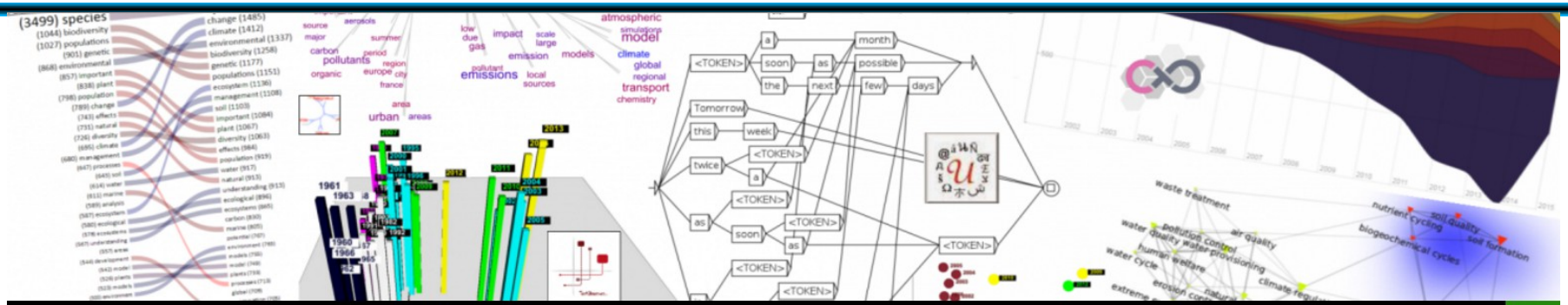
Exemple d'exploration : articles scientifiques

Test de l'effet d'un paramètre : l'année de publication



Tushar Agarwal, Philippe Gambette, David Morrison (2016) *Who is Who in Phylogenetic Networks: Articles, Authors and Programs*.
https://hal-upec-upem.archives-ouvertes.fr/hal-01376483

Logiciels d'analyse textuelle à Université Paris-Est



https://eclavit.hypotheses.org/

Logiciels développés à Université Paris-Est :

- Unitex (LIGM, http://www-igm.univ-mlv.fr/~unitex/) : annotation de textes, extraction d'informations par recherche de patrons grammaticaux ou lexicaux
- Cortext (LISIS, http://www.cortext.net/) : analyses textométriques sur le web
- TextObserver (CEDITEC, http://textopol.u-pec.fr/textobserver/) : analyses textométriques avec interactivité et mise à jour dynamique
- TreeCloud (LIGM, http://www.treecloud.org) : arbres de mots

Formation aux outils de textométrie

http://textopol.u-pec.fr

Web

http://ceditec.u-pec.fr
http://textopol.u-pec.fr

Contact

jean-marc.leblanc@u-pec.fr

Intervenants

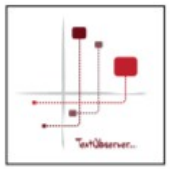
Jean-Marc Leblanc, Philippe Gambette, Claude Martineau, Emilie Née, Marie Pérès

Localisation

Salle multimédia, I2-317 / i2 306
bâtiment i - Campus centre (CMC)
61 avenue du Général de Gaulle
94010 Créteil Cedex

Horaire

samedi 10h-13h
6 séances - 20 heures



Tutoriels, manuel, téléchargement :
http://textopol.u-pec.fr/textobserver/

Stages de formation à TextObserver

Préparation de corpus, prise en main sur les corpus des participants.

En semaine, deux séances (dates à préciser).

PROGRAMME 2019-2020

16 novembre 2019 : Introduction - Approches textométriques des discours.

J.M. Leblanc :
Présentation et typologie pratique de logiciels standard et d'analyseurs de données textuelles. Options théoriques, principes méthodologiques, limites interprétatives.

Initiation à Lexico 3 / 5 :
- Fonctions documentaires, décomptes statistiques, modèles probabilistes.
- Distributions statistiques, distributions linguistiques.
- Analyse factorielle des correspondances, spécificités, fréquences...

14 décembre 2019 : Présentation et prise en main de TextObserver [10h-16h]

J.M. Leblanc :
- Expliciter l'analyse factorielle des correspondances
- Analyser la variation lexicométrique
- Introduction aux opérations de catégorisation
- Recension de corpus et balisage semi-automatisé : présentation de la base Textopol

18 janvier 2020 : Des corpus textuels aux corpus multimodaux (annoter, catégoriser, étiqueter, visualiser, interpréter).

J.M. Leblanc, M. Pérès :
- Transformer des textes pour les soumettre à des traitements automatisés. Repérer les régularités d'un document, extraire des motifs textuels, concaténer des fichiers.
- Prise en main de quelques catégoriseurs, évaluateurs, étiqueteurs (cordial, treetager, tropes)
- Base de données textuelles et outils de constitution et de balisage de corpus.
- Outils de visualisation, de caractérisation de corpus : Gephi, R, Xlstat, Textstat.

29 février 2019 : De la lexicométrie au traitement automatique des langues (TAL) [10h-16h]

P. Gambette (LIGM-MLV) : Les nuages arborés dans TextObserver et Treecloud.
C. Martineau (LIGM-MLV) : Présentation et prise en main du logiciel UNITEX.

PROGRAMME 2019-2020

11 avril 2020 : De la textométrie à l'analyse des données, quels outils pour quels usages?

J.M. Leblanc :
Cooccurrences généralisées et mondes lexicaux : analyses comparées Alceste et Iramuteq.
Ontologies et mondes sémantiques (Tropes, Alceste, Astartex)
Quantifier les données en sciences sociales : Prospero, Nvivo (sous réserve)

Corrélations et causalités interprétatives. Expérimentations, distance intertextuelle et voisinages.

J.M. Leblanc :
Distances, cooccurrences, voisinage
Présentation et prise en main d'Hyperbase : de la lexicométrie à la stylométrie.
Présentation d'Hyperbase en ligne.

16 mai 2020 : Exploration textométrique sur une base annotée - Prise en main du Trameur

E. Née (Céditec-UPEC) :
- i-trameur
- La logique Trameur
- Création d'une base étiquetée et intervention sur l'annotation (annotation manuelle, ajout de niveaux d'annotation).
- Présentation de quelques fonctionnalités textométriques classiques sur une base annotée : cooccurrences spécifiques, patrons, cartes des sections

Séances spécifiques

Certaines séances spécifiques sont organisées sur la journée entière (10h-16h).
Le samedi après-midi sera consacré aux questions des participants (sur rendez-vous).

S'inscrire à la liste de diffusion de Textopol

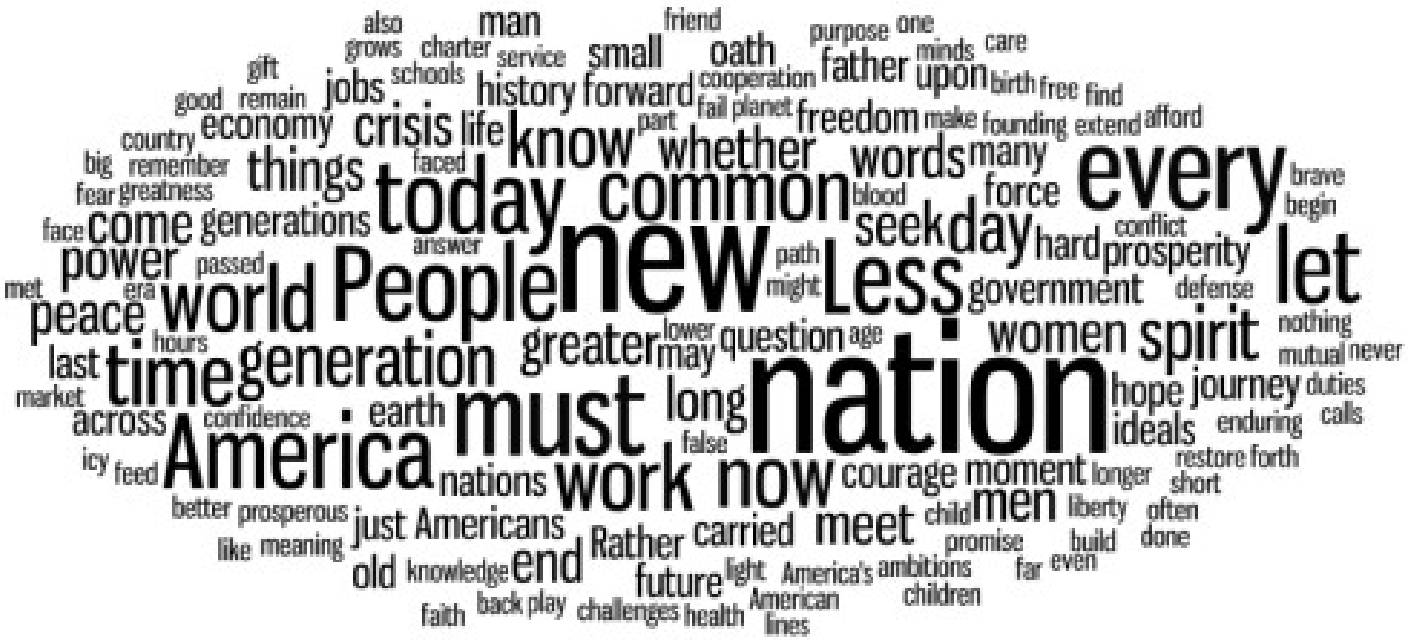
https://listes.u-pec.fr/wws/subscribe/textopol

http://textopol.u-pec.fr/index.php/category/enseignement/seminaire-doctoral/2019-20/

Arbres de mots

Le « nuage arboré », une information double

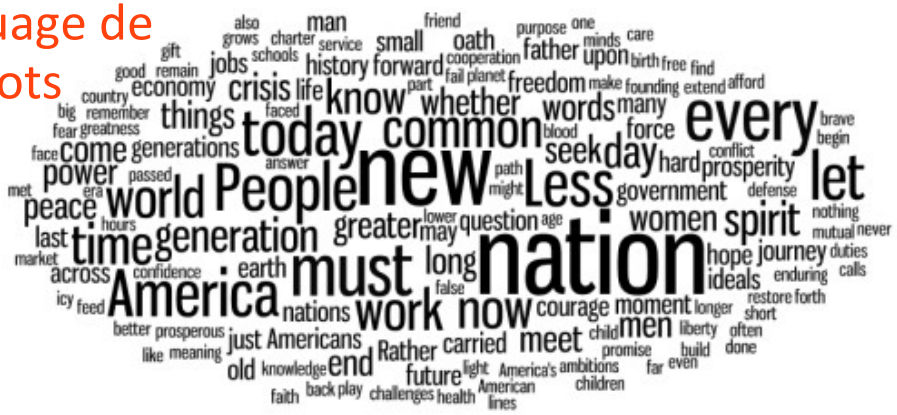
nuage de mots



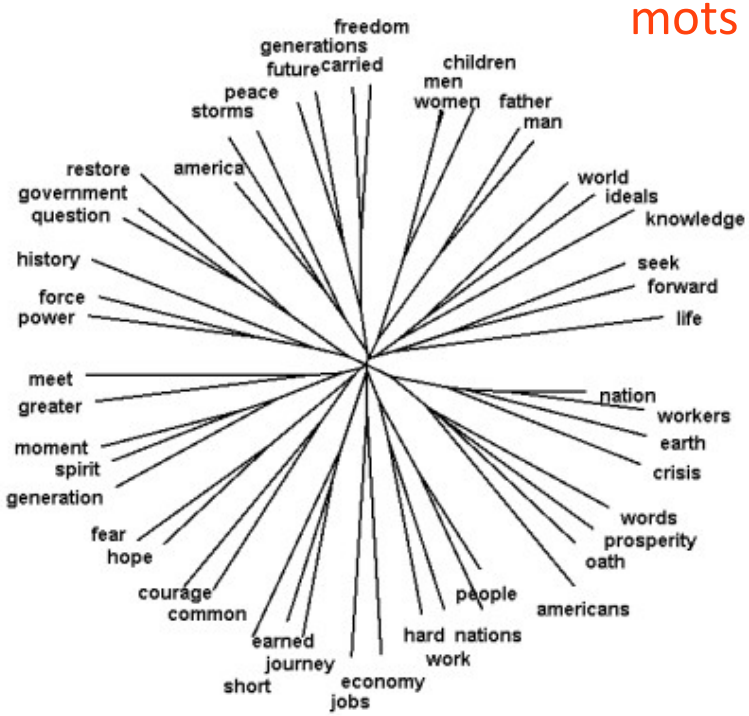
Discours inaugural de Barack Obama en 2008, Wordle

Le « nuage arboré », une information double

nuage de mots



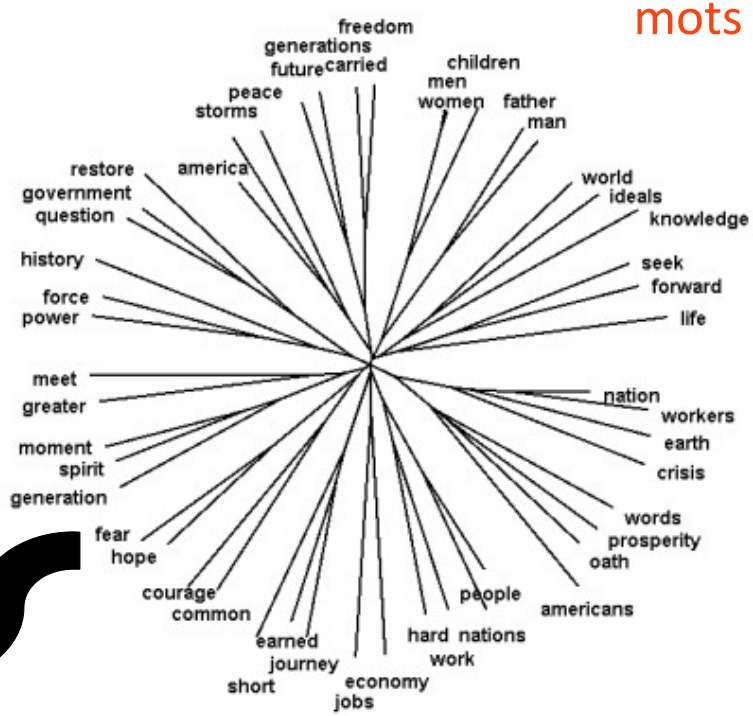
arbre de mots

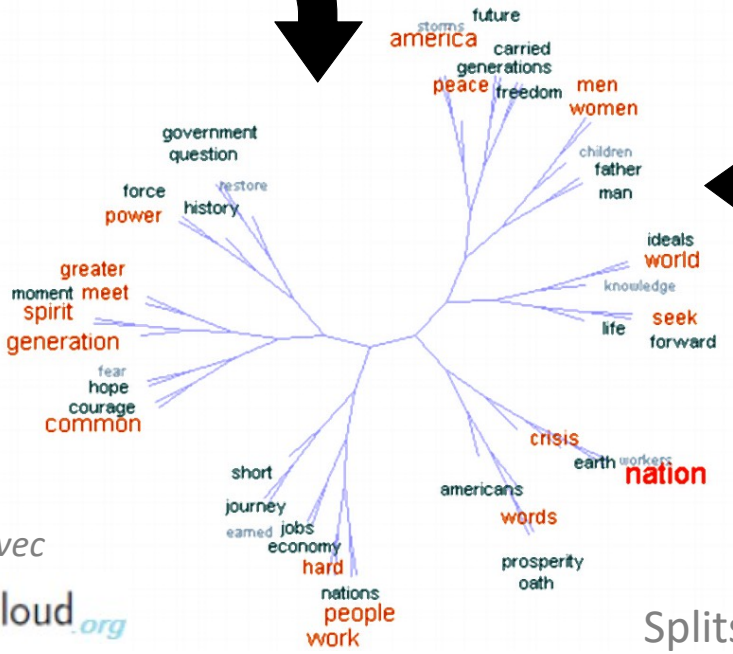


Le « nuage arboré », une information double

nuage de mots

arbre de mots





Discours inaugural de Barack Obama

construit avec

TreeCloud.org

SplitsTree4

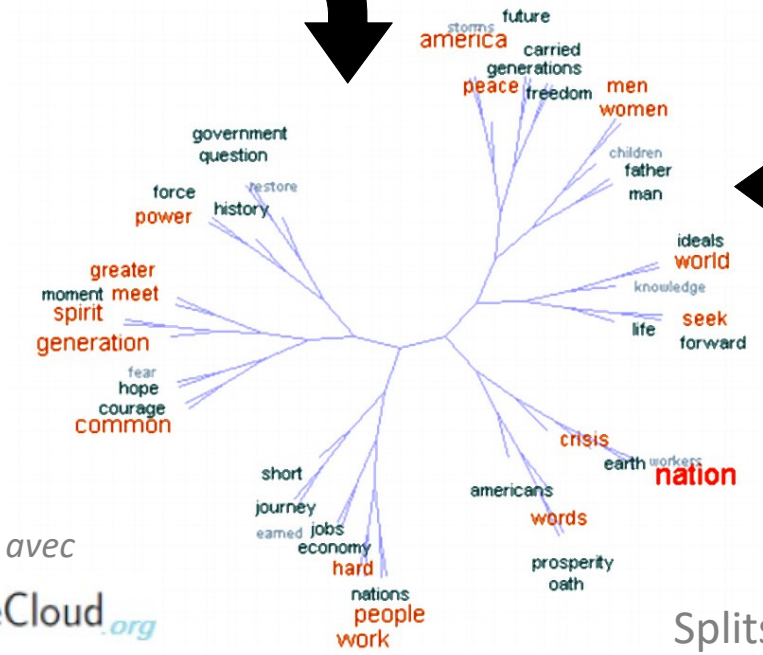
SplitsTree : Huson & Bryant, *Bioinformatics*, 2006

TreeCloud : Gambette & Véronis, *IFCS'09*

Le « nuage arboré », une information double

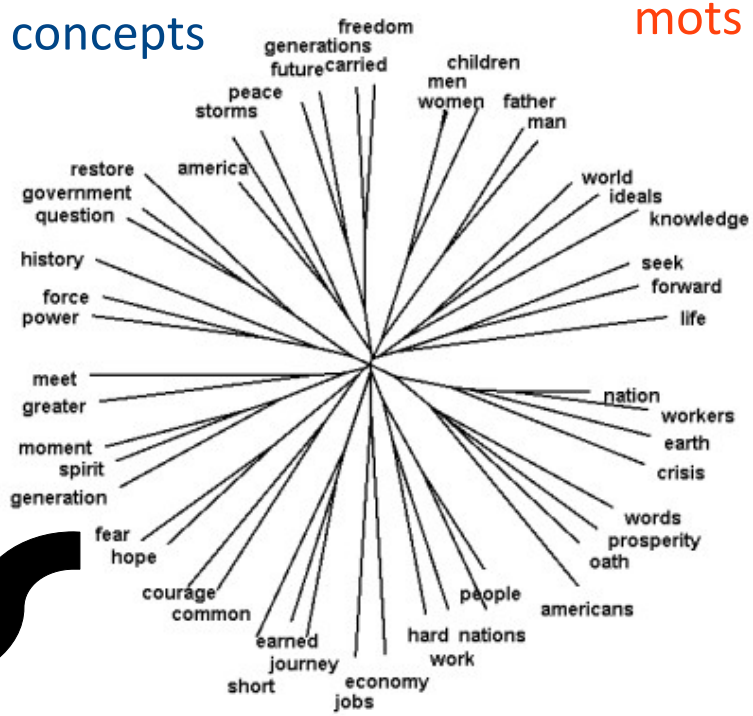
nuage de mots

hiérarchie des mots



hiérarchie des concepts

arbre de mots



Discours inaugural de Barack Obama

construit avec

TreeCloud.org

SplitsTree4

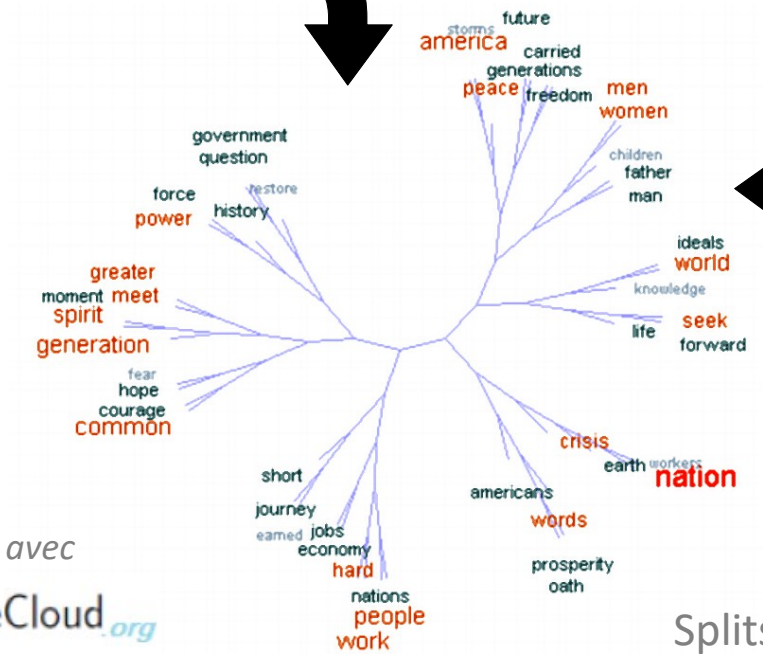
SplitsTree : Huson & Bryant, *Bioinformatics*, 2006

TreeCloud : Gambette & Véronis, *IFCS'09*

Le « nuage arboré », une information double

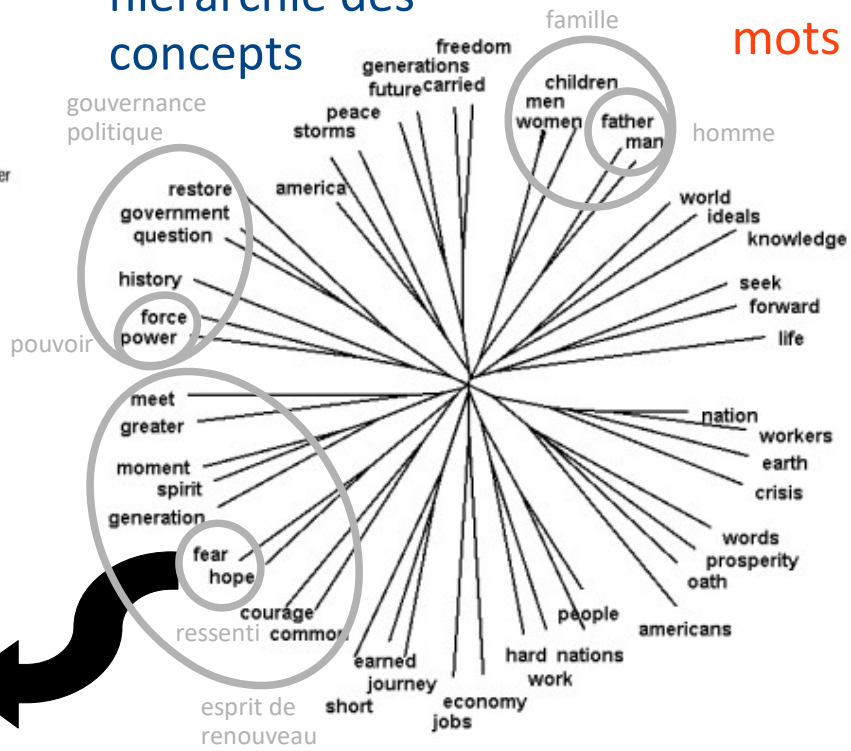
nuage de mots

hiérarchie des mots



hiérarchie des concepts

arbre de mots



Discours inaugural de Barack Obama

construit avec

TreeCloud.org

SplitsTree4

SplitsTree : Huson & Bryant, *Bioinformatics*, 2006

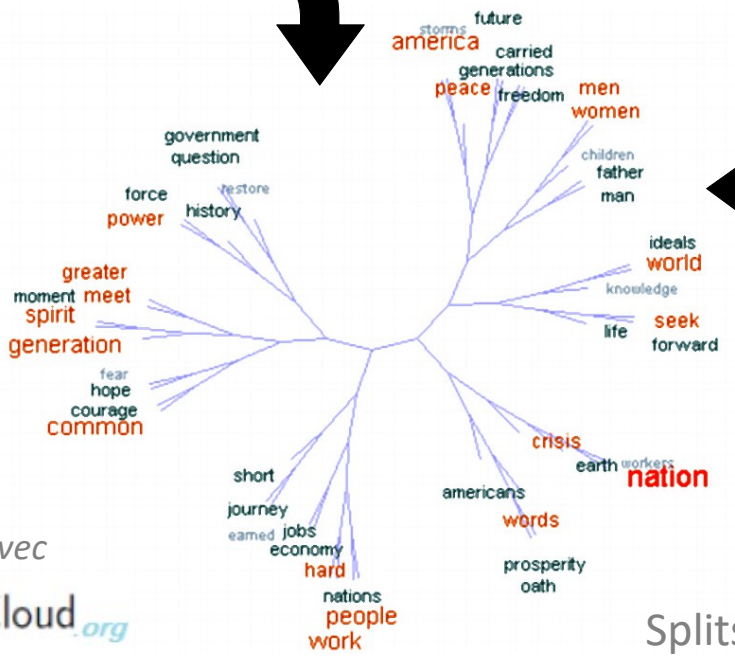
TreeCloud : Gambette & Véronis, *IFCS'09*

