


Master Intelligence stratégique, analyse des risques et territoires
07/11/2016 – IFIS (Serris)

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

Philippe Gambette

LIGM
Université Paris-Est
Marne-la-Vallée



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

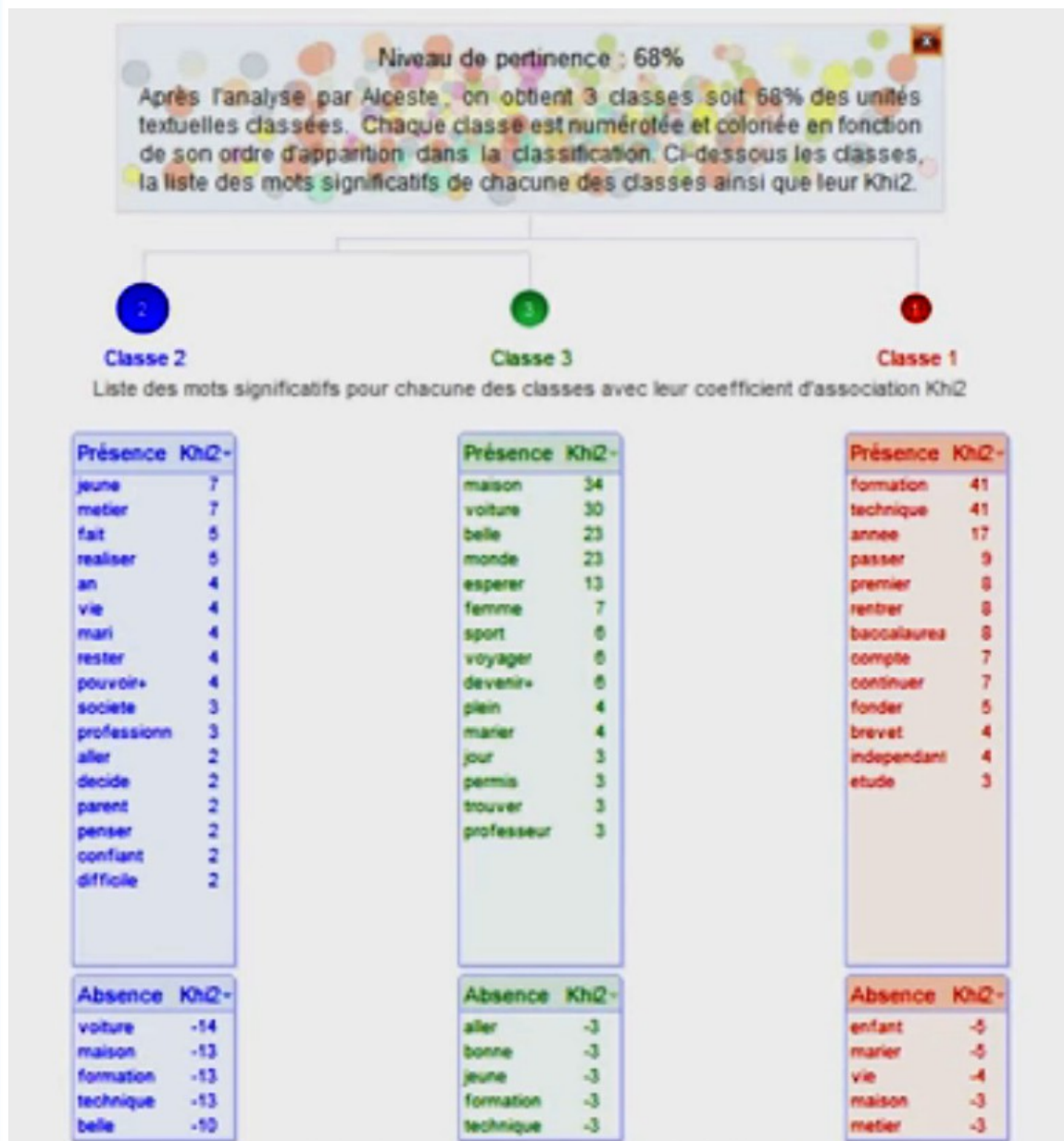
Alceste (depuis 1983)



Société IMAGE

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

La méthode Alceste



Répartition des phrases du texte en différentes classes

→ vocabulaire de chaque classe

-
- Etape 1**
 - Lecture du corpus et extraction des formes
 - Catégorisation et lemmatisation des formes
 - Calcul des dictionnaires des formes réduites
 - Etape 2**
 - Définition des unités textuelles du corpus
 - Construction des tableaux de données
 - Classification Descendante Hiérarchique
 - Etape 3**
 - Définition et sélection des classes à retenir
 - Présences et absences des formes
 - Analyse Factorielle des Correspondances
 - Etape 4**
 - Sélection des unités textuelles par classe
 - Segments répétés des classes et du corpus
 - Classification Ascendante Hiérarchique
 - Etape 5**
 - Réseaux de proximité de formes
 - Cartographies du corpus en unités textuelles
 - Courbes d'accroissement du vocabulaire
 - Classement des individus et des variables
 - Création des rapports détaillé et de synthèse

Quelques logiciels de textométrie

Hyperbase (depuis 1989)



