

Master Intelligence stratégique, analyse des risques et territoires
20/02/2020 – IFIS (Serris)

Analyse de données textuelles en pratique : logiciels de textométrie et arbres de mots

Philippe Gambette

LIGM
Université Gustave Eiffel



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

- Prétraitements de textes
- Prétraitements de textes obtenus par OCR
- Extraction de contextes avec Voyant Tools

A vertical blue bar with a gradient is located on the left side of the slide.

Panorama des logiciels de textométrie

Quelques logiciels de textométrie

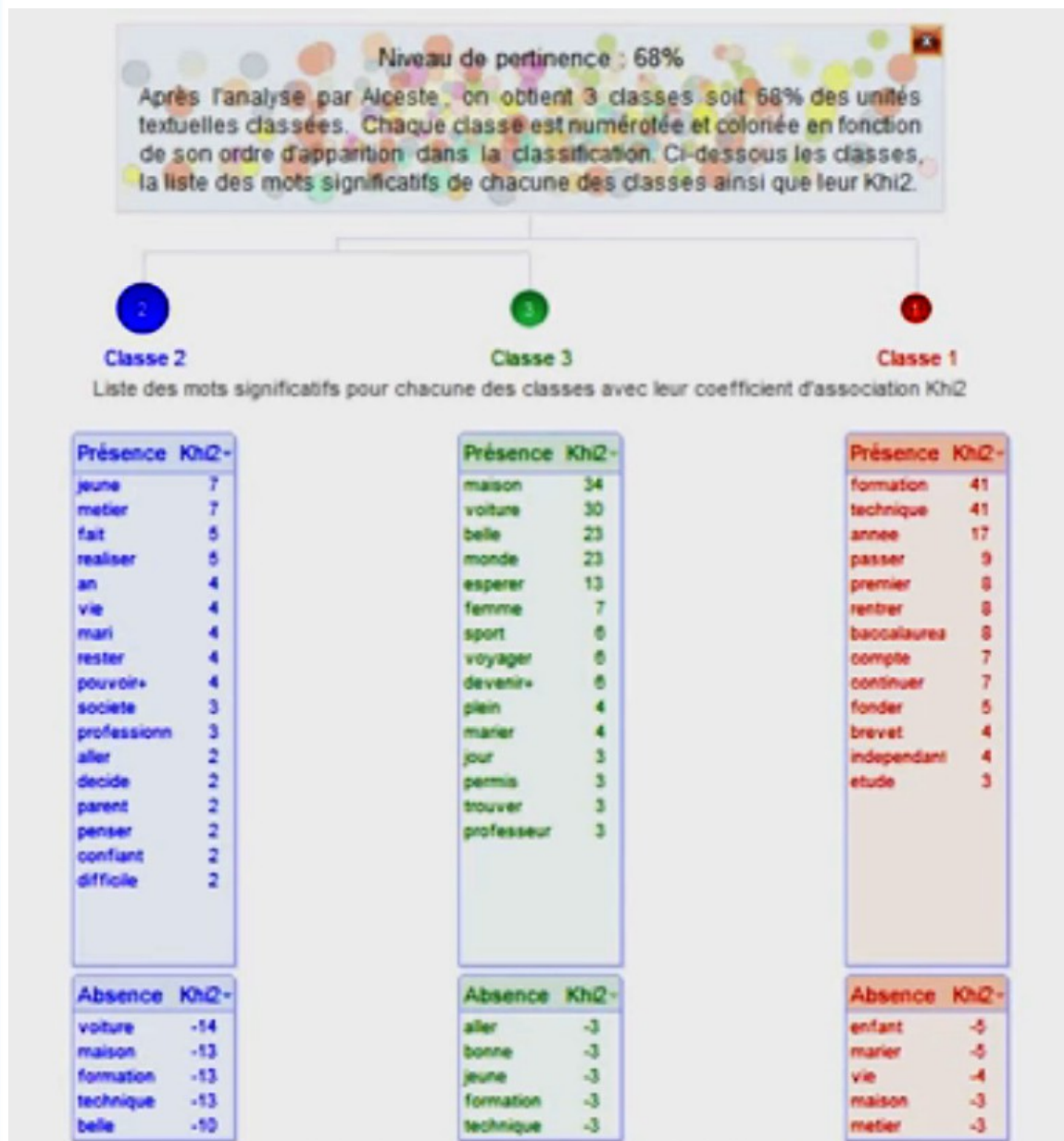
Alceste (depuis 1983)



Société IMAGE

<http://www.image-zafar.com/Logiciel.html>

La méthode Alceste



Répartition des phrases du texte en différentes classes

→ vocabulaire de chaque classe



Quelques logiciels de textométrie

Hyperbase (depuis 1989)



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

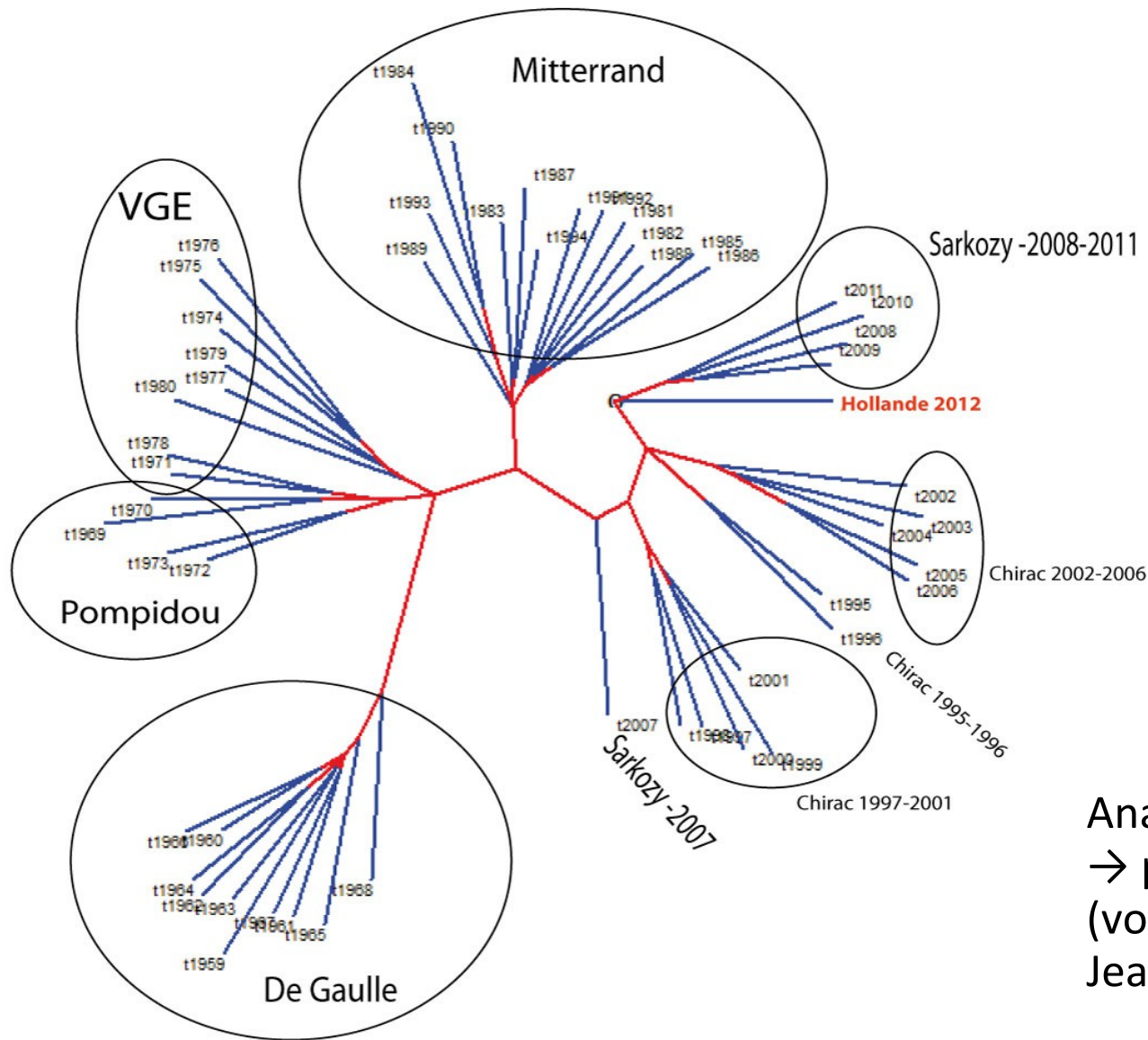
Alceste (depuis 1983)



Société IMAGE

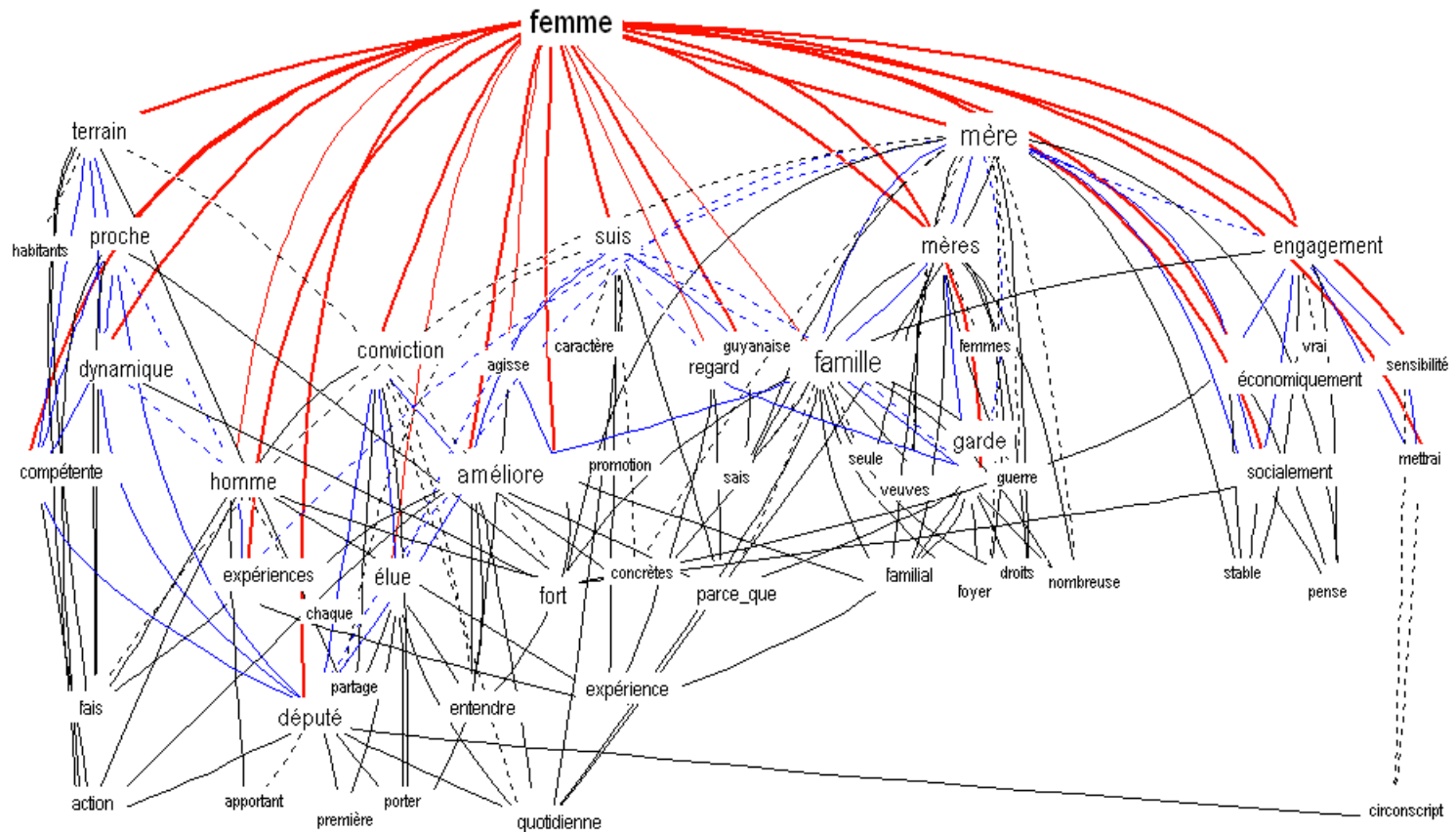
<http://www.image-zafar.com/Logiciel.html>

Hyperbase



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

Hyperbase



Réseau de cooccurents d'un mot

→ graphique des co-occurents directs et indirects du mot-pôle « femme » dans le corpus des Professions de foi de candidates

Magali Guaresi (2014) L'approche co-occurentielle, 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

The screenshot shows the Lexico3 software interface. The title bar reads "Lexico3 - [Section - Délimiteurs : \$ - vue n°1]". The menu bar includes "Fichier", "Traitement", and "Fenêtre". The toolbar contains various icons for file operations and editing. The main window is divided into several panes:

- Navigation**: Includes "Rapport", "Dictionnaire", and "Segments répétés".
- Segments répétés**: A table listing segments and their frequencies.
- Grid**: A grid of 600 small cards representing sections, with the number "9" and "<Aucune>" displayed above it.
- Section details**: A pane showing the selected section "la grande colère" and its occurrence.

Lg	Segment	Frq
2	la constitution	40
2	la contre	39
3	la convention a	15
2	la convention	187
2	la cour	11
2	la crête	11
2	la danse	12
2	la dernière	15
2	la disette	10
2	la famine	14
2	la fête	12
2	la fin	22
2	la force	16
2	la foudre	13
2	la garce	10
2	la gloire	17
6	la grande colère du *père *duch...	54
3	la grande colère	55
6	la grande joie du *père *duchesne	36
4	la grande joie du	37
2	la grande	113
3	la guerre civile	32
2	la guerre	99
2	la guillotine	32
2	la journée	17
2	la justice	13
2	la langue	10
4	la liberté et l	12
3	la liberté et	14
2	la liberté	202
2	la linotte	34
2	la loi	38
3	la louve autrichienne	19
2	la louve	21
2	la main	80
2	la même	19
2	la misère	20
2	la moitié	13
2	la montagne	22
2	la mort	43

Section : la grande douleur du *père *duchesne au sujet de la mort de *marat assassiné à coups de couteau par une garce du *calvados , dont l' évêque *fauchet était le directeur . ses bons avis aux braves *sans - culottes pour qu' ils se tiennent sur leurs gardes , attendu qu' il y a dans *paris plusieurs milliers de tondues de la *vendée qui ont la patte graissée pour égorger tous les bons citoyens . <edito=1>\$

Occurrence :

Section

C:\Program Files\Lexico 3\Duchn.par

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/>

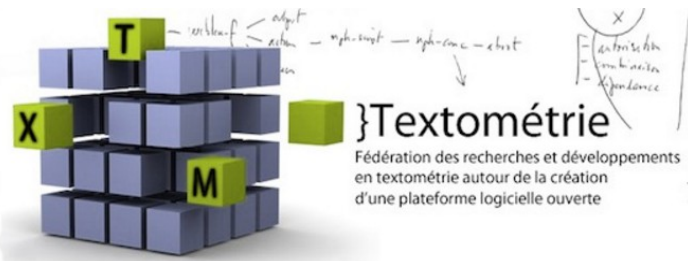
Alceste (depuis 1983)



Société IMAGE

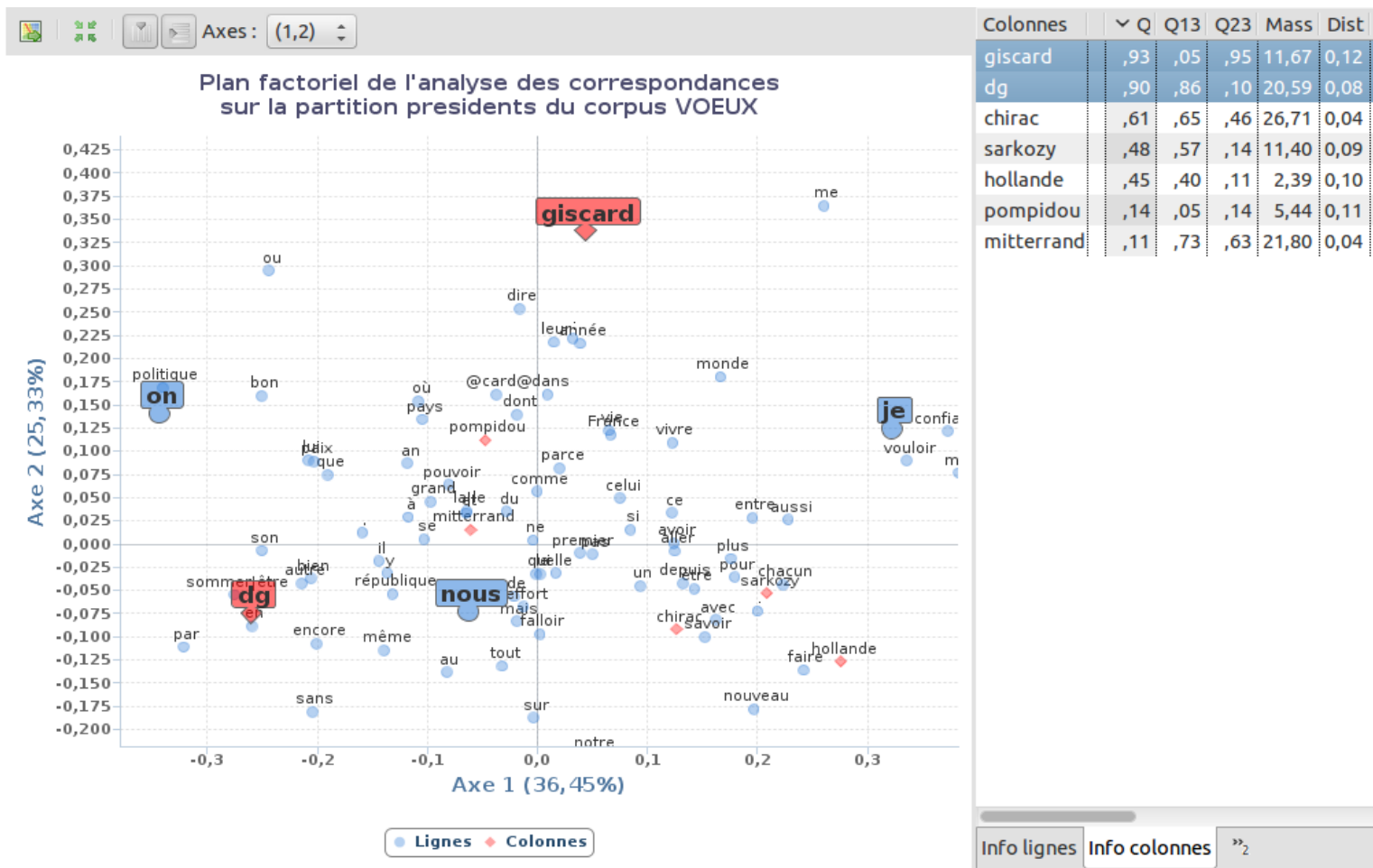
<http://www.image-zafar.com/Logiciel.html>

TXM (depuis 2008)



ENS Lyon

<http://textometrie.ens-lyon.fr/>



Analyse factorielle des correspondances

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

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)



}Textométrie
Fédération des recherches et développements
en textométrie autour de la création
d'une plateforme logicielle ouverte

ENS Lyon

<http://textometrie.ens-lyon.fr/>

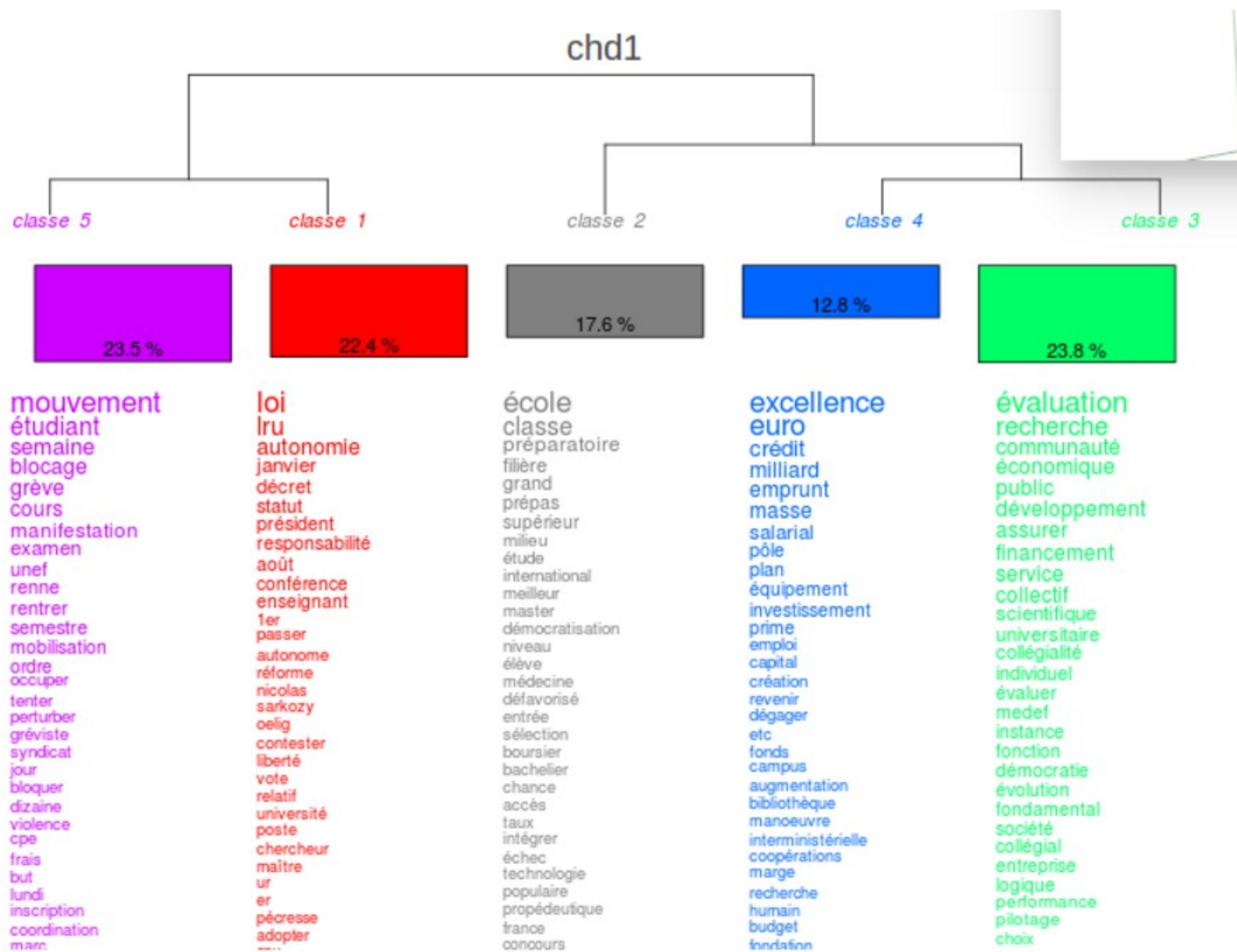
Iramuteq (depuis 2009)