Le « nuage arboré », une information double

nuage de mots

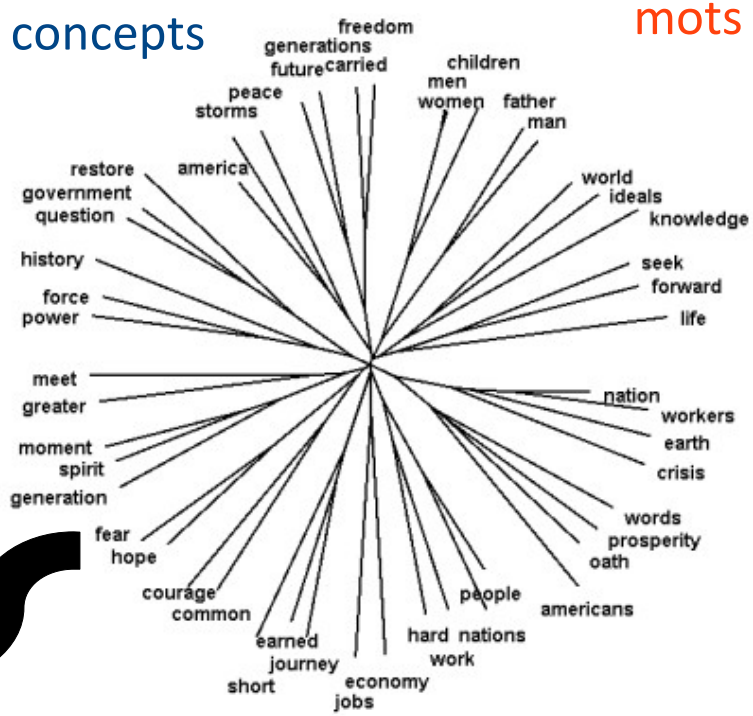
hiérarchie des mots

occurrences



hiérarchie des concepts

arbre de mots



cooccurrences

Discours inaugural de Barack Obama

construit avec

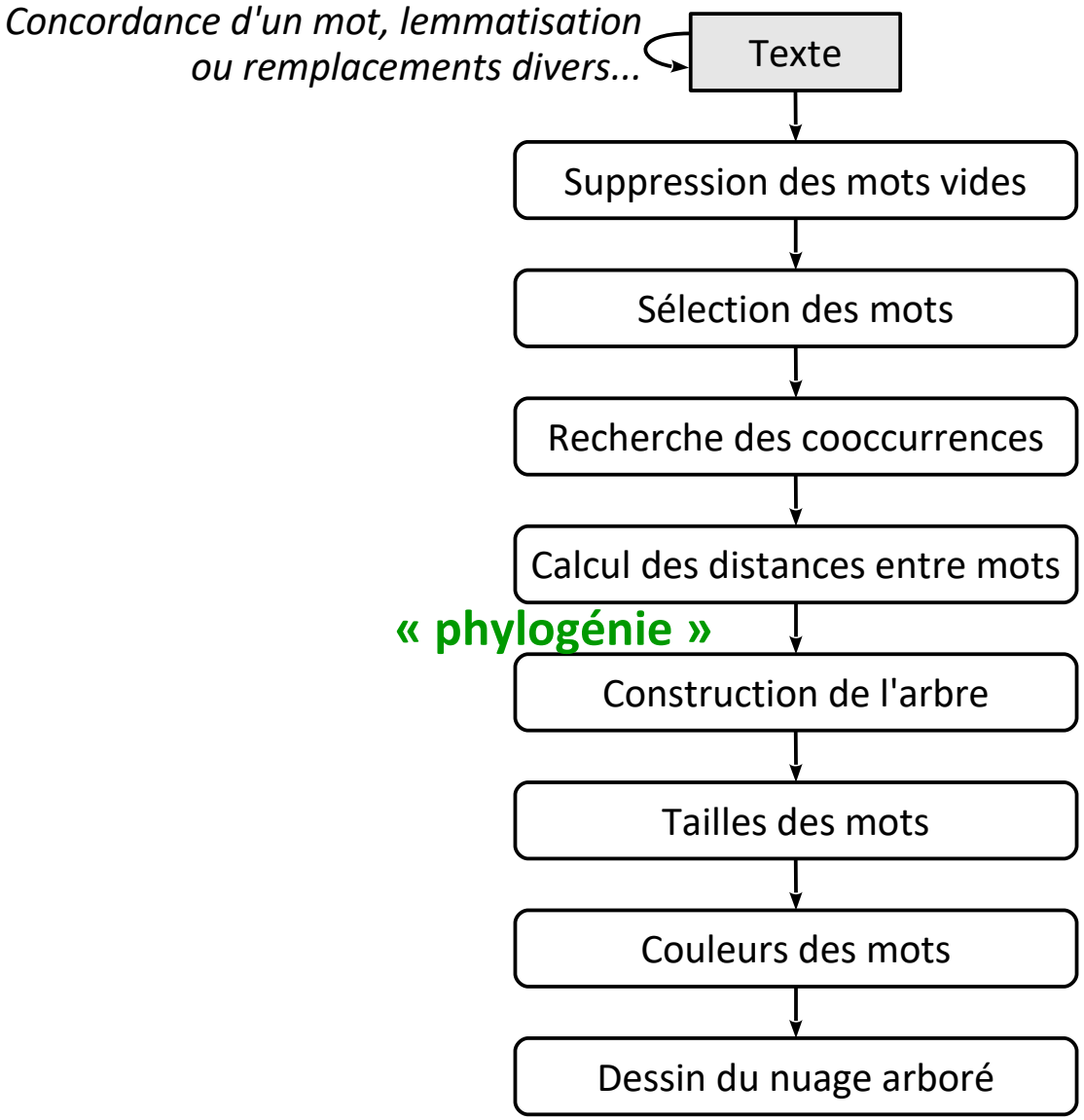
TreeCloud.org

SplitsTree4

SplitsTree : Huson & Bryant, *Bioinformatics*, 2006

TreeCloud : Gambette & Véronis, *IFCS'09*

Processus de construction



Proposé dans la version téléchargeable de TreeCloud

antidico anglais, français

n mots les plus fréquents, mots apparaissant plus de n fois, ou liste personnalisée

Fenêtre de cooccurrence paramétrée par taille et pas de glissement, ou caractère séparateur

12 formules de distance de cooccurrence

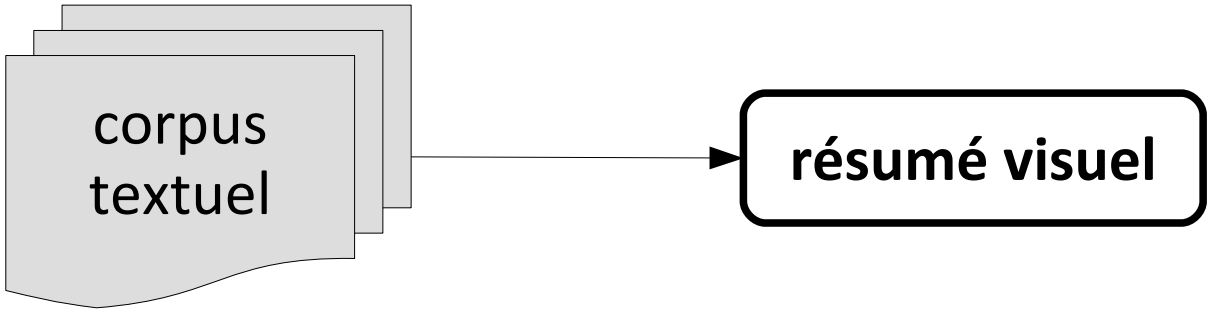
Appel transparent au logiciel SplitsTree

Fréquences ou valeurs personnalisées

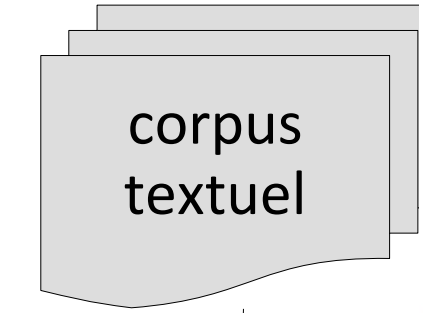
Fréquences, chronologie, dispersion, ciblées sur la cooccurrence d'un mot, ou valeurs personnalisées

Appel transparent au logiciel SplitsTree ou Dendroscope

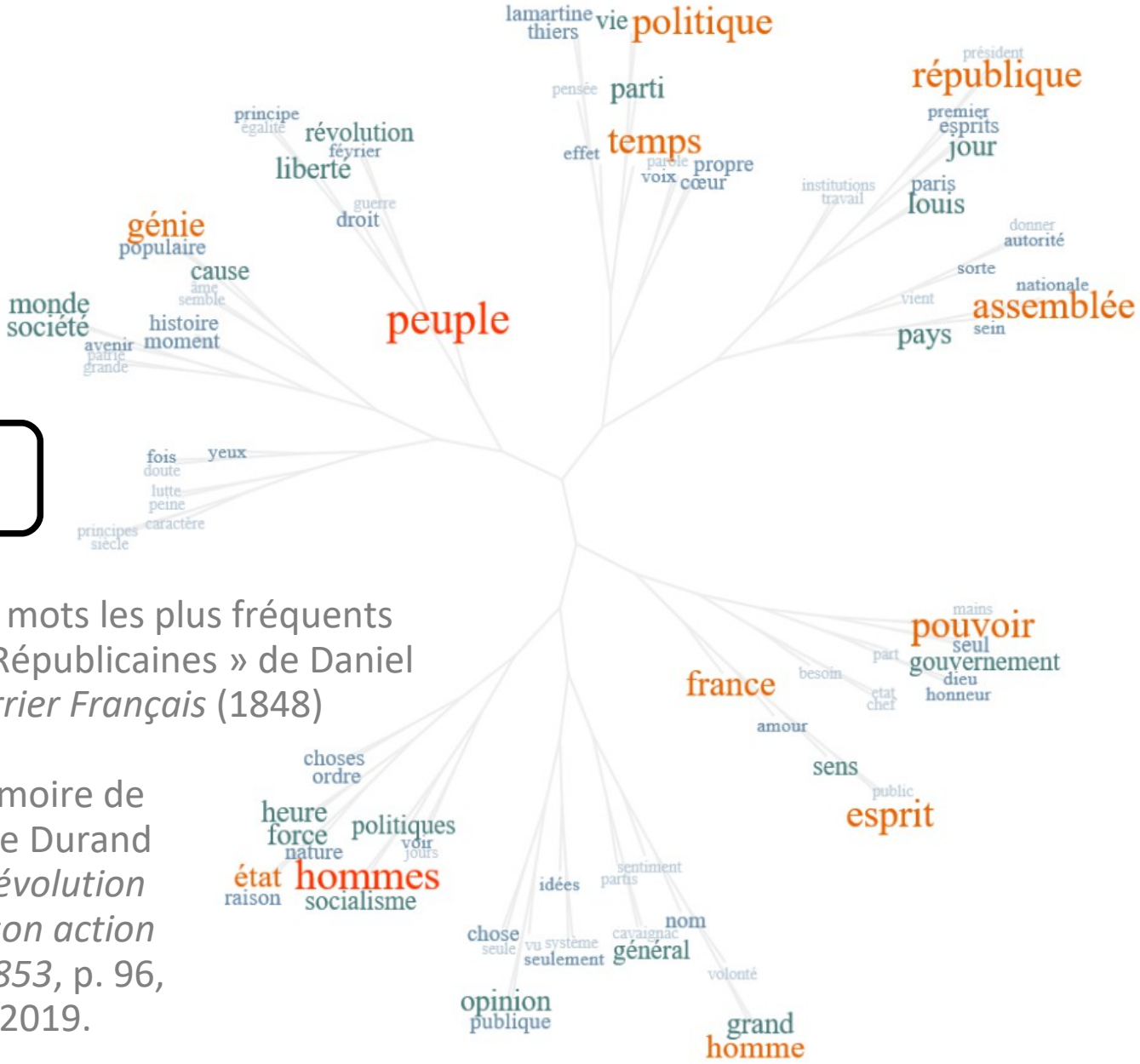
Le « nuage arboré », pour quoi faire ?



Le « nuage arboré », pour quoi faire ?



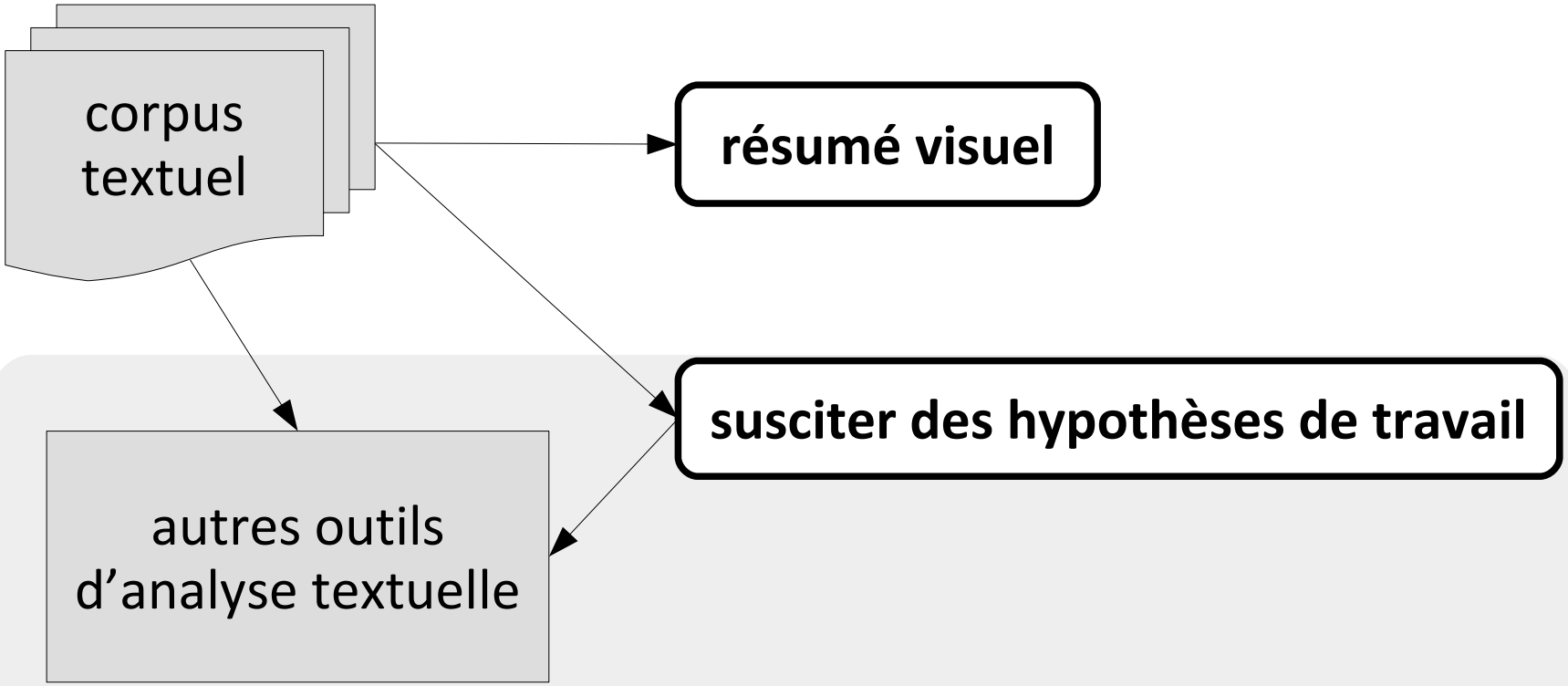
résumé visuel



Nuage arboré des100 mots les plus fréquents dans les 18 « Lettres Républicaines » de Daniel Stern publiées le *Courrier Français* (1848)

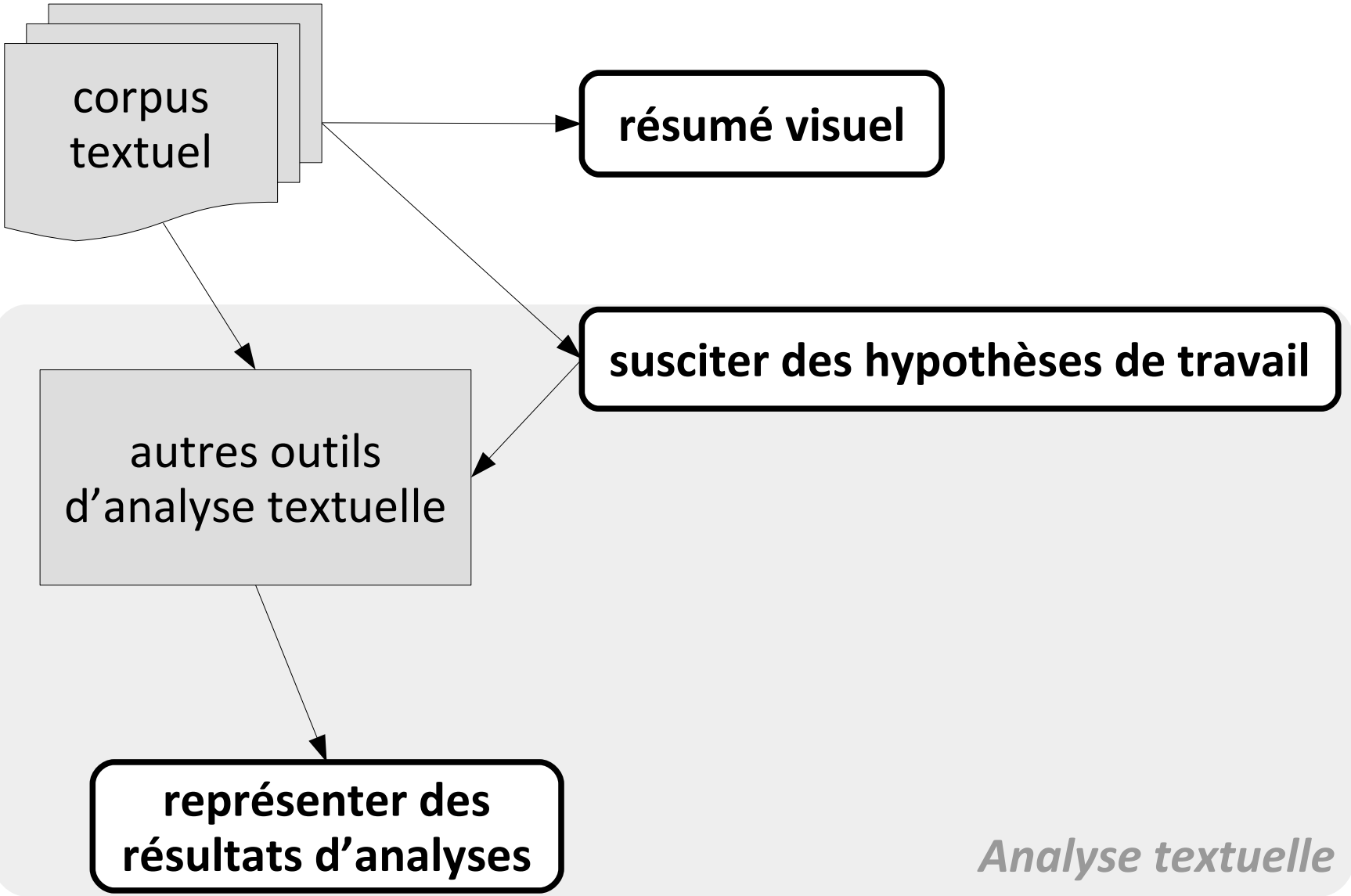
Figure extraite du mémoire de master 2 de Constance Durand *Marie d'Agoult et la révolution de 1848. Sa pensée, son action et son œuvre, 1847-1853*, p. 96, Sorbonne Université, 2019.

Le « nuage arboré », pour quoi faire ?



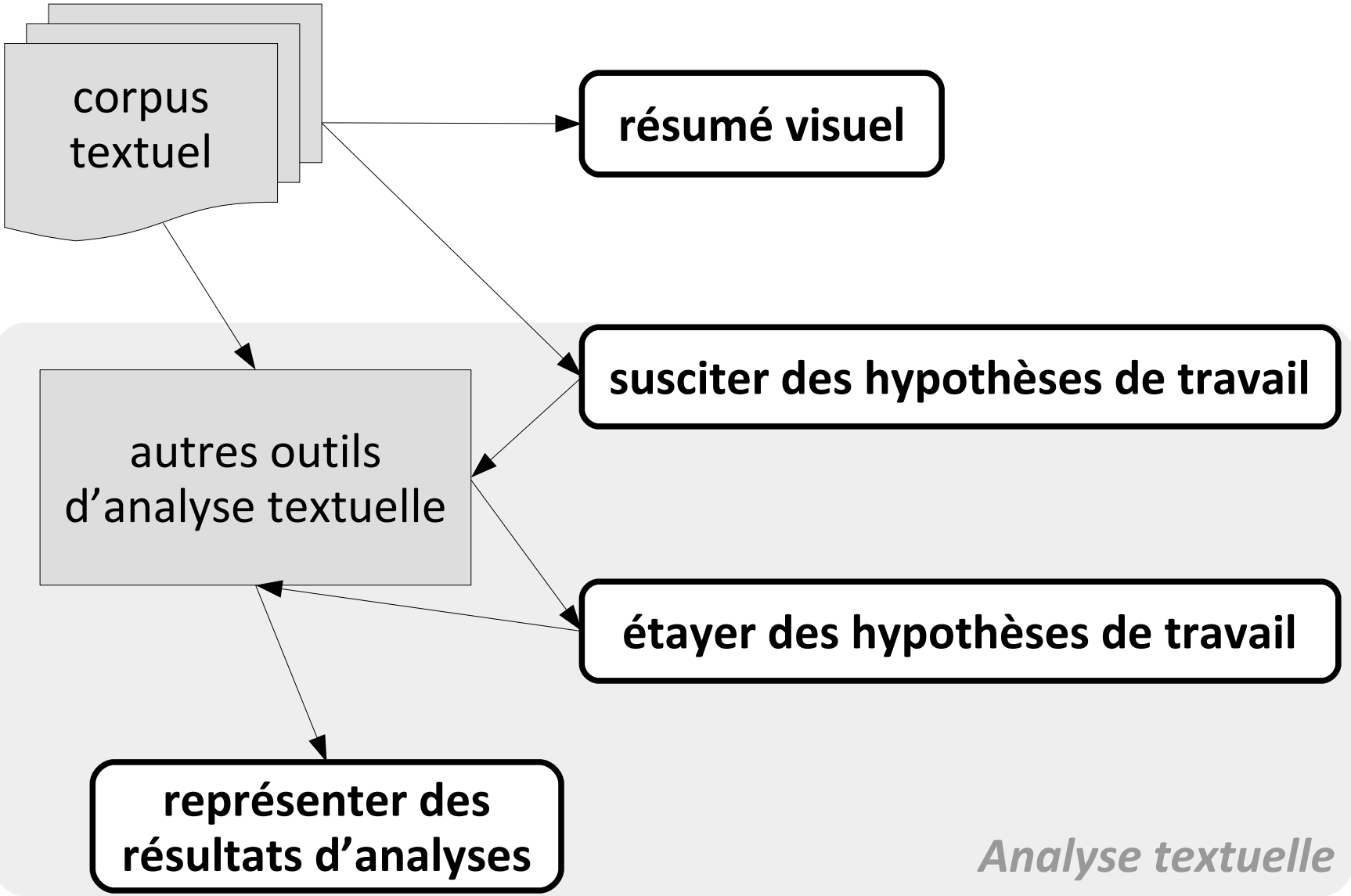
Amstutz & Gambette, *JADT 2010*

Le « nuage arboré », pour quoi faire ?



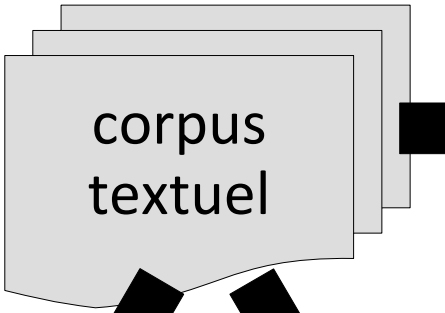
Amstutz & Gambette, *JADT 2010*

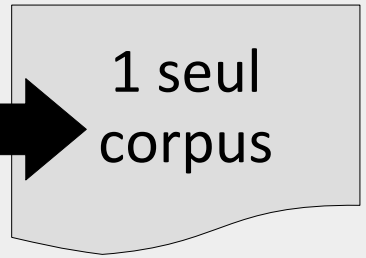
Le « nuage arboré », pour quoi faire ?



Amstutz & Gambette, *JADT 2010*

Exploration de corpus avec TreeCloud

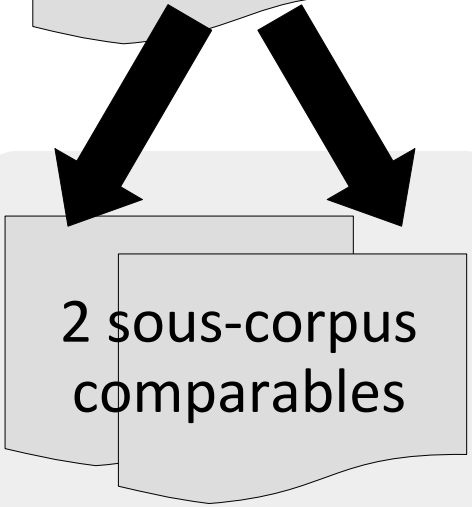




- Interpréter les regroupements
- Examiner les voisinages des mots fréquents dans l'arbre
- Examiner le voisinage de mots ciblés dans l'arbre («acteurs», verbes, etc.)

Si le corpus est assez grand :

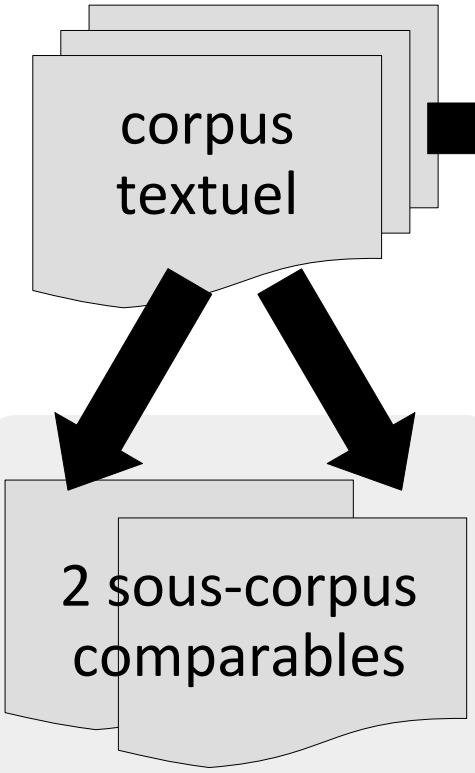
- Identifier les mots plutôt au début ou à la fin du corpus
- extraire le contexte dans le corpus de certains mots ciblés et utiliser ces données comme corpus.



- Identifier des regroupements communs ou différents
- Comparer la fréquence des mots fréquents dans les deux arbres
- Comparer (quanti+quali) les regroupements dans les deux arbres
- Construire le nuage arboré des mots spécifiques de chaque sous-corpus

Travail en cours avec Edna Hernandez

Exploration de corpus avec TreeCloud



1 seul corpus

- **Interpréter les regroupements**
- Examiner les voisinages des mots fréquents dans l'arbre
- Examiner le voisinage de mots ciblés dans l'arbre («acteurs», verbes, etc.)

Si le corpus est assez grand :

- Identifier les mots plutôt au début ou à la fin du corpus
- extraire le contexte dans le corpus de certains mots ciblés et utiliser ces données comme corpus.