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

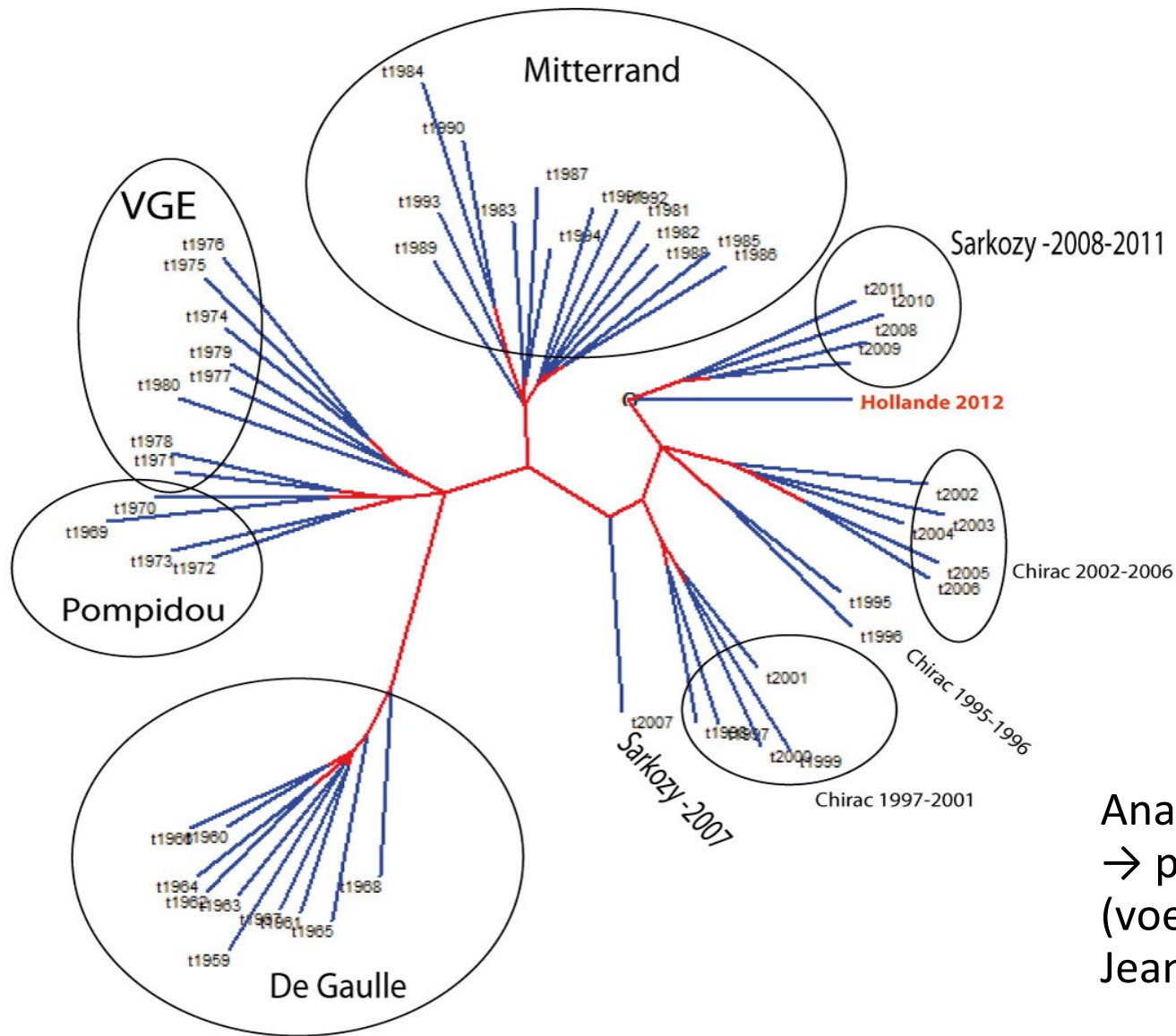
Alceste (depuis 1983)