Université de Toulouse

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

Iramuteq



Iramuteq

IRaMuTeQ 0.7 alpha 2

Historique

- Corpus textuel
 - lru4_M2R
 - Sublru4_test
 - Sublru4_test
 - Sublru4_test
 - Sublru4_test
 - souscorpuscl
 - reforme
 - lru4_test
 - lru4_stat_1
 - lru4_spec
 - lru4_alces
 - lru4_cluste
 - lru4_simitx
 - lru4_stat_2
 - lru4_corpus
 - lru4_corpus
 - gaymariage
 - mpt_europre
 - discours_cor
 - discoursTXM
 - corpus_from
 - corpus_from
 - discoursfrom
 - Subjgauche
 - Sublru4_for
 - q101112
 - jgauche
 - lru4_classe5
 - etudiant
 - lru4_FFS
 - sab_thema1
 - figaro
 - lru4_ESS
 - ASPA051116

Classification - lru4_test x

AFC x

AFC Facteur Graphe 3D

num eff. s.t. eff. tota

1 Classe 1	2 Classe 2
318/1419	249/1419
22.41%	17.55%

num eff. s.t. eff. tota

num	eff.	s.t.	eff. tota
0	49	66	
1	95	204	
2	24	28	
3	18	19	
4	25	32	
5	30	44	
6	15	15	
7	21	26	
8	18	22	
9	19	24	
10	19	24	

RGL device 1 [Focus]

facteur 2 - 26.25%

facteur 3 - 21.72%

facteur 1 - 26.25%

classe 5

classe 4

classe 3

classe 2

classe 1

chd1

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/>

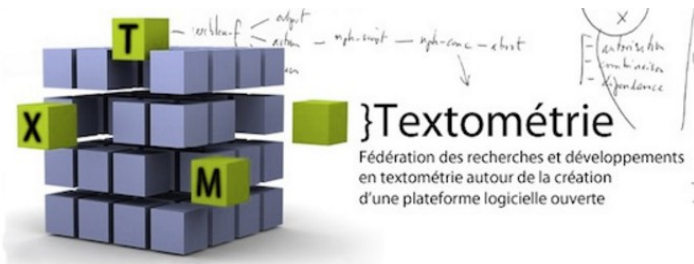
Alceste (depuis 1983)



Société IMAGE

<http://www.image-zafar.com/Logiciel.html>

TXM (depuis 2008)



}Textométrie
Fédération des recherches et développements
en textométrie autour de la création
d'une plateforme logicielle ouverte

ENS Lyon

<http://textometrie.ens-lyon.fr/>

Iramuteq (depuis 2009)



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 t_0, t_1, \dots, t_k 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 t_0 and its secondary phylogenetic trees are t_1, \dots, t_k .

<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

Exemple d'exploration : articles scientifiques

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

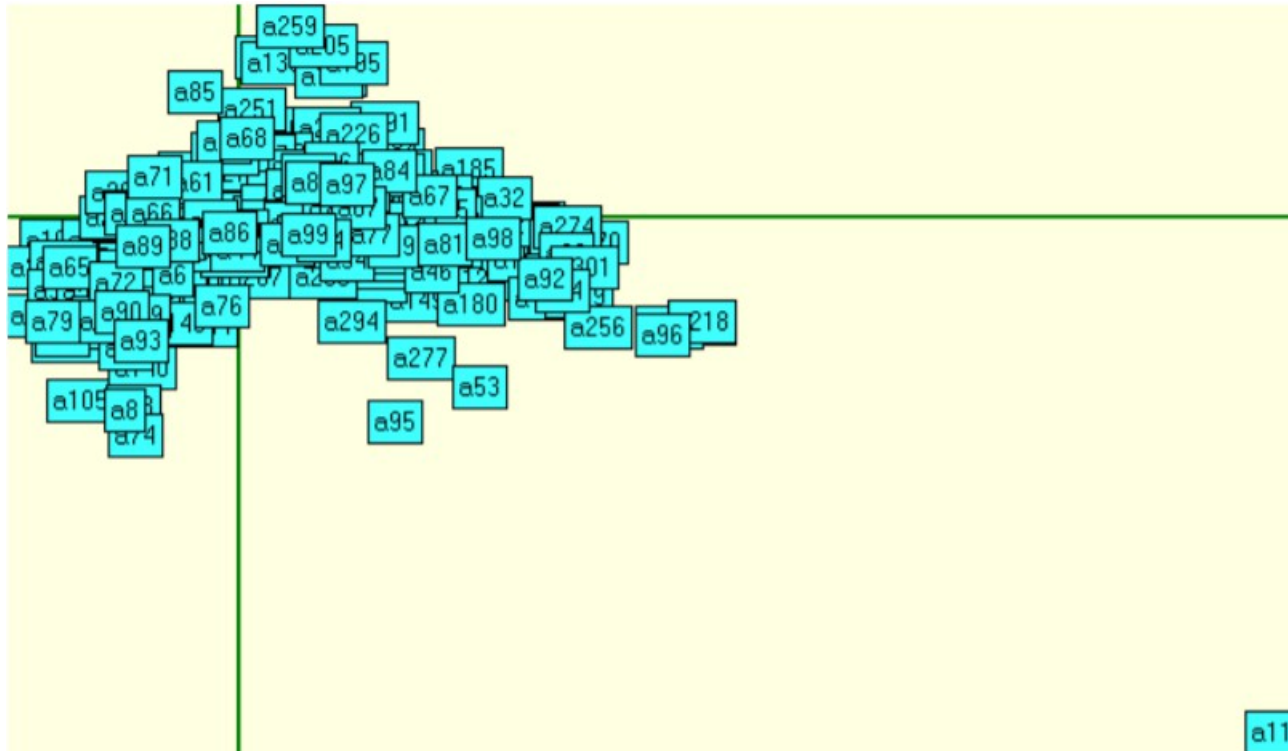
<annee=2015> <type=article> <doc=a1> background: lateral, or horizontal, gene transfers are a type of asymmetric evolutionary events where genetic material is transferred from one species to another. in this paper we consider lgt networks, a general model of phylogenetic networks with lateral gene transfers 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 t_0, t_1, \dots, t_k 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 t_0 and its secondary phylogenetic trees are t_1, \dots, t_k .

<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

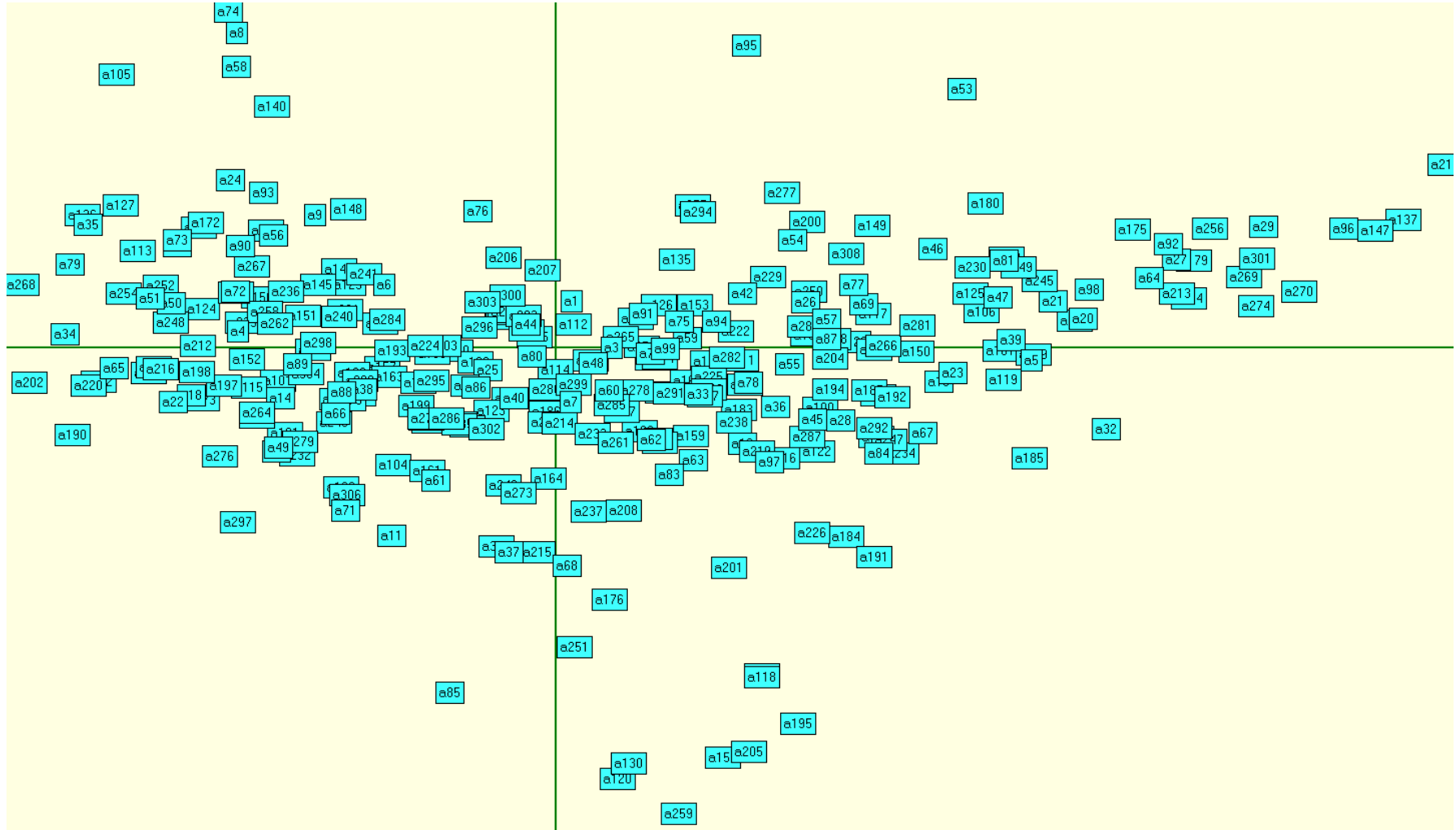
Exemple d'exploration : articles scientifiques

Etape 2) Analyse factorielle des correspondances



Exemple d'exploration : articles scientifiques

Etape 2) Analyse factorielle des correspondances (sans a110)



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
...

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
...

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
...

Attention aux
conclusions
tirées : tests
multiples...

Déclaration de
2016 de l'AAS
sur l'usage des
p-values

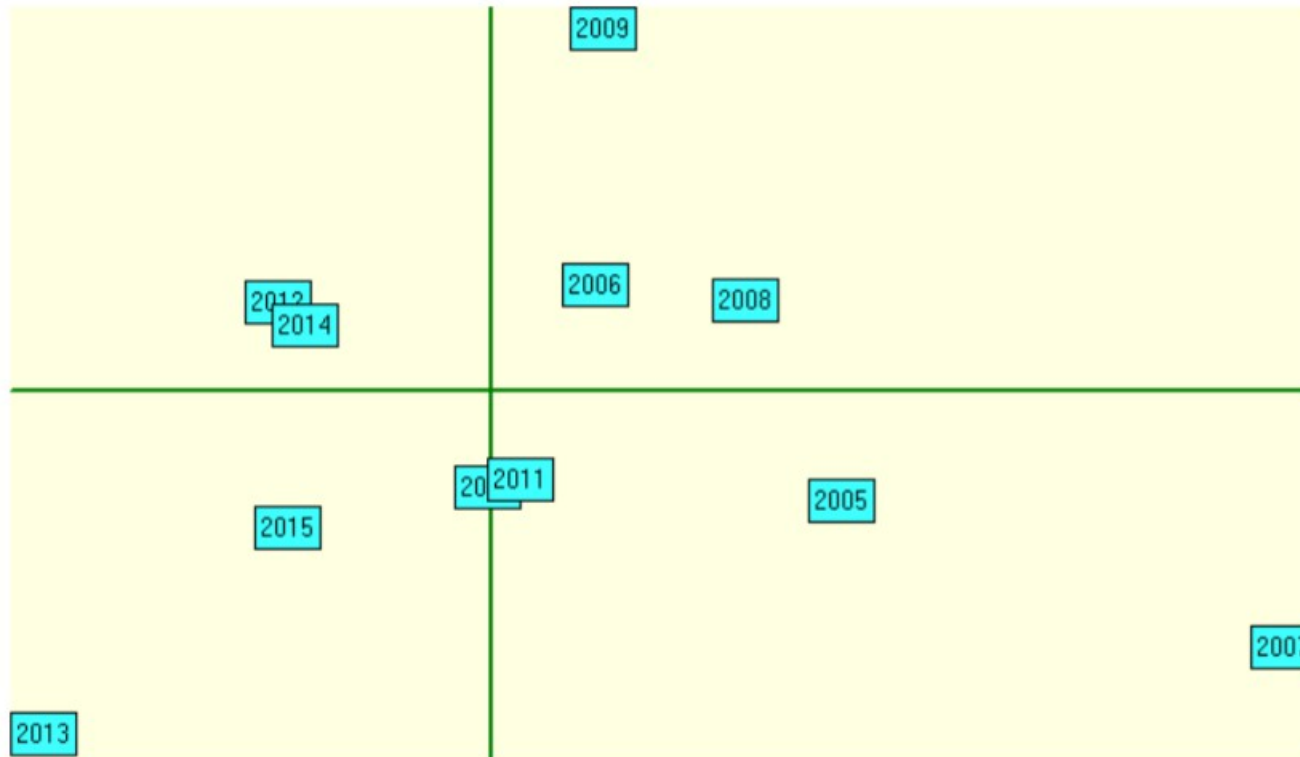
Exemple d'exploration : articles scientifiques

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 \hat{a} 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
high 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

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>

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

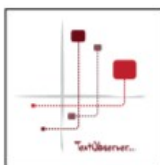
Jean-Marc Leblanc, Philippe Gambette, Claude Martineau, Emilie Née, Marie Pérès

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).

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, Xistat, 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/www/subscribe/textopol>

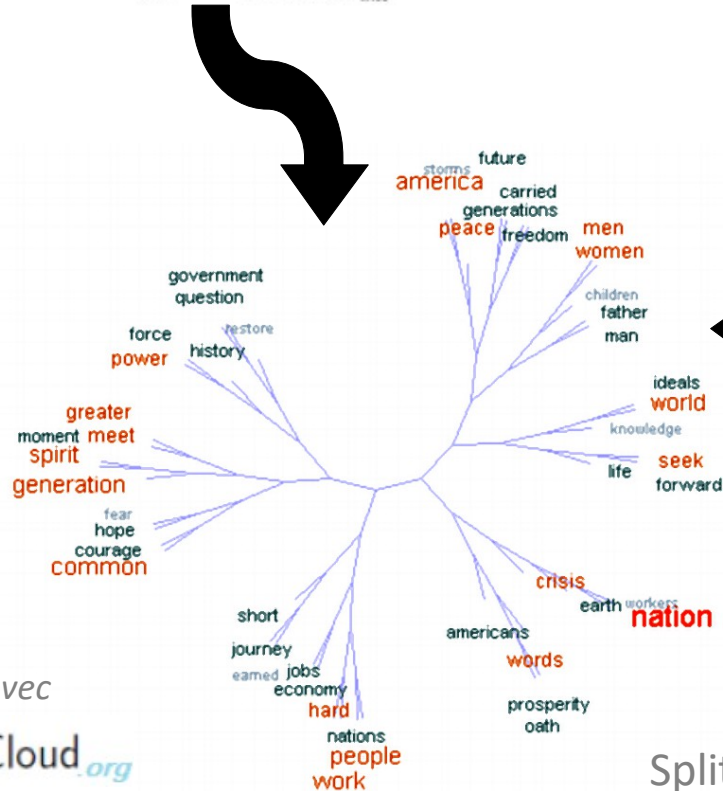
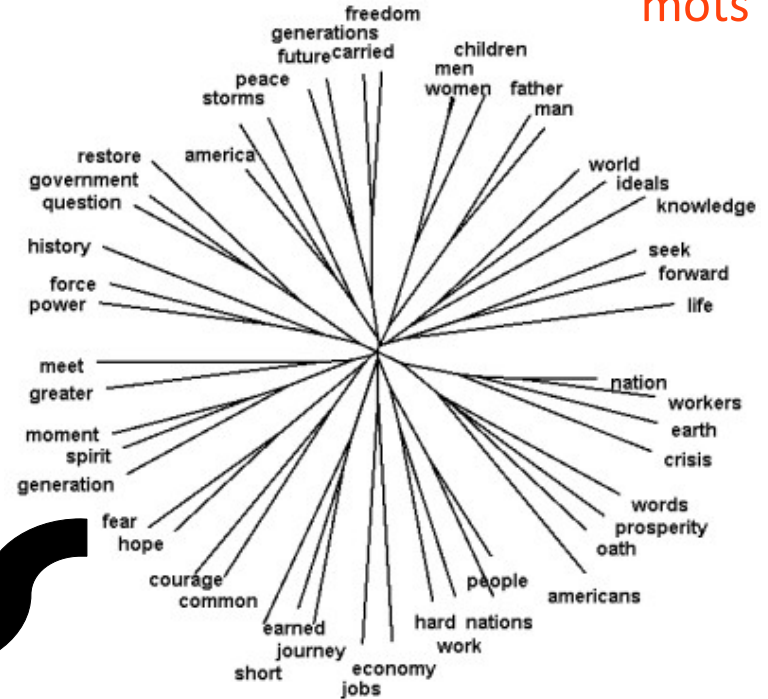
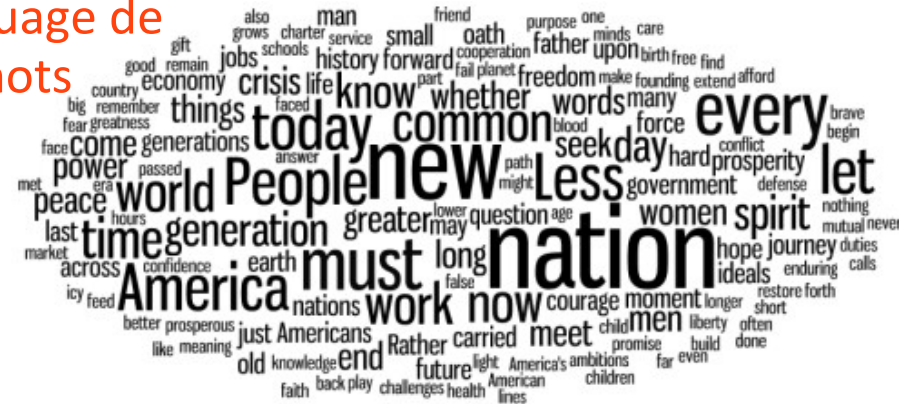
PROGRAMME 2019-2020

Arbres de mots

Le « nuage arboré », une information double

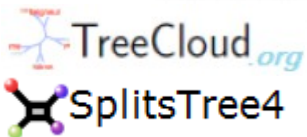
nuage de mots

arbre de mots



Discours inaugural de Barack Obama

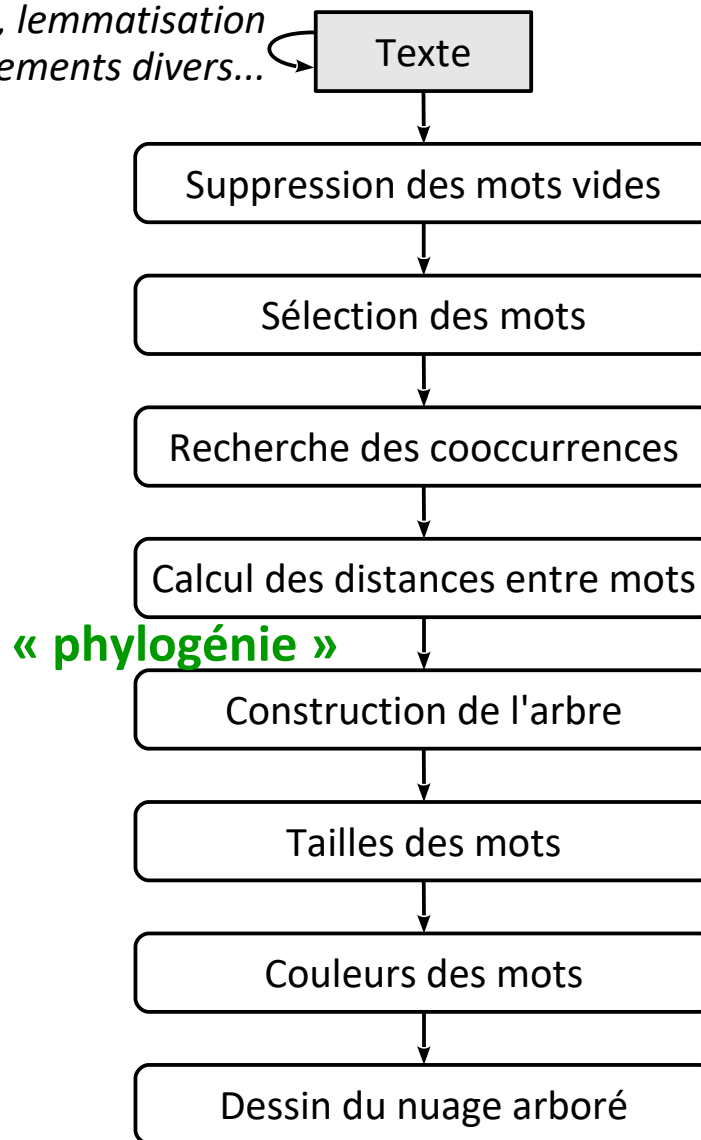
construit avec



SplitsTree : Huson & Bryant, *Bioinformatics*, 2006
TreeCloud : Gambette & Véronis, *IFCS'09*