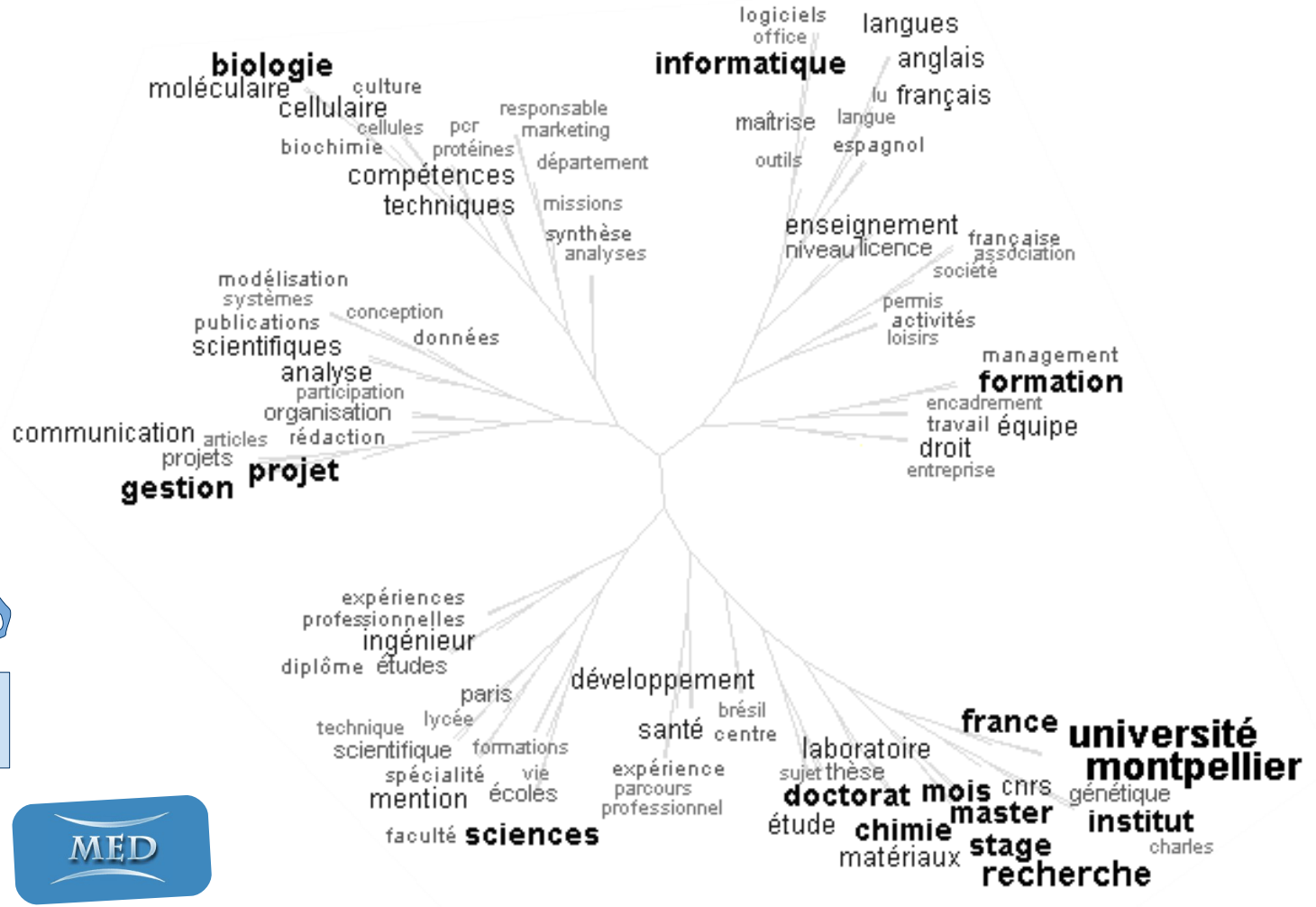
- Identifier des regroupements communs ou différents
- Comparer la fréquence des mots fréquents dans les deux arbres
- Comparer (quanti+quali) les regroupements dans les deux arbres
- Construire le nuage arboré des mots spécifiques de chaque sous-corpus

Travail en cours avec Edna Hernandez

Méthode : interpréter les regroupements

Dessiner des « patates »

Corpus : une centaine de CV soumis à une rencontre docteurs-entreprises



Rencontre Docteurs & Entreprises

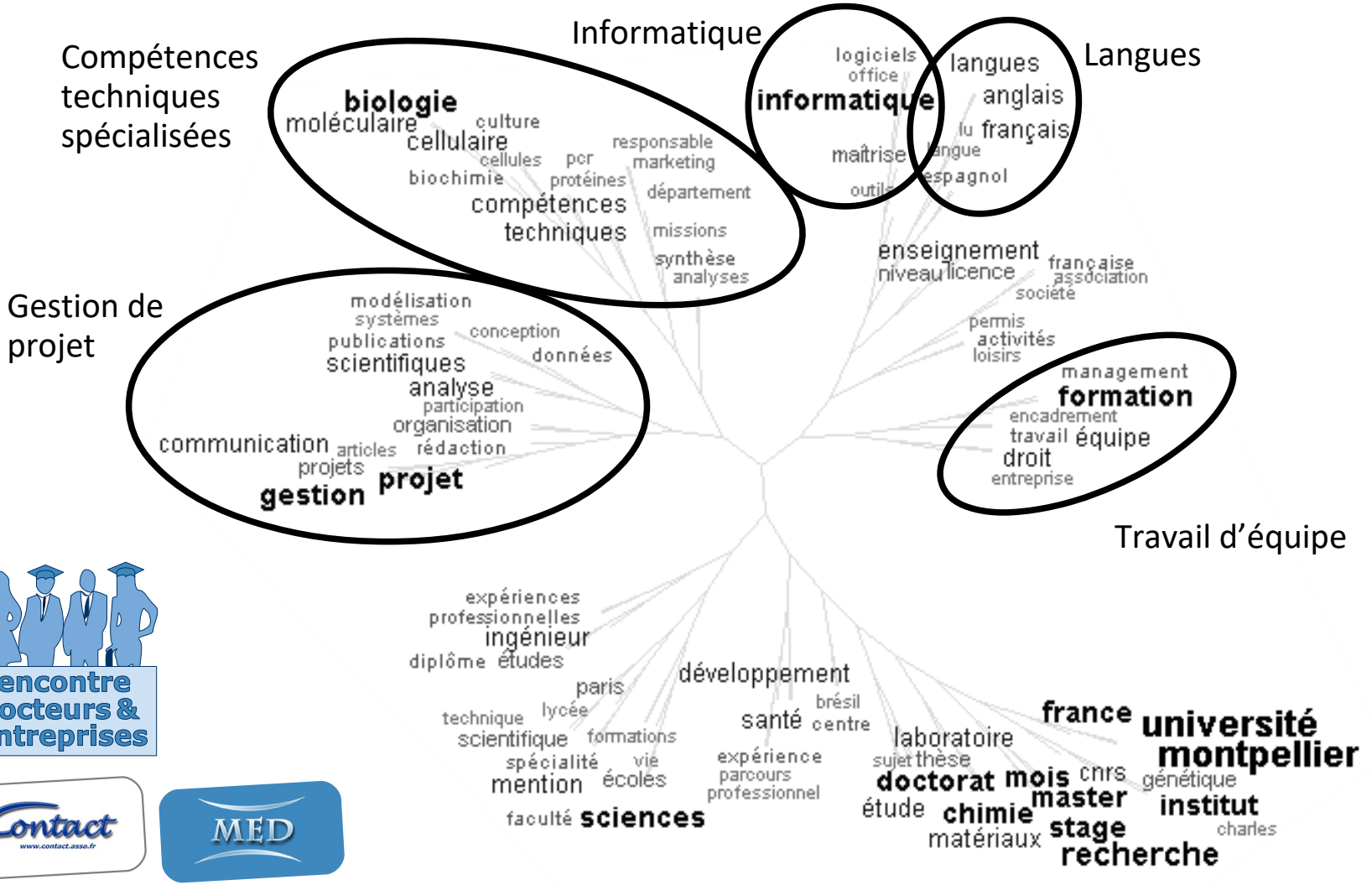
Contact www.contact.asso.fr

MED

Méthode : interpréter les regroupements

Dessiner des « patates »

Corpus : une centaine de CV soumis à une rencontre docteurs-entreprises

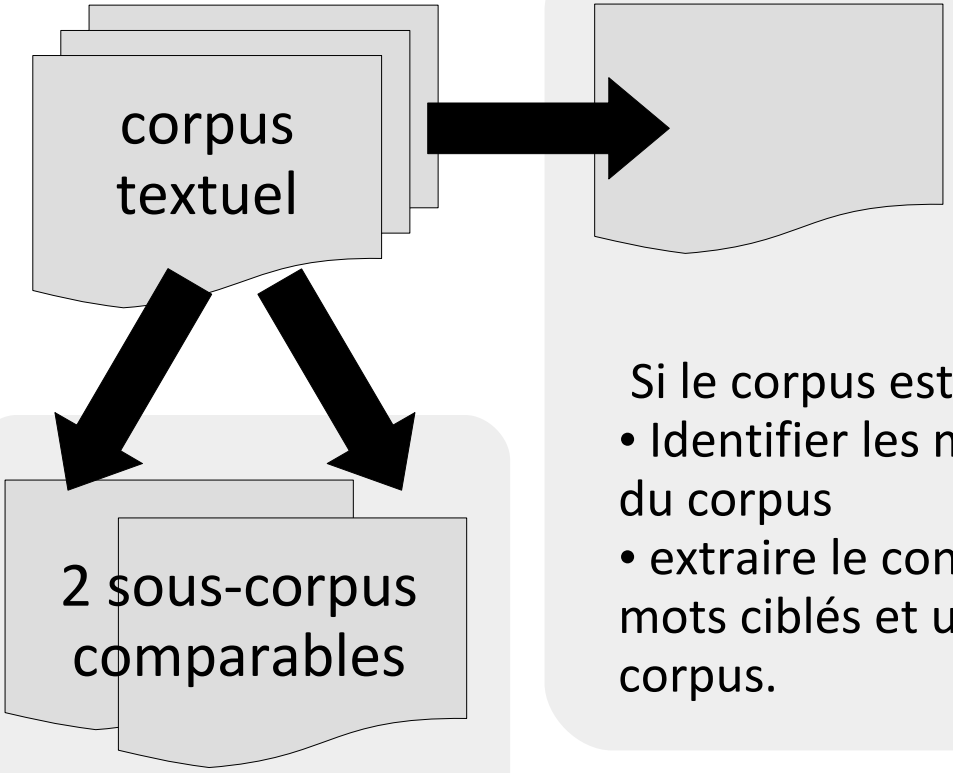


Rencontre Docteurs & Entreprises

Contact
www.contact.asso.fr

MED

Exploration de corpus avec TreeCloud



- Interpréter les regroupements
- **Examiner les voisinages des mots fréquents dans l'arbre**
- Examiner le voisinage de mots ciblés dans l'arbre («acteurs», verbes, etc.)

Si le corpus est assez grand :

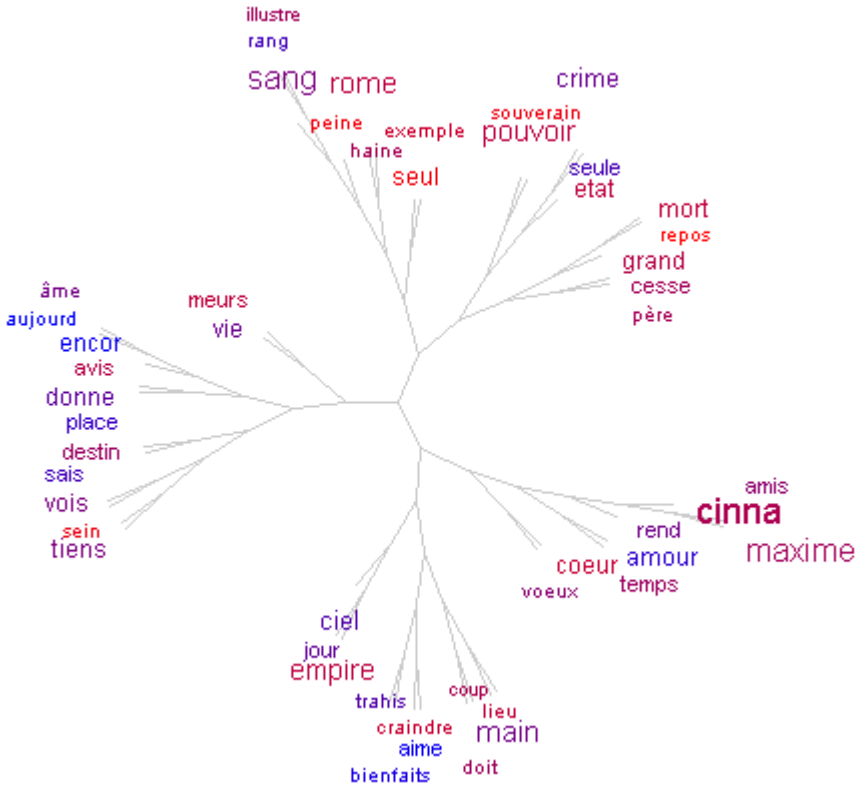
- Identifier les mots plutôt au début ou à la fin du corpus
- extraire le contexte dans le corpus de certains mots ciblés et utiliser ces données comme corpus.

- Identifier des regroupements communs ou différents
- Comparer la fréquence des mots fréquents dans les deux arbres
- Comparer (quanti+quali) les regroupements dans les deux arbres
- Construire le nuage arboré des mots spécifiques de chaque sous-corpus

Travail en cours avec Edna Hernandez

Méthode : voisinage des mots fréquents

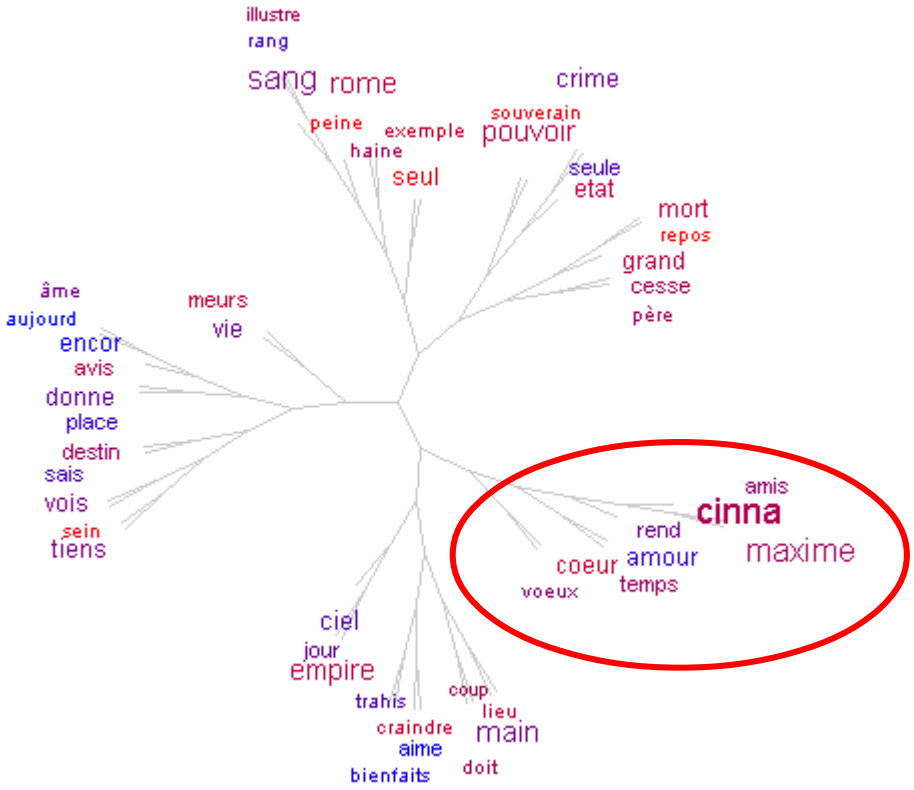
Amstutz & Gambette,
JADT 2010



Nuage arboré des 50 mots les plus fréquents des paroles d'Auguste dans Cinna

Méthode : voisinage des mots fréquents

Amstutz & Gambette,
JADT 2010

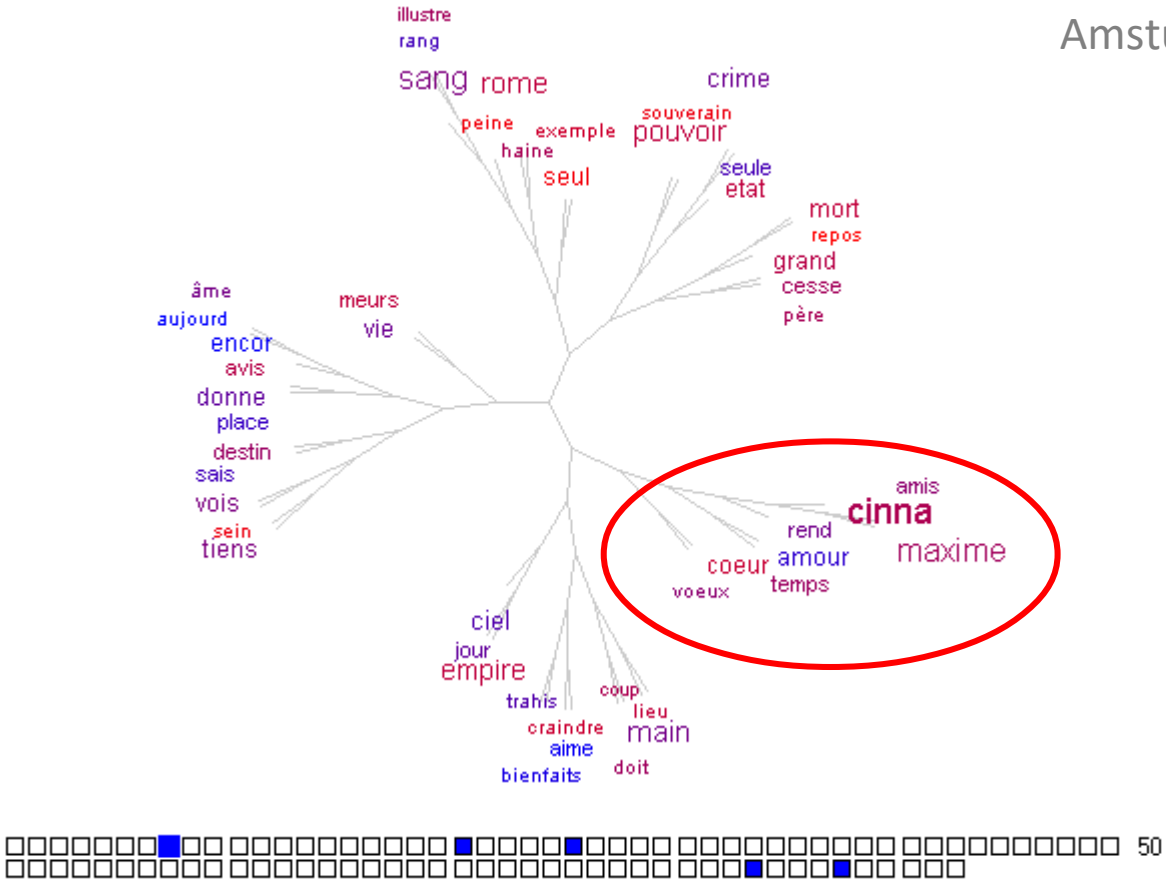


Nuage arboré des 50 mots les plus fréquents des paroles d'Auguste dans Cinna

Pour manipuler facilement des pièces de théâtre dans TreeCloud, chargement dans un fichier tableur (exemples de fichiers Open Office pour *Cinna* et *Othon* de Corneille sur http://theatre.treecloud.org) : possibilité de filtrer les lignes (répliques) en fonction de la valeur dans une colonne donnée → sélectionner un acte, une scène, un personnage.

Méthode : voisinage des mots fréquents

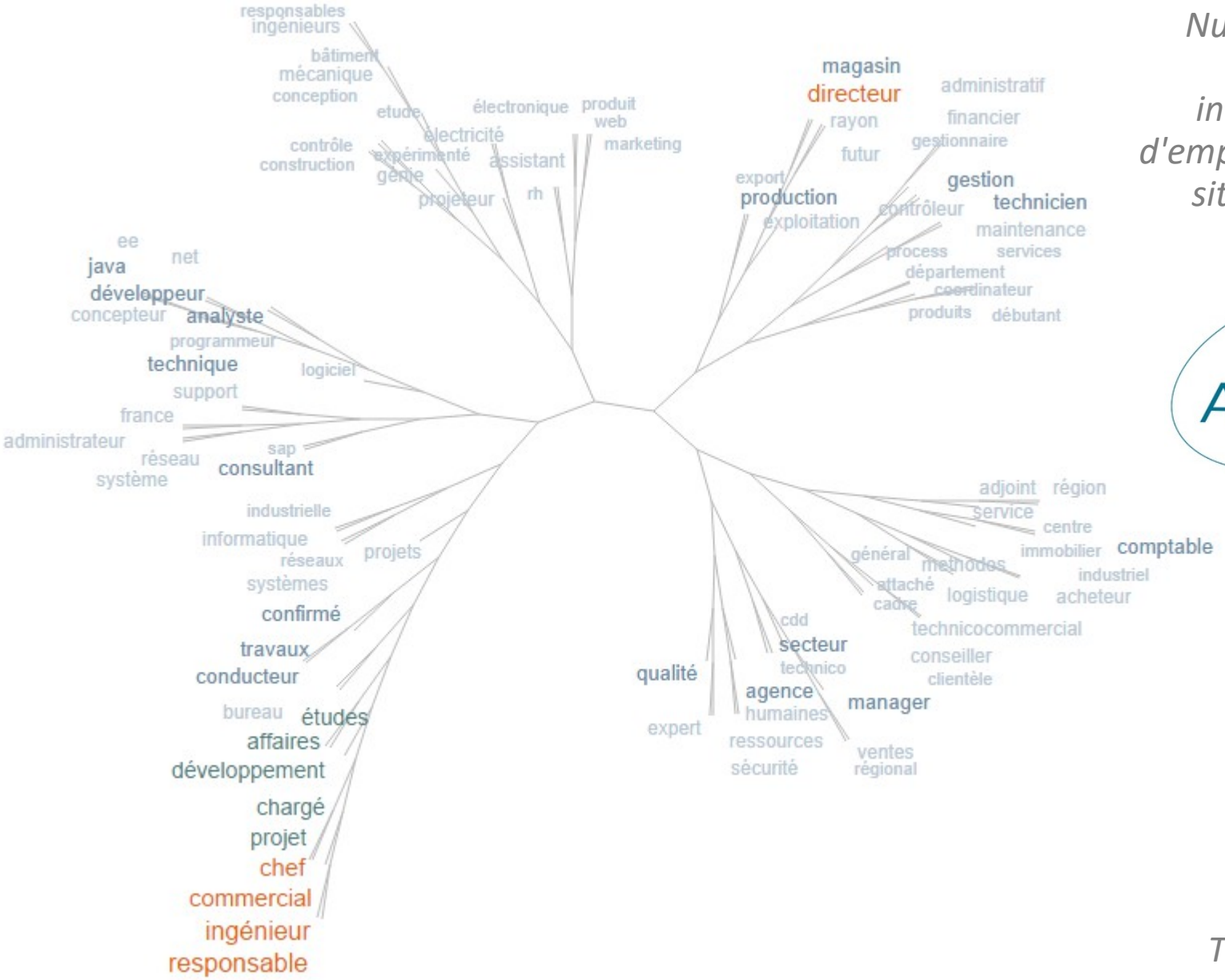
Amstutz & Gambette, JADT 2010



Carte des sections Lexico3 et contextes de « amis » dans les paroles d'Auguste dans Cinna.

1. *Voilà, mes chers **amis**, ce qui me met en peine.*
2. *Quoi ! mes plus chers **amis** ! quoi ! Cinna ! quoi ! Maxime !*
3. *Reprenez le pouvoir que vous m'avez commis, Si donnant des sujets il ôte les **amis***
4. *Soyons **amis**, Cinna, c'est moi qui t'en convie*
5. *Il nous a trahis tous ; mais ce qu'il a commis Vous conserve innocents, et me rend mes **amis**.*

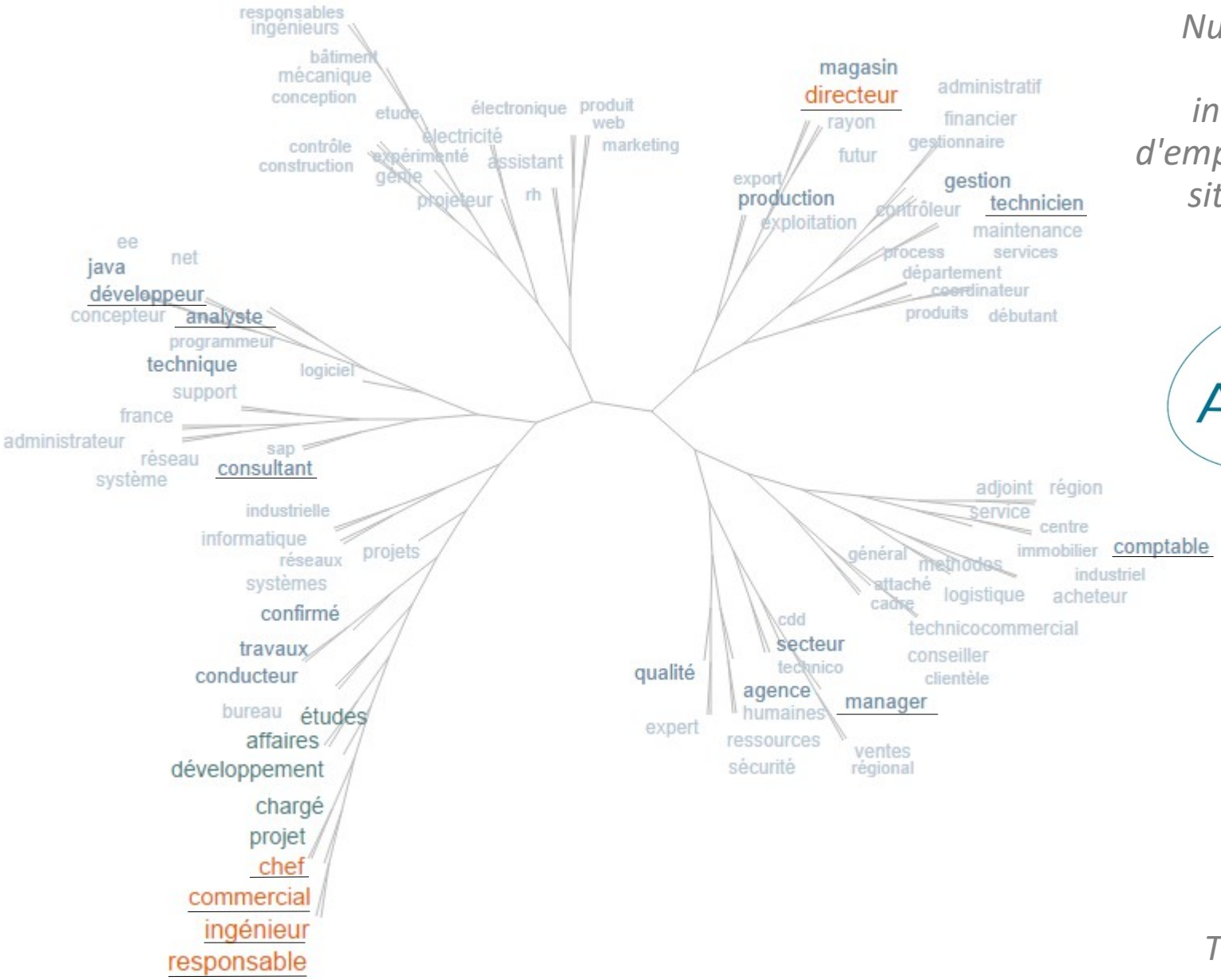
Méthode : voisinage des mots fréquents



Nuage arboré de plus de 4800 intitulés d'offres d'emploi extraits du site de l'APEC en avril 2011.