Société IMAGE

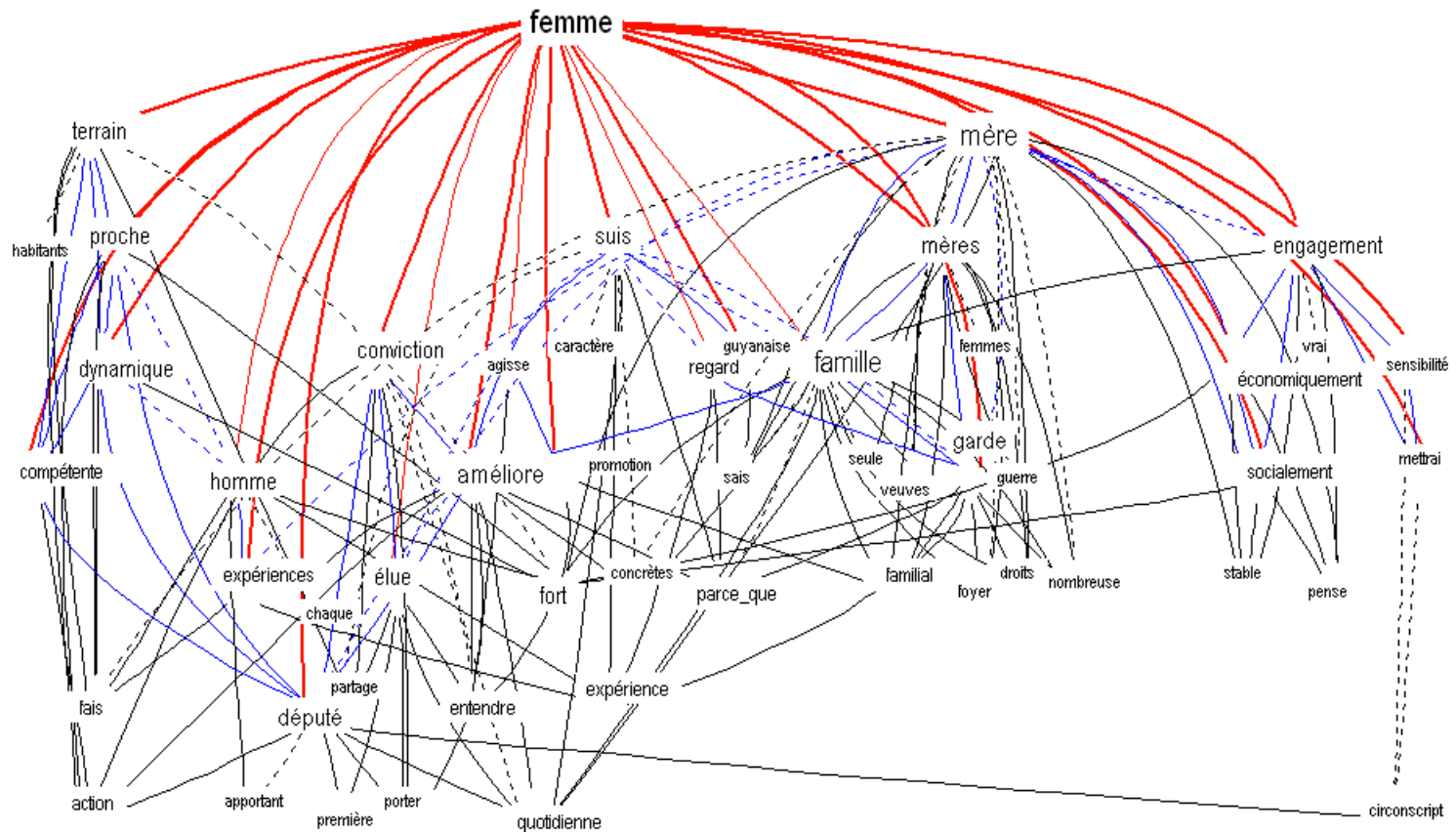
<http://www.image-zafar.com/Logiciel.html>

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 three panes:

- Left Pane:** A table listing segments with their length (Lg) and frequency (Frq). The segment "la grande colère" is highlighted.
- Top Center Pane:** A grid of 60 columns and 60 rows of small squares, representing a section card. The grid is mostly empty, with some squares filled with grey.
- Bottom Center Pane:** A text area containing a paragraph of text, likely a definition or description of the selected segment.

Table of Segments:

| 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 Card Grid:

The grid shows a 60x60 layout of small squares. The rightmost column of squares is numbered from 50 to 600. The grid is mostly empty, with some squares filled with grey, indicating the presence of sections.

Section Card Text:

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>\$

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/>

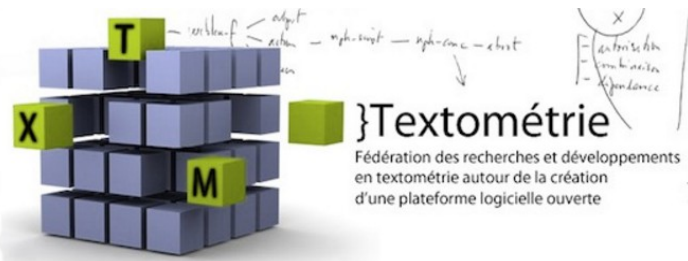
Alceste (depuis 1983)



Société IMAGE

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

TXM (depuis 2008)