Processus de construction

*Concordance d'un mot, lemmatisation
ou remplacements divers...*



**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

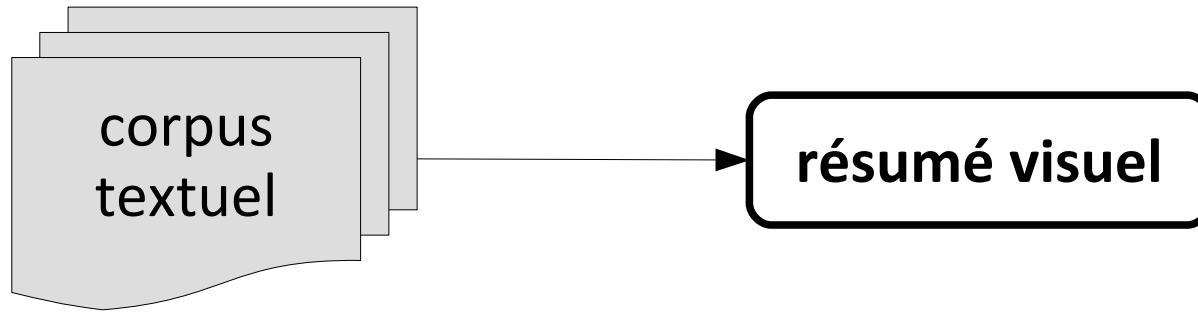
*Appel transparent au logiciel
SplitsTree*

Fréquences ou valeurs personnalisées

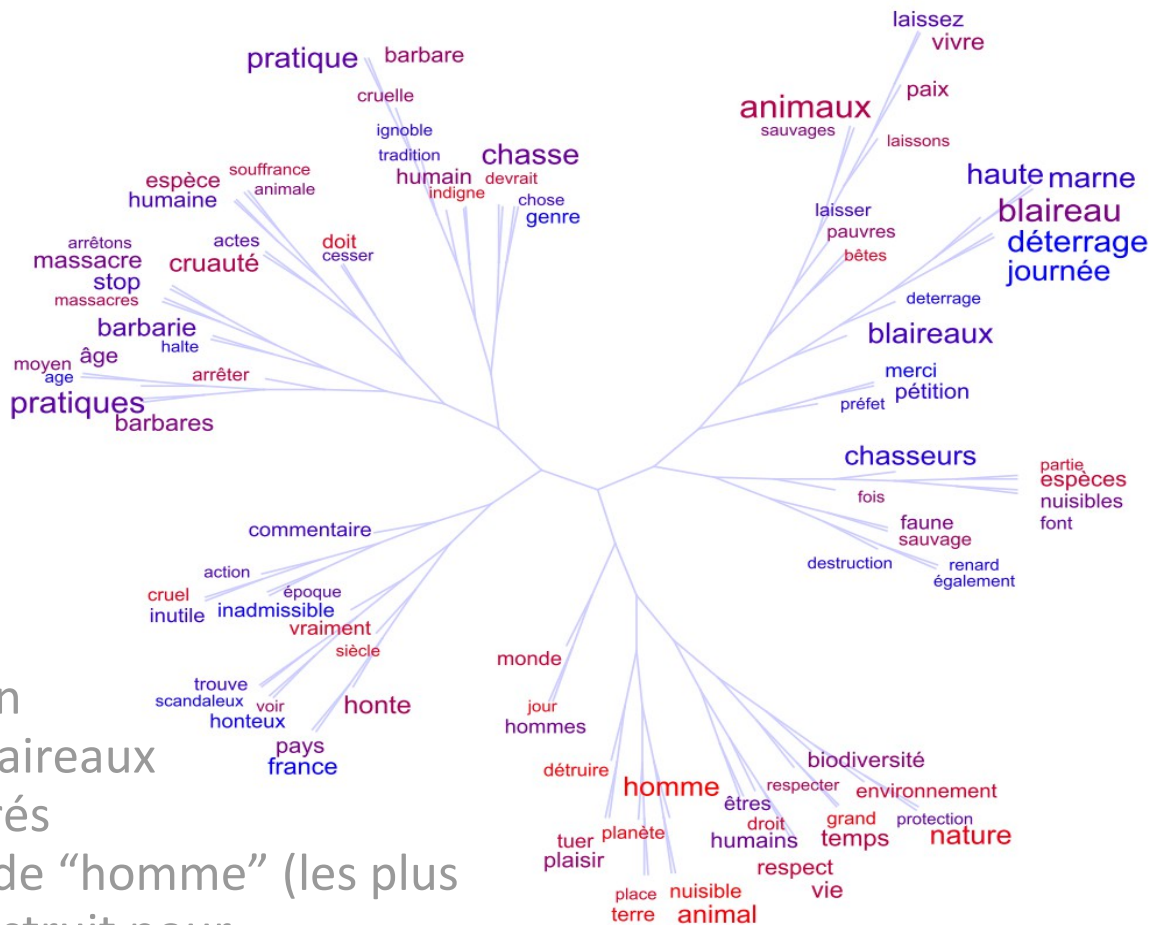
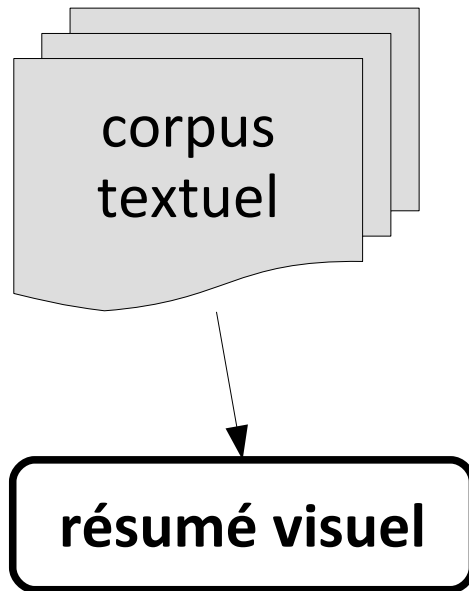
*Fréquences, chronologie, dispersion,
ciblées sur la cooccurrence d'un mot,
ou valeurs personnalisées*

*Appel transparent au logiciel SplitsTree
ou Dendroscope*

Le « nuage arboré », pour quoi faire ?

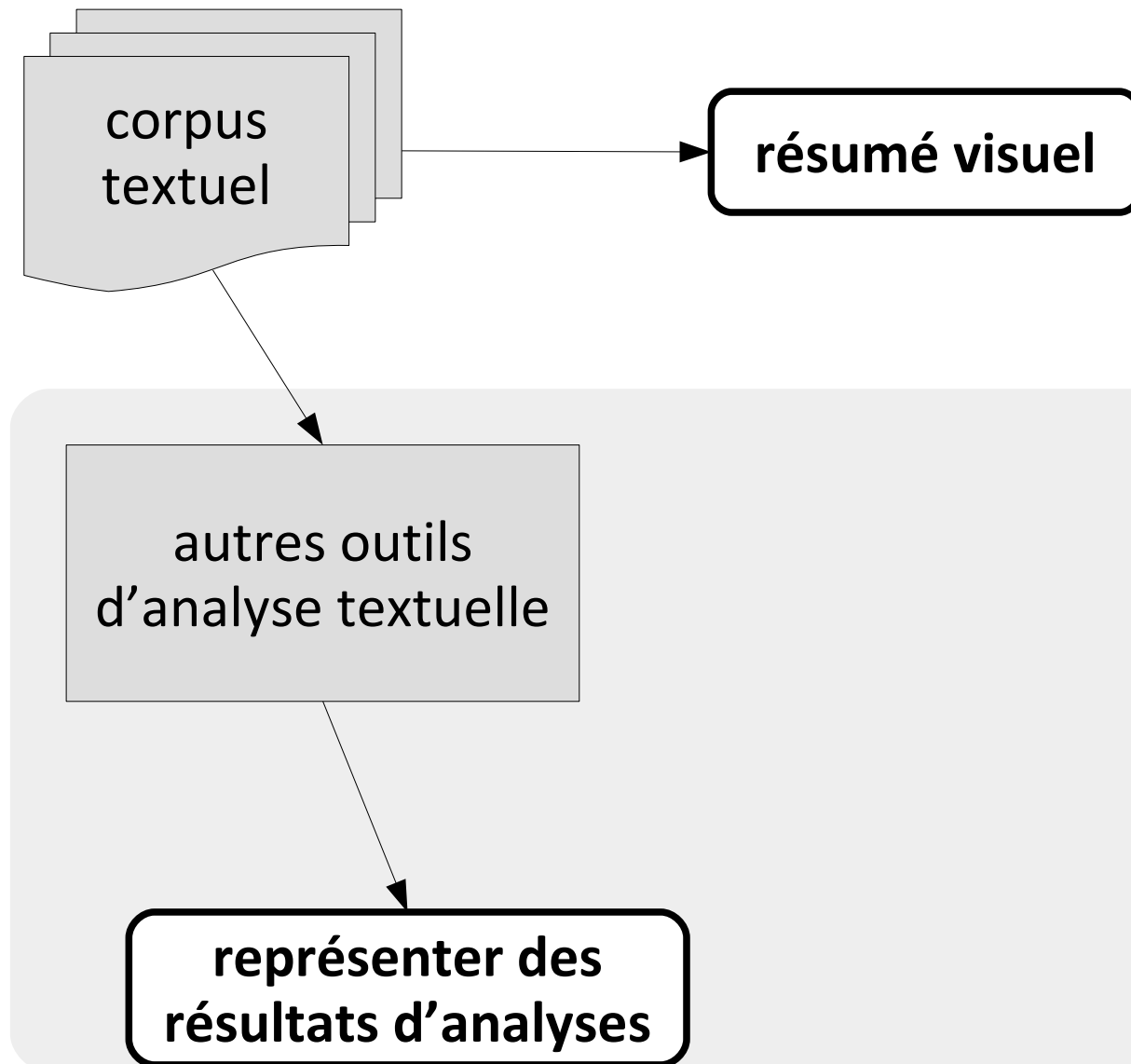


Le « nuage arboré », pour quoi faire ?



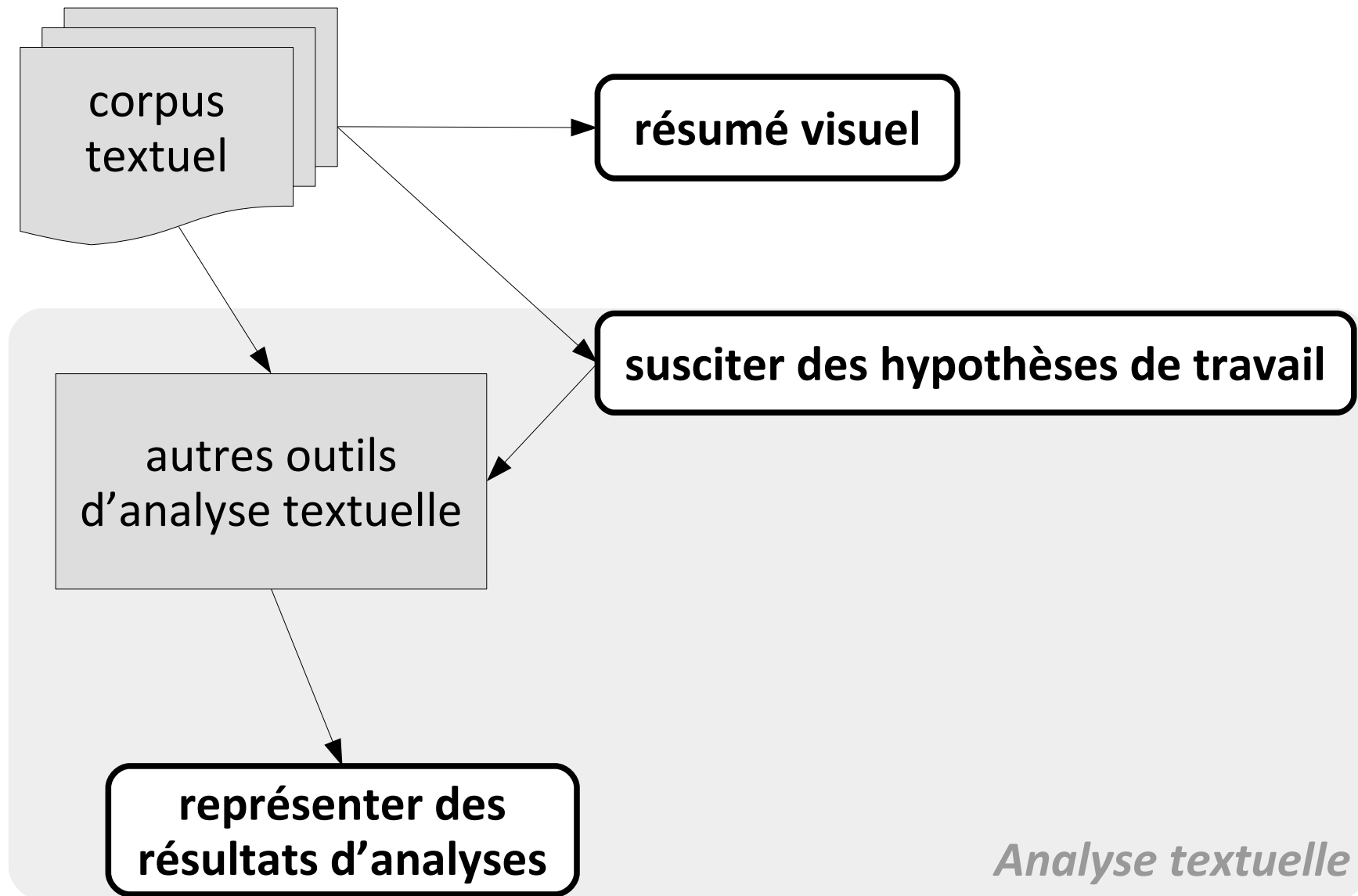
Nuage arboré des 100 mots les plus fréquents des commentaires d'une pétition contre l'extermination de blaireaux sur *lapetition.be*, mots colorés en fonction de la proximité de "homme" (les plus cooccurrents en rouge), construit pour R. Matuszewicz et M. Legris-Revel (projet ANR APPEL)

Le « nuage arboré », pour quoi faire ?



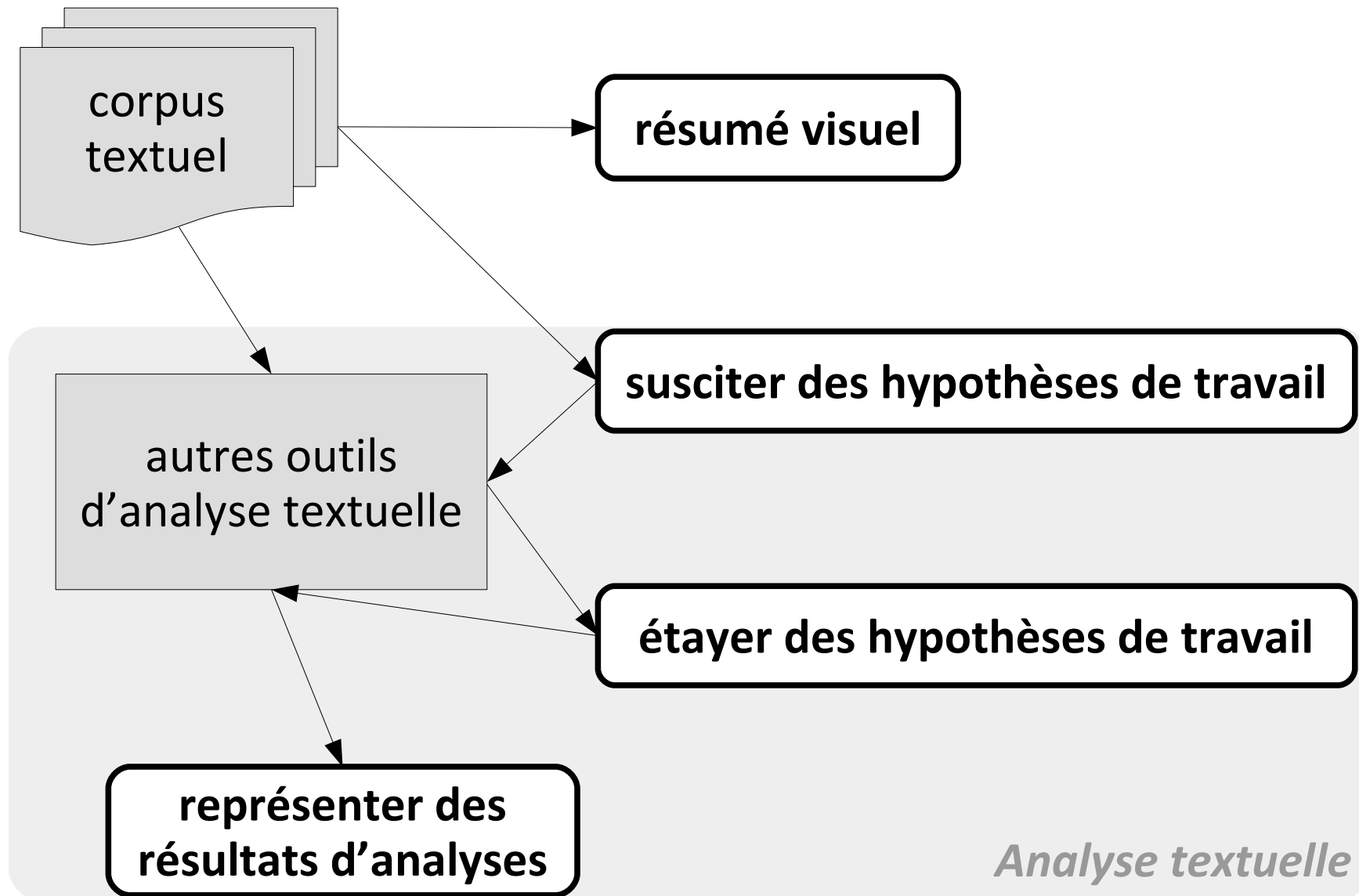
Analyse textuelle

Le « nuage arboré », pour quoi faire ?



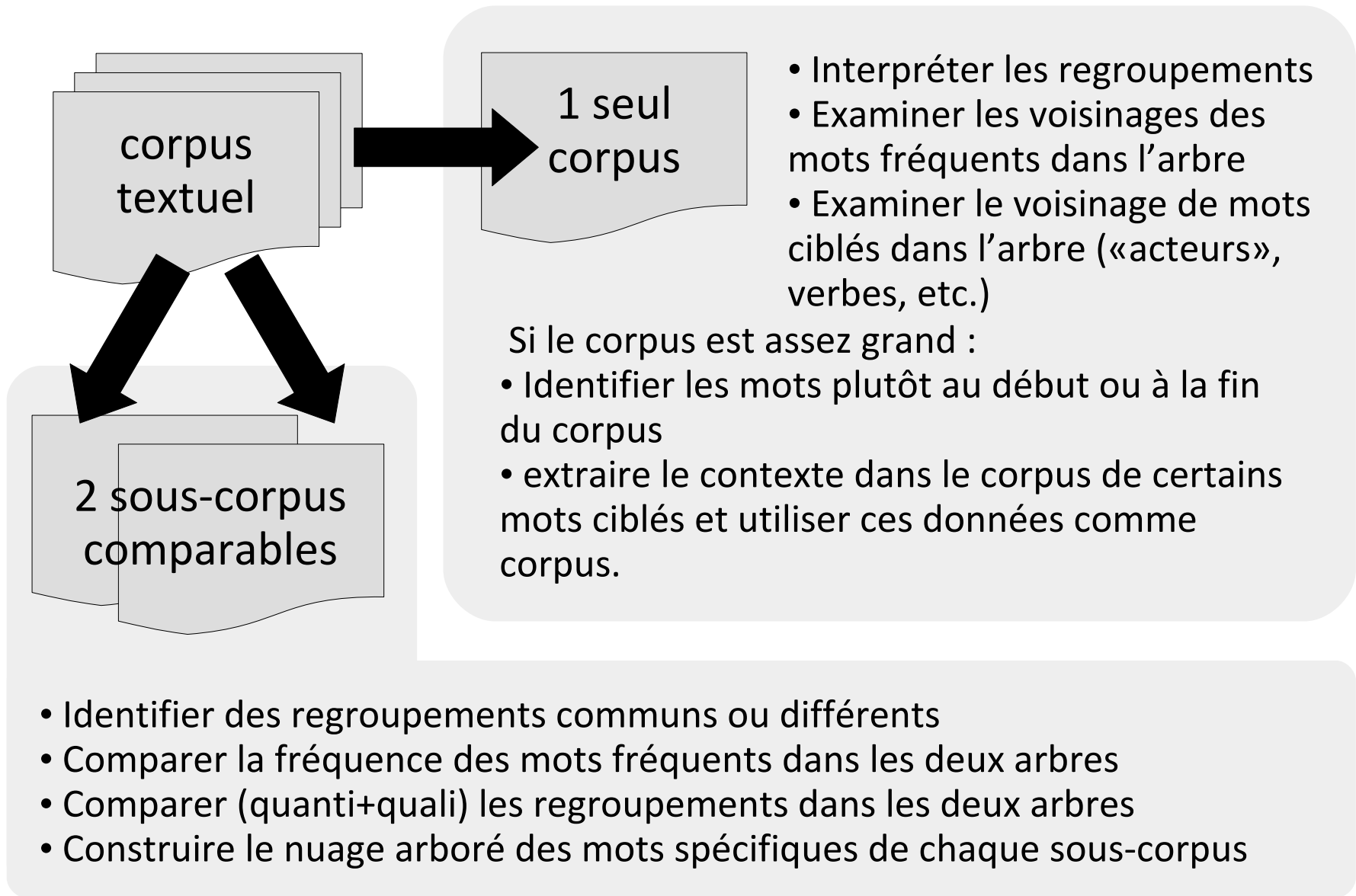
Analyse textuelle

Le « nuage arboré », pour quoi faire ?