Travail de 2011 avec Paola Salle

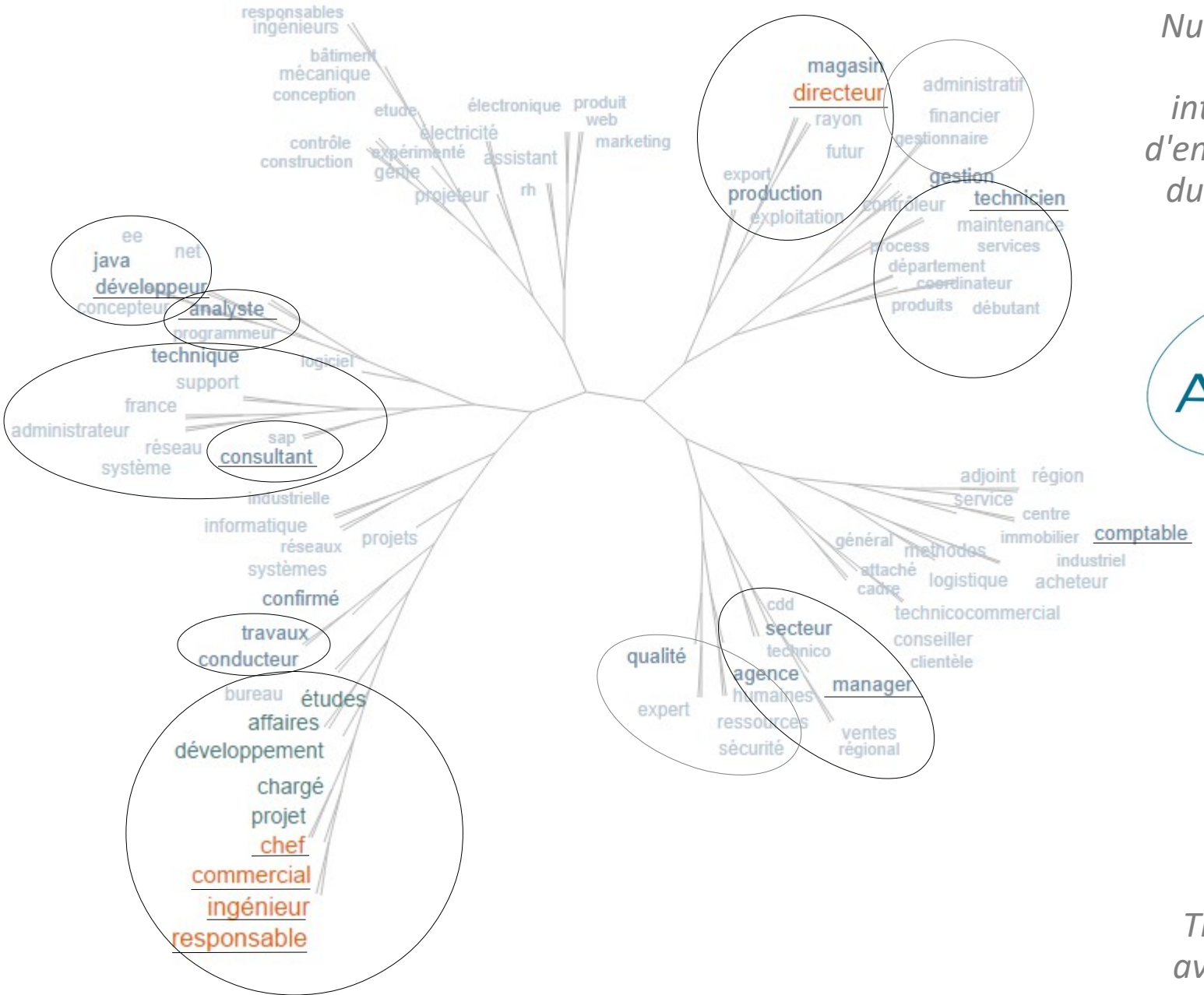
Méthode : voisinage des mots fréquents



Nuage arboré de plus de 4800 intitulés d'offres d'emploi extraits du site de l'APEC en avril 2011.

Travail de 2011 avec Paola Salle

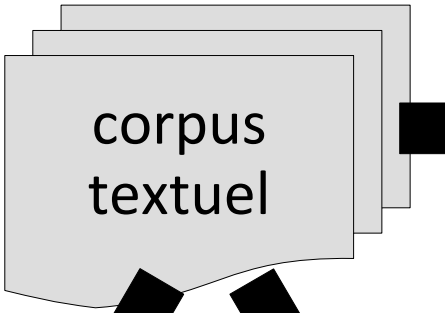
Méthode : voisinage des mots fréquents

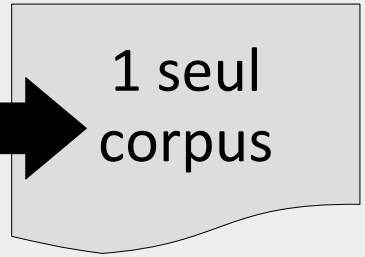


Nuage arboré de plus de 4800 intitulés d'offres d'emploi extraites du site de l'APEC en avril 2011.

Travail de 2011 avec Paola Salle

Exploration de corpus avec TreeCloud

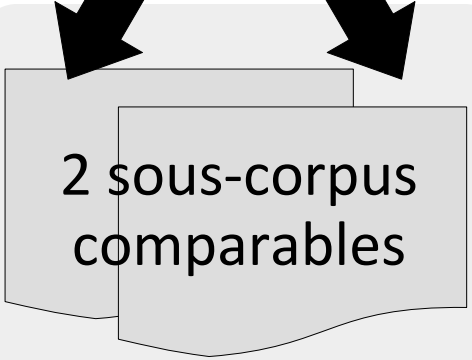




- Interpréter les regroupements
- Examiner les voisinages des mots fréquents dans l'arbre
- **Examiner le voisinage de mots ciblés dans l'arbre («acteurs», verbes, etc.)**

Si le corpus est assez grand :

- Identifier les mots plutôt au début ou à la fin du corpus
- extraire le contexte dans le corpus de certains mots ciblés et utiliser ces données comme corpus.



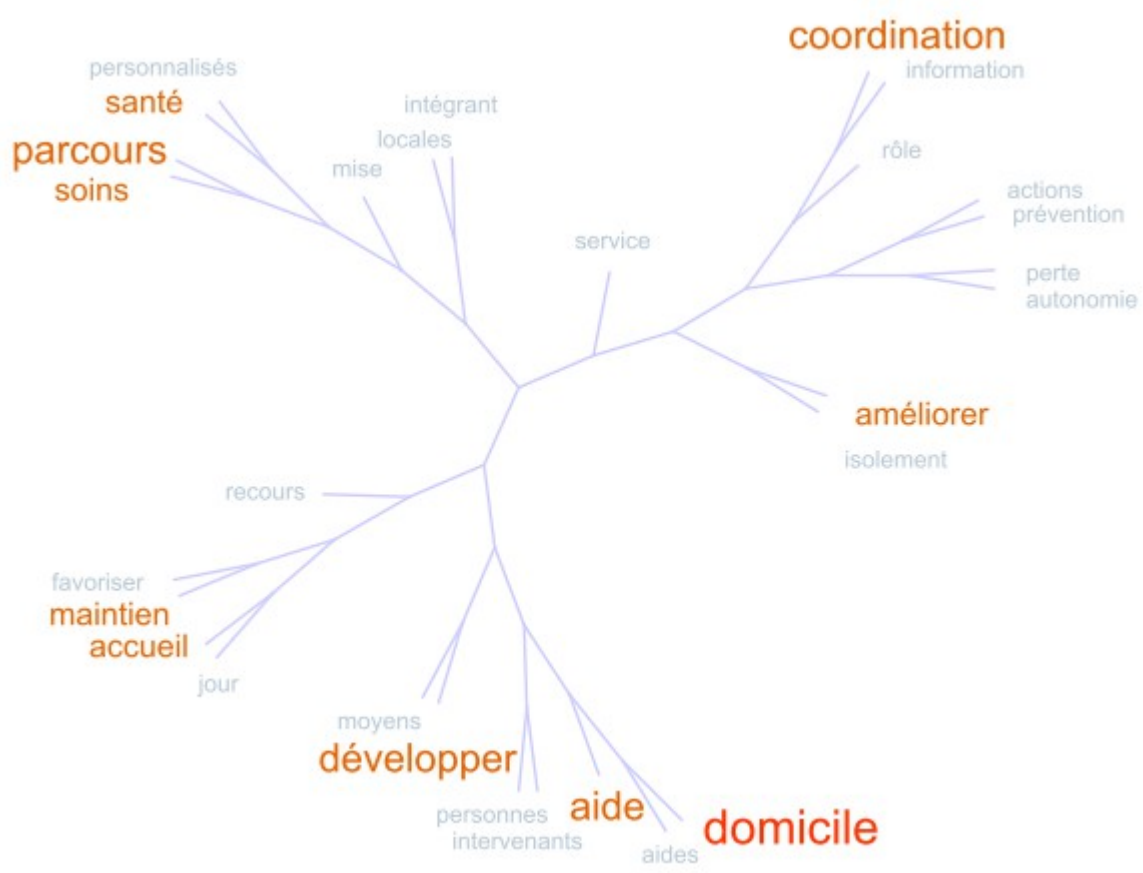
- Identifier des regroupements communs ou différents
- Comparer la fréquence des mots fréquents dans les deux arbres
- Comparer (quanti+quali) les regroupements dans les deux arbres
- Construire le nuage arboré des mots spécifiques de chaque sous-corpus

Travail en cours avec Edna Hernandez

Méthode : voisinage des verbes

Corpus : réponses à des questions ouvertes à des professionnels de la santé sur le parcours de santé des personnes âgées dans les Alpes de Haute-Provence

Suggestions d'améliorations :

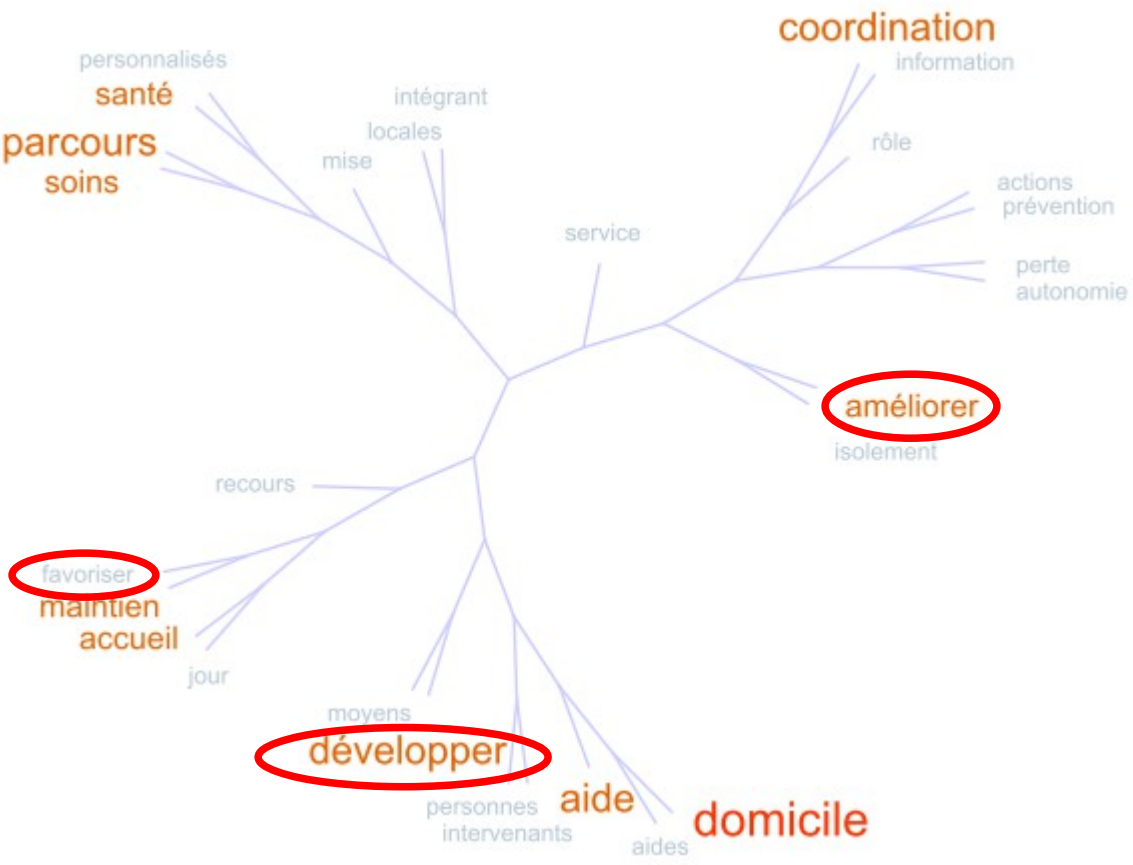


P. Grenier-Tisserand et
l'équipe projet de la DT 04

Méthode : voisinage des verbes

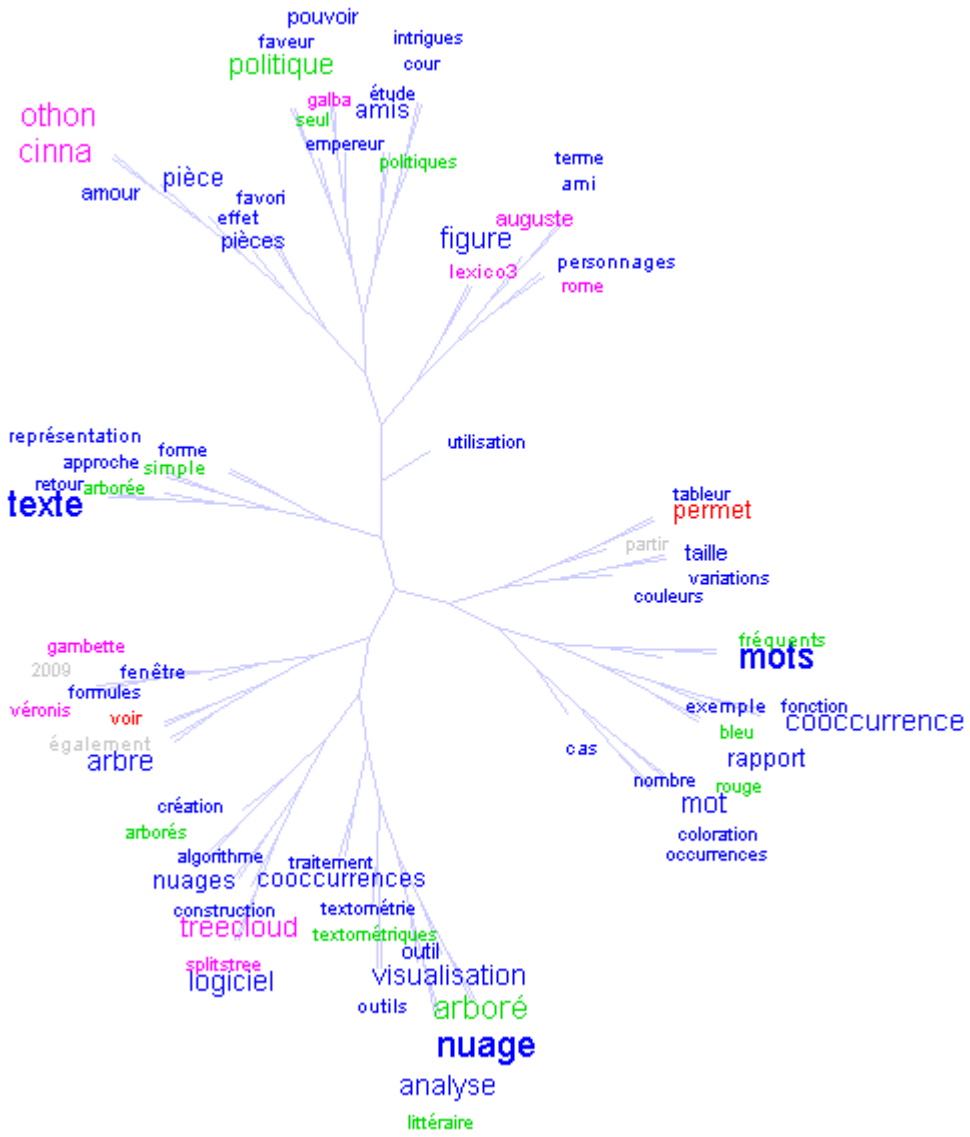
Corpus : réponses à des questions ouvertes à des professionnels de la santé sur le parcours de santé des personnes âgées dans les Alpes de Haute-Provence

Suggestions d'améliorations :



P. Grenier-Tisserand et
l'équipe projet de la DT 04

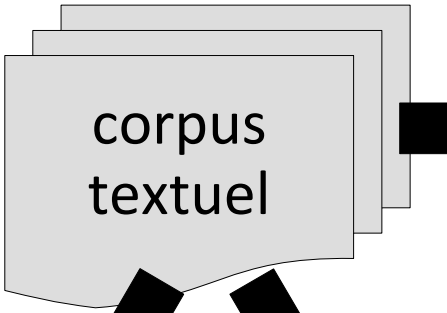
Perspective : coloration grammaticale

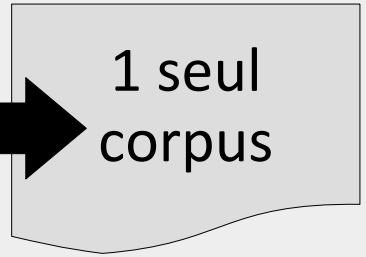


noms
adjectifs
verbes
noms propres

Nuage arboré des mots apparaissant 5 fois ou plus dans l'article d'Amstutz & Gambette, JADT 2010, distance Liddell, fenêtre de 20 mots, coloration personnalisée à partir d'un étiquetage TreeTagger

Exploration de corpus avec TreeCloud

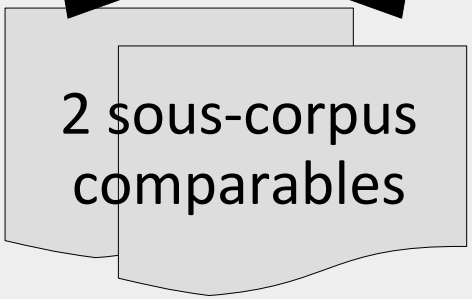




- Interpréter les regroupements
- Examiner les voisinages des mots fréquents dans l'arbre
- Examiner le voisinage de mots ciblés dans l'arbre («acteurs», verbes, etc.)

Si le corpus est assez grand :

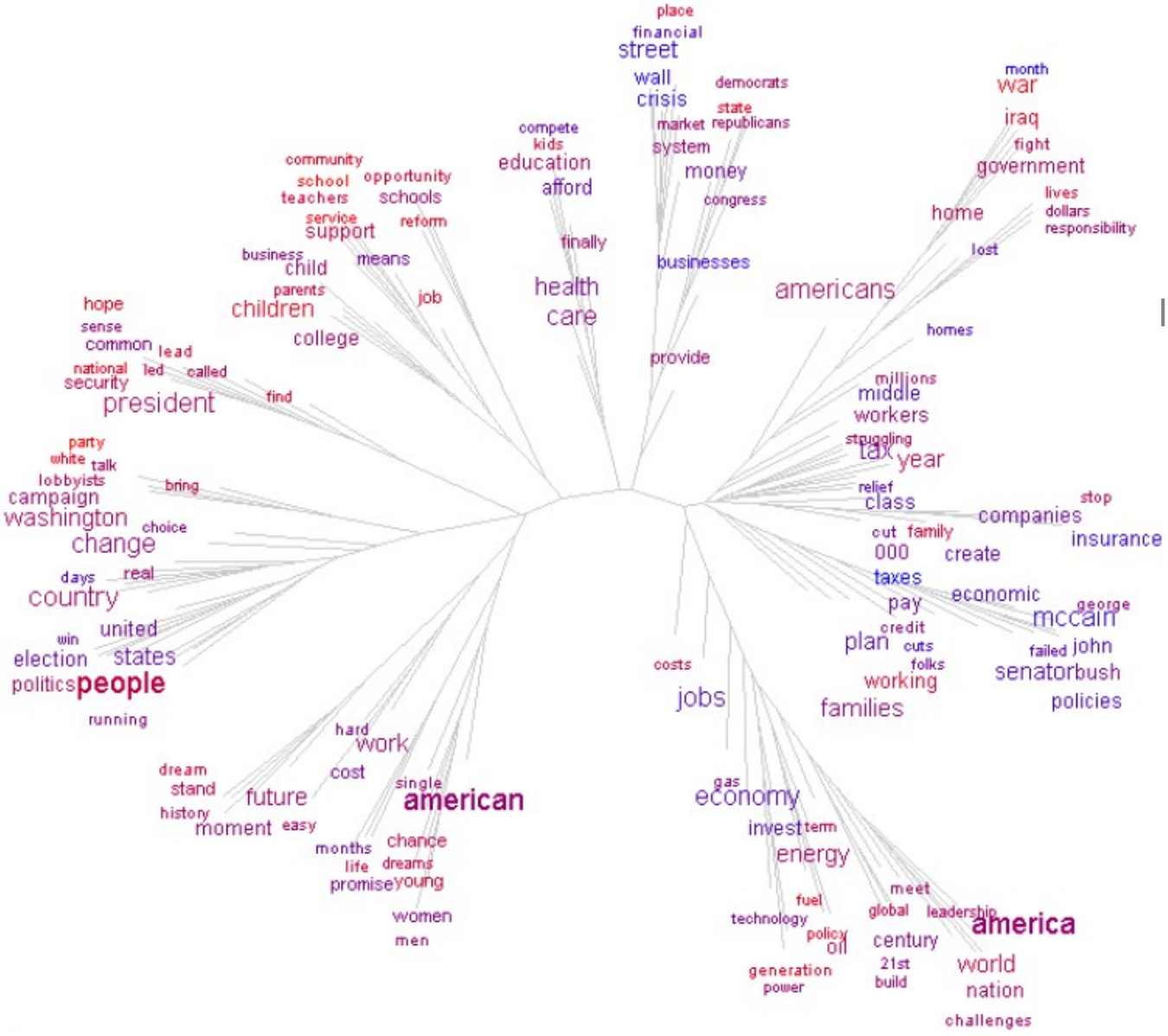
- **Identifier les mots plutôt au début ou à la fin du corpus**
- extraire le contexte dans le corpus de certains mots ciblés et utiliser ces données comme corpus.



- Identifier des regroupements communs ou différents
- Comparer la fréquence des mots fréquents dans les deux arbres
- Comparer (quanti+quali) les regroupements dans les deux arbres
- Construire le nuage arboré des mots spécifiques de chaque sous-corpus

Travail en cours avec Edna Hernandez

Méthode : mots au début ou à la fin

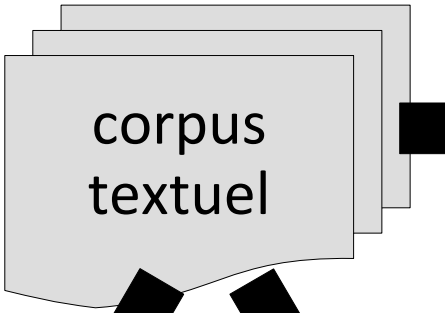


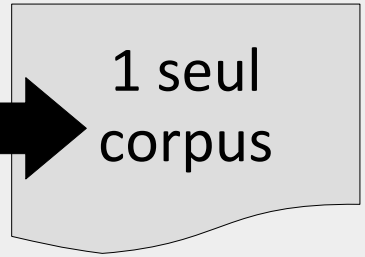
Nuage arboré de l'ensemble des discours de campagne de 2008 de Barack Obama, coloration chronologique

début de la campagne
fin de la campagne

Gambette & Véronis, IFCS 2009

Exploration de corpus avec TreeCloud

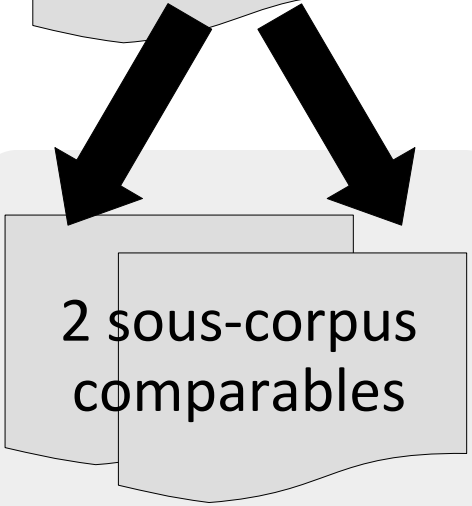




- Interpréter les regroupements
- Examiner les voisinages des mots fréquents dans l'arbre
- Examiner le voisinage de mots ciblés dans l'arbre («acteurs», verbes, etc.)