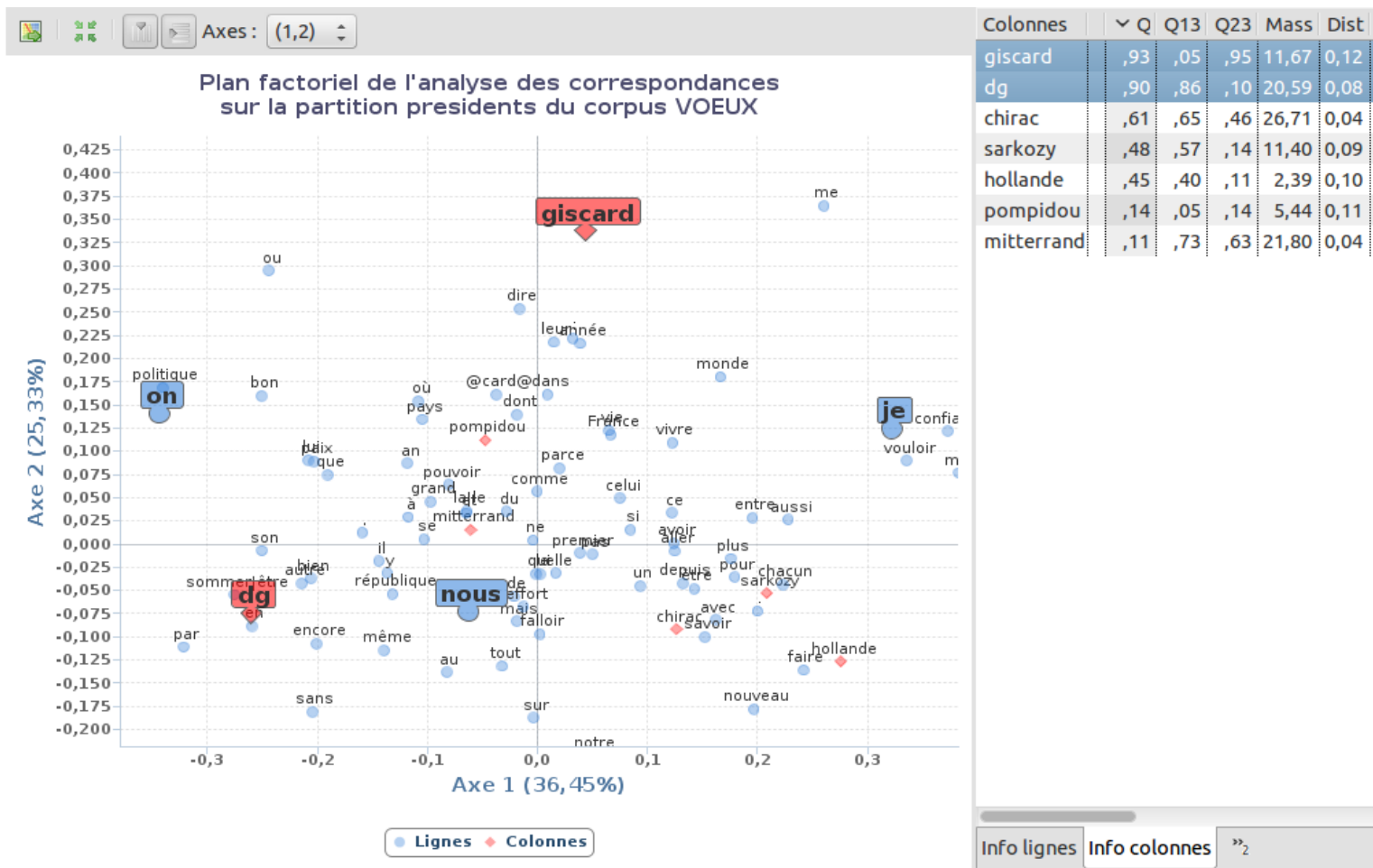


}Textométrie

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

ENS Lyon

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



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)



ENS Lyon

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

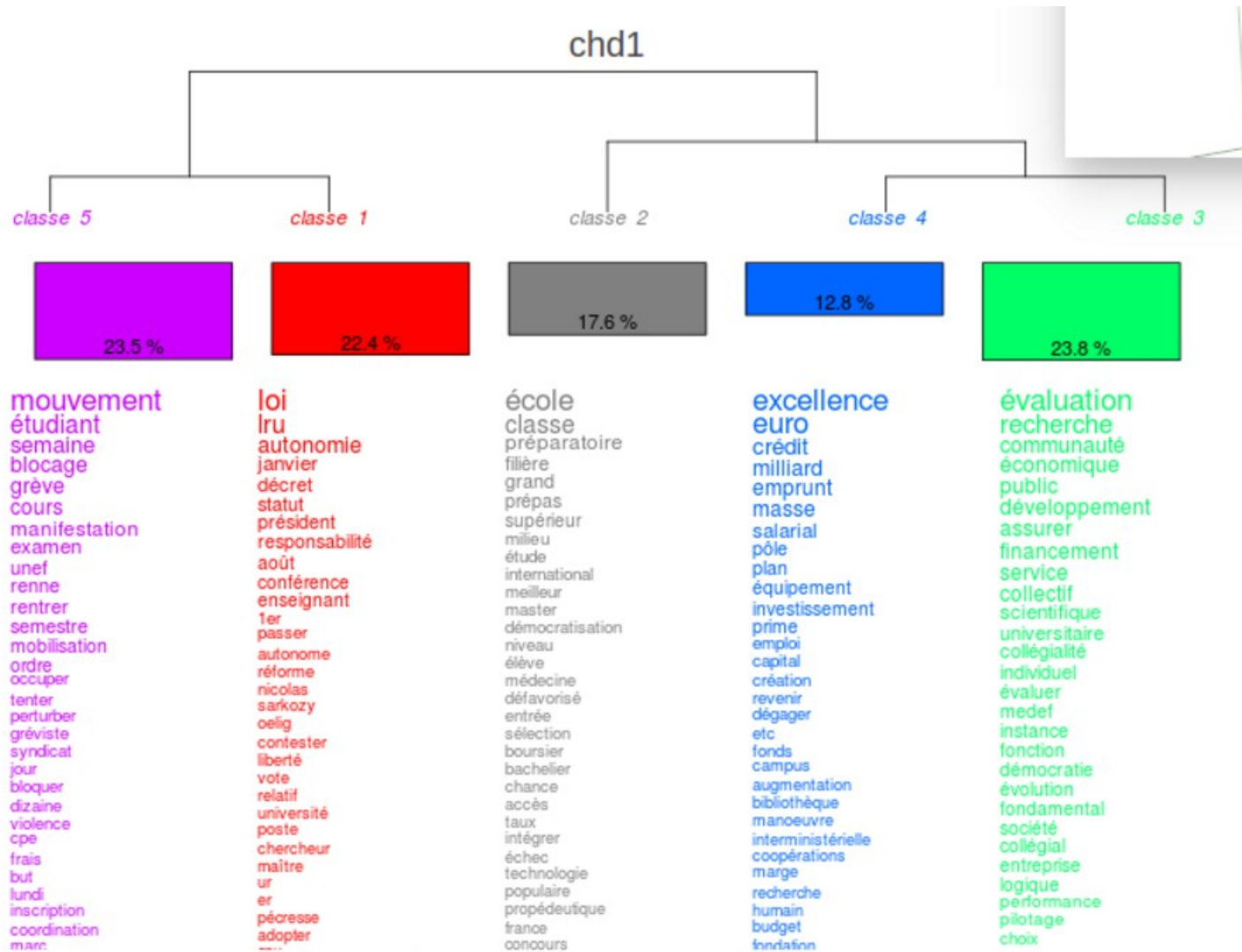
Iramuteq (depuis 2009)