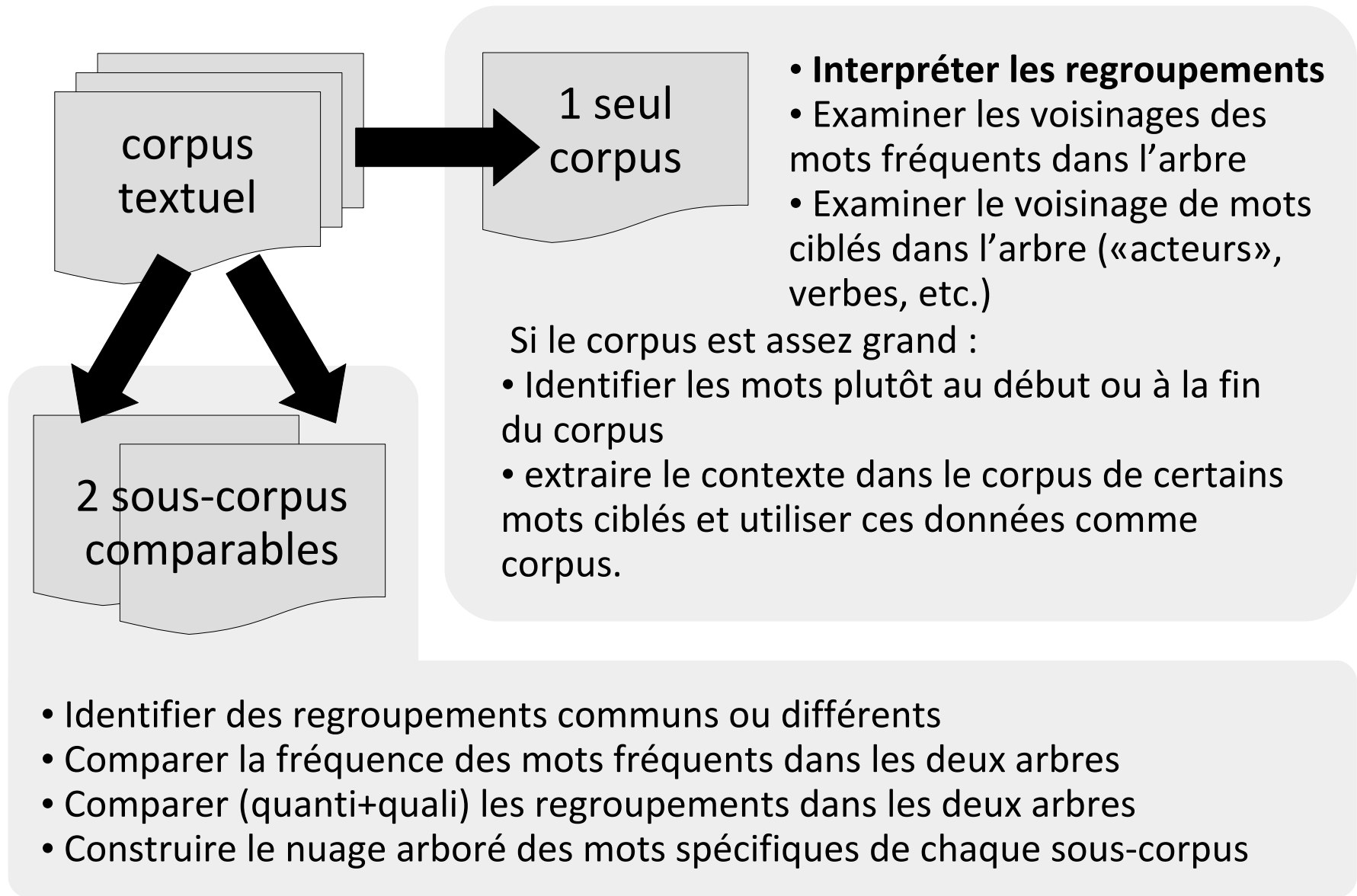


Analyse textuelle

Exploration de corpus avec TreeCloud



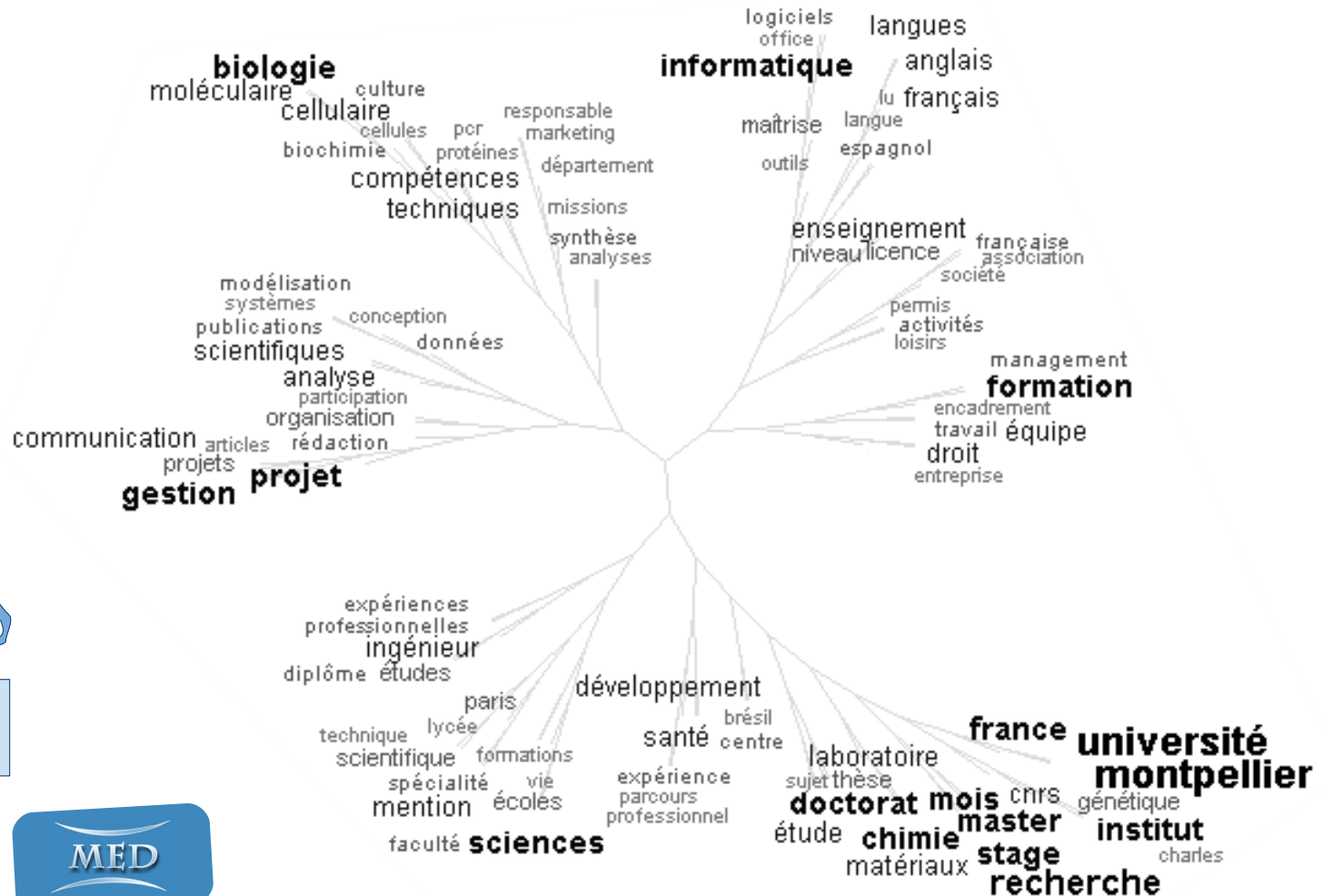
Exploration de corpus avec TreeCloud



Méthode : interpréter les regroupements

Dessiner des « patates »

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



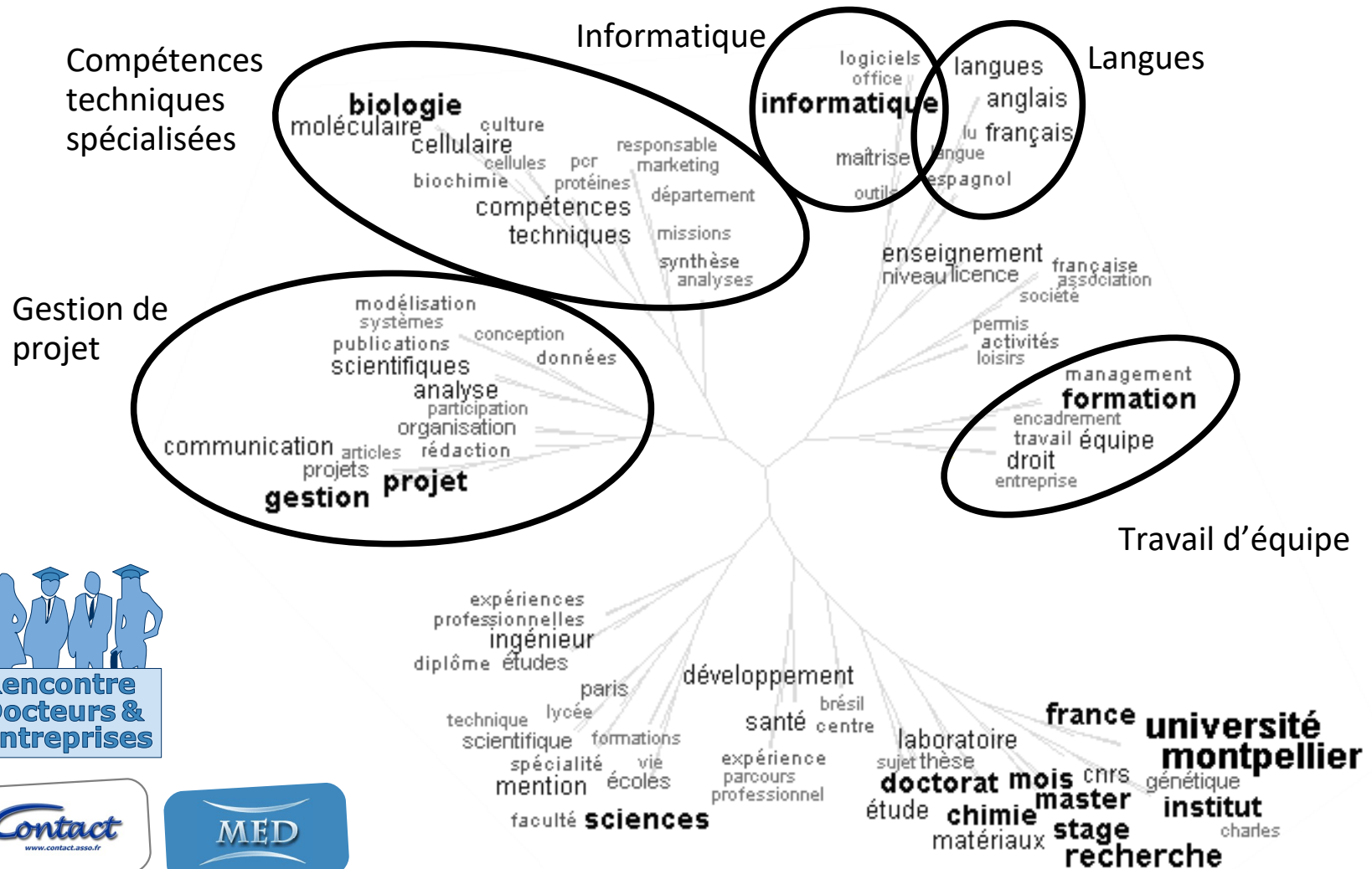
Rencontre
Docteurs &
Entreprises



Méthode : interpréter les regroupements

Dessiner des « patates »

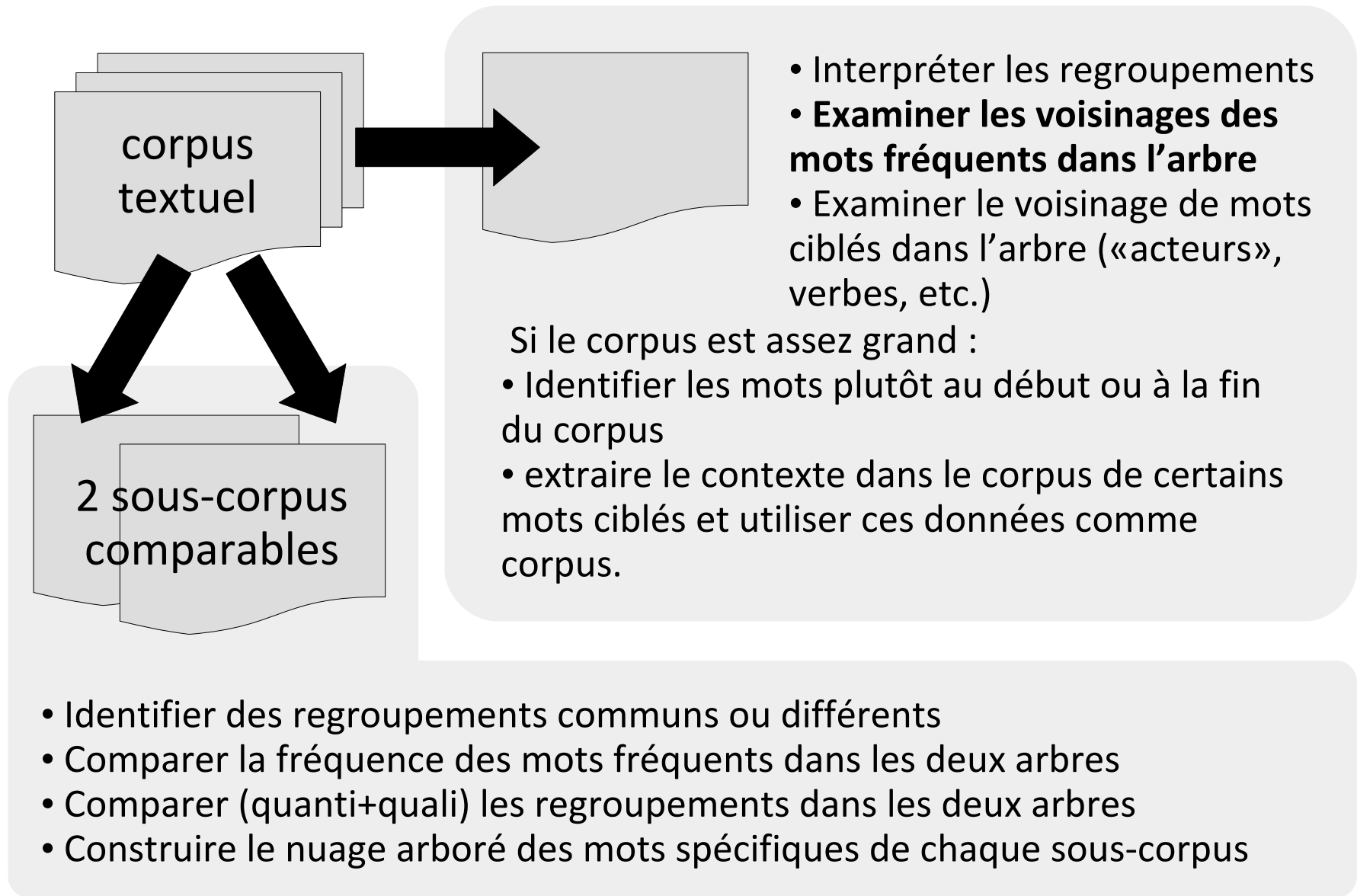
Corpus : une centaine de CV soumis à une rencontre docteurs-entreprises



Rencontre
Docteurs &
Entreprises

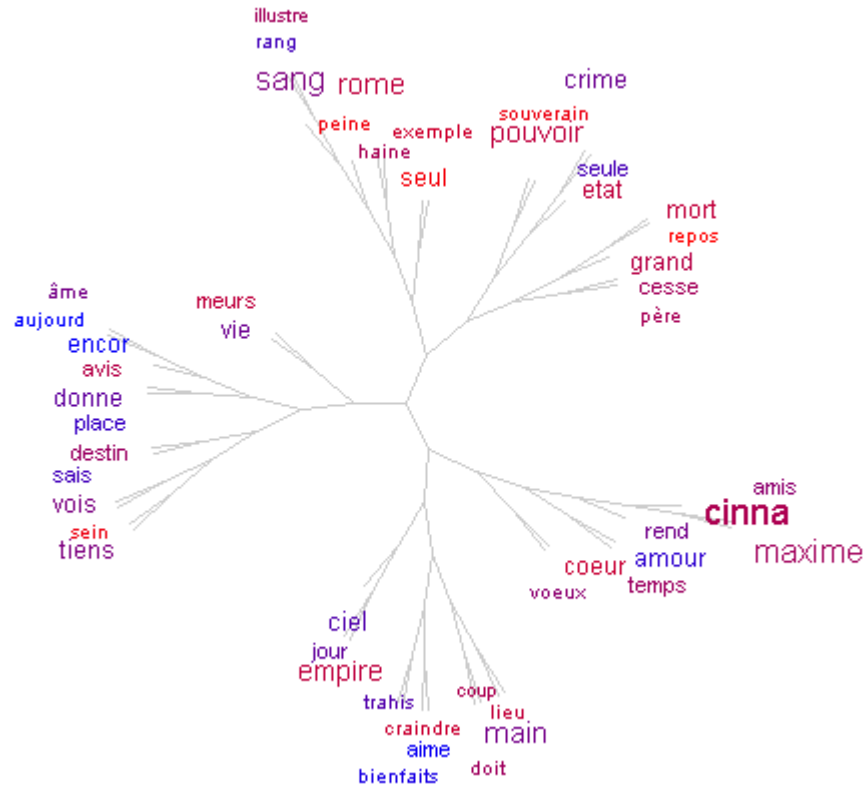


Exploration de corpus avec TreeCloud



Méthode : voisinage des mots fréquents

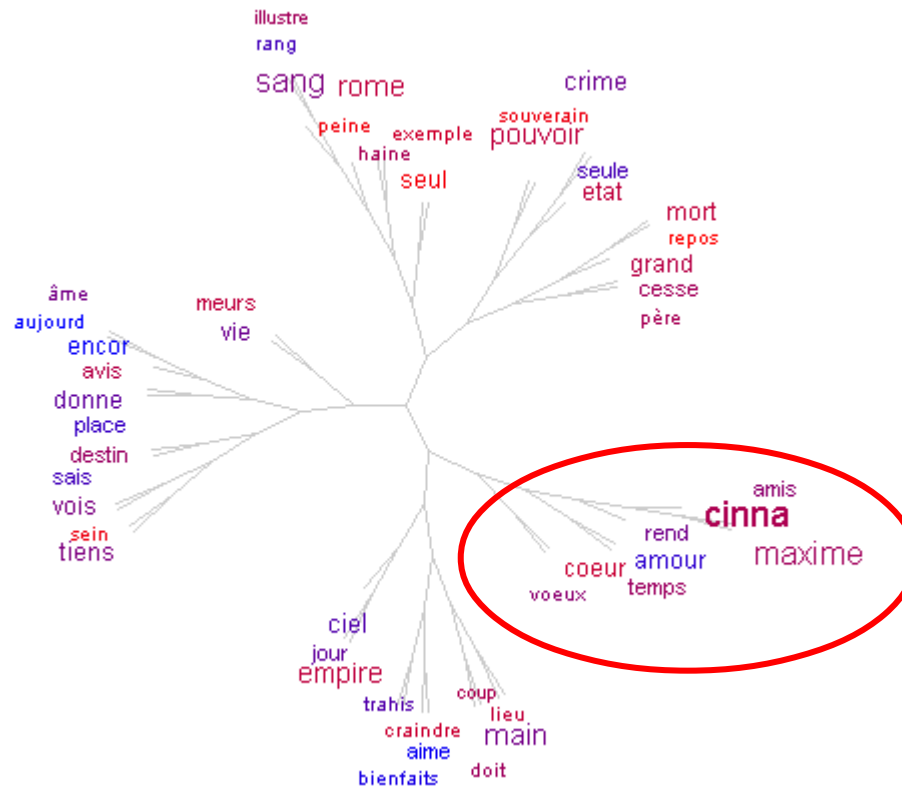
Amstutz & Gambette,
JADT 2010



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

Méthode : voisinage des mots fréquents

Amstutz & Gambette,
JADT 2010

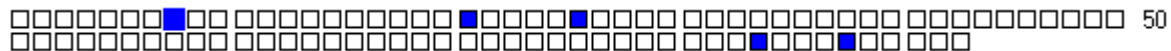
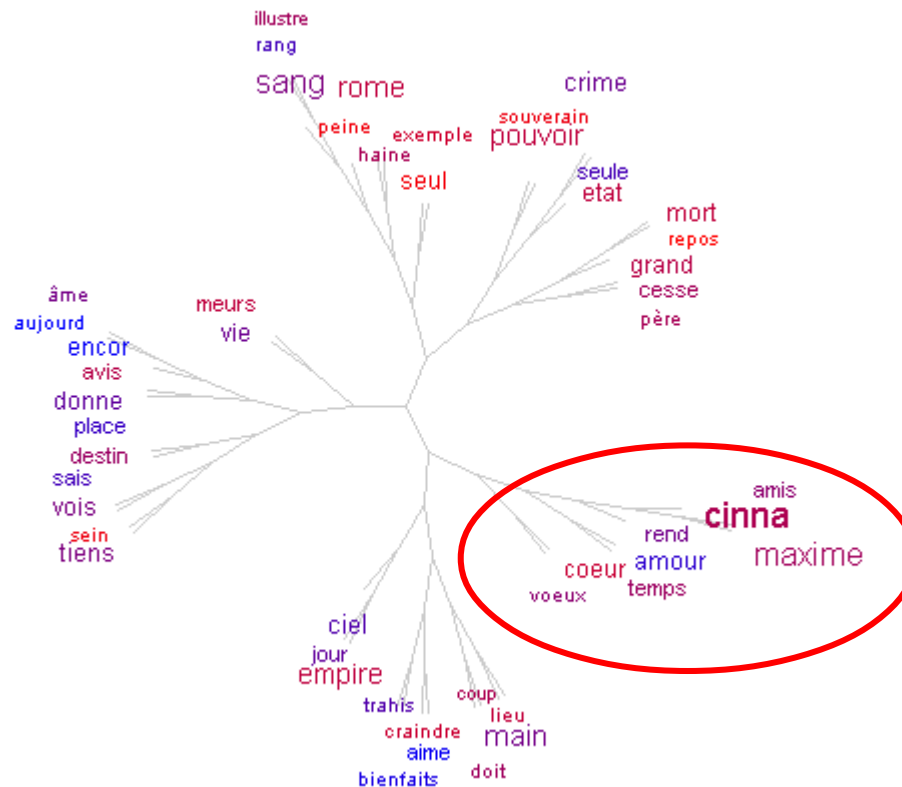