Si le corpus est assez grand :

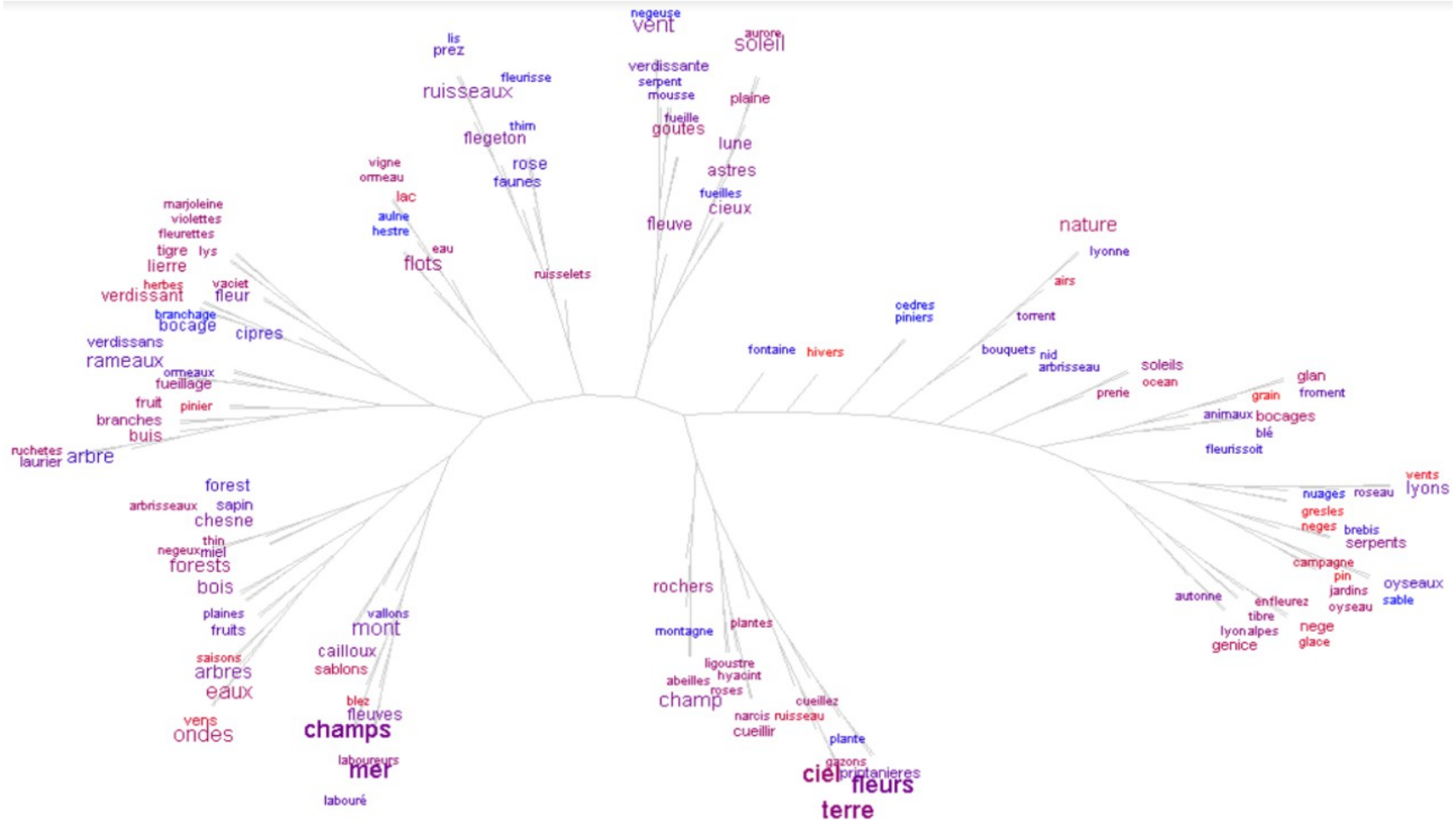
- Identifier les mots plutôt au début ou à la fin du corpus
- **extraire le contexte dans le corpus de certains mots ciblés et utiliser ces données comme corpus.**



- Identifier des regroupements communs ou différents
- Comparer la fréquence des mots fréquents dans les deux arbres
- Comparer (quanti+quali) les regroupements dans les deux arbres
- Construire le nuage arboré des mots spécifiques de chaque sous-corpus

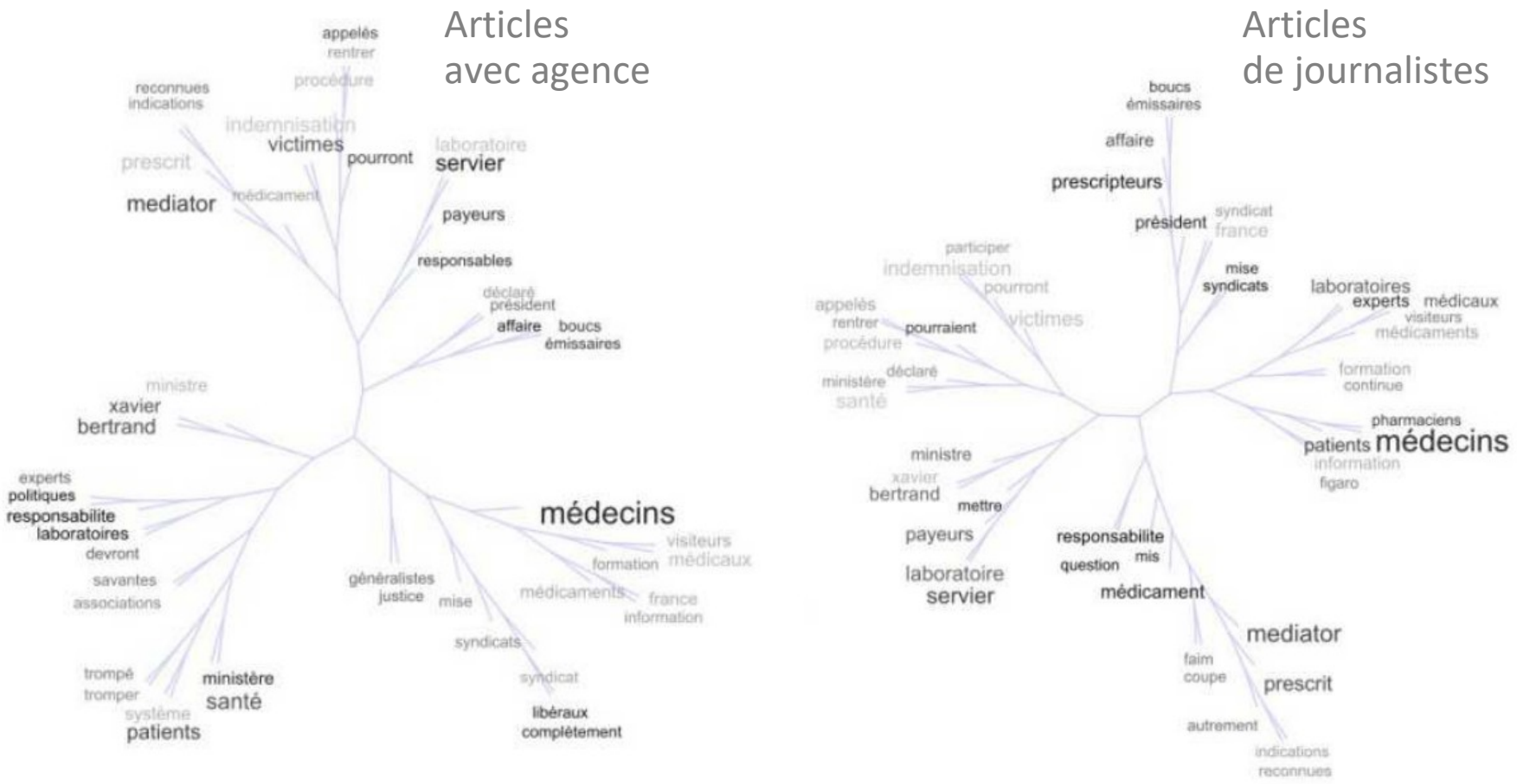
Travail en cours avec Edna Hernandez

Nuages arborés du voisinage des mots de la nature



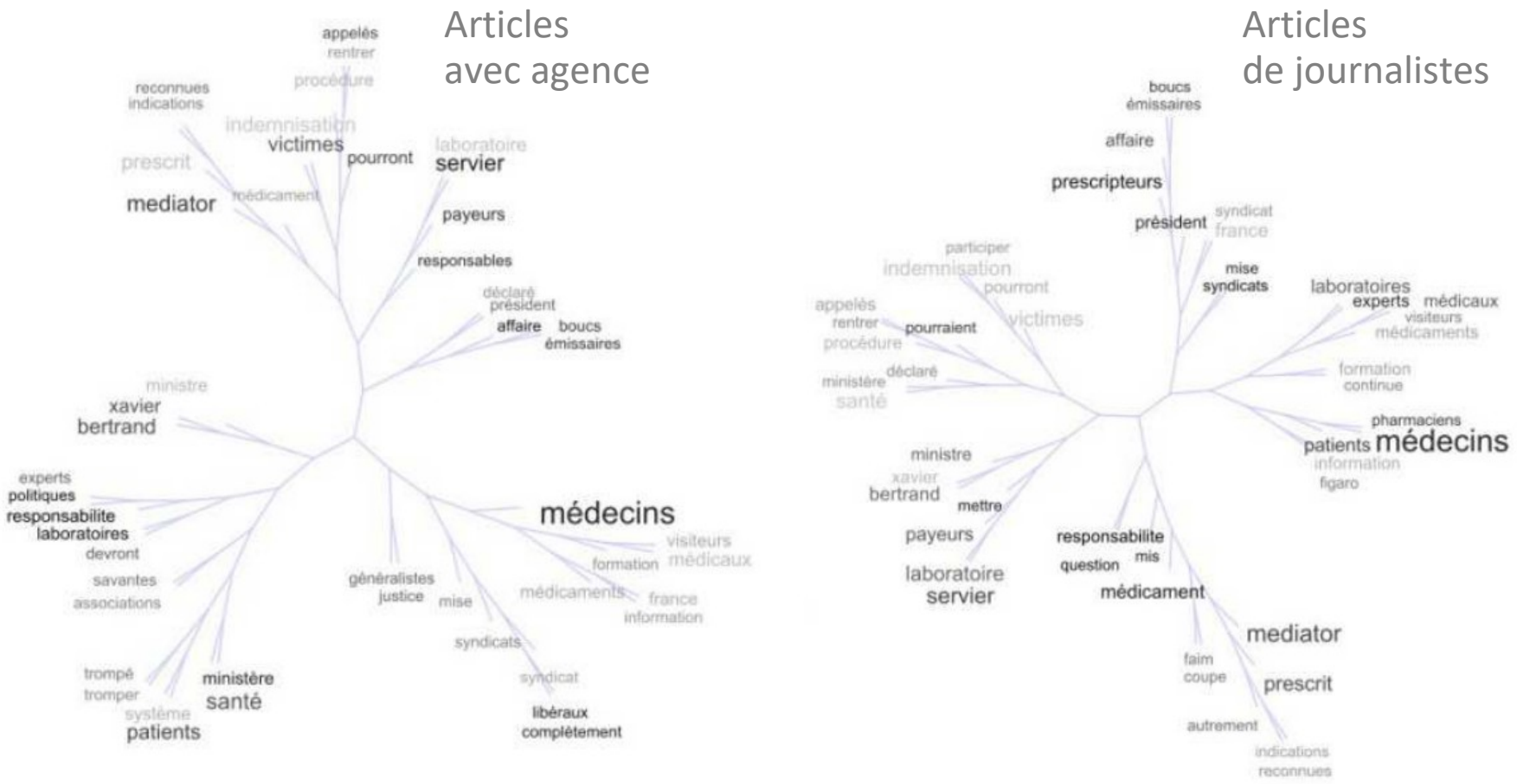
Nuage arboré des voisinages des termes liés à la nature dans *Le Ravissement de Proserpine* de Catherine Des Roches (Laura Borie)

Nuages arborés des contextes de « médecins »



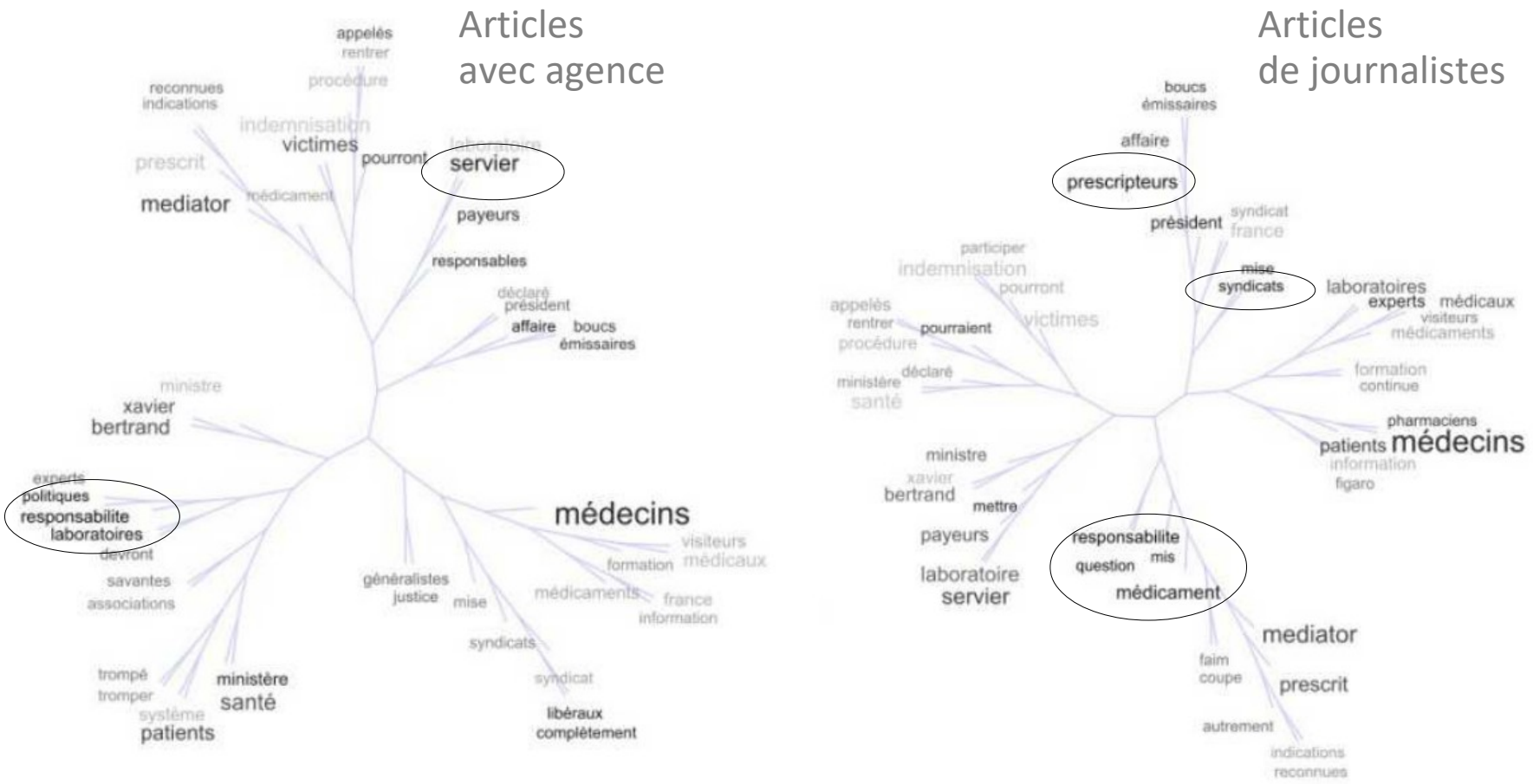
Nuage arboré des 50 mots les plus fréquents des contextes (10 mots avant et 10 mots après) du mot médecins dans le sous-corpus des articles sur le Mediator, colorés par le degré de cooccurrence avec le mot responsabilités (en noir pour les mots les plus cooccurrents), construit par TreeCloud avec la formule Liddell, et des fenêtres glissantes de 20 mots

Nuages arborés des contextes de « médecins »



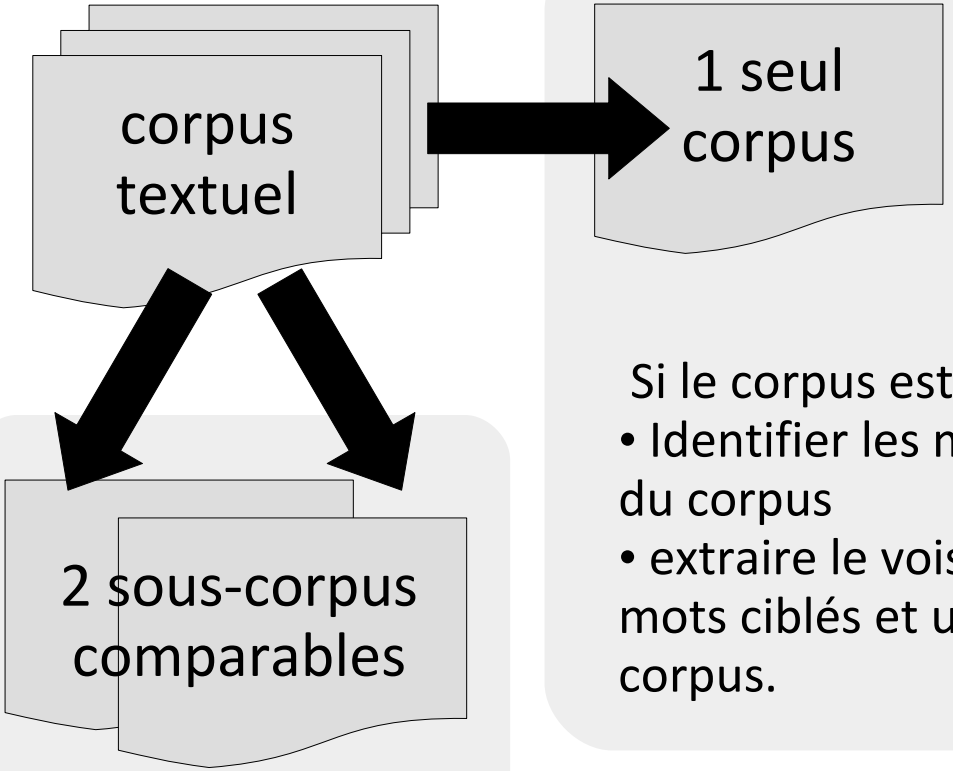
Nuage arboré des 50 mots les plus fréquents des contextes (10 mots avant et 10 mots après) du mot médecins dans le sous-corpus des articles sur le Mediator, colorés par le degré de cooccurrence avec le mot responsabilités (en noir pour les mots les plus cooccurrents), construit par TreeCloud avec la formule Liddell, et des fenêtres glissantes de 20 mots

Nuages arborés des contextes de « médecins »



Nuage arboré des 50 mots les plus fréquents des contextes (10 mots avant et 10 mots après) du mot médecins dans le sous-corpus des articles sur le Mediator, colorés par le degré de cooccurrence avec le mot responsabilités (en noir pour les mots les plus cooccurrents), construit par TreeCloud avec la formule Liddell, et des fenêtres glissantes de 20 mots

Exploration de corpus avec TreeCloud



- Interpréter les regroupements
- Examiner les voisinages des mots fréquents dans l'arbre
- Examiner le voisinage de mots ciblés dans l'arbre («acteurs», verbes, etc.)

Si le corpus est assez grand :

- Identifier les mots plutôt au début ou à la fin du corpus
- extraire le voisinage dans le corpus de certains mots ciblés et utiliser ces données comme corpus.

- **Identifier des regroupements communs ou différents**
- Comparer la fréquence des mots fréquents dans les deux arbres
- Comparer (quanti+quali) les regroupements dans les deux arbres
- Construire le nuage arboré des mots spécifiques de chaque sous-corpus

Travail en cours avec Edna Hernandez

Illustration sur le corpus Mediator

Comparer les articles d'agences et articles de journalistes

Corpus : 595 articles d'agences contre 1496 articles de journalistes de 2011 évoquant l'affaire du Mediator dans la presse française.

Ensemble des articles

Gambette & Martinez, *Texto!*, 2013

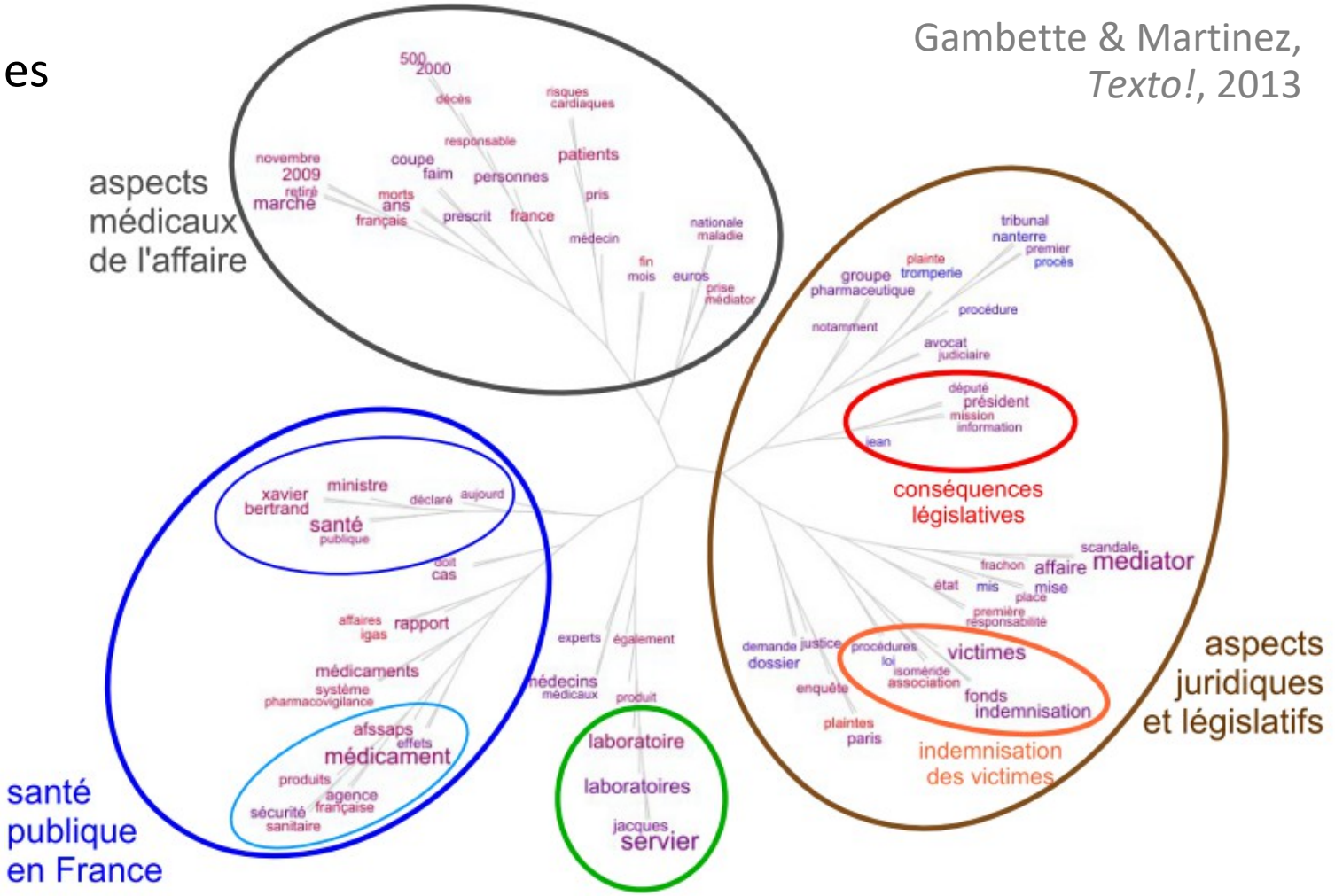
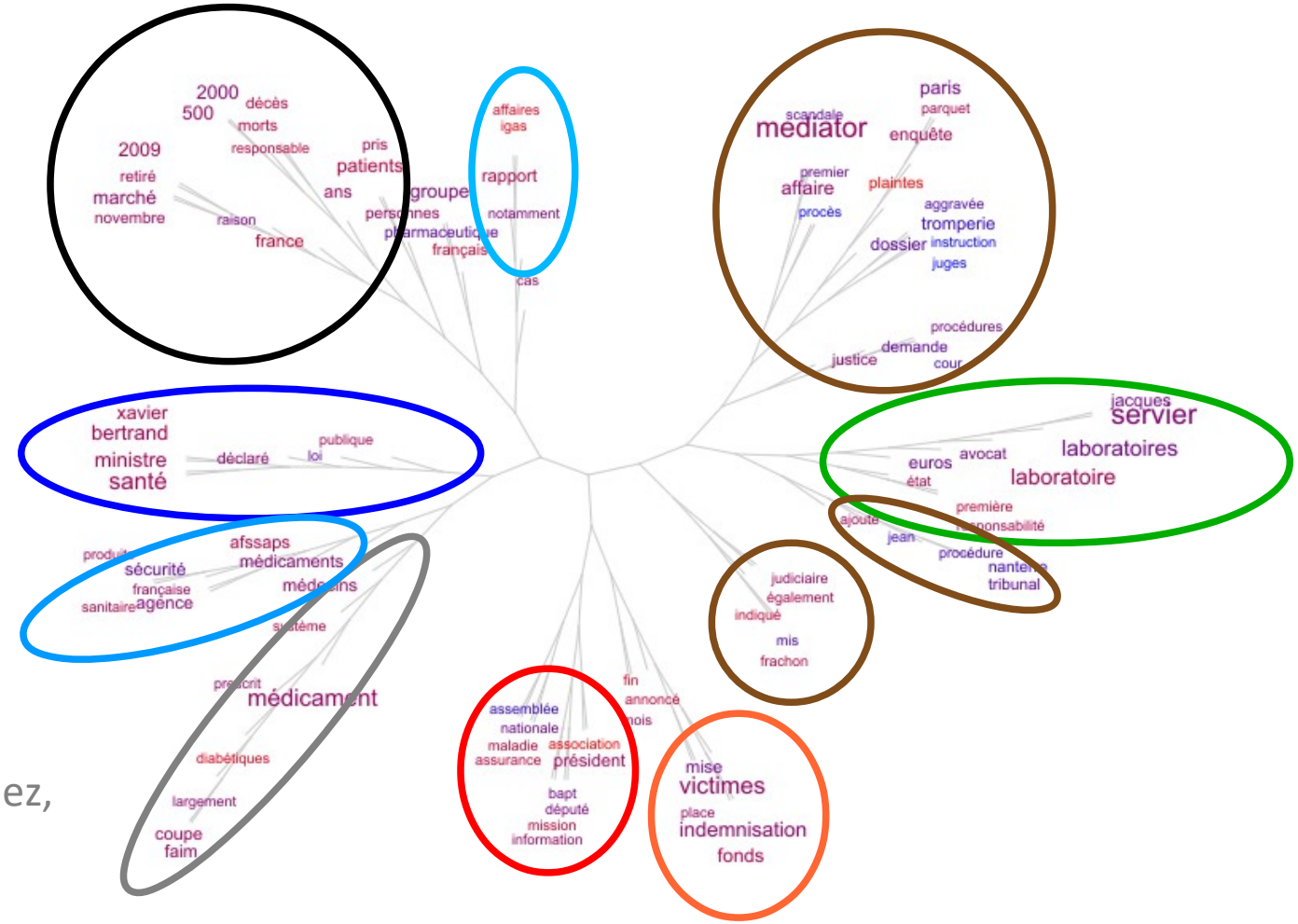


Illustration sur le corpus Mediator

Comparer les articles d'agences et articles de journalistes

Corpus : 595 articles d'agences contre 1496 articles de journalistes de 2011 évoquant l'affaire du Mediator dans la presse française.