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%

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)

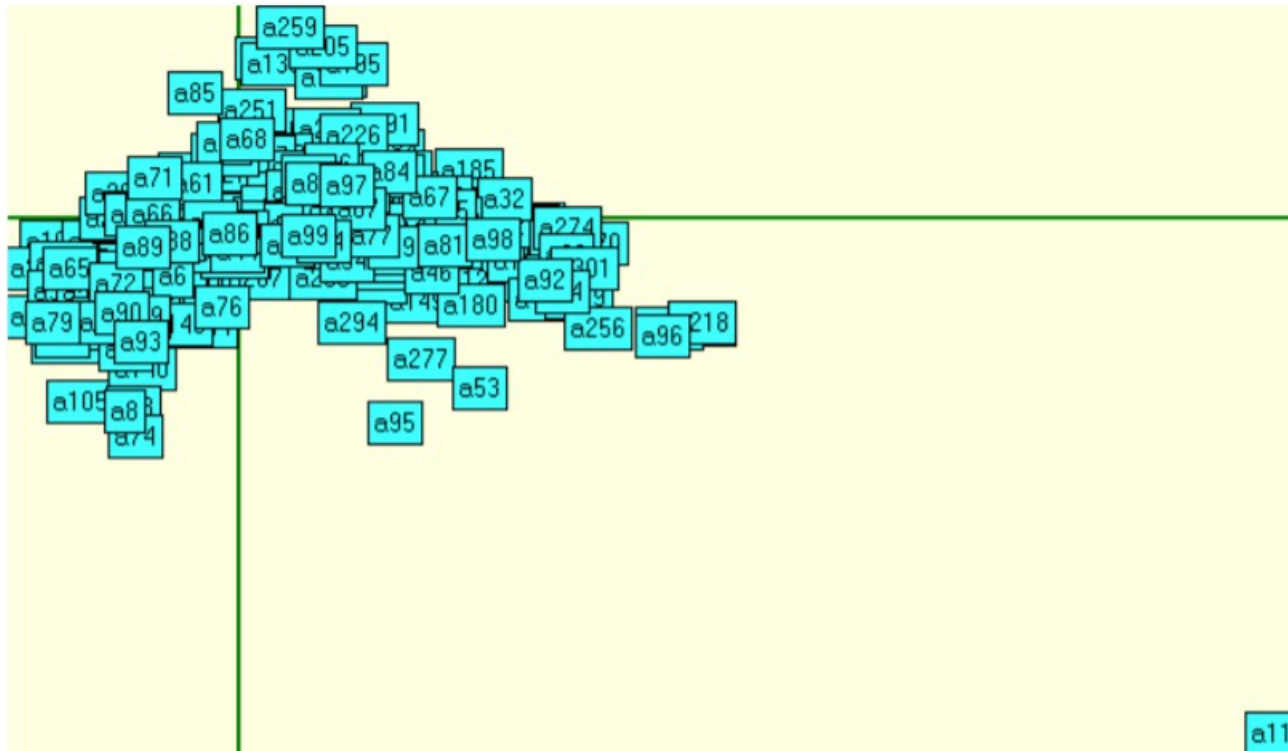
<annee=2015> <type=article> <doc=a1> background: lateral, or horizontal, gene transfers are a type of asymmetric evolutionary events where genetic material is transferred from one species to another. in this paper we consider lgt networks, a general model of phylogenetic networks with lateral gene transfers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

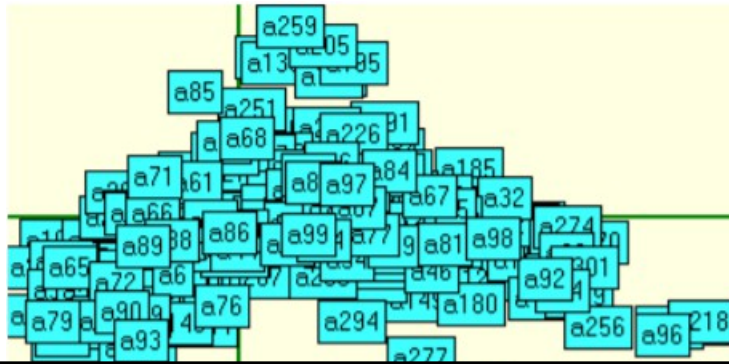
Exemple d'exploration : articles scientifiques