Nuage arboré des 50 mots les plus fréquents des paroles d'Auguste dans Cinna

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

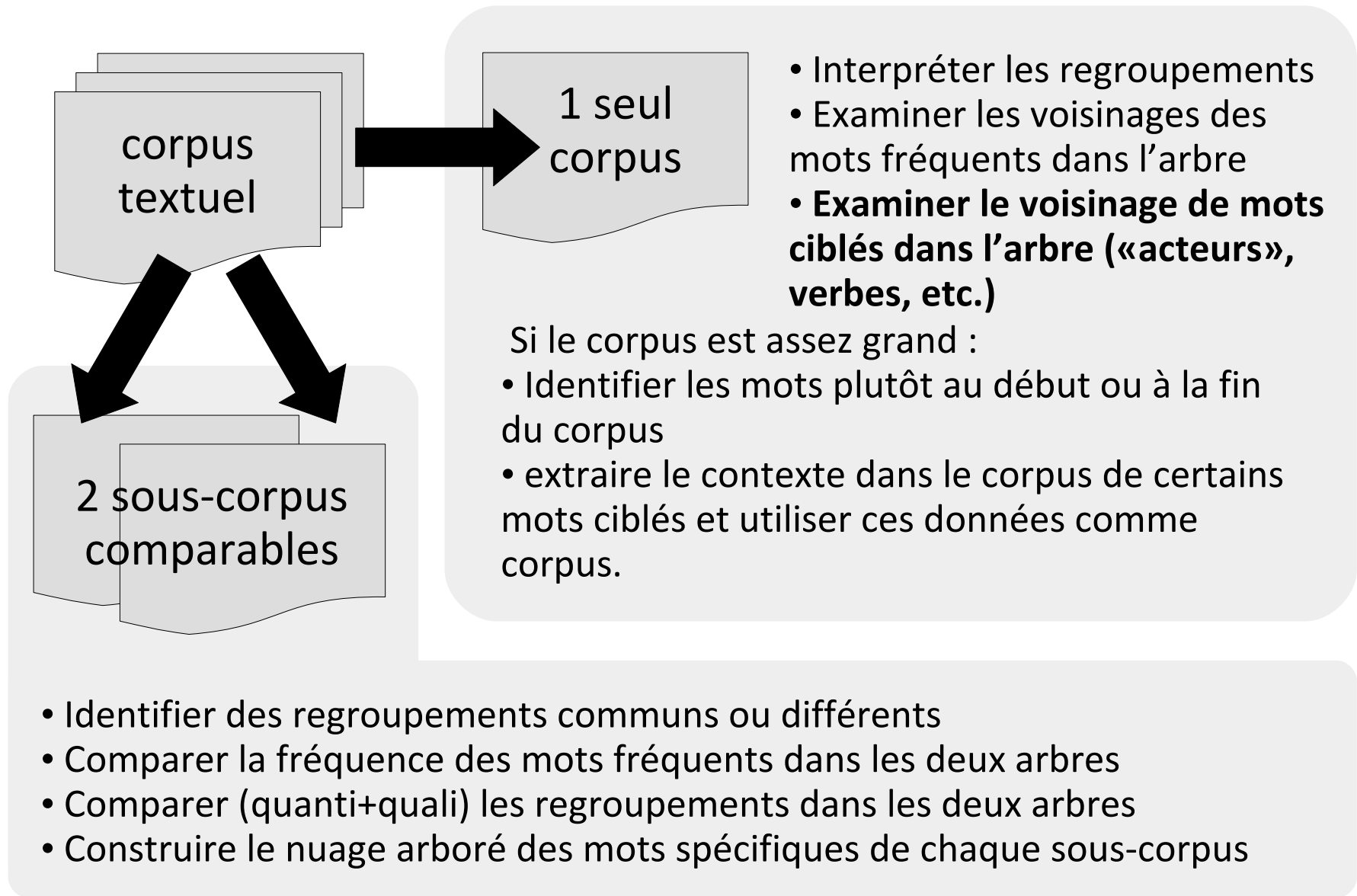
Amstutz & Gambette,
JADT 2010



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**.

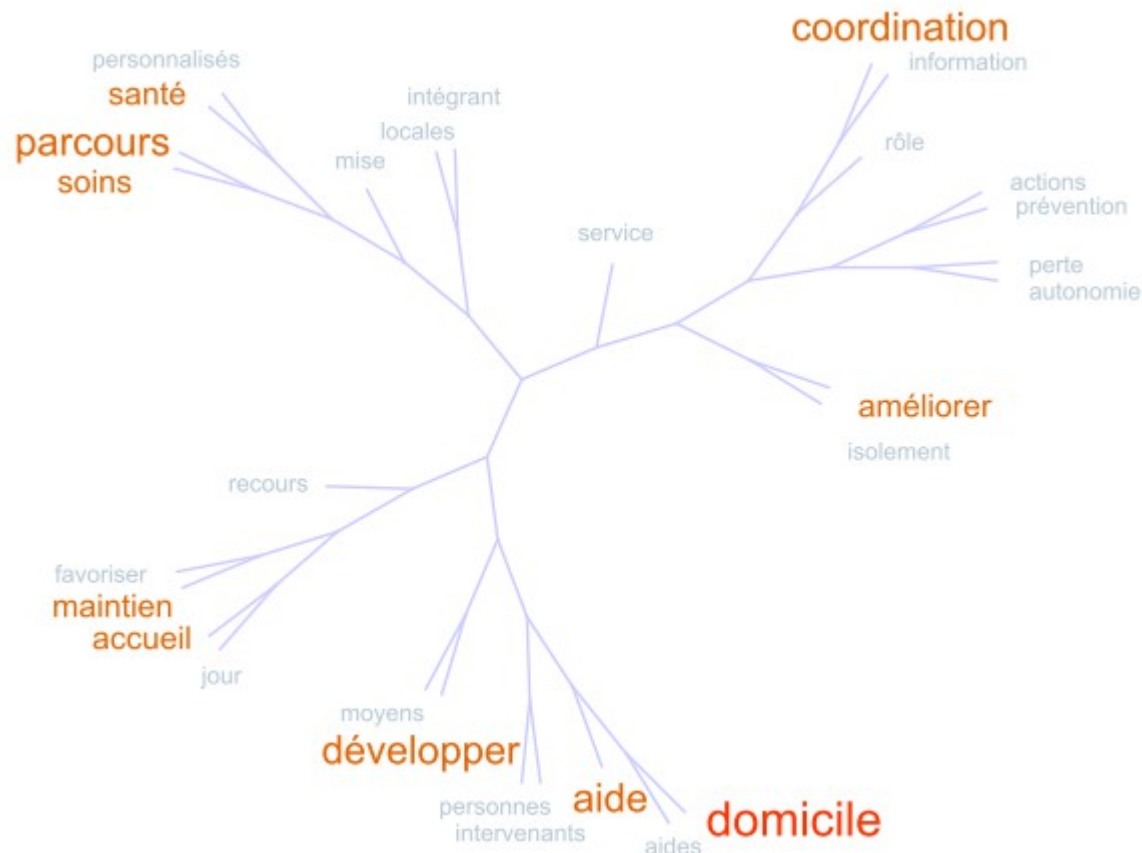
Exploration de corpus avec TreeCloud



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

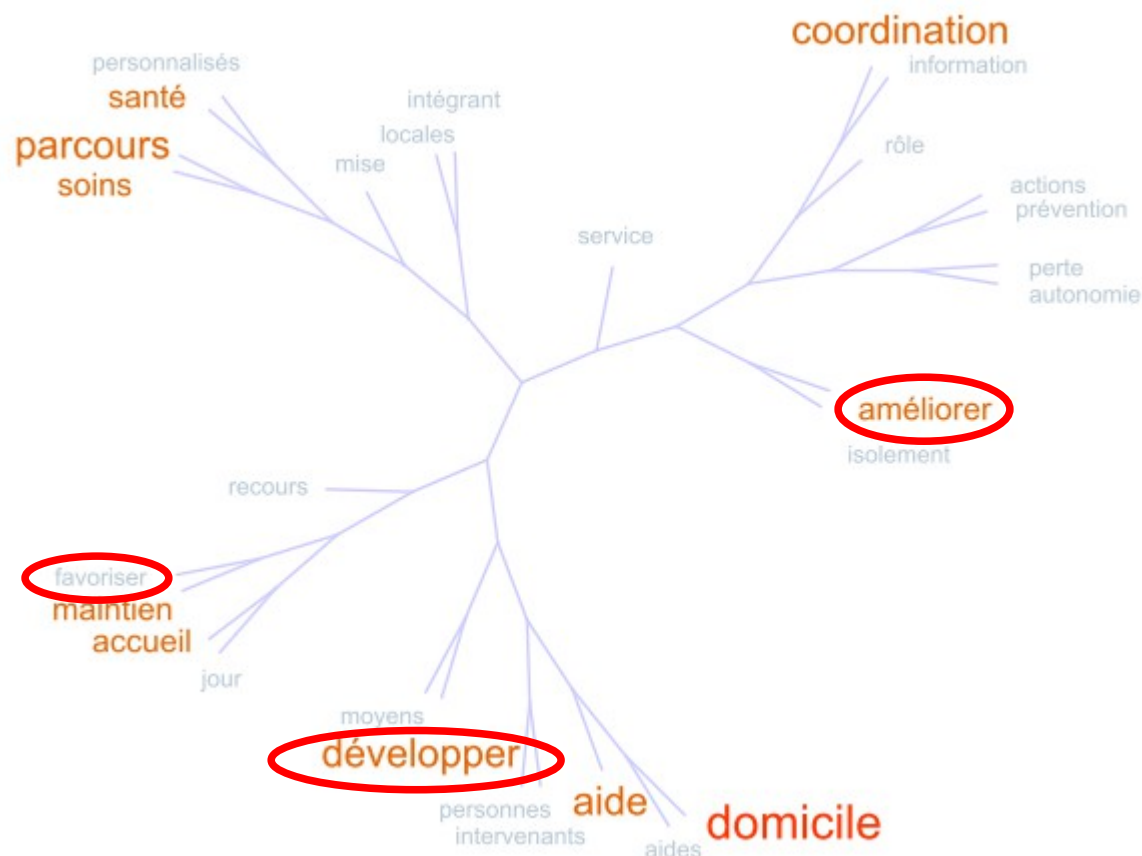
Suggestions d'améliorations :



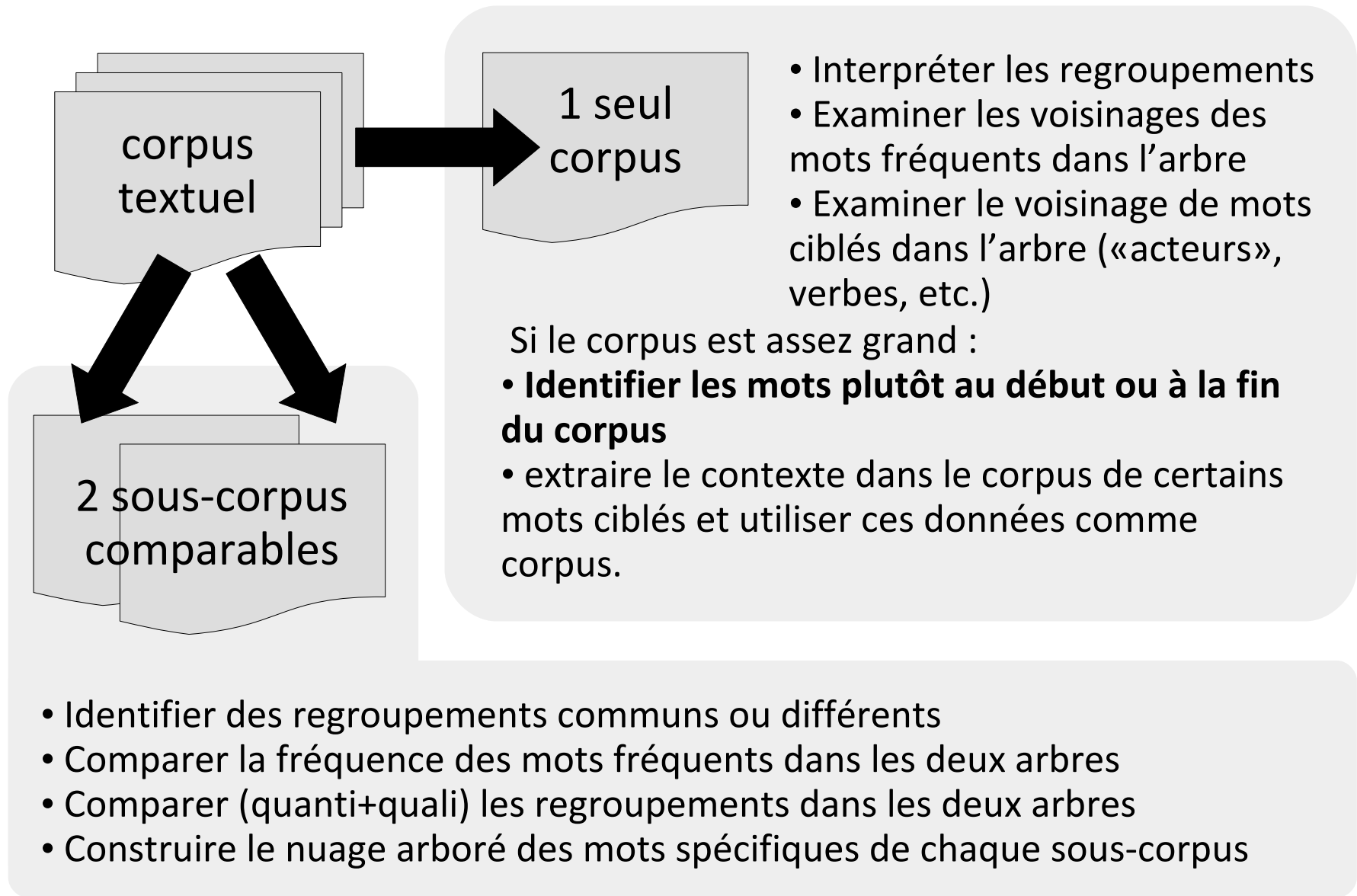
Méthode : voisinage des verbes

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

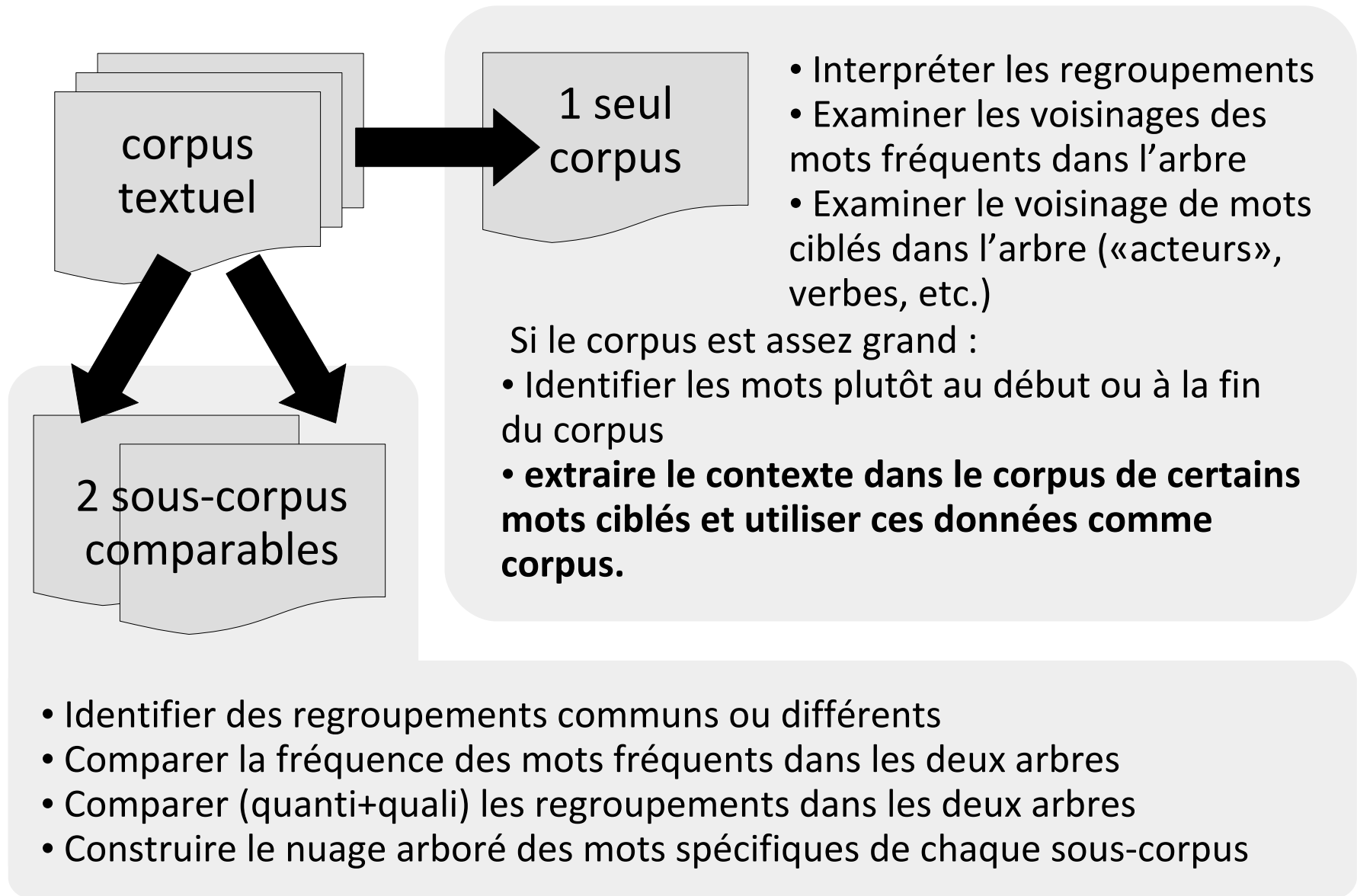
Suggestions d'améliorations :



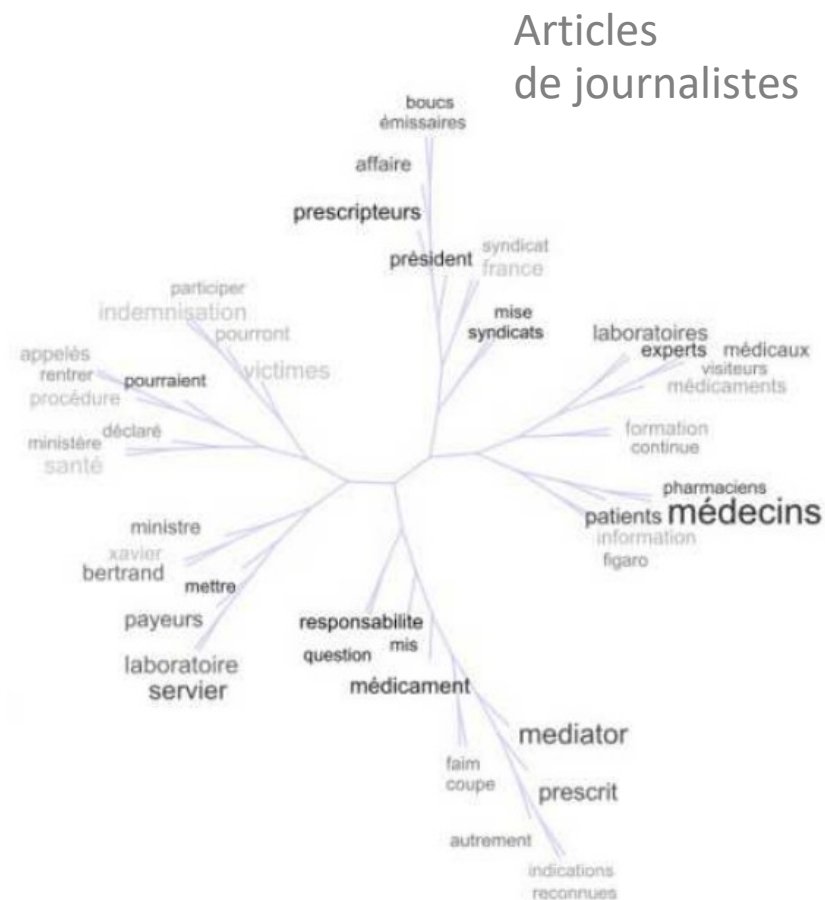
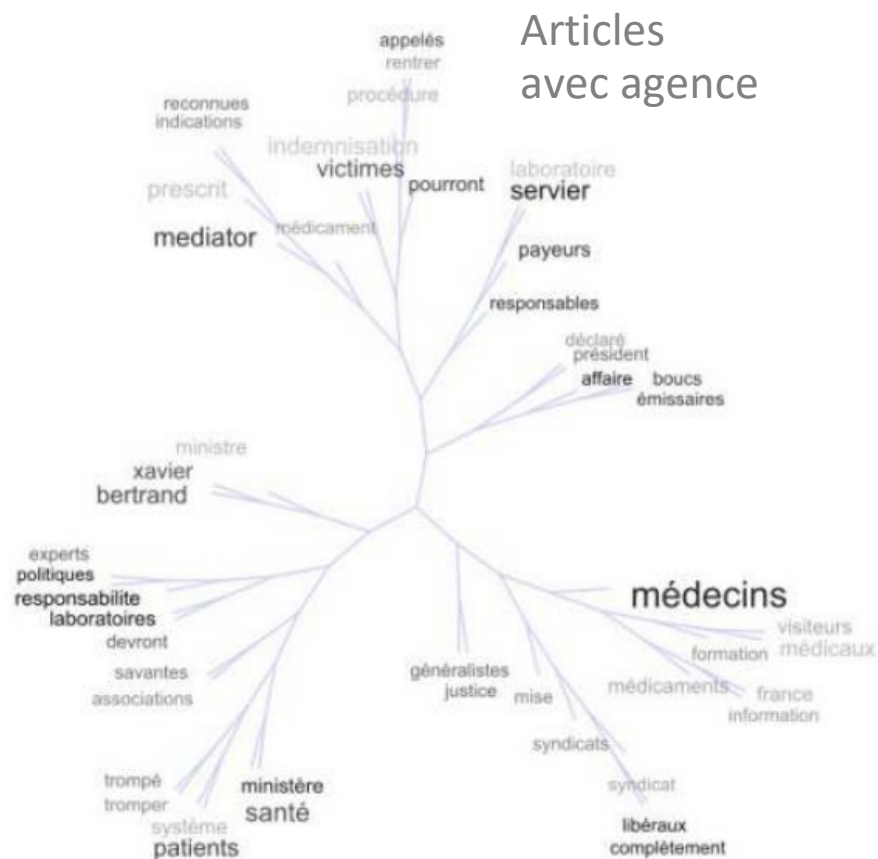
Exploration de corpus avec TreeCloud