Articles d'agences



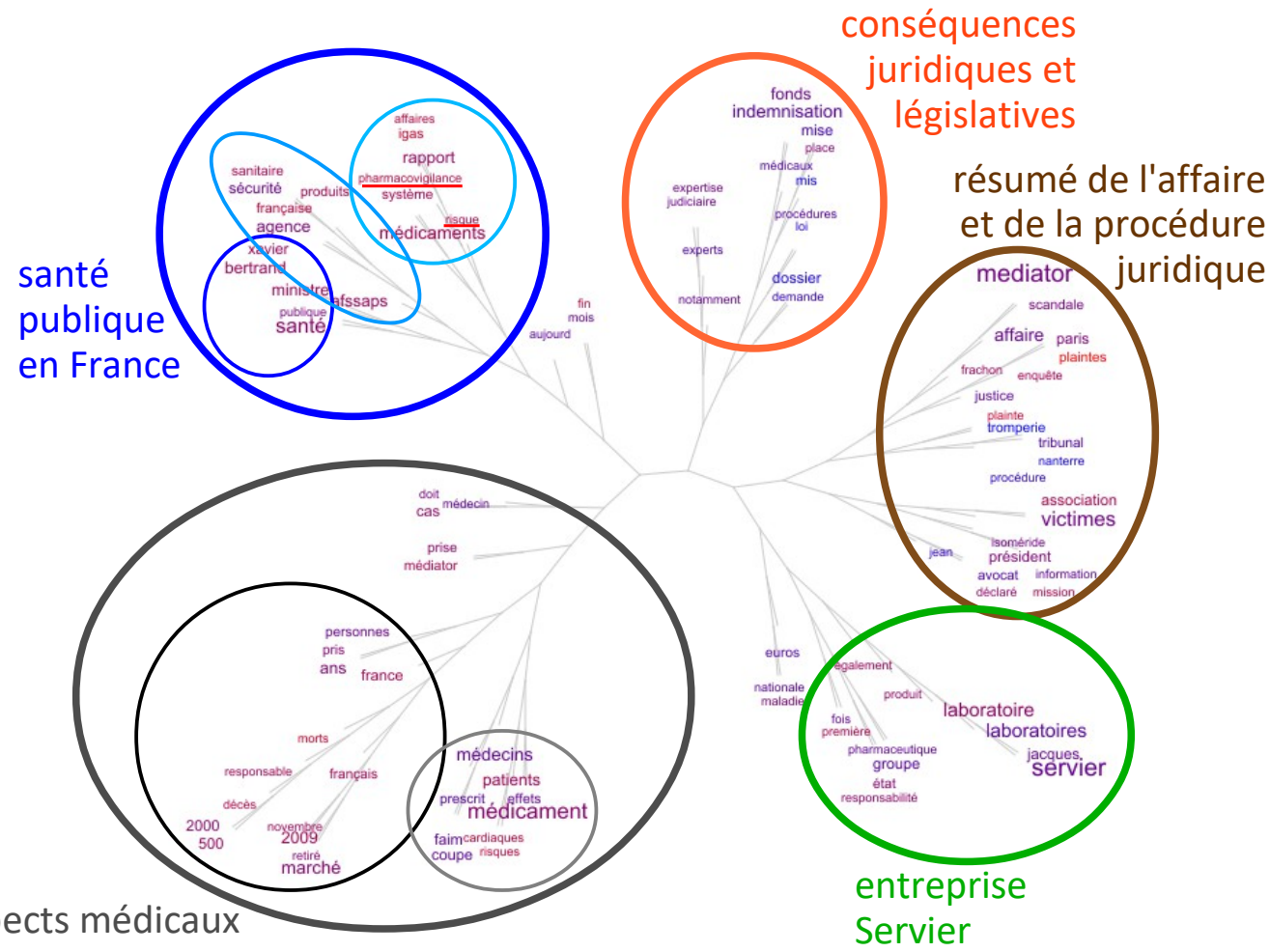
Gambette & Martinez, *Texto!*, 2013

Illustration sur le corpus Mediator

Comparer les articles d'agences et articles de journalistes

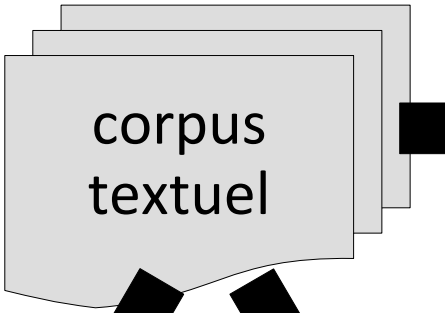
Corpus : 595 articles d'agences contre 1496 articles de journalistes de 2011 évoquant l'affaire du Mediator dans la presse française.

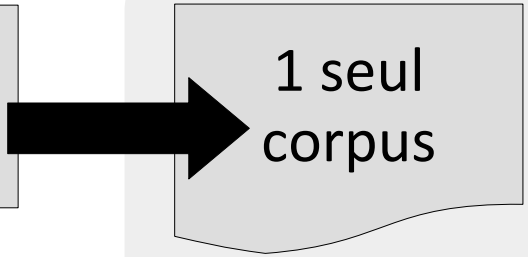
Articles
de journalistes



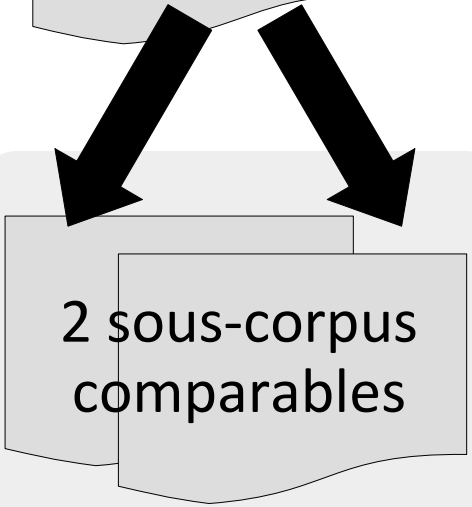
Gambette & Martinez,
Texto!, 2013

Exploration de corpus avec TreeCloud





- Interpréter les regroupements
- Examiner les voisinages des mots fréquents dans l'arbre
- Examiner le voisinage de mots ciblés dans l'arbre («acteurs», verbes, etc.)



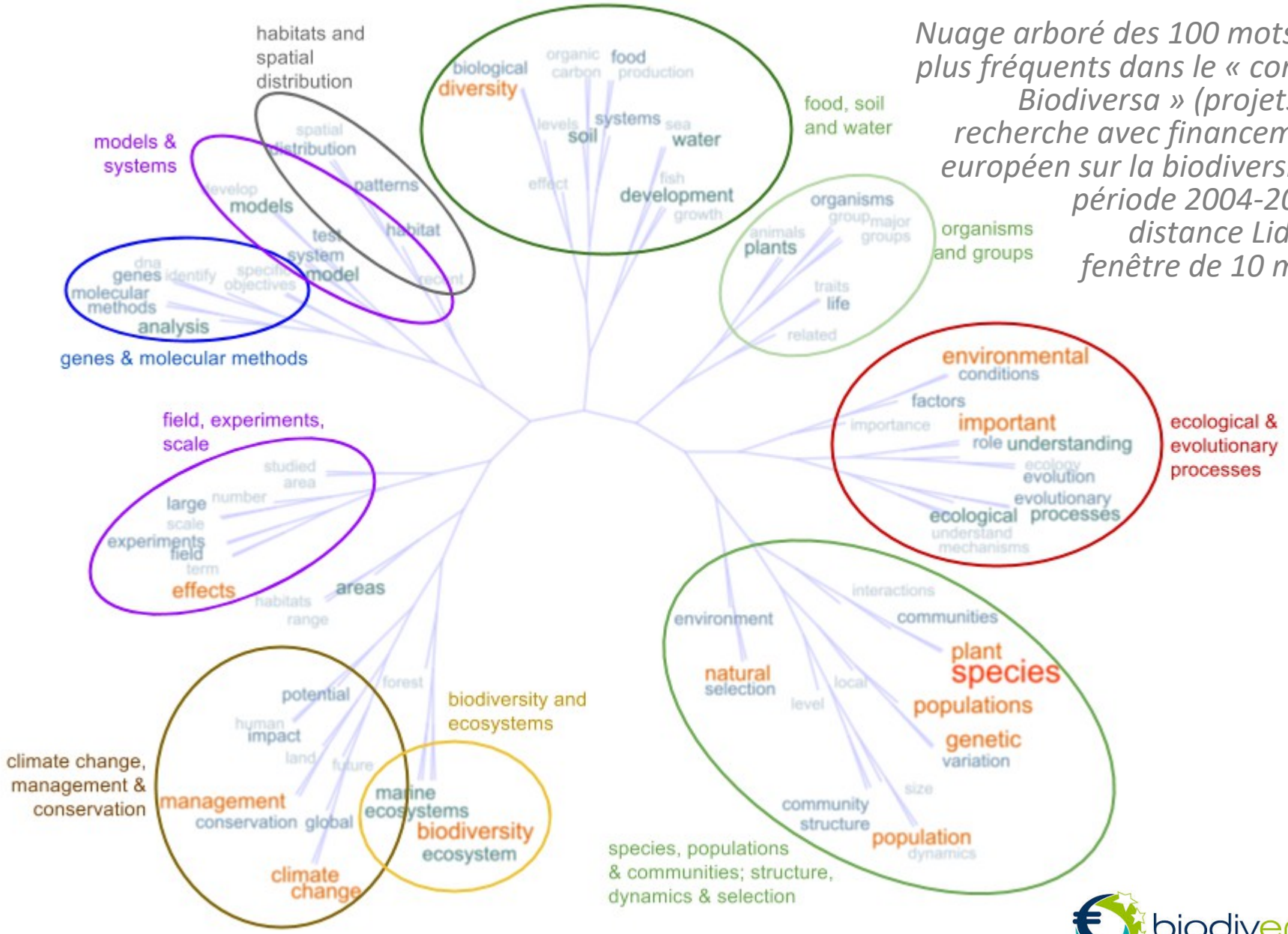
Si le corpus est assez grand :

- Identifier les mots plutôt au début ou à la fin du corpus
- extraire le voisinage dans le corpus de certains mots ciblés et utiliser ces données comme corpus.

- Identifier des regroupements communs ou différents
- Comparer la fréquence des mots fréquents dans les deux arbres
- **Comparer (quanti+quali) les regroupements dans les deux arbres**
- Construire le nuage arboré des mots spécifiques de chaque sous-corpus

Travail en cours avec Edna Hernandez

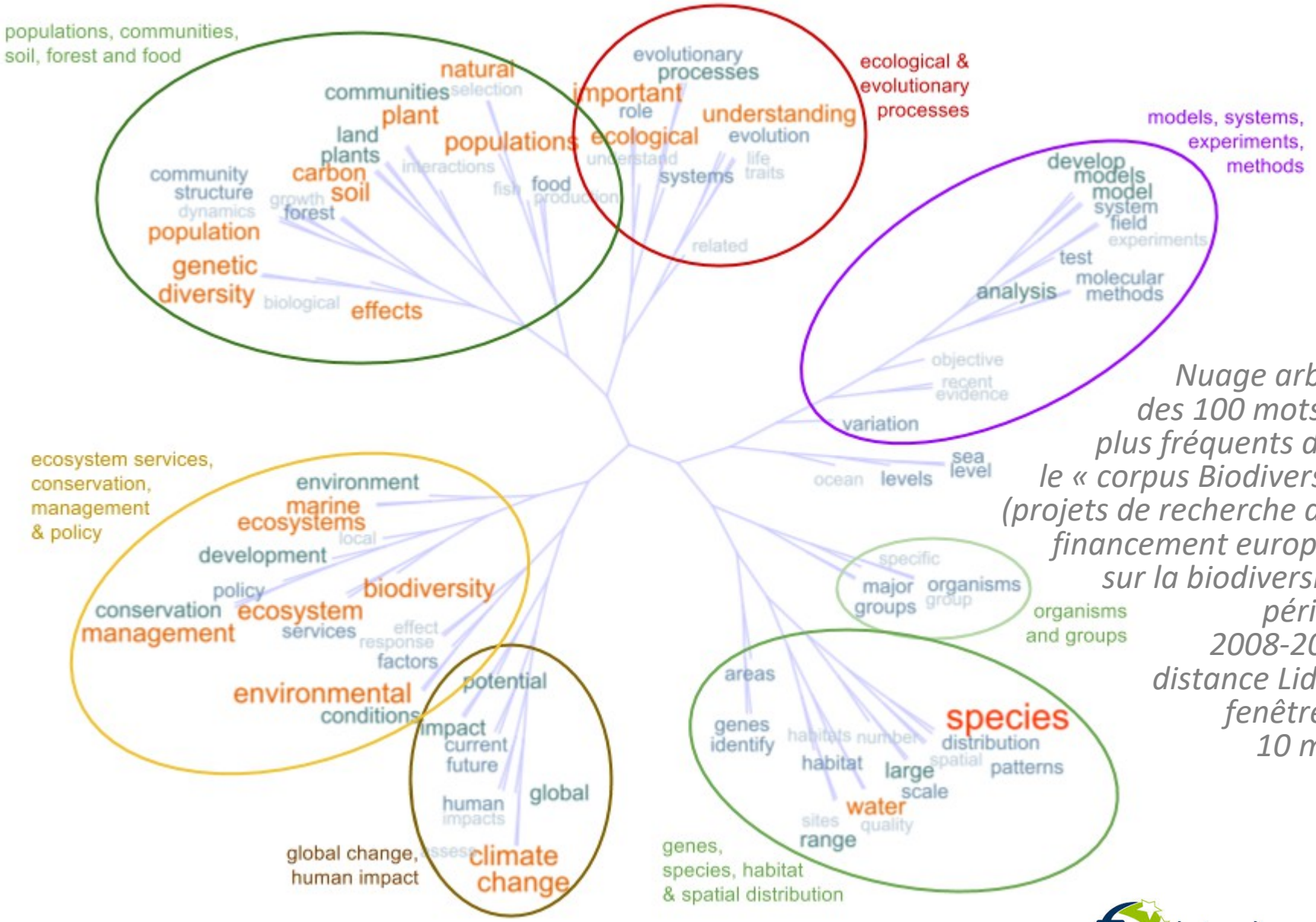
Méthode : comparaison de voisinages dans l'arbre



Nuage arboré des 100 mots les plus fréquents dans le « corpus Biodiversa » (projets de recherche avec financement européen sur la biodiversité), période 2004-2007, distance Liddel, fenêtre de 10 mots

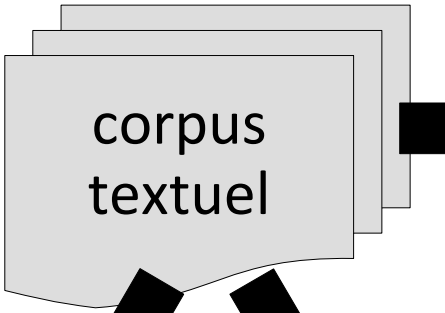
biodiversa

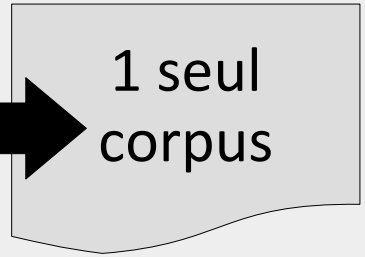
Méthode : comparaison de voisinages dans l'arbre



Nuage arboré des 100 mots les plus fréquents dans le « corpus Biodiversa » (projets de recherche avec financement européen sur la biodiversité), période 2008-2011, distance Liddel, fenêtre de 10 mots

Exploration de corpus avec TreeCloud

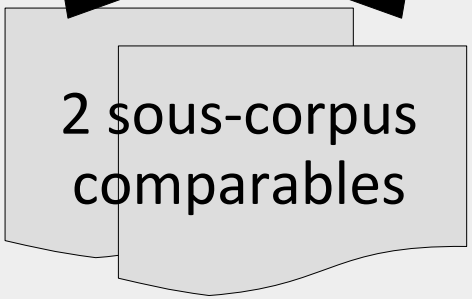




- Interpréter les regroupements
- Examiner les voisinages des mots fréquents dans l'arbre
- Examiner le voisinage de mots ciblés dans l'arbre («acteurs», verbes, etc.)

Si le corpus est assez grand :

- Identifier les mots plutôt au début ou à la fin du corpus
- extraire le voisinage dans le corpus de certains mots ciblés et utiliser ces données comme corpus.

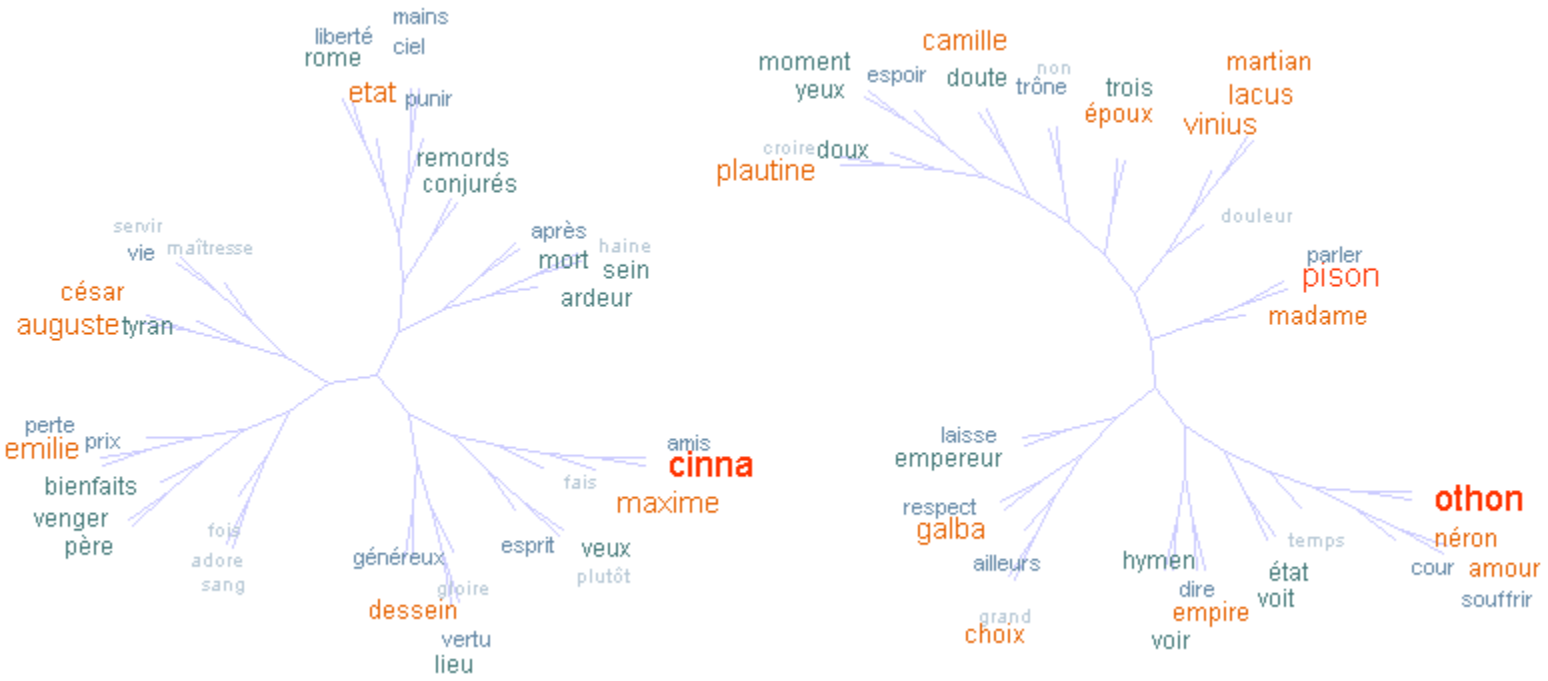


- Identifier des regroupements communs ou différents
- Comparer la fréquence des mots fréquents dans les deux arbres
- Comparer (quanti+quali) les regroupements dans les deux arbres
- **Construire le nuage arboré des mots spécifiques de chaque sous-corpus**

Travail en cours avec Edna Hernandez

Méthode : comparaison des spécifiques

Amstutz & Gambette,
JADT 2010

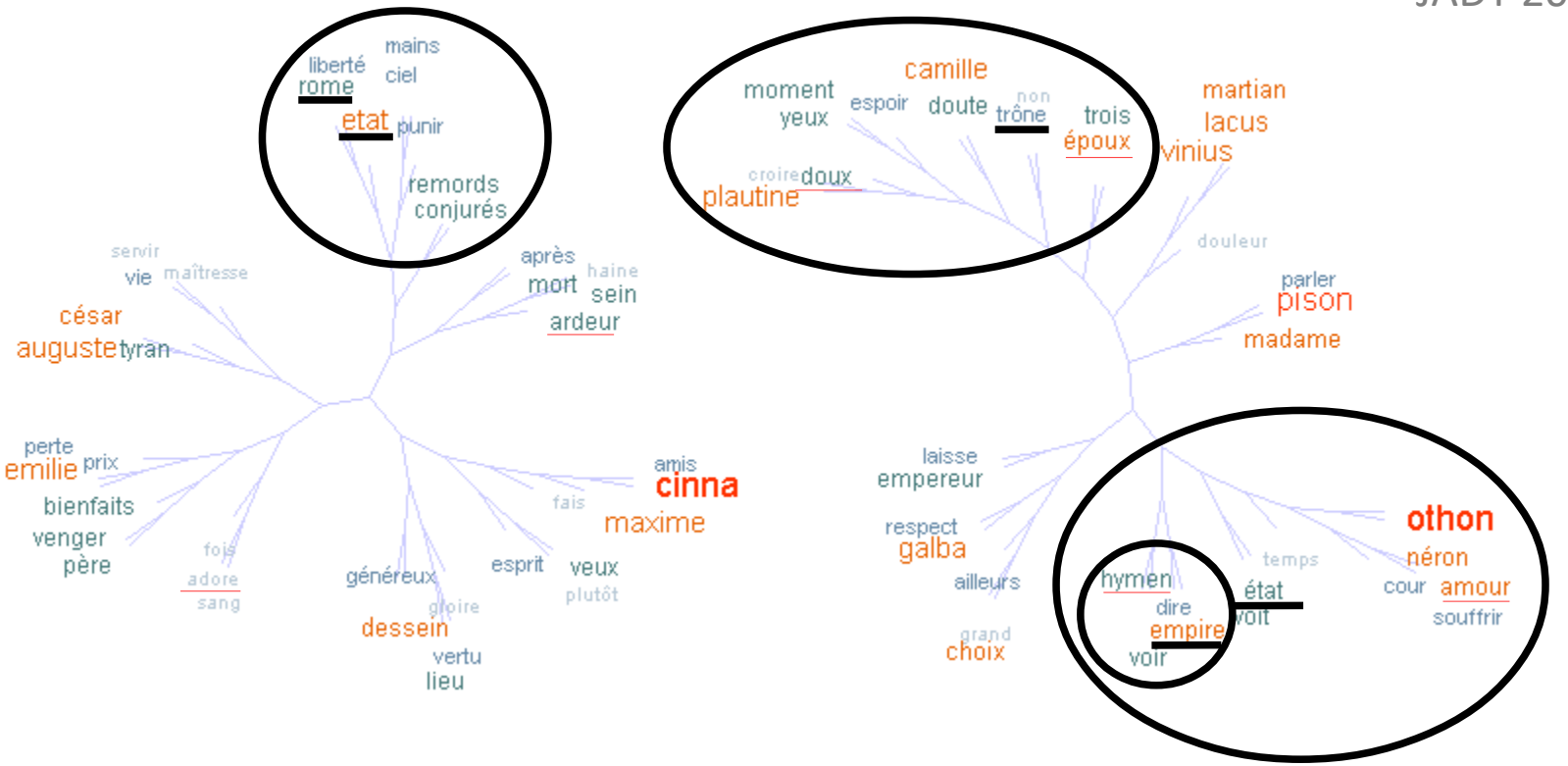


*Nuages arborés des **mots spécifiques** de Cinna et Othon, dimensionnés et colorés d'après leur spécificité calculée dans Lexico3.*

Quels moyens au service de la cause politique ?

Méthode : comparaison des spécifiques

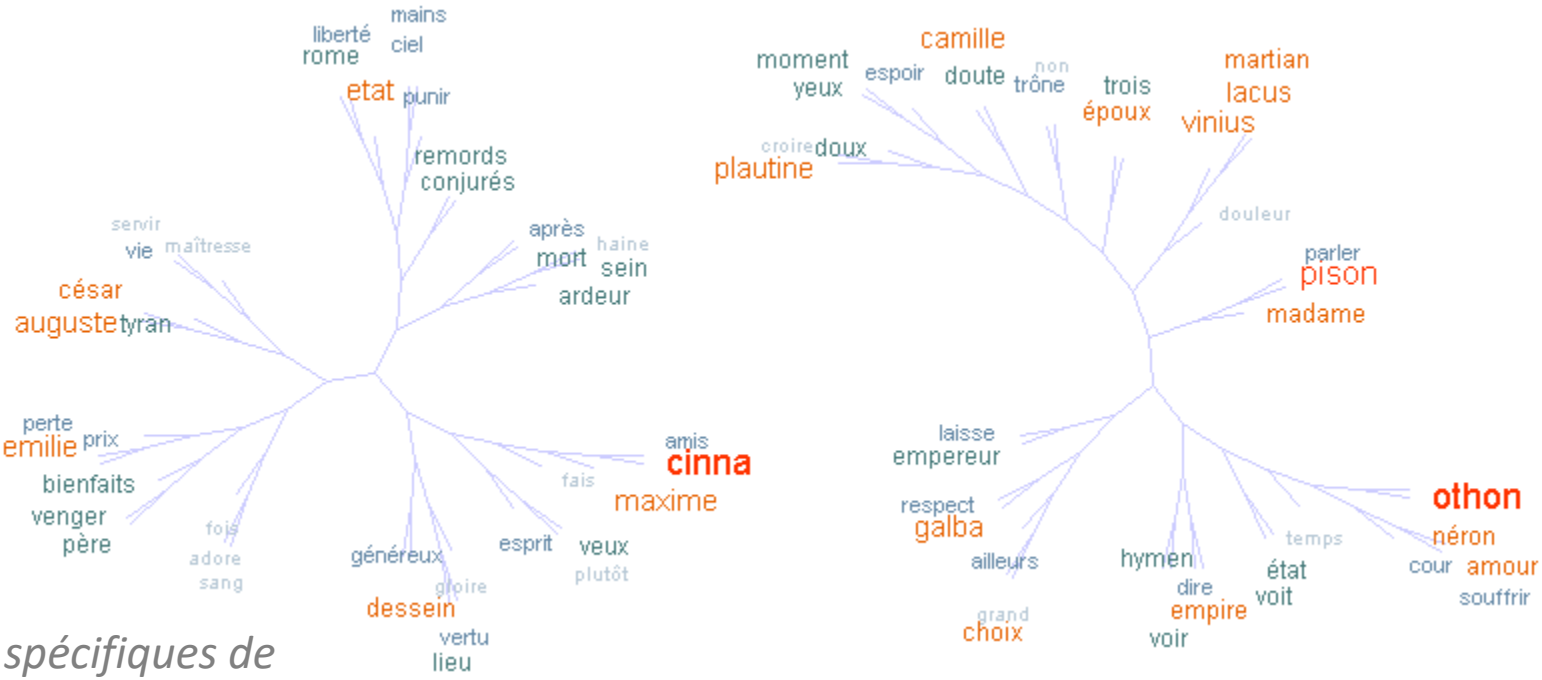
Amstutz & Gambette,
JADT 2010



*Nuages arborés des **mots spécifiques** de Cinna et Othon, dimensionnés et colorés d'après leur spécificité calculée dans Lexico3.*

Quels moyens au service de la cause politique ?