Etape 2) Analyse factorielle des correspondances



Exemple d'exploration : articles scientifiques

Etape 2) Analyse factorielle des correspondances

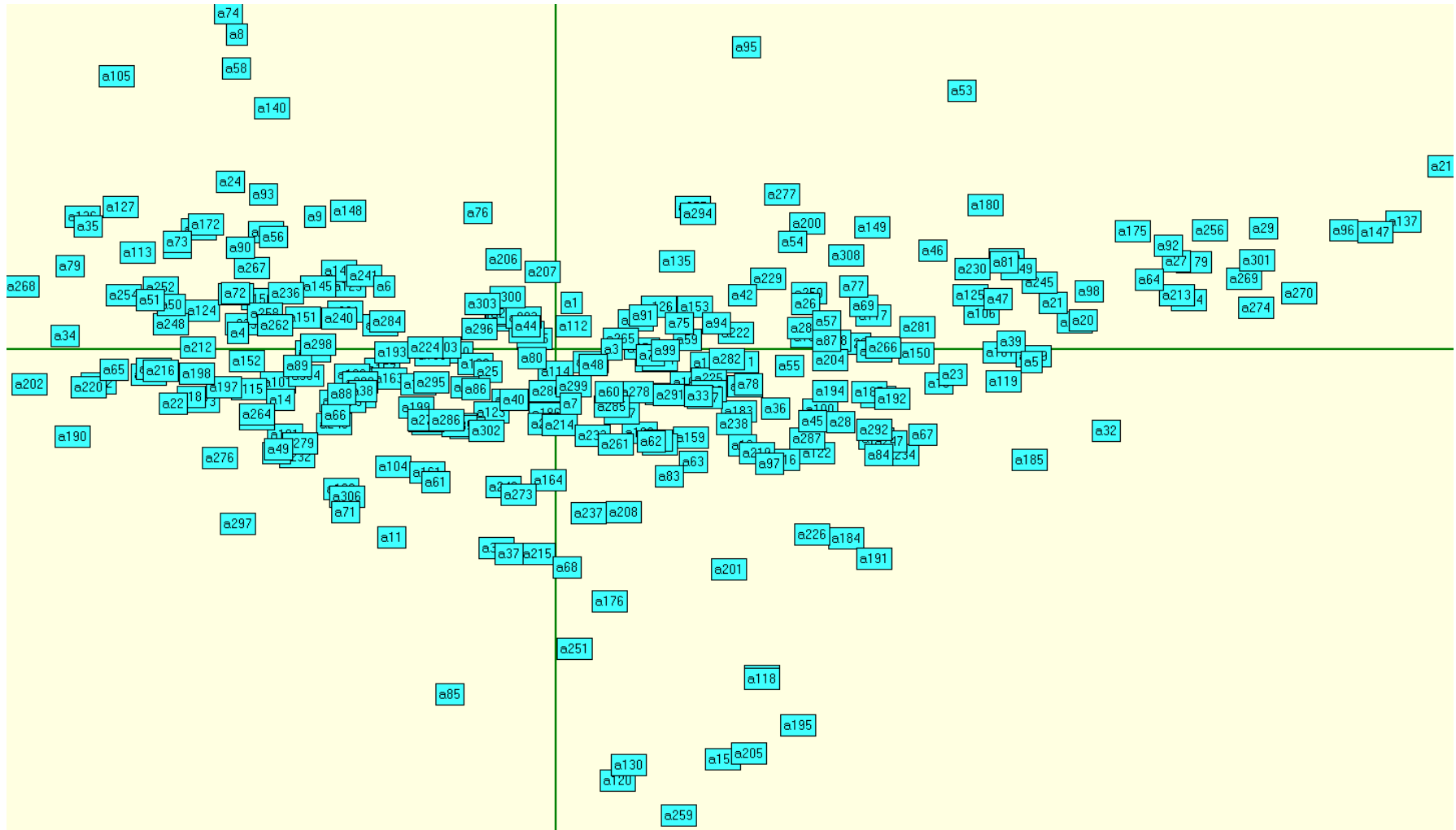


we study the asymptotic behavior of a new type of maximization recurrence, defined as follows. let k be a positive integer and $p_k(x)$ a polynomial of degree k satisfying $p_k(0) = 0$. define $a_0 = 0$ and for $n \geq 1$, let $a_n = \max_{0 \leq i < n} \{a_{i+n} + p_k(i/n)\}$. we prove that $\lim_{n \rightarrow \infty} a_n/n = \sup\{p_k(x)/(1-x^k) : 0 \leq x < 1\}$. we also consider two closely related maximization recurrences s_n and s'_n , defined as $s_0 = s'_0 = 0$, and for $n \geq 1$, $s_n = \max_{0 \leq i < n} \{s_i + i(n-i)(n-i-1)/2\}$ and $s'_n = \max_{0 \leq i < n} \{s'_i + (3n-i) + 2i(2n-i) + (n-i)(2i)\}$. we prove that $\lim_{n \rightarrow \infty} s'_n/3(3n) = 2(\sqrt{3}-1)/3 \approx 0.488033\dots$, resolving an open problem from bioinformatics about rooted triplets consistency in phylogenetic networks.