Exploration de corpus avec TreeCloud

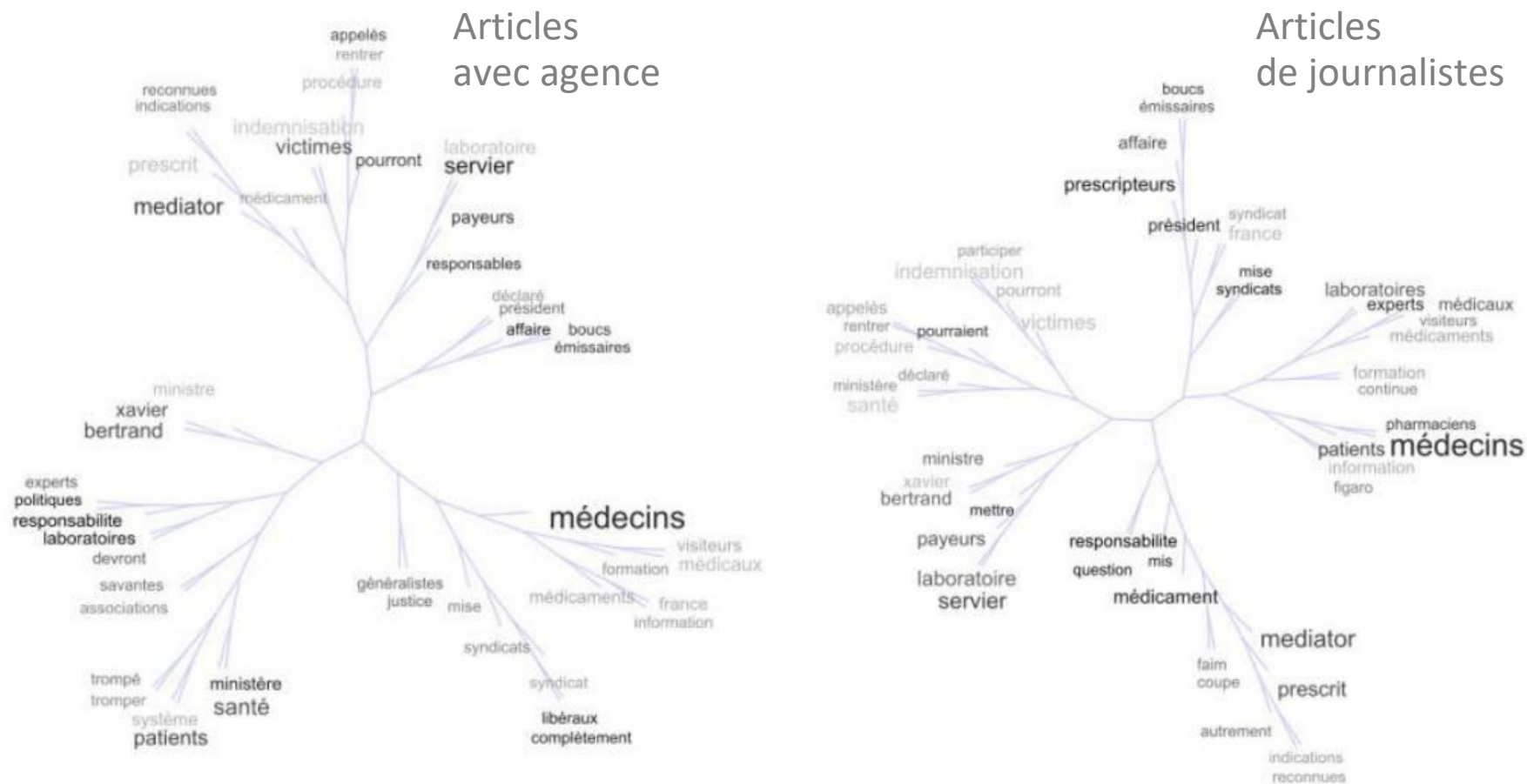


Nuages arborés des contextes de « médecins »



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 cooccurents), construit par TreeCloud avec la formule Liddell, et des fenêtres glissantes de 20 mots

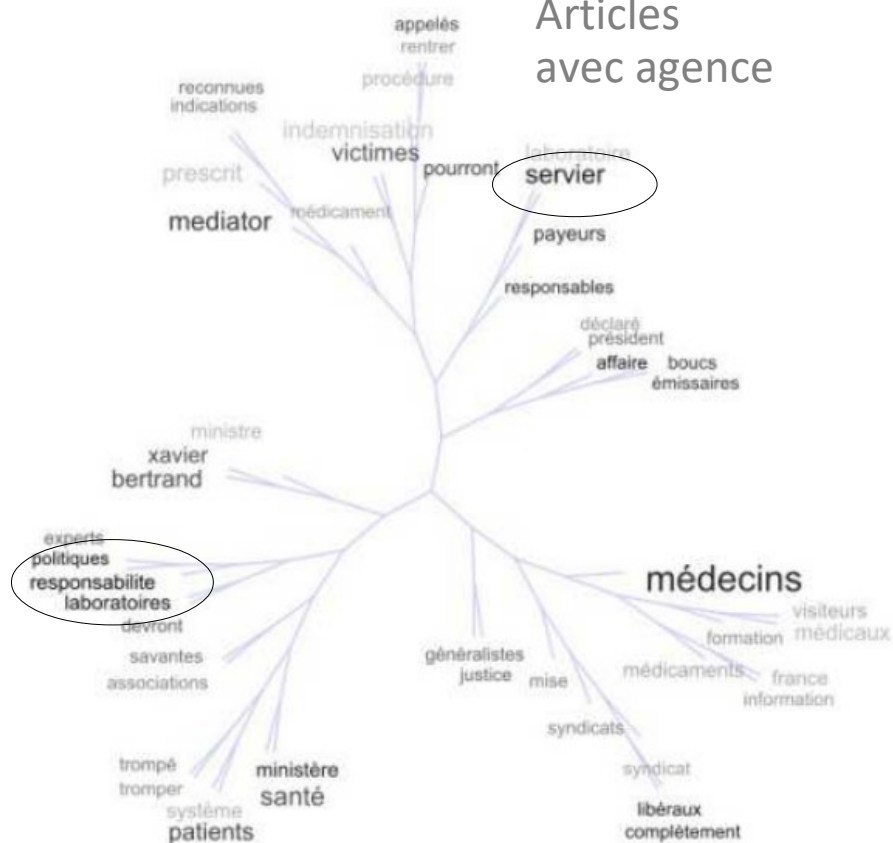
Nuages arborés des contextes de « médecins »



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 cooccurents), construit par TreeCloud avec la formule Liddell, et des fenêtres glissantes de 20 mots

Nuages arborés des contextes de « médecins »

Articles avec agence



Articles de journalistes



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 cooccurents), construit par TreeCloud avec la formule Liddell, et des fenêtres glissantes de 20 mots

Exploration de corpus avec TreeCloud

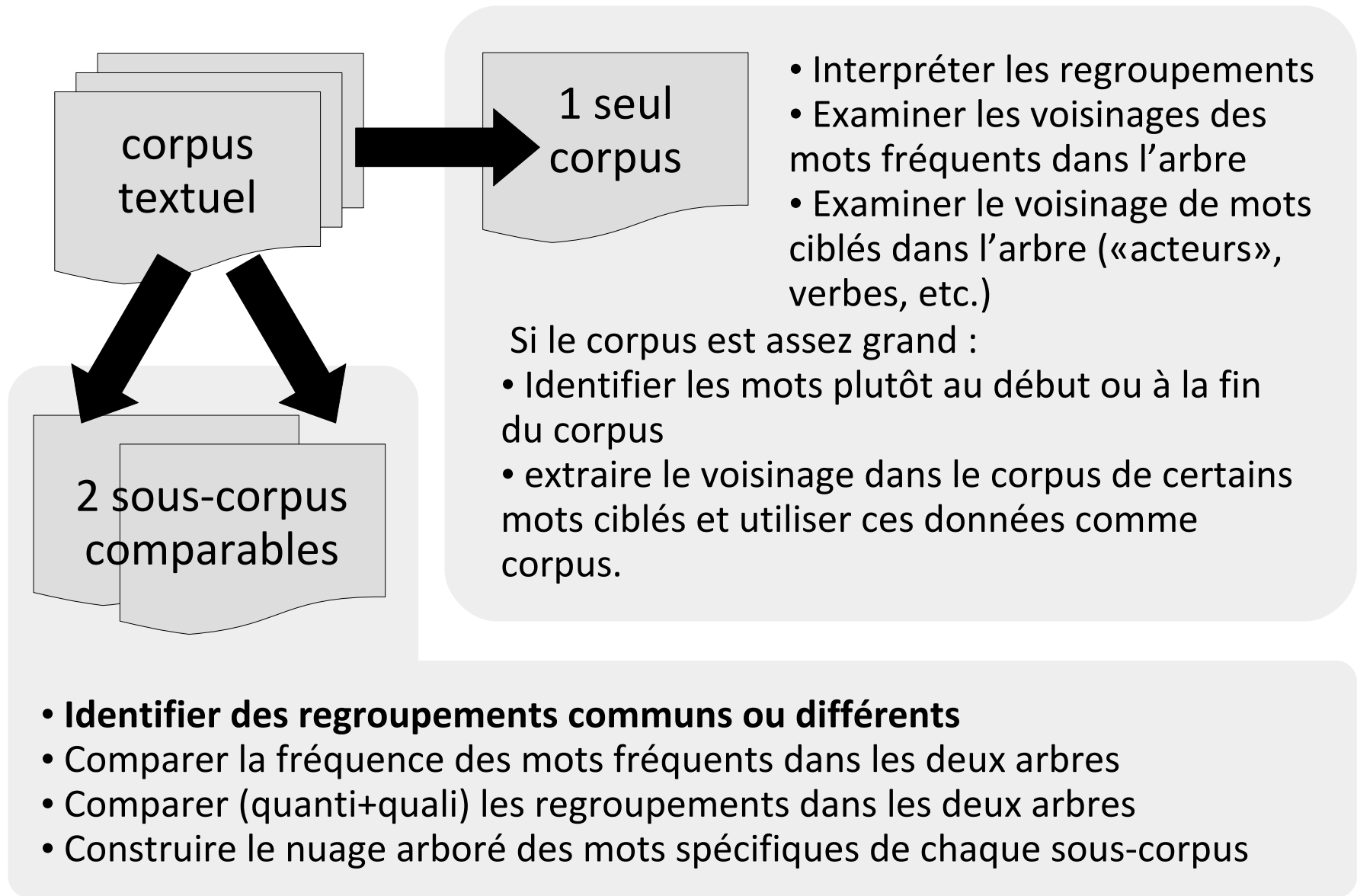


Illustration sur le corpus Mediator

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

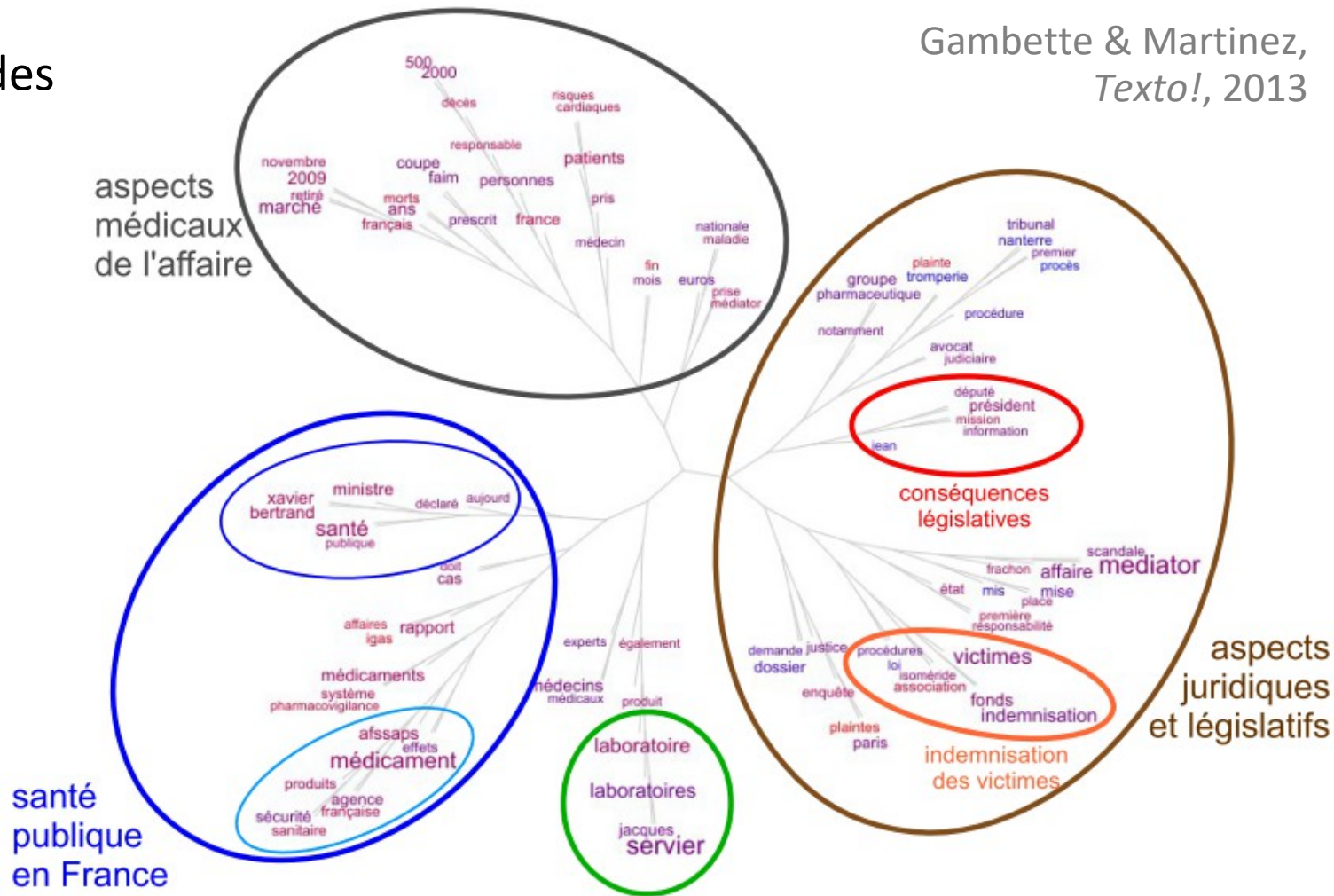
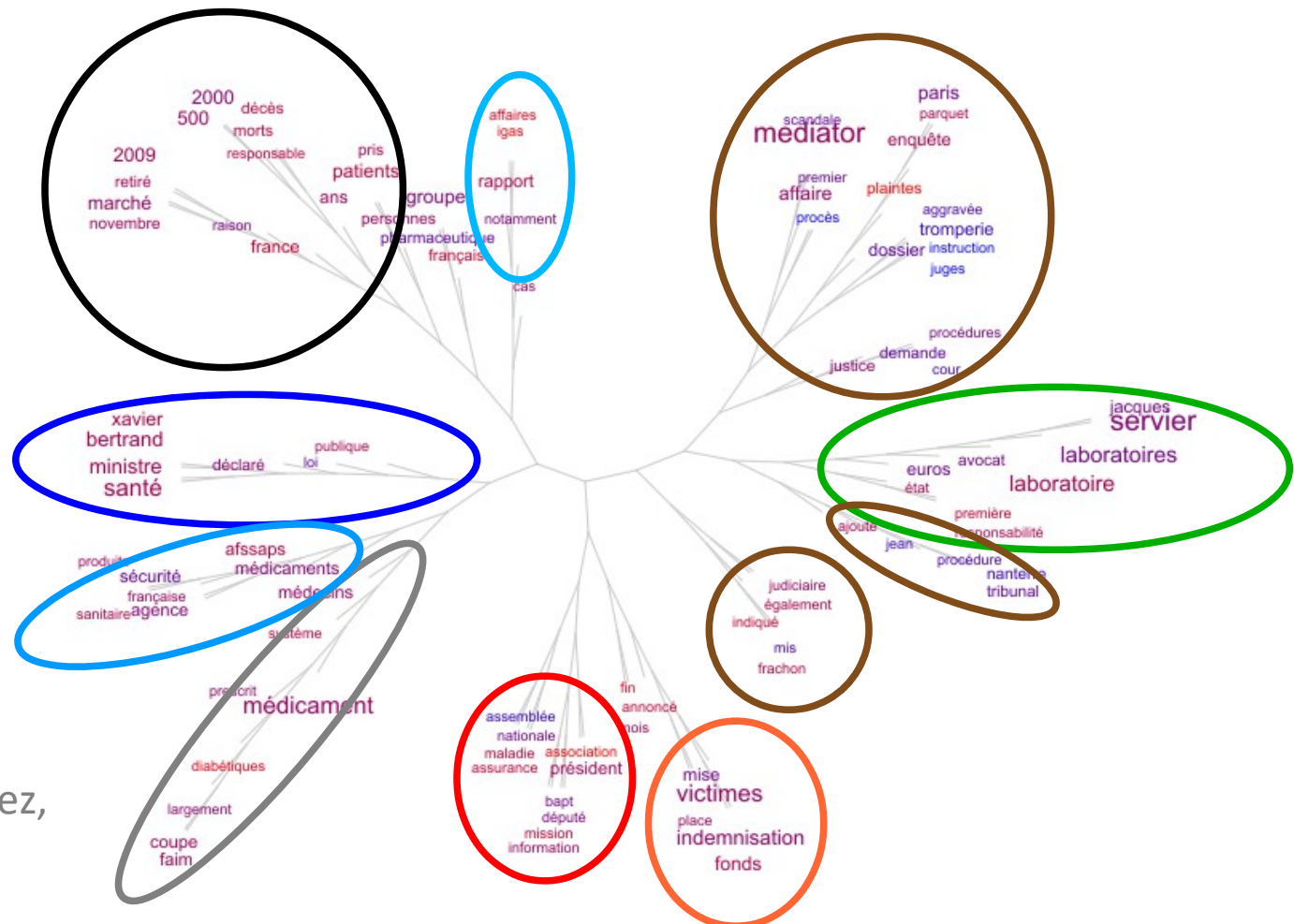


Illustration sur le corpus Mediator

Comparer les articles d'agences et articles de journalistes

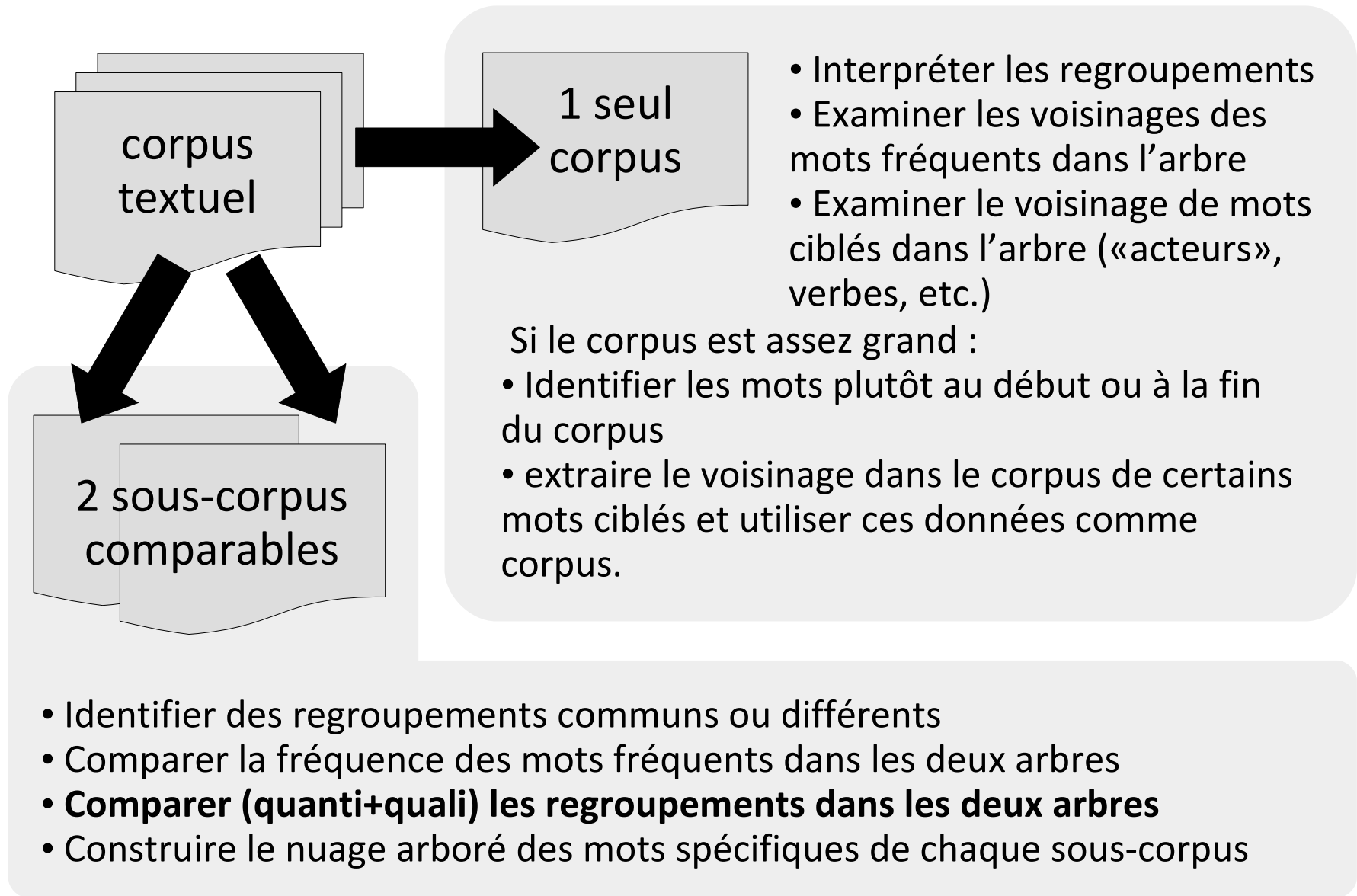
Corpus : 595 articles d'agences contre 1496 articles de journalistes de 2011 évoquant l'affaire du Mediator dans la presse française.