Méthode : comparaison des spécifiques



mots spécifiques de Cinna et Othon d'après Lexico3

	Cinna	*Othon*
Lieu du pouvoir et objet de la confrontation entre les personnages	Rome (« liberté »)	Empire (« trône »)
Souverain en place	tyran	Empereur
Membres du corps politique	amis	maîtres / seigneurs
Moyens au service de la cause politique	gloire	amour matrimonial (« amour », « hymen », « choix »)
Caractérisation de la pièce	Pièce de FONDATION	Pièce de SUCCESSION DYNASTIQUE

Arbres phylogénétiques et arbres de mots

ESPÈCES

Séquences ADN

Données sur les feuilles

MOTS

Position des mots

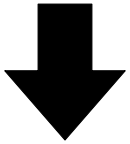
Arbres phylogénétiques et arbres de mots

ESPÈCES

Séquences ADN

Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

Données sur les feuilles



Distances entre les feuilles

	A	B	C	D
A	0	2	5	6
B	2	0	5	6
C	5	5	0	3
D	6	6	3	0

MOTS

Position des mots

Distances fondées sur la cooccurrence entre les deux mots

Arbres phylogénétiques et arbres de mots

ESPÈCES

Séquences ADN

Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

Données sur les feuilles

Distances entre les feuilles

	A	B	C	D
A	0	**2**	5	6
B	**2**	0	5	6
C	5	5	0	3
D	6	6	3	0

classification hiérarchique ascendante
algorithme UPGMA

Arbre

MOTS

Position des mots

Distances fondées sur la cooccurrence entre les deux mots

Algorithme UPGMA : https://fr.wikipedia.org/wiki/Unweighted_pair_group_method_with_arithmetic_mean

Arbres phylogénétiques et arbres de mots

ESPÈCES

Séquences ADN

Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

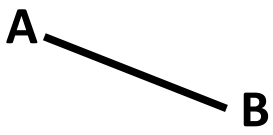
Données sur les feuilles

Distances entre les feuilles

	A+B	C	D
A+B	0	**5**	**6**
C	**5**	0	3
D	**6**	3	0

classification hiérarchique ascendante
algorithme UPGMA

Arbre



MOTS

Position des mots

Distances fondées sur la cooccurrence entre les deux mots

Algorithme UPGMA : https://fr.wikipedia.org/wiki/Unweighted_pair_group_method_with_arithmetic_mean

Arbres phylogénétiques et arbres de mots

ESPÈCES

Séquences ADN

Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

Données sur les feuilles

Distances entre les feuilles

	A+B	C	D
A+B	0	5	6
C	5	0	**3**
D	6	**3**	0

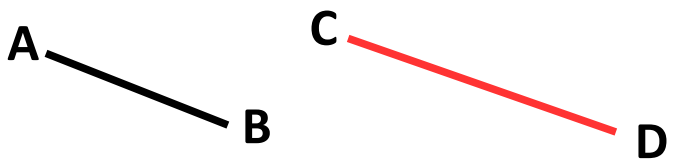
MOTS

Position des mots

Distances fondées sur la cooccurrence entre les deux mots

classification hiérarchique ascendante
algorithme UPGMA

Arbre



Algorithme UPGMA : https://fr.wikipedia.org/wiki/Unweighted_pair_group_method_with_arithmetic_mean

Arbres phylogénétiques et arbres de mots

ESPÈCES

Séquences ADN

Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

Données sur les feuilles

Distances entre les feuilles

	A+B	C+D
A+B	0	**5,5**
C+D	**5,5**	0

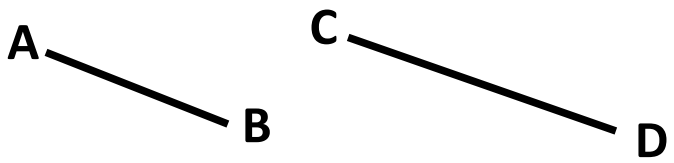
MOTS

Position des mots

Distances fondées sur la cooccurrence entre les deux mots

classification hiérarchique ascendante
algorithme UPGMA

Arbre



Algorithme UPGMA : https://fr.wikipedia.org/wiki/Unweighted_pair_group_method_with_arithmetic_mean

Arbres phylogénétiques et arbres de mots

ESPÈCES

Séquences ADN

Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

Données sur les feuilles

Distances entre les feuilles

	A+B	C+D
A+B	0	**5,5**
C+D	**5,5**	0

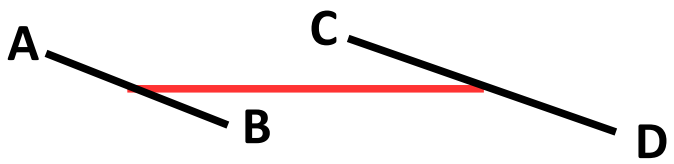
MOTS

Position des mots

Distances fondées sur la cooccurrence entre les deux mots

classification hiérarchique ascendante
algorithme UPGMA

Arbre



Algorithme UPGMA : https://fr.wikipedia.org/wiki/Unweighted_pair_group_method_with_arithmetic_mean

Arbres phylogénétiques et arbres de mots

ESPÈCES

Séquences ADN

Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

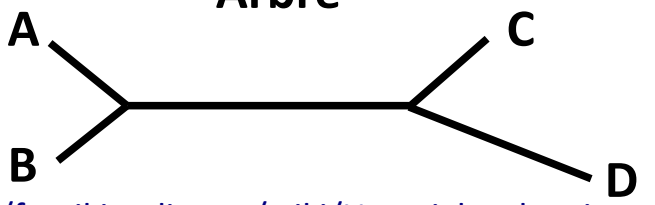
Données sur les feuilles

Distances entre les feuilles

	A	B	C	D
A	0	2	5	6
B	2	0	5	6
C	5	5	0	3
D	6	6	3	0

classification hiérarchique ascendante
algorithme UPGMA

Arbre



MOTS

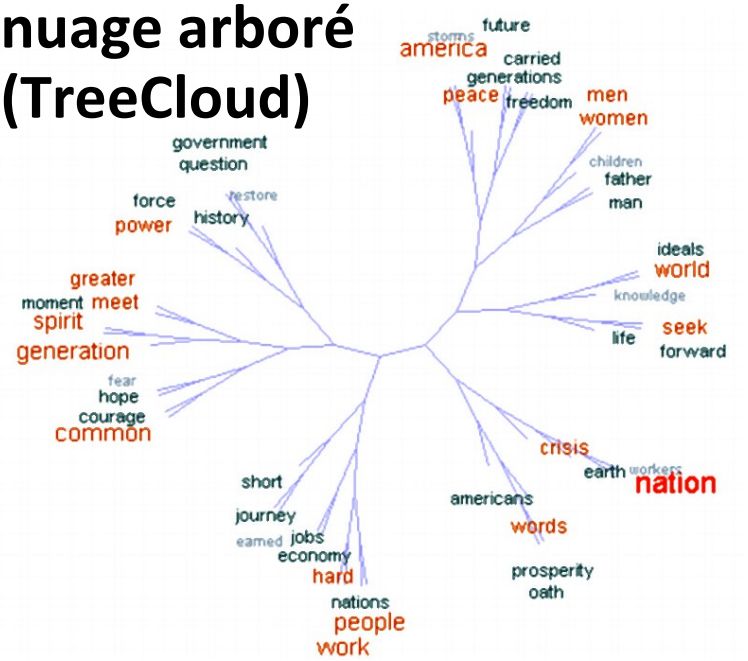
Position des mots

Distances fondées sur la cooccurrence entre les deux mots

Algorithme UPGMA : https://fr.wikipedia.org/wiki/Unweighted_pair_group_method_with_arithmetic_mean

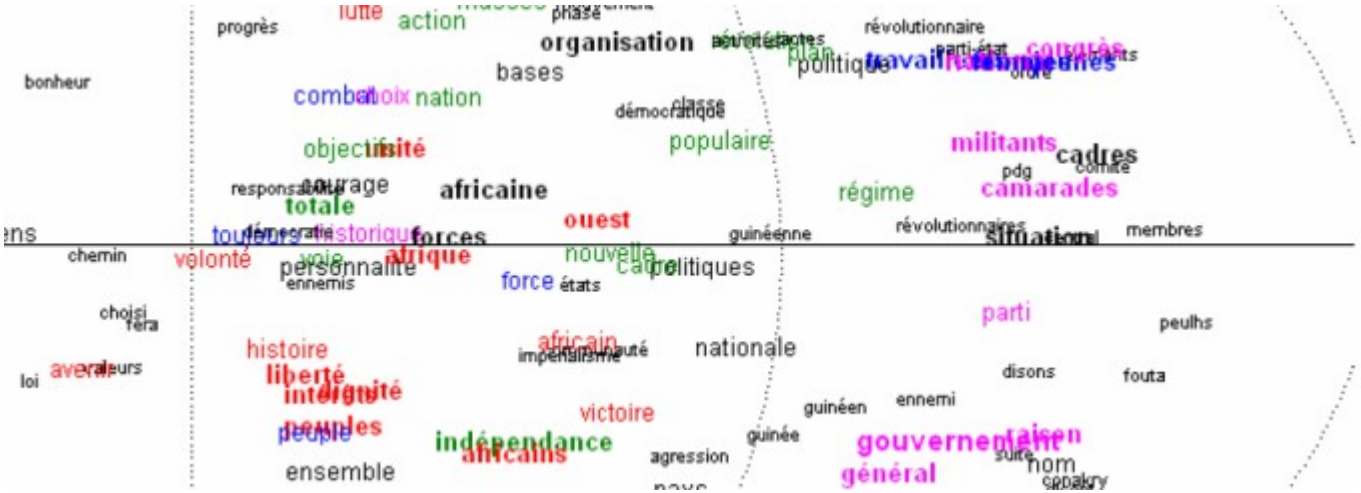
Comparaison avec d'autres visualisations

nuage arboré (TreeCloud)



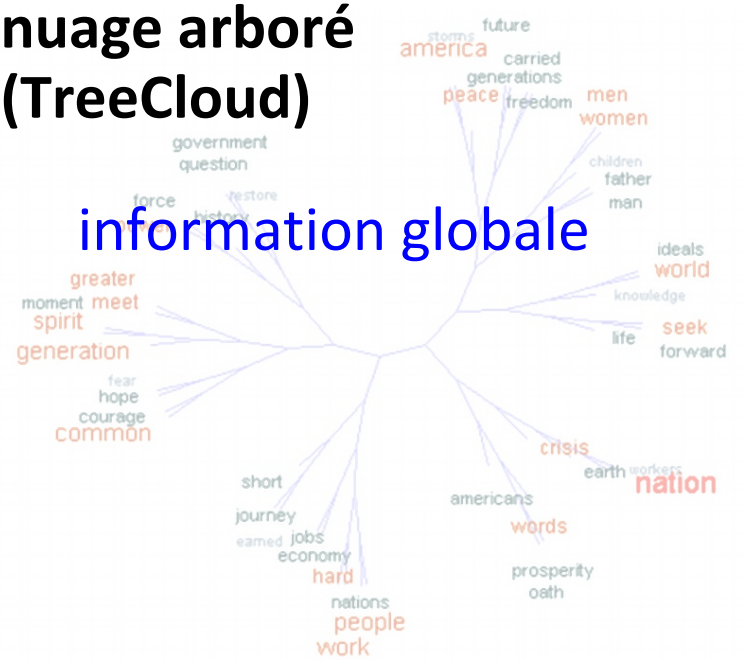
réseau de mots (PhraseNet d'IBM ManyEyes, Tropes)

projection des mots (Astartex)



Comparaison avec d'autres visualisations

nuage arboré (TreeCloud)

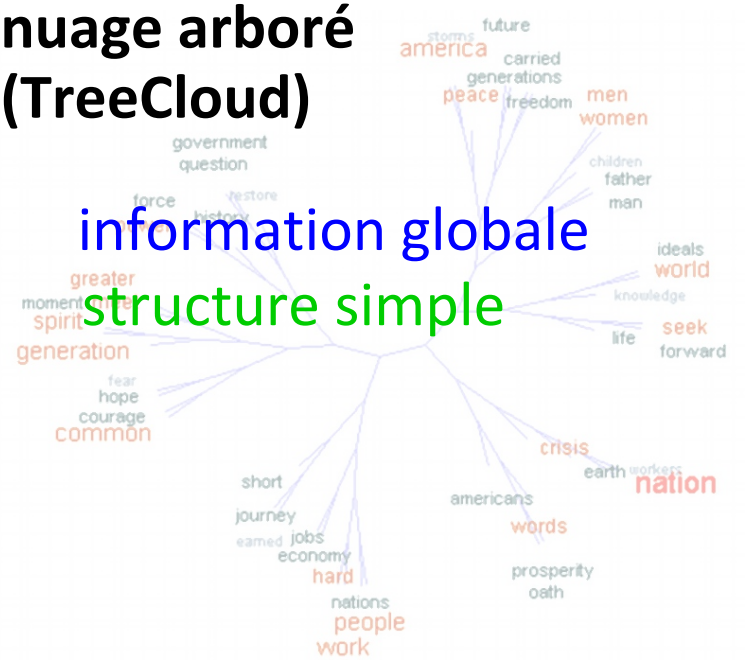


réseau de mots (PhraseNet d'IBM ManyEyes, Tropes)

projection des mots (Astartex)

Comparaison avec d'autres visualisations

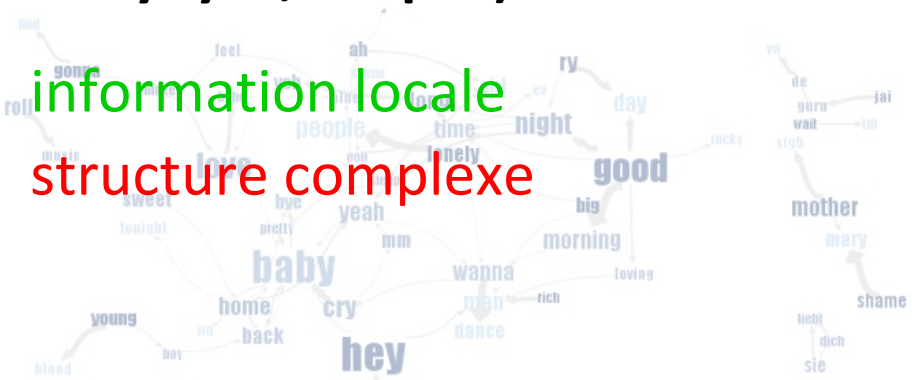
nuage arboré (TreeCloud)



information globale

structure simple

réseau de mots (PhraseNet d'IBM ManyEyes, Tropes)



information locale

structure complexe

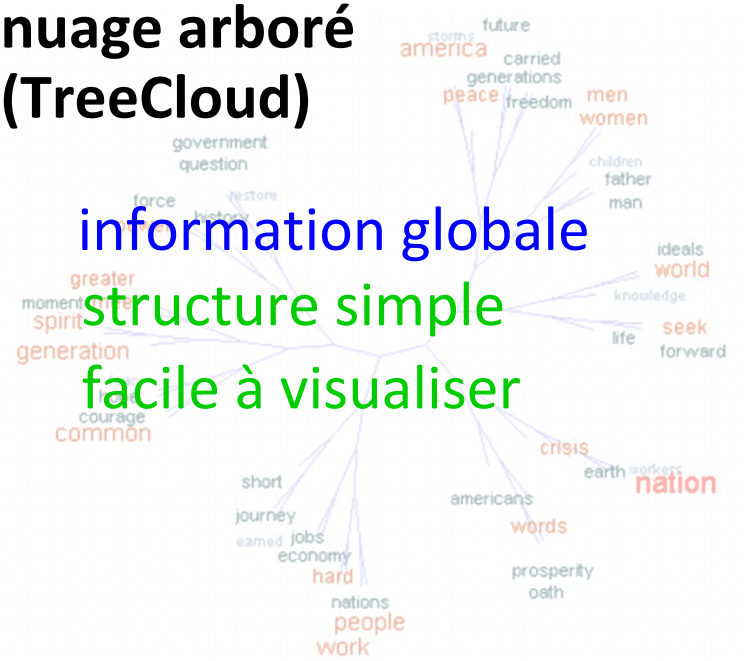
projection des mots (Astartex)

information globale

structure complexe

Comparaison avec d'autres visualisations

nuage arboré (TreeCloud)



information globale

structure simple

facile à visualiser

réseau de mots (PhraseNet d'IBM ManyEyes, Tropes)

information locale

structure complexe

difficile à visualiser

projection des mots (Astartex)

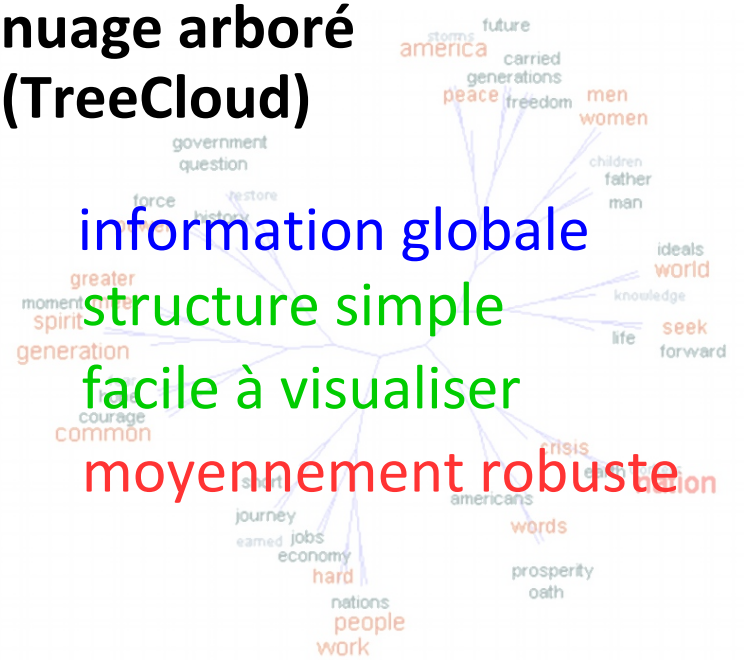
information globale

structure complexe

facile à visualiser mais chevauchements

Comparaison avec d'autres visualisations

nuage arboré (TreeCloud)



information globale

structure simple

facile à visualiser

moyennement robuste

réseau de mots (PhraseNet d'IBM ManyEyes, Tropes)

information locale

structure complexe

difficile à visualiser

robuste

projection des mots (Astartex)

information globale

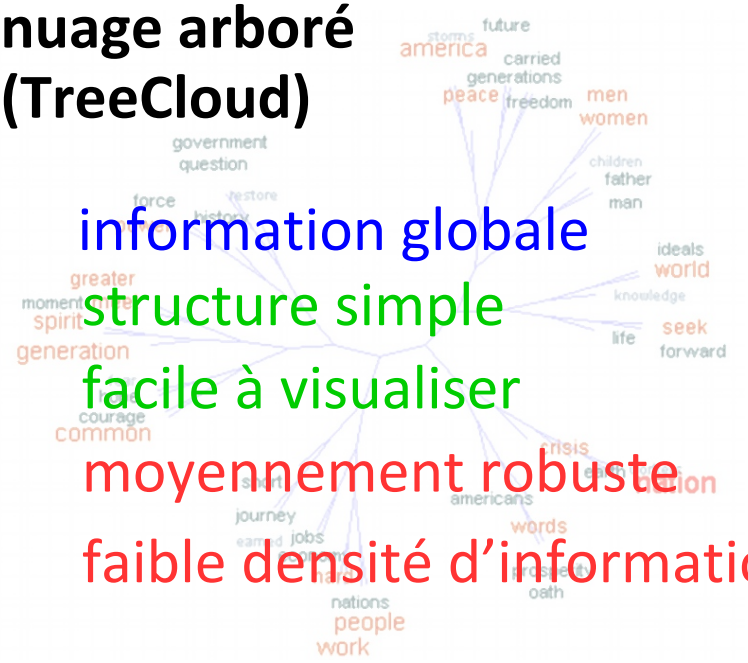
structure complexe

facile à visualiser mais chevauchements

robuste

Comparaison avec d'autres visualisations

nuage arboré (TreeCloud)



information globale

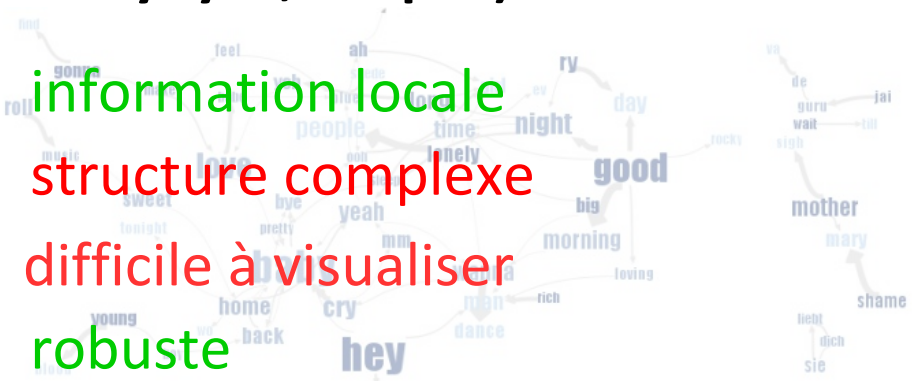
structure simple

facile à visualiser

moyennement robuste

faible densité d'information

réseau de mots (PhraseNet d'IBM ManyEyes, Tropes)



information locale

structure complexe

difficile à visualiser

robuste

forte densité d'information

projection des mots (Astartex, AFC sur Lexico, TextObserver, etc.)

information globale

structure complexe

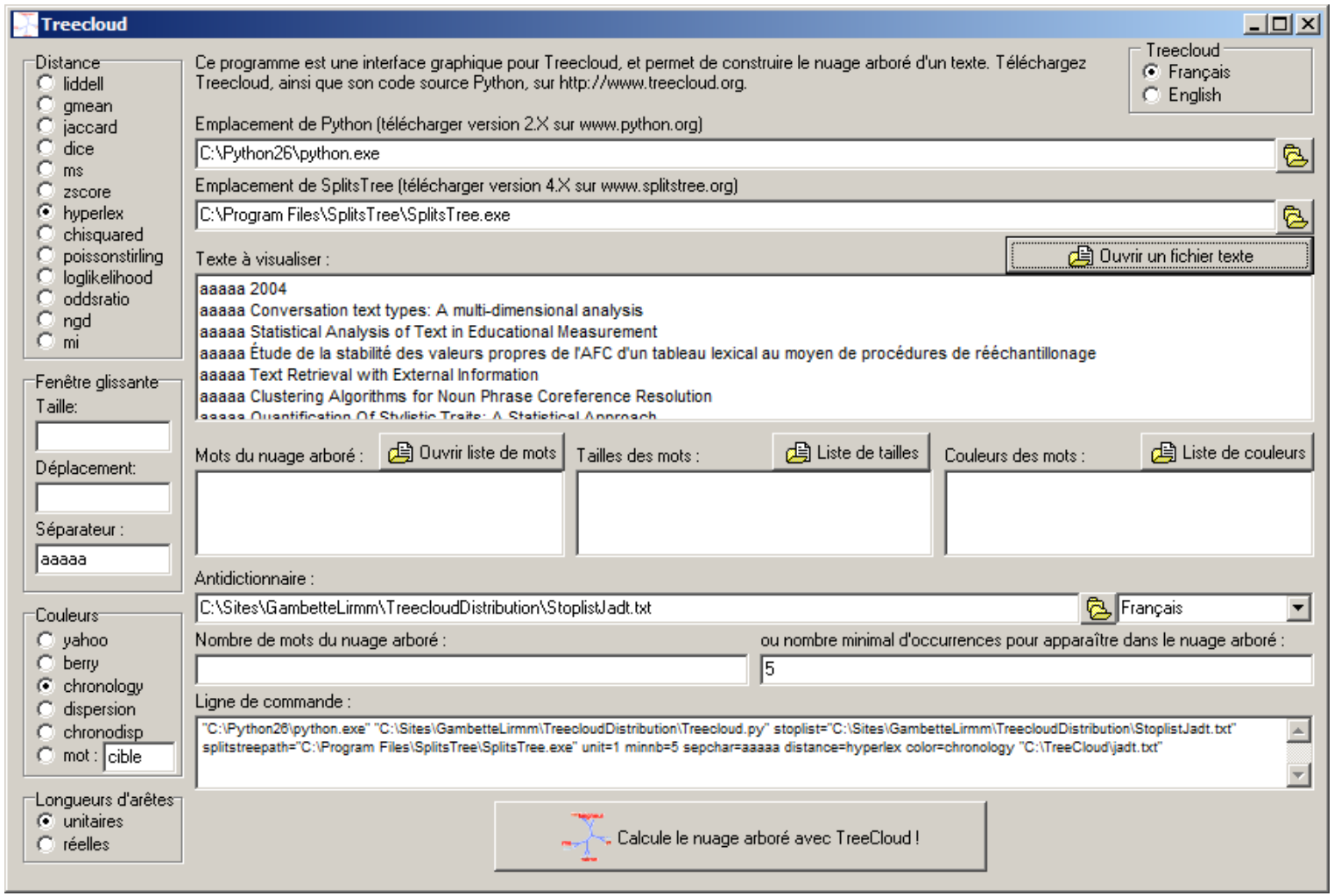
facile à visualiser mais chevauchements

robuste

forte densité d'information

Implémentations

Logiciel libre TreeCloud (Python/Delphi) + SplitsTree (Java)