a111

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 |

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 |

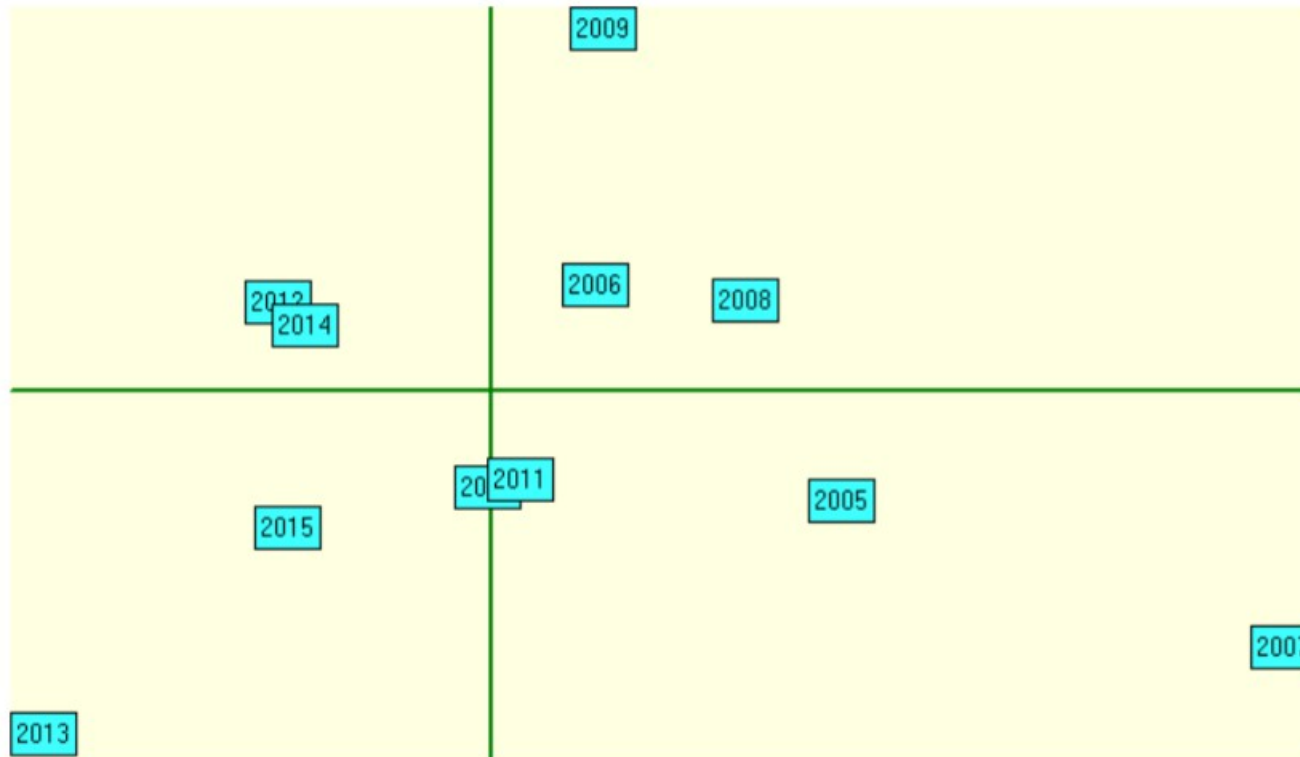
Exemple d'exploration : articles scientifiques

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
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

Exemple d'exploration : articles scientifiques

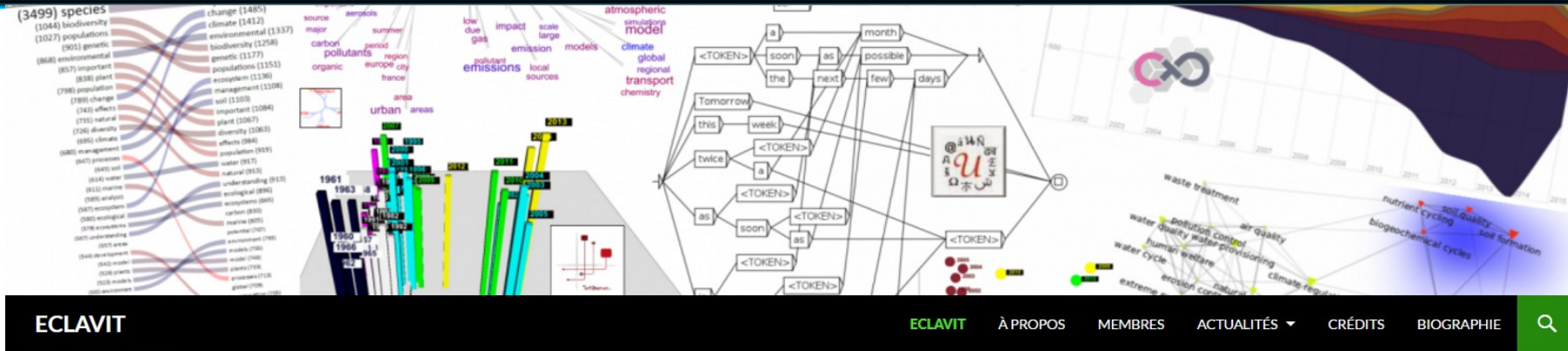
Test de l'effet d'un paramètre : l'année de publication



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



ECLAVIT

Extraction, classification et visualisation de données textuelles, mutualisation de méthodes et interopérabilité d'outils textuels existants

Recherche...

ECLAVIT À PROPOS MEMBRES ACTUALITÉS CRÉDITS BIOGRAPHIE

<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

Localisation

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

Horaire

samedi 10h-13h
6 séances - 20 heures



Stages de formation à TextObserver

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

En semaine, deux séances (dates à préciser).

5 novembre 2016 : Introduction - Approches textométriques des discours.

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...

10 décembre 2016 : Présentation et prise en main de TextObserver [10h-16h]

- Expliciter l'analyse factorielle des correspondances
- Analyser la variation lexicométrique
- Introduction aux opérations de catégorisation
<http://textopol.u-pec.fr/textobserver>
- Recension de corpus et balisage semi-automatisé : présentation de la base Textopol

21 janvier 2017 : Des corpus textuels aux corpus multimodaux (annoter, catégoriser, étiqueter, visualiser, interpréter).

- Transformer des textes pour les soumettre à des traitements automatisés. Repérer les régularités d'un document, extraire des motifs textuels, concaténer des fichiers.
- Forme graphique, catégories morphosyntaxiques, univers sémantiques.
- 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.