Articles
d'agences

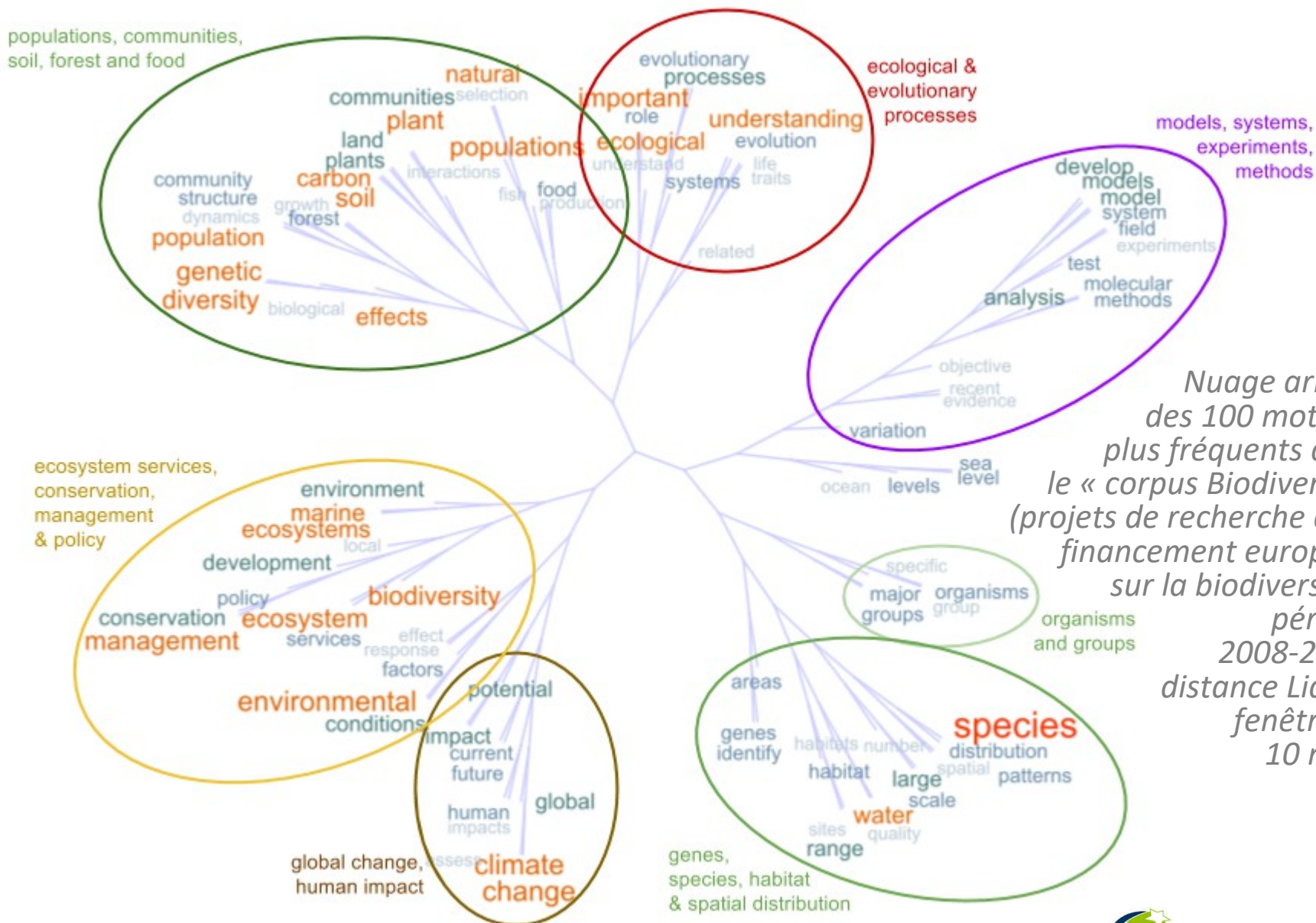


Gambette & Martinez,
Texto!, 2013

Exploration de corpus avec TreeCloud

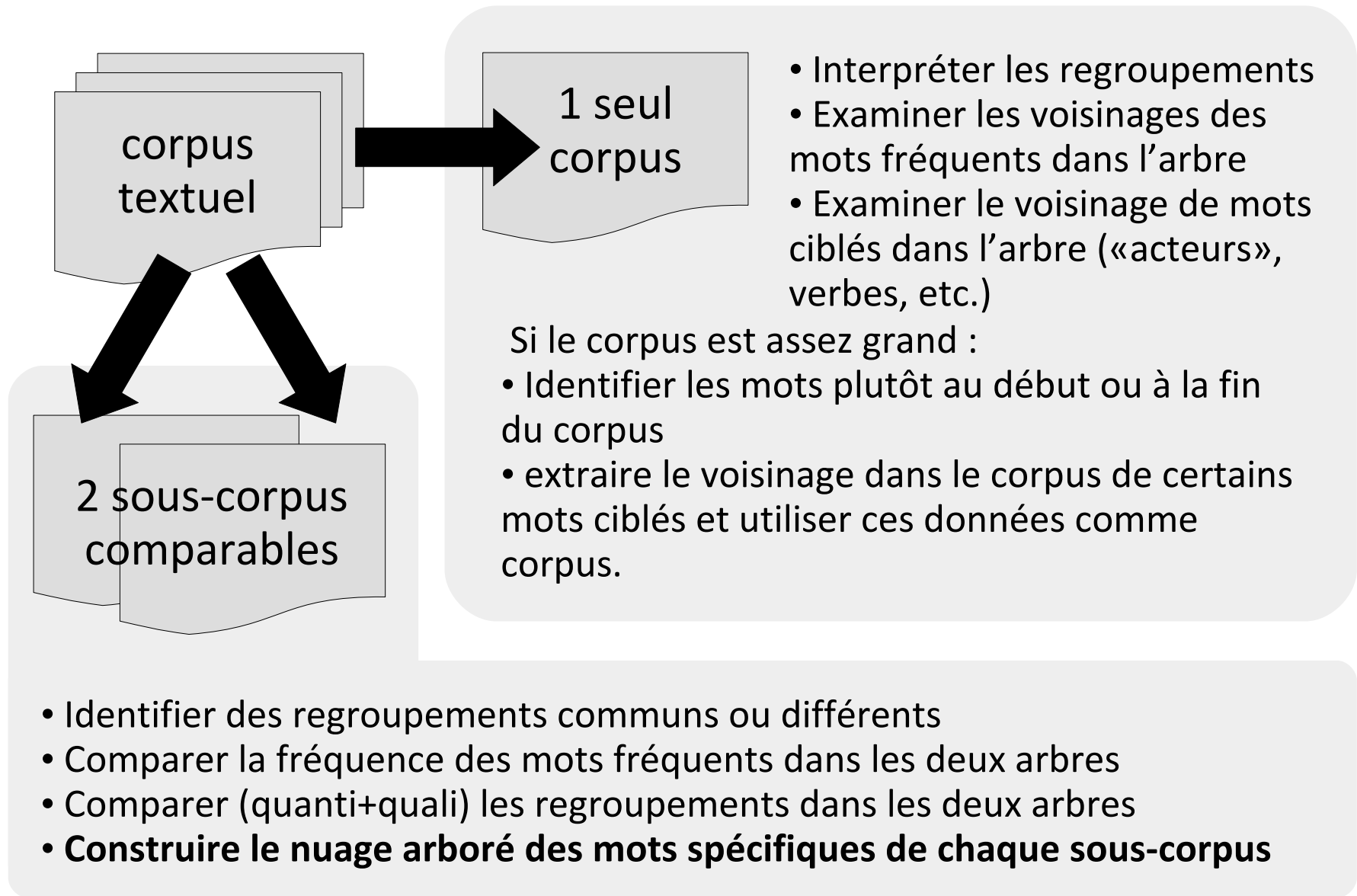


Méthode : comparaison de voisinages dans l'arbre



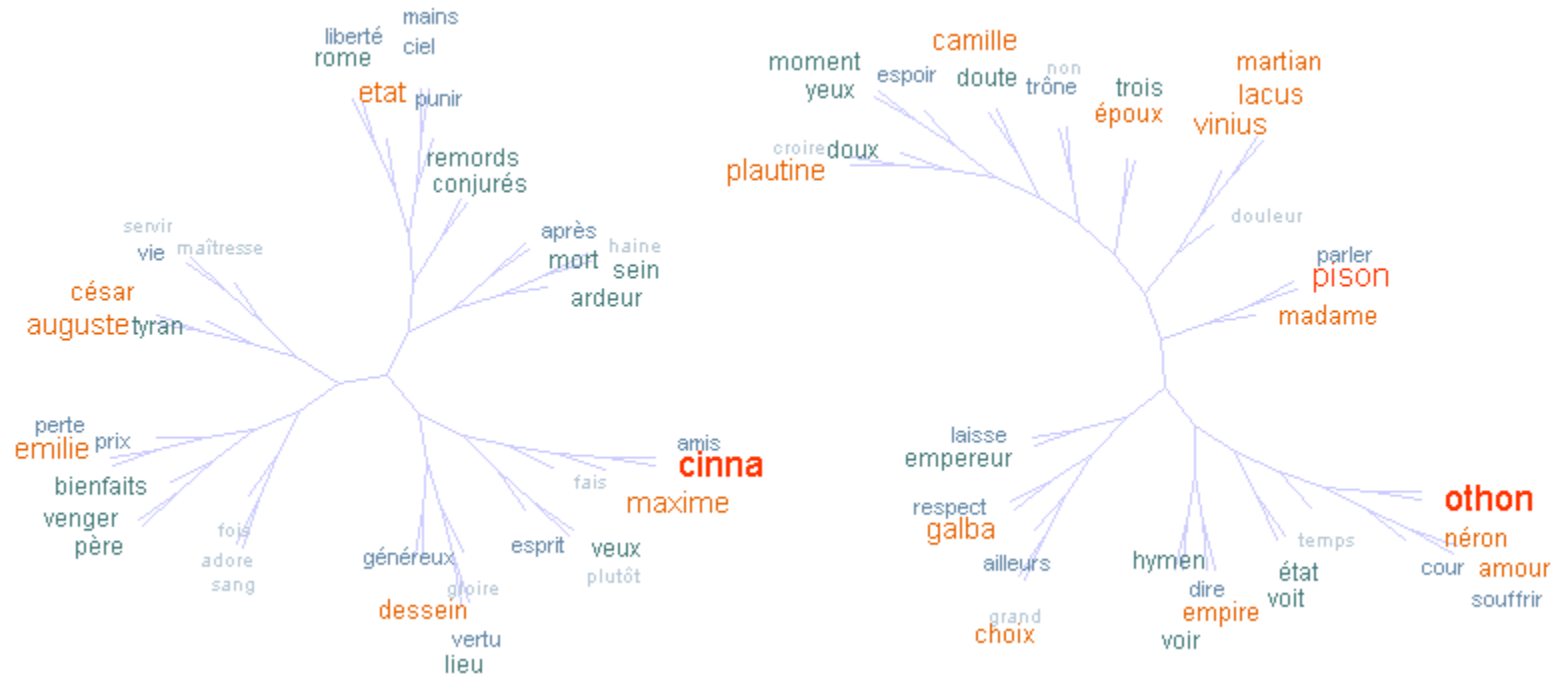
Nuage arboré des 100 mots les plus fréquents dans le « corpus Biodiversa » (projets de recherche avec financement européen sur la biodiversité), période 2008-2011, distance Liddel, fenêtre de 10 mots

Exploration de corpus avec TreeCloud



Méthode : comparaison des spécifiques

Amstutz & Gambette,
JADT 2010



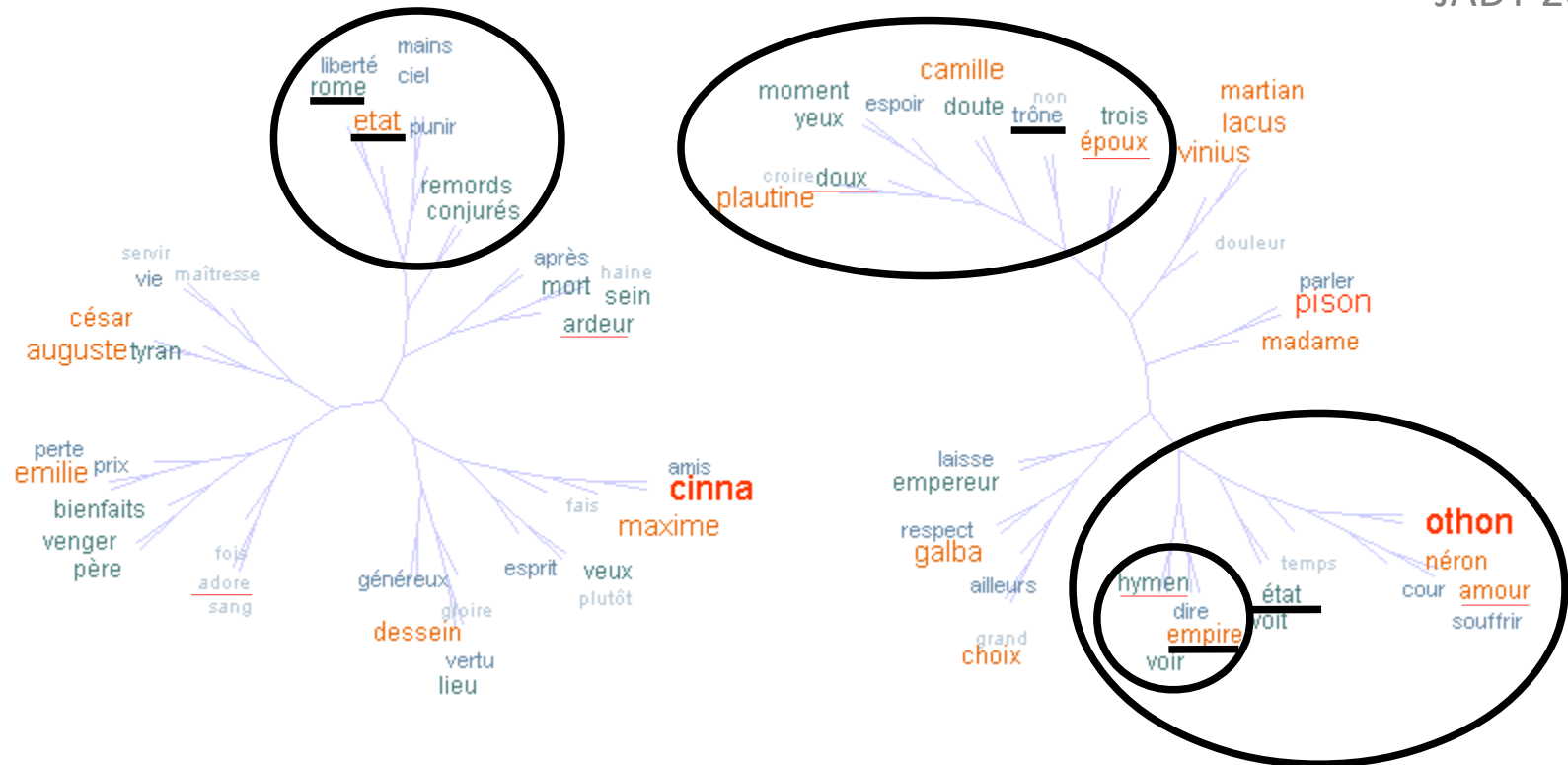
*Nuages arborés des **mots spécifiques** de Cinna et Othon, dimensionnés et colorés d'après leur spécificité calculée dans Lexico3.*



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 ?

Arbres phylogénétiques et arbres de mots

ESPÈCES

Séquences ADN

Données sur
les feuilles

MOTS

Position des mots

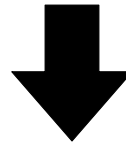
Arbres phylogénétiques et arbres de mots

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

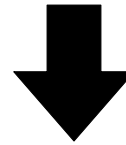
Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

MOTS

Position des mots

Distances fondées sur la cooccurrence entre les deux mots

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



Arbres phylogénétiques et arbres de mots

ESPÈCES

Séquences ADN

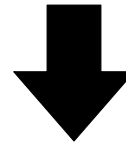
Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

MOTS

Position des mots

Distances fondées sur la cooccurrence entre les deux mots

Données sur les feuilles



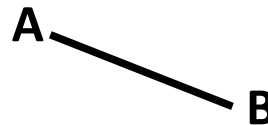
Distances entre les feuilles

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



classification hiérarchique ascendante
algorithme UPGMA

Arbre



Arbres phylogénétiques et arbres de mots

ESPÈCES

Séquences ADN

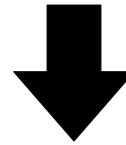
Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

MOTS

Position des mots

Distances fondées sur la cooccurrence entre les deux mots

Données sur les feuilles



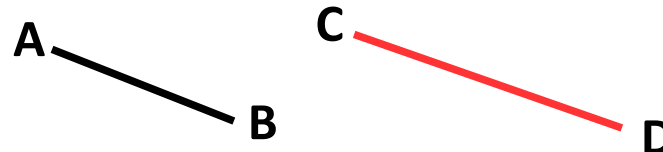
Distances entre les feuilles

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



classification hiérarchique ascendante
algorithme UPGMA

Arbre



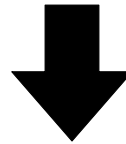
Arbres phylogénétiques et arbres de mots

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

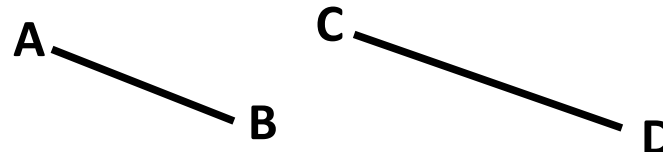
Position des mots

Distances fondées sur la cooccurrence entre les deux mots



classification hiérarchique ascendante
algorithme UPGMA

Arbre



Arbres phylogénétiques et arbres de mots

ESPÈCES

Séquences ADN

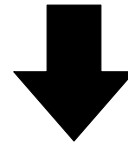
Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

MOTS

Position des mots

Distances fondées sur la cooccurrence entre les deux mots

Données sur les feuilles



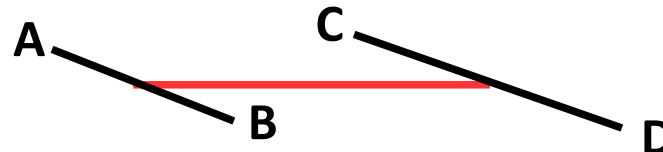
Distances entre les feuilles

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



classification hiérarchique ascendante
algorithme UPGMA

Arbre



Arbres phylogénétiques et arbres de mots

ESPÈCES

Séquences ADN

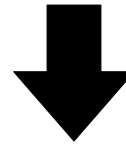
Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

MOTS

Position des mots

Distances fondées sur la cooccurrence entre les deux mots

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



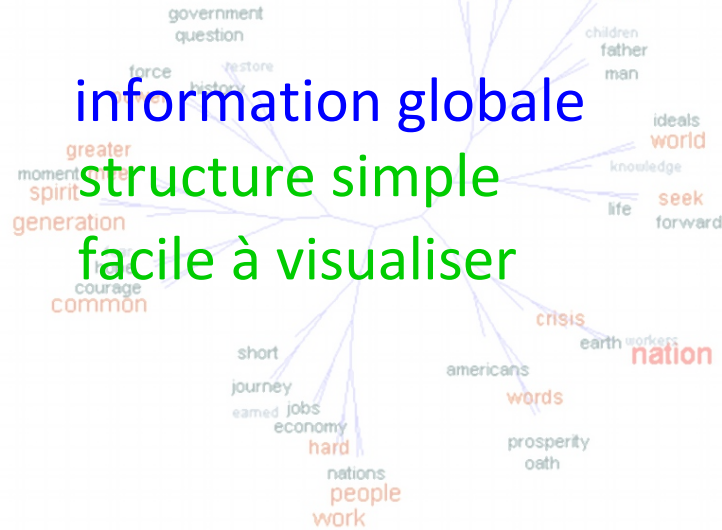
classification hiérarchique ascendante
algorithme UPGMA

Arbre



Comparaison avec d'autres visualisations

nuage arboré (TreeCloud)