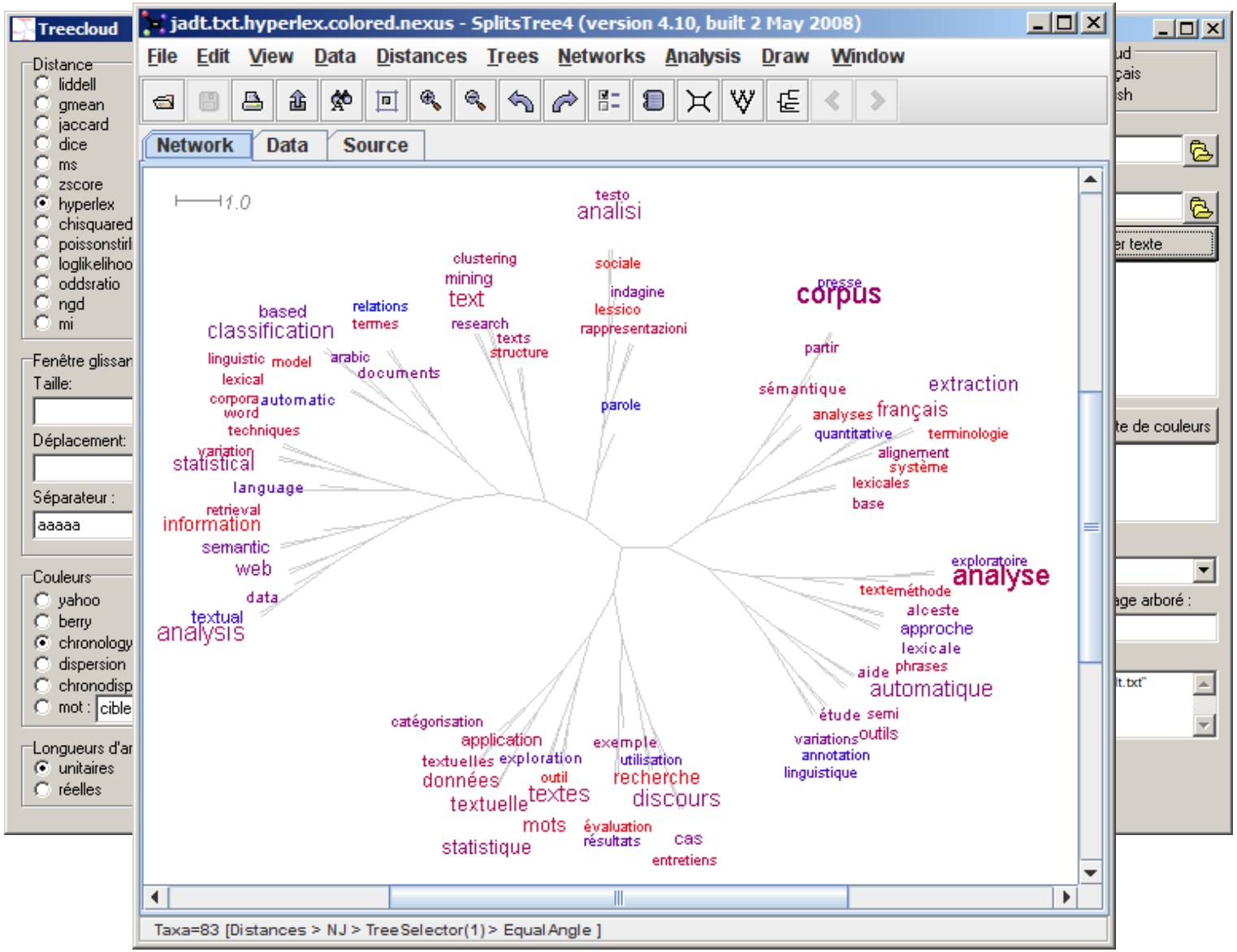


www.treecloud.org

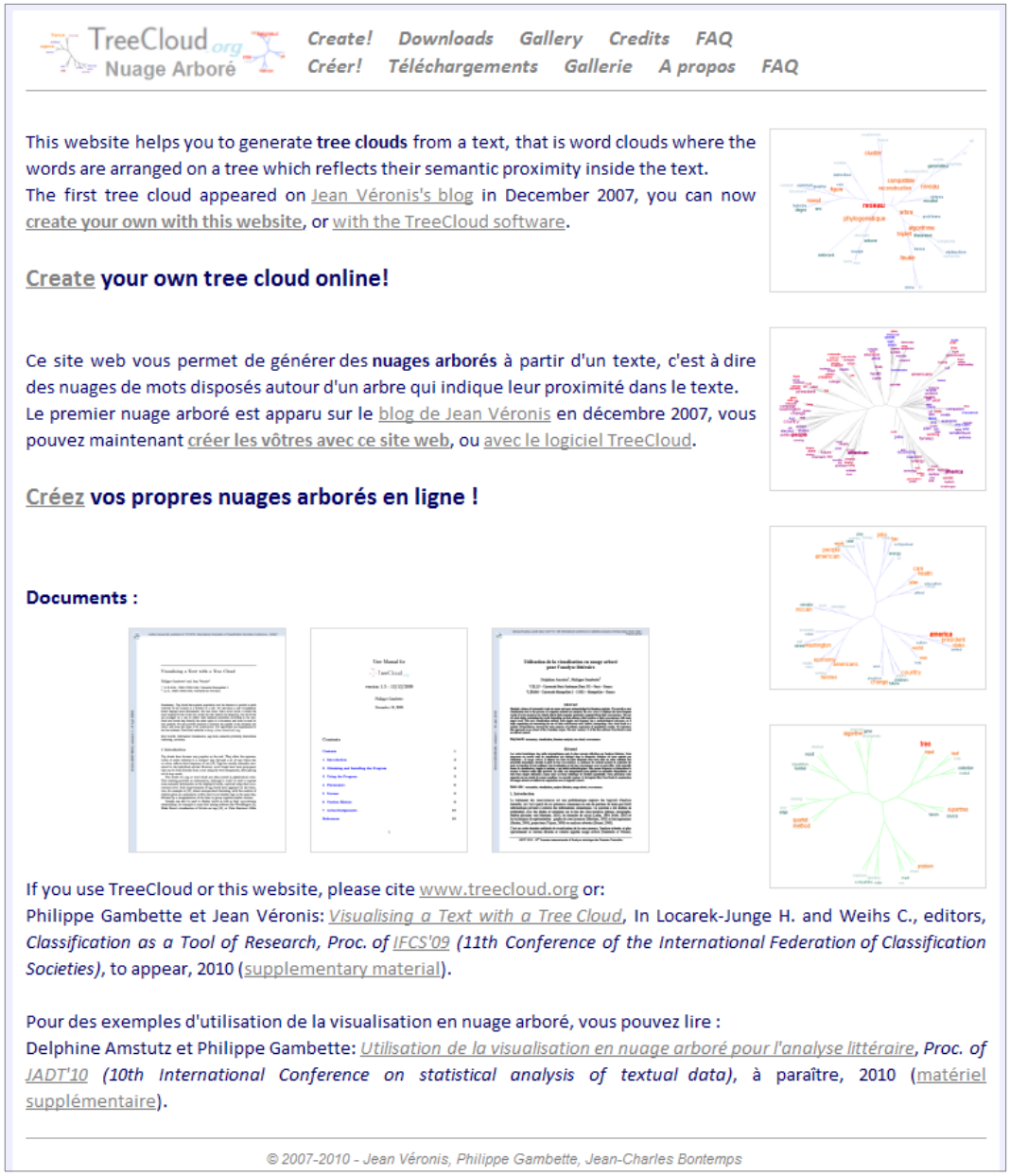
www.splitstree.org

Implémentations

Logiciel libre TreeCloud (Python/Delphi) + SplitsTree (Java)



Interface web



www.treecloud.org

Interface basée sur le logiciel libre NuageArboré de Jean-Charles Bontemps, en C, CGI/Python, et JavaScript.

http://sourceforge.net/projects/nuagearbor/

Développements supplémentaires avec d3.js par Deepak Srinivas

Interface web

www.treecloud.org

TreeCloud org
Nuage Arboré

Create! *Downloads* *Gallery* *Credits* *FAQ*
Créer! *Téléchargements* *Gallerie* *A propos* *FAQ*

This website helps you to generate **tree clo**
words are arranged on a tree which reflects
The first tree cloud appeared on Jean Vé
create your own with this website, or with t

Create **your own tree cloud online**

Ce site web vous permet de générer des r
des nuages de mots disposés autour d'un ar
Le premier nuage arboré est apparu sur le
pouvez maintenant créer les vôtres avec ce

Créez **vos propres nuages arborés**

Documents :

If you use TreeCloud or this website, please
Philippe Gambette et Jean Véronis: *Visual*
Classification as a Tool of Research, Proc. c
Societies), to appear, 2010 (supplementary r

Pour des exemples d'utilisation de la visual
Delphine Amstutz et Philippe Gambette: *Ut*
JADT'10 (10th International Conference
supplémentaire).

TreeCloud org
Nuage Arboré

Créer! *Téléchargements* *Gallerie* *A propos* *FAQ*

Créez vos propres nuages arborés !

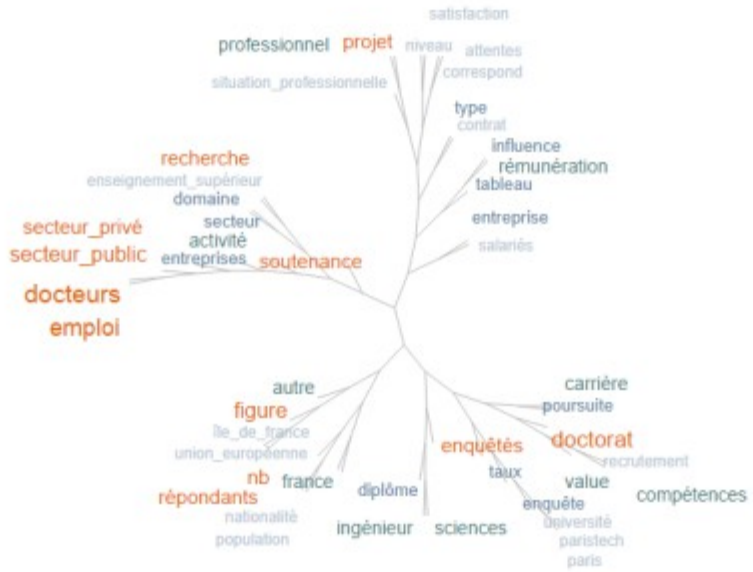
Collez votre texte dans le cadre ci-dessous, puis cliquez sur *Envoyer* ! Attention, l'utilisateur suivant verra votre texte quand il se connectera au site, si vous ne voulez pas faire apparaître vos textes, installez plutôt TreeCloud sur votre machine.

Texte :

[Texte extrait de http://www.adoc-tm.com/2013rapport.pdf]

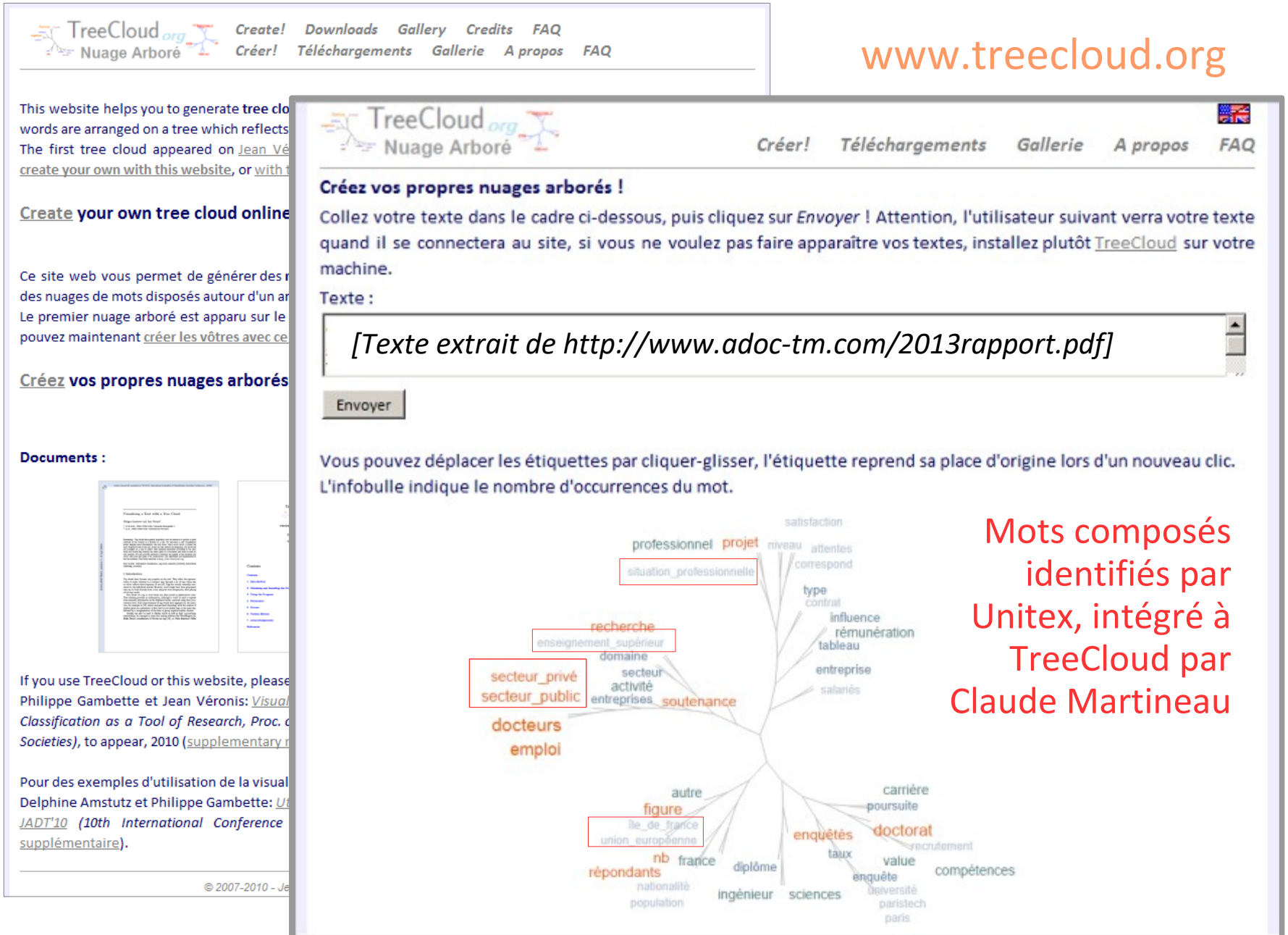
Envoyer

Vous pouvez déplacer les étiquettes par cliquer-glisser, l'étiquette reprend sa place d'origine lors d'un nouveau clic. L'infobulle indique le nombre d'occurrences du mot.



Interface web

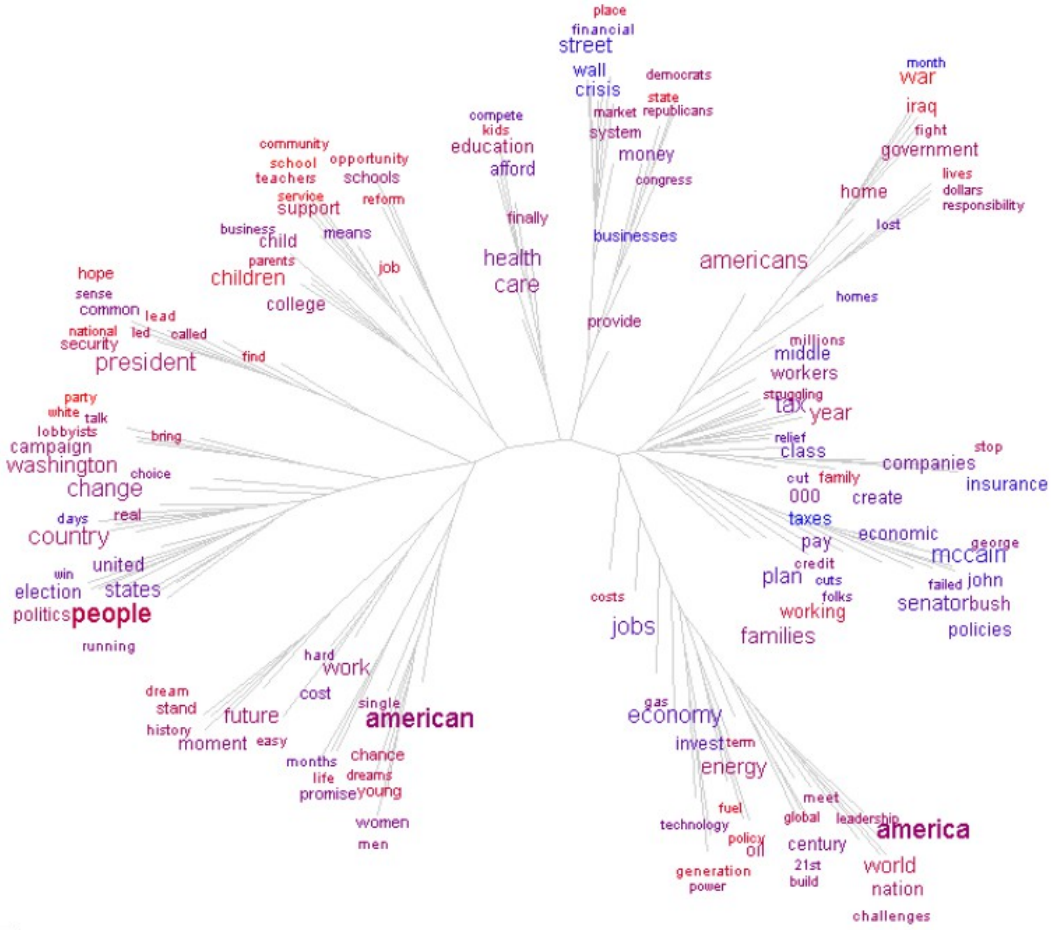
www.treecloud.org



Mots composés identifiés par Unitex, intégré à TreeCloud par Claude Martineau

Temps d'exécution

Limites sur la taille du corpus pour utiliser TreeCloud ?



30 secondes pour la construction du nuage arboré de l'ensemble des discours de campagne de Barack Obama (>300 000 mots)

Implémentations

Version téléchargeable

Logiciel libre TreeCloud (Python/Delphi) + SplitsTree (Java) :

- Tutoriel, manuel d'utilisation
- Coloration de mots personnalisée
- Tailles de mots personnalisée
- Calcul des cooccurrences par blocs délimités par un séparateur

Version en ligne sur TreeCloud.org

- Intégration d'Unitex et réimplémentations par Claude Martineau
- Suppression des mots vides par Unitex
- Filtrage par nature grammaticale avec Unitex
- Reconnaissance de mots composés par Unitex

Implémentations dans d'autres outils

Version dans TextObserver

• intégrée par Yacine Ouchène

• à partir d'une implémentation en Java (Aleksandra Chaschina, projet Google Summer of Code 2016 pour Unitex) : https://github.com/aleksandrachasch/treecloud

Formation à TextObserver samedi 16/11 et 14/12 à Créteil (10h-13h, Jean-Marc Leblanc) https://tinyurl.com/textopol2019

• Expliciter l'analyse factorielle des correspondances

• Analyser la variation lexicométrique

• Introduction aux opérations de catégorisation

• Recension de corpus et balisage semi-automatisé : présentation de la base Textopol

http://textopol.u-pec.fr/textobserver/

Références (*treecloud.org*)

Philippe Gambette, Jean Véronis (2009)
Visualising a Text with a Tree Cloud, *IFCS'09, Studies in Classification, Data Analysis, and Knowledge Organization* 40, p. 561-570
http://www.slideshare.net/PhilippeGambette/visualising-a-text-with-a-tree-cloud

Delphine Amstutz & Philippe Gambette (2010)
Utilisation de la visualisation en nuage arboré pour l'analyse littéraire, JADT'10 (Proceedings of the 10th International Conference on statistical analysis of textual data), Statistical Analysis of Textual Data, p. 227-238
http://www.slideshare.net/PhilippeGambette/utilisation-de-la-visualisation-en-nuage-arbor-pour-lanalyse-littraire

Philippe Gambette, Nuria Gala & Alexis Nasr (2012)
Longueur de branches et arbres de mots, *Corpus* 11:129-146
http://www.slideshare.net/PhilippeGambette/longueur-de-branches-et-arbres-de-mots

William Martinez & Philippe Gambette (2013)
L'affaire du Médiator au prisme de la textométrie, *Texto!* XVIII(4)
http://www.revue-texto.net/index.php?id=3318

Philippe Gambette, Hilde Eggermont & Xavier Le Roux (2014)
Temporal and geographical trends in the type of biodiversity research funded on a competitive basis in European countries, *rapport BiodivERsa*
http://www.biodiversa.org/700/download

Philippe Gambette et Nadège Lechevrel (2016)
Une approche textométrique pour étudier la transmission des savoirs biologiques au XIXe siècle, *Nouvelles perspectives en sciences sociales*, Prise de parole (Ontario, Canada), 2016, 12 (1), pp.221-253
https://hal-upec-upem.archives-ouvertes.fr/hal-01408455

Claude Martineau (2017)
TreeCloud, Unitex: une synergie accrue, colloque ECLAVIT, Extraction, classification et visualisation de données textuelles
https://hal-upec-upem.archives-ouvertes.fr/hal-01702091/fr

Recueil et prétraitement de corpus

Océrisation de texte

OCR = optical character recognition (reconnaissance automatique de caractères)

- Logiciel libre Tesseract
- Logiciel gratuit PDF X-Change Viewer
- Logiciel payant Abby FineReader

Tutoriel rédigé et testé avec Jonathan Barkate :

https://redocparisest.wordpress.com/2017/02/19/numerisation-et-comparaison-douvrages/

Prétraitements de textes obtenus par OCR

Sources de textes :

- numérisation + reconnaissance automatique de caractères (OCR) :
  - Gallica : http://gallica.bnf.fr (estimation du taux d'OCR)
  - Internet Archive : https://archive.org/
  - HathiTrust Digital Library : https://www.hathitrust.org/
  - Google Books : http://books.google.fr
  - autres bibliothèques européennes : https://www.europeana.eu/portal/fr
- relus par des humains :
  - Wikisource francophone : http://fr.wikisource.org
  - Projet Gutenberg : https://www.gutenberg.org
  - Les classiques des sciences sociales : http://classiques.uqac.ca/
  - moteur de recherche parmi plusieurs sites web d'e-books gratuits : http://noslivres.net

Prétraitements de textes obtenus par OCR

Astuce pour récupérer le texte complet d'un ouvrage découpé en chapitres sur Wikisource :

https://fr.wikisource.org/wiki/Les_Travailleurs_de_la_mer

• Récupérer le nom du fichier en cliquant sur l'onglet "**Source**" : https://fr.wikisource.org/wiki/Livre:Hugo_-_Les_Travailleurs_de_la_mer_Tome_II_(1892).djvu

• Cliquer sur l'onglet "**Modifier**"

• Remplacer le code de la page par

```
<pages index="Hugo_-_Les_Travailleurs_de_la_mer_Tome_I_(1891).djvu" from=1 to=433 />
```

→ **1** : page de début de l'ouvrage

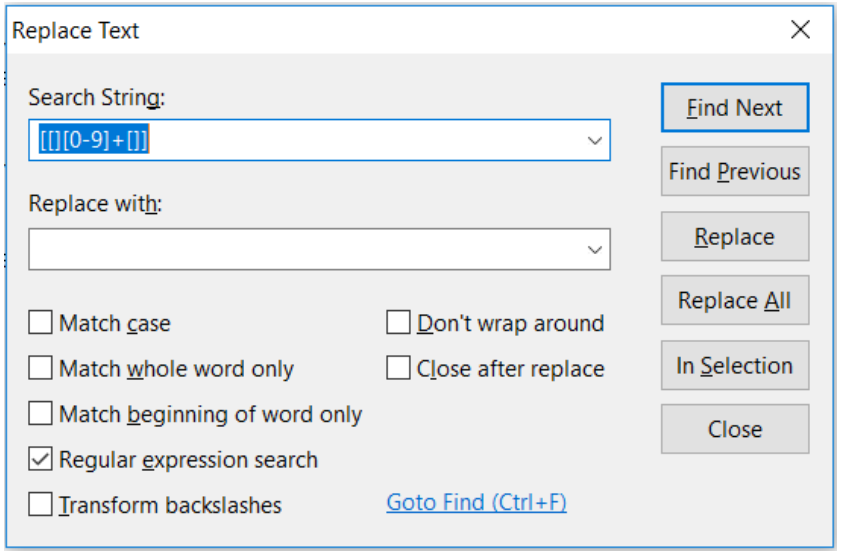
→ **433** : page de fin de l'ouvrage

• Cliquer sur "**Prévisualiser**" (attention **NE PAS cliquer** sur "Publier les modifications)

Prétraitements de textes obtenus par OCR

Rechercher/remplacer (par exemple avec Notepad2) :

- remplacer les apostrophes courbes : remplacer "’" par "'"
- supprimer les références aux notes de fin de texte, en utilisant les “expressions régulières”, remplacer "[[][0-9]+[]]" par "" : un caractère "[" suivi d'un chiffre, éventuellement répété, suivi d'un caractère "]".



Prétraitements de textes obtenus par OCR

Aide automatique à la détection des césures de fin de ligne :

coupeCésure

http://igm.univ-mlv.fr/~gambette/text-processing/coupeCesure/

Code couleur :

- mot dont la césure a été supprimée car trouvé en entier dans le dictionnaire
- mot pour lequel le trait d'union a été gardé

Texte obtenu après remplacements...

Quand vous y ferez quelque réflexion, je crois
que vous trouverez que j'ai raison, et que si je fusse retournée , je rendois mon voyage inutile par être trop court. Pour mon fils et sa femme, ils sont ravis de passer ici jusqu'au carême avec moi : en ce temps-là j'irai à Rennes par complaisance pour eux, parce que ce
temps est plus triste que l'hiver à la campagne : peut-être que ce projet changera, il ne faut point voir de si loin. Ge qui est sûr, ma fille, c'est que l'air d'ici est fort
bon; vous lui faites tort de le croire mauvais. Il fait
depuis plus de deux mois le plus beau temps du monde,
des chaleurs dans la canicule, un mois de septembre

Prétraitements de textes obtenus par OCR

Aide automatique à la dissimilation entre “i” et “j” et entre “u” et “v” : **ijuv**

http://igm.univ-mlv.fr/~gambette/ijuv/

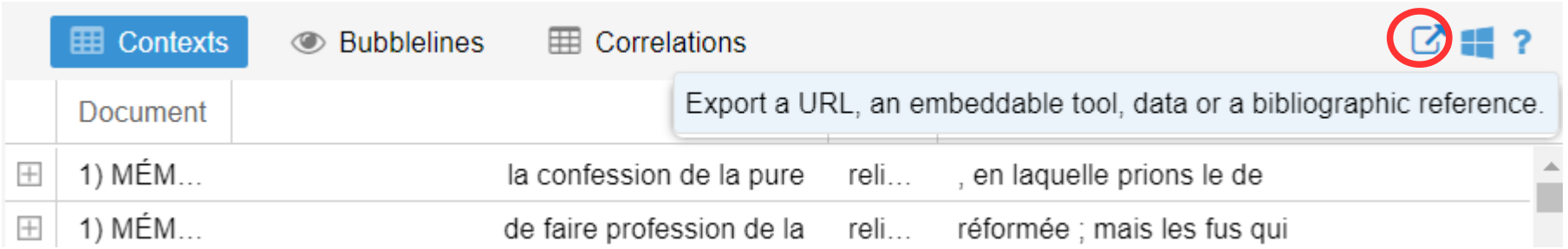
Texte obtenu après remplacements...

Dieu nous gard' tous, autant gros que menus,
Petits et grans, bien soyez-vous venus.
Long temps y a, au moins comme il me semble,
Qu'ici n'y eut autant de peuple ensemble.
Que pleust à Dieu que toutes les semaines
Nous peussions voir les Eglises si pleines.

Orça messieurs, et nous dames honnestes,
le nous suppli d'entendre mes requestes.
le nous requier nous taire seulement.
Comment? dira quelcune uoirement,
le ne sauroy' ni ne uoudroy' avec.
Or si faut-il pourtant clorre le bec.
Ou nous et moy avons peine perdue,
Moy de parler, et nous d'estre venue.

Extraction de contextes avec VoyantTools

- Charger le corpus dans **http://voyant-tools.org** (ex. : *Mémoires* de Mme Duplessis-Mornay, cf. mémoire de Mallory Thiebaud)
- Charger les **contextes** de “religion”, par exemple, dans le cadre en bas à droite, puis exporter avec le bouton entouré en rouge :



- Dans la **fenêtre d'export**, choisir “Export Current Data”, “export current data as tab separated values (text)”
- Coller dans un **document tableur**
- Sélectionner uniquement les contextes gauches, droits, ou les deux, pour les charger dans TreeCloud.