25 février 2017 : 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*.

25 mars 2017 : Corrélations et causalités interprétatives. Expérimentations, distance intertextuelle et voisinages.

Distances, cooccurrences, voisinage
Présentation et prise en main d'*Hyperbase* : de la lexicométrie à la stylogrammétrie.
Présentation d'*Hyperbase* en ligne.

22 avril 2017 : De la textométrie à l'analyse des données, quels outils pour quels usages?

Cooccurrences généralisées et mondes lexicaux : analyses comparées *Alceste* et *Iramuteq*.
Outils de dépouillement d'enquêtes et de formulaires : *Sphinx*, *Modalisa*.
Ontologies et mondes sémantiques (*Tropes*, *Alceste*, *Astartex*)
Quantifier les données en sciences sociales : *Prospero*, *Nvivo*

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).

Courant 2016 - 2017 :

3e journée de présentation des outils de traitements de corpus textuels

Gramlab/Unitex, Treecloud, TextObserver/WebObserver, Cortext...
Echanges, présentations, ateliers. Outils développés par les équipes de recherche de Paris-Est.
Le programme détaillé sera mis en ligne sur Textopol.

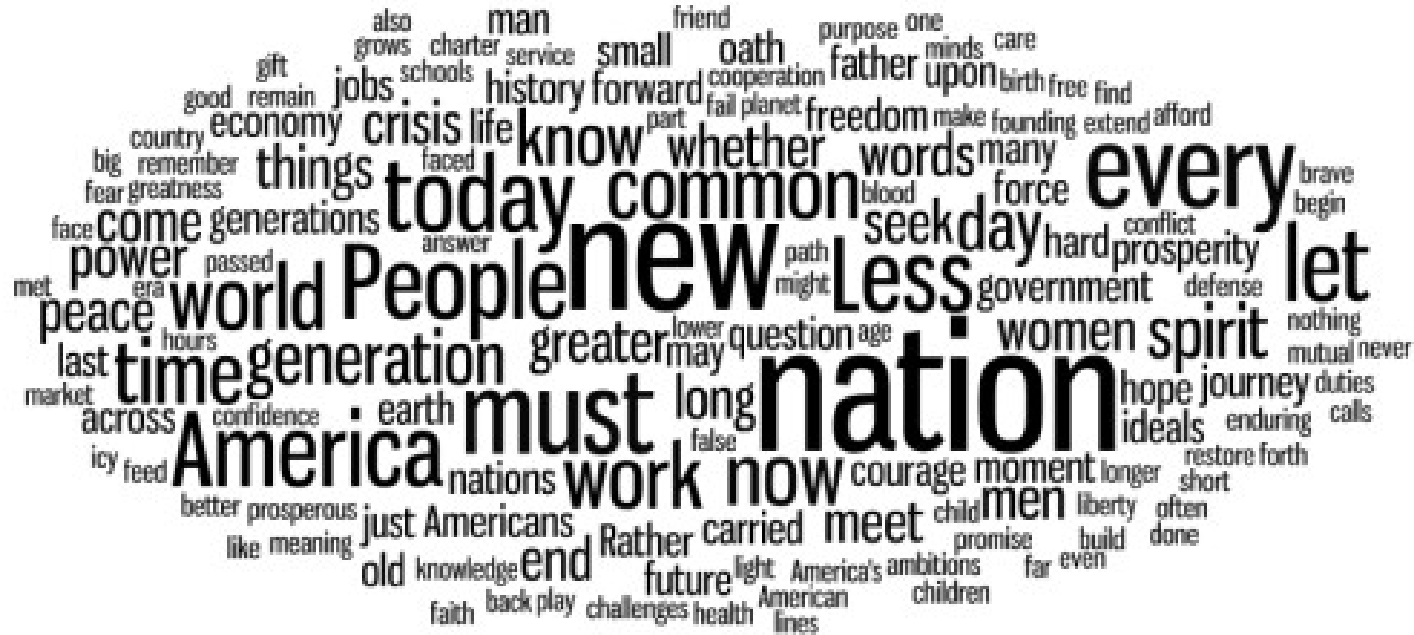
PROGRAMME 2015-2016

PROGRAMME 2015-2016

Arbres de mots

Le « nuage arboré », une information double

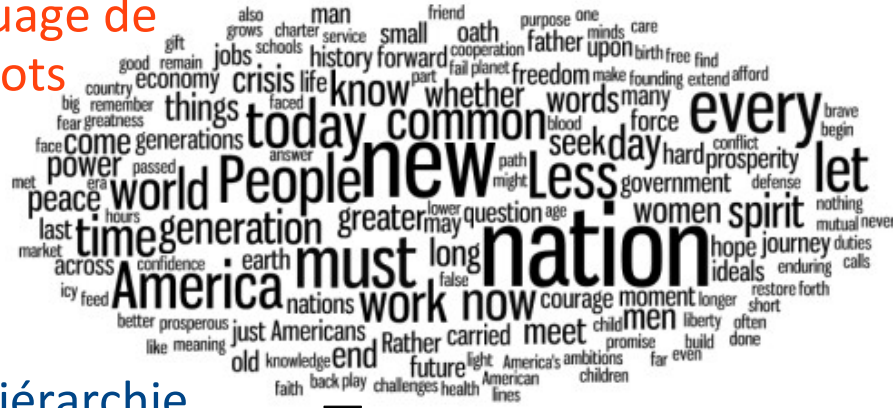
nuage de
mots