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

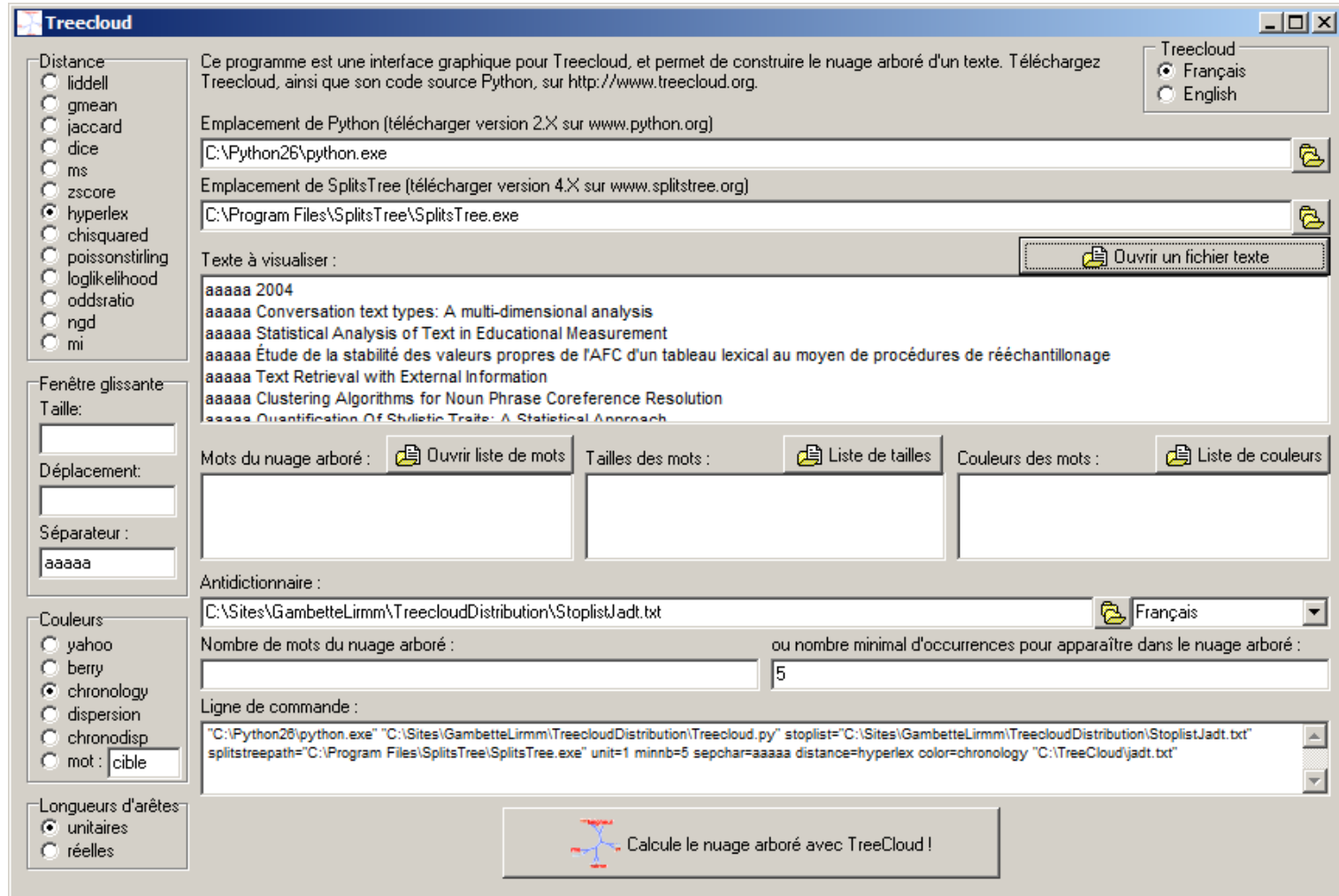


projection des mots (Astartex)



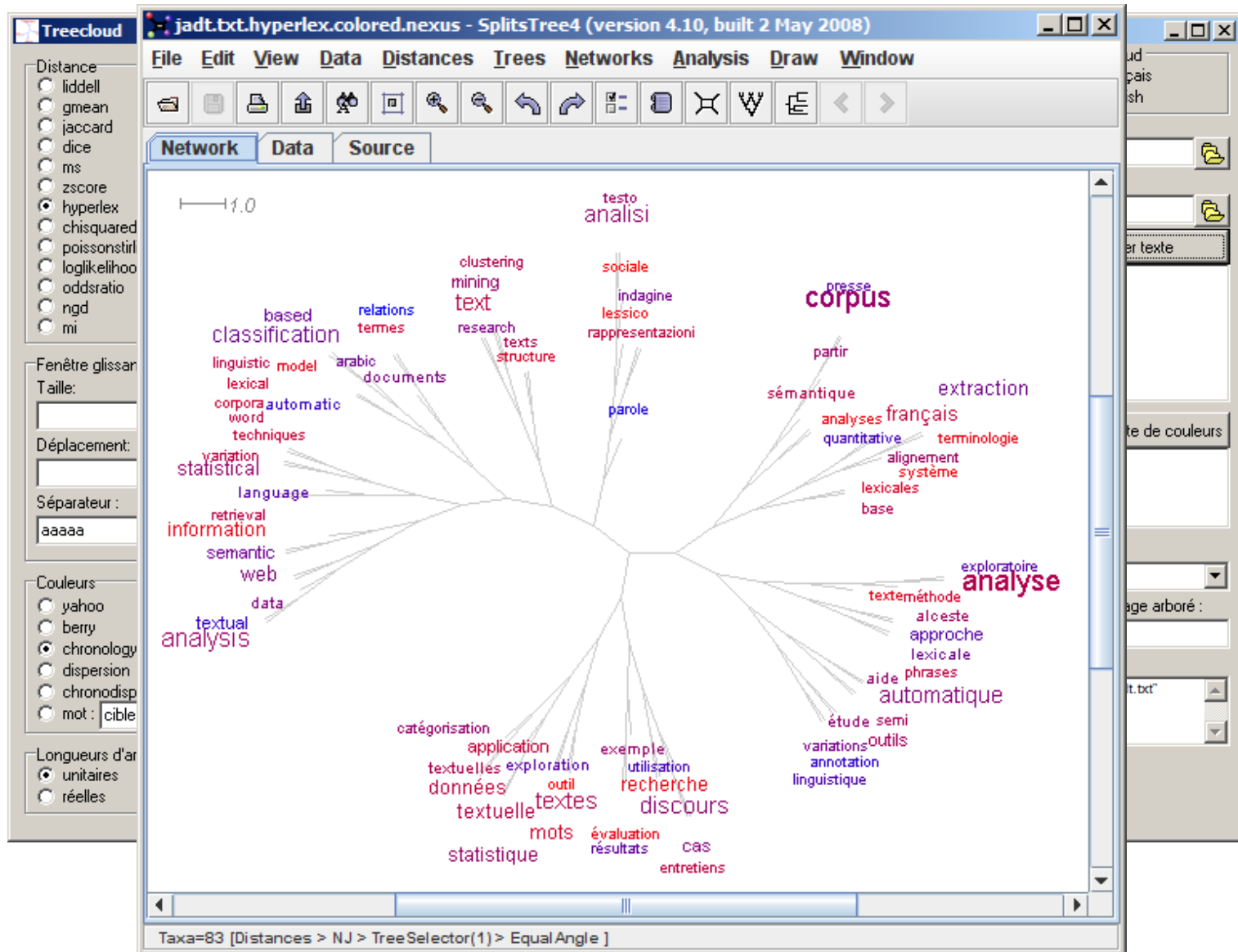
Implémentations

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



Implémentations

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



Interface web



Create! Downloads Gallery Credits FAQ
Créer! Téléchargements Galerie A propos FAQ

www.treecloud.org

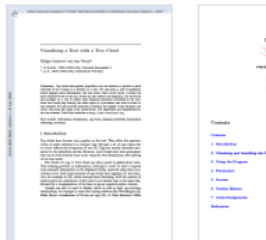
This website helps you to generate tree cloud words are arranged on a tree which reflects The first tree cloud appeared on [Jean Véronis](#) create your own with this website, or with t

Create your own tree cloud online

Ce site web vous permet de générer 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. of Societies\)](#), to appear, 2010 ([supplementary r](#)

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



Créer! Téléchargements Galerie A propos FAQ

Créez vos propres nuages arborés !

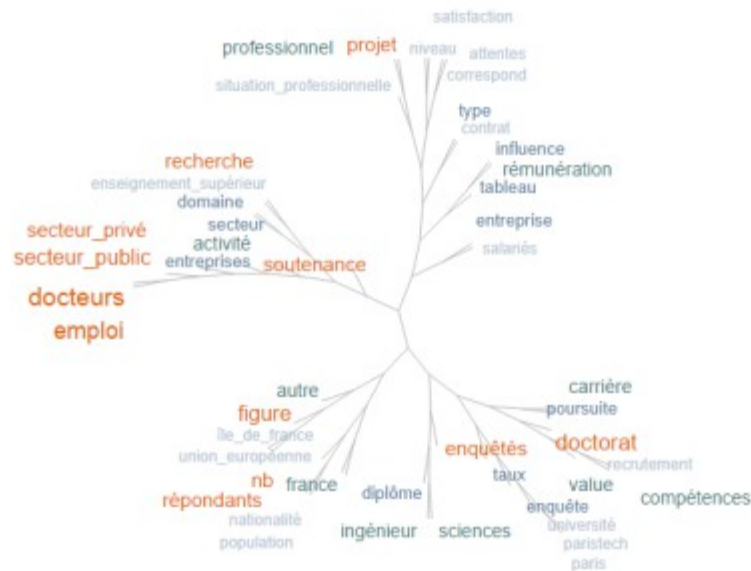
Collez votre texte dans le cadre ci-dessous, puis cliquez sur *Envoyer* ! Attention, l'utilisateur suivant verra votre texte quand il se connectera au site, si vous ne voulez pas faire apparaître vos textes, installez plutôt [TreeCloud](#) sur votre machine.

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



Create! Downloads Gallery Credits FAQ
Créer! Téléchargements Galerie A propos FAQ

This website helps you to generate tree clouds from a text, that is word clouds where the

words are arranged on a tree which reflects their semantic

The first tree cloud appeared on [Jean Véronis's blog](#)
[create your own with this website](#), or [with the TreeCloud](#)

Create your own tree cloud online!

Ce site web vous permet de générer des nuages arborescents
des nuages de mots disposés autour d'un arbre qui indique
Le premier nuage arboré est apparu sur le [blog de Jean Véronis](#)
vous pouvez maintenant [créer les vôtres avec ce site web](#), [ou avec ce logiciel](#)

Créez vos propres nuages arborés en ligne

Documents :

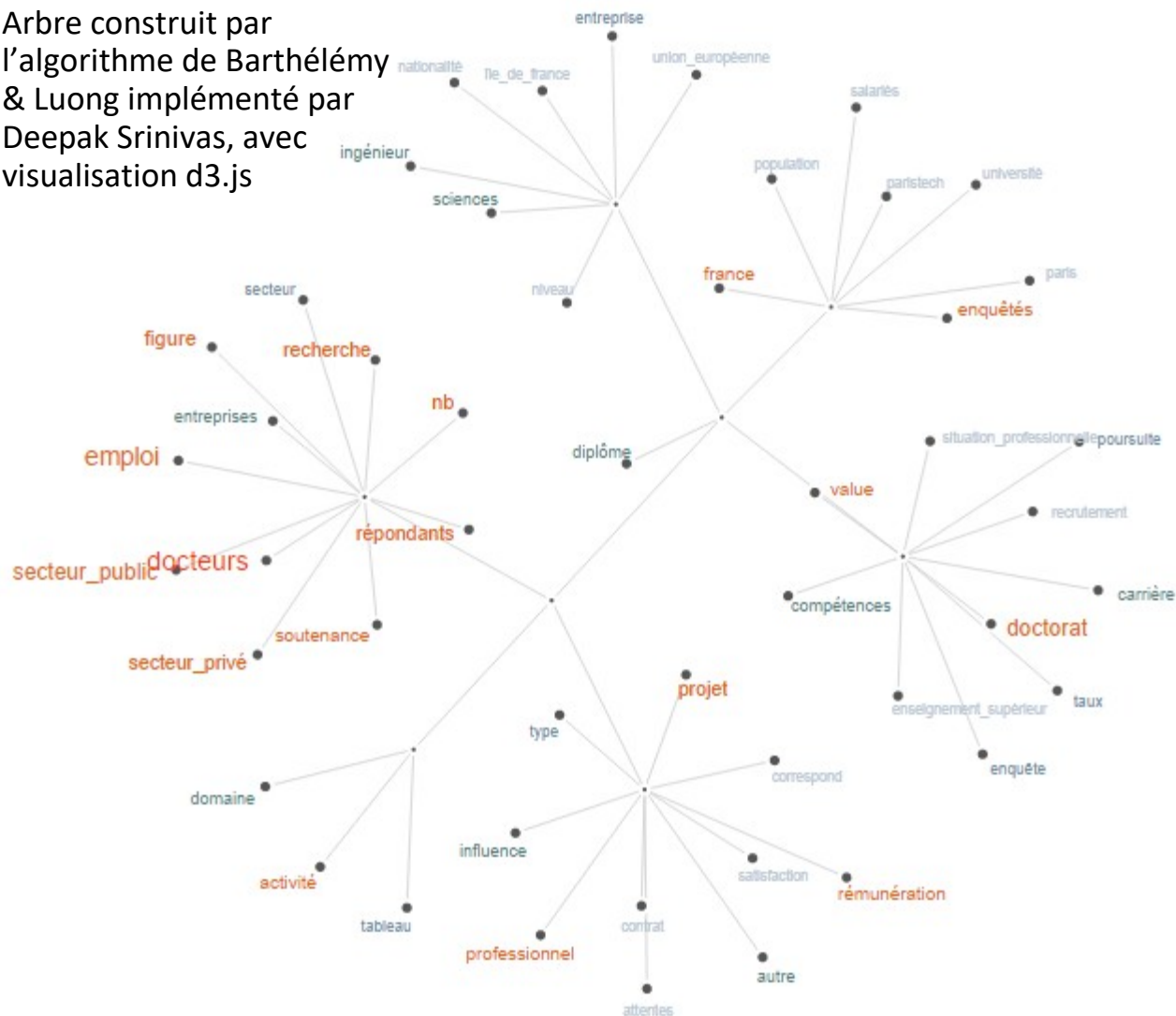


If you use TreeCloud or this website, please cite [www.treecloud.org](#)
Philippe Gambette et Jean Véronis: *Visualising a Text Classification as a Tool of Research*, Proc. of *IFCS'09* (10th International Conference on Intelligent and Flexible Societies), to appear, 2010 ([supplementary material](#)).

Pour des exemples d'utilisation de la visualisation en nuage arboré, voir
Delphine Amstutz et Philippe Gambette: *Utilisation de JADT'10* (10th International Conference on statistical data analysis), to appear, 2010 ([supplémentaire](#)).

www.treecloud.org

Arbre construit par l'algorithme de Barthélémy & Luong implémenté par Deepak Srinivas, avec visualisation d3.js



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 dans le séminaire doctoral de Jean-Marc Leblanc <http://textopol.u-pec.fr/?cat=103>

- 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 Médiateur 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.archives-ouvertes.fr/hal-01702091v1>

Philippe Gambette, Tita Kyriacopoulou, Nadège Lechevrel, Claude Martineau (2018).

Anatomie, animaux, vocabulaire de la vivisection : Construire des ressources lexicales pour visualiser une thématique dans un corpus littéraire. Gisèle Séginger. *Animalhumanité - Expérimentation et fiction : l'animalité au cœur du vivant*, LISAA, pp.223-231.

<https://hal.archives-ouvertes.fr/hal-01609198>

Recueil et prétraitement de corpus

Prétraitements des extractions d'Europresse

→ consulter et copier ou télécharger [ce document tableur partagé](#)

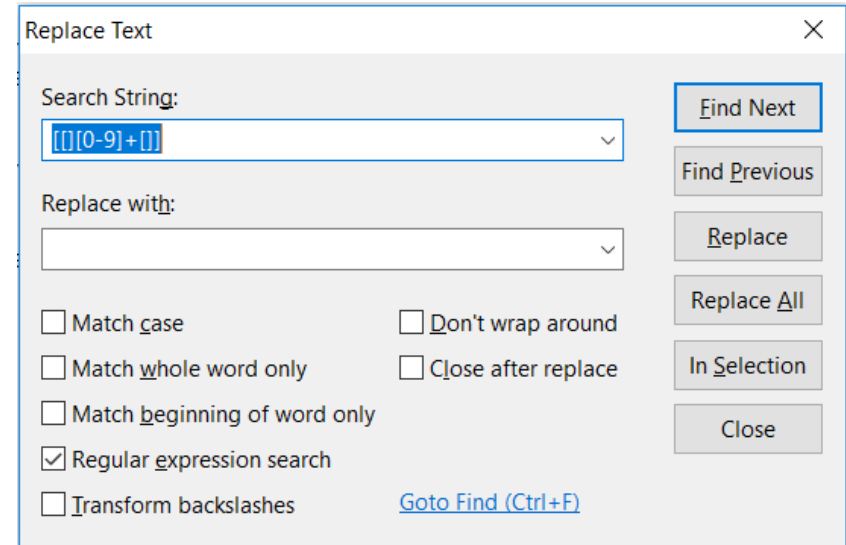
fx | =SI(ET(B542="Article";D542=1);E542;"")

	A	B	C	D	E	F	G	H	I	J
1	texte corrigé	propagation déb	repérage début	lignes à ignorer	texte extrait du PDF					
542	modélisation par	Article			1 modélisation par éléments finis du					
543	rachis (colonne v	Article			1 rachis (colonne vertébrale); Evasym, sur					
544	le vieillissement,	Article			1 le vieillissement, le handicap, l'orthopédie et la traumatologie.					
545		Article			0 Documents sauvegardés par Jeudi 20 février 2020 à 7 h 49					
546		9 Article			1 9					
547		Article			0 Ce document est réservé à l'usage exclusif de l'utilisateur désigné par UNIVERSITE-DE-MARNE-					
548		Article			0 faire l'objet d'aucune autre utilisation ou diffusion auprès de tiers. • Tous droits réservés • Service					
549			Fin		1 Nom de la source					
550					1 Décideurs Magazine (Site web)					
551					1 Type de source					
552					1 Presse • Presse Web					
553					1 Périodicité					
554					1 En continu					
555					1 Couverture géographique					
556					1 Nationale					
557					1 Provenance					
558					1 Paris, Ile-de-France, France					
559					1 mercredi 29 janv					
560		Article	Début		0 Décideurs Magazine (Site web) • 2888 mots					
561	La veille urbaine	Article			1 La veille urbaine du 29 janvier 2020					
562	L	Article			1 L					
563	es 50 premières	Article			1 es 50 premières villes sélectionnées pour accueillir un service des finances publiques, la					
564	présentation de l	Article			1 présentation de l'ordonnance relative à					
565	la réécriture des	Article			1 la réécriture des règles de construction,					
566	le siège de Chau	Article			1 le siège de Chaumet qui change de propriétaire, Vincent Ego promu DG de					
567	Cogedim... Décic	Article			1 Cogedim... Décideurs vous propose une					
568	synthèse des act	Article			1 synthèse des actualités immobilières et					

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 :

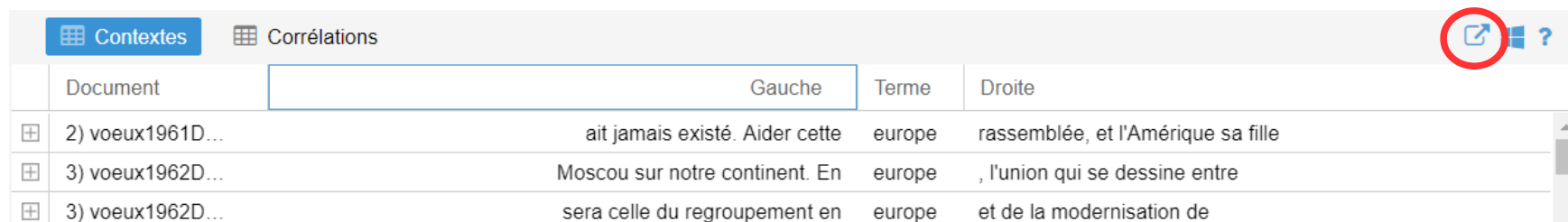
- mot dont la césure a été supprimée car trouvé en entier dans le dictionnaire
- mot pour lequel le trait d'union a été gardé

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

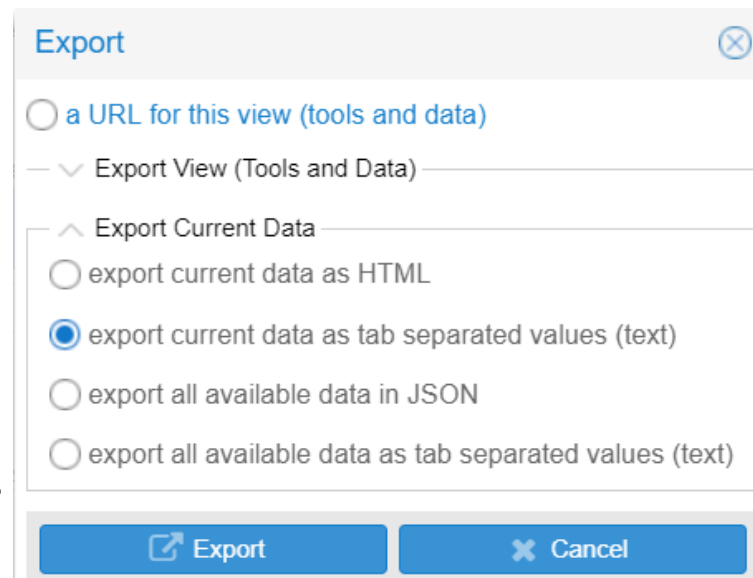
Extraction de contextes avec VoyantTools

- Charger le corpus dans <http://voyant-tools.org> (ex. : [corpus des vœux présidentiels](#), de Jean-Marc Leblanc ; sélection de plusieurs fichiers txt sur le disque dur avec le bouton « Charger »)
- Charger les **contextes** de “religion”, par exemple, dans le cadre en bas à droite, puis exporter avec le bouton entouré en rouge :



Document	Gauche	Terme	Droite
2) vœux1961D...	ait jamais existé. Aider cette	europe	rassemblée, et l'Amérique sa fille
3) vœux1962D...	Moscou sur notre continent. En	europe	, l'union qui se dessine entre
3) vœux1962D...	sera celle du regroupement en	europe	et de la modernisation de

- 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.
- Visualiser les fréquences d'une liste de mots
VisuLexique



Export

a URL for this view (tools and data)

— Export View (Tools and Data) —

^ Export Current Data

export current data as HTML

export current data as tab separated values (text)

export all available data in JSON

export all available data as tab separated values (text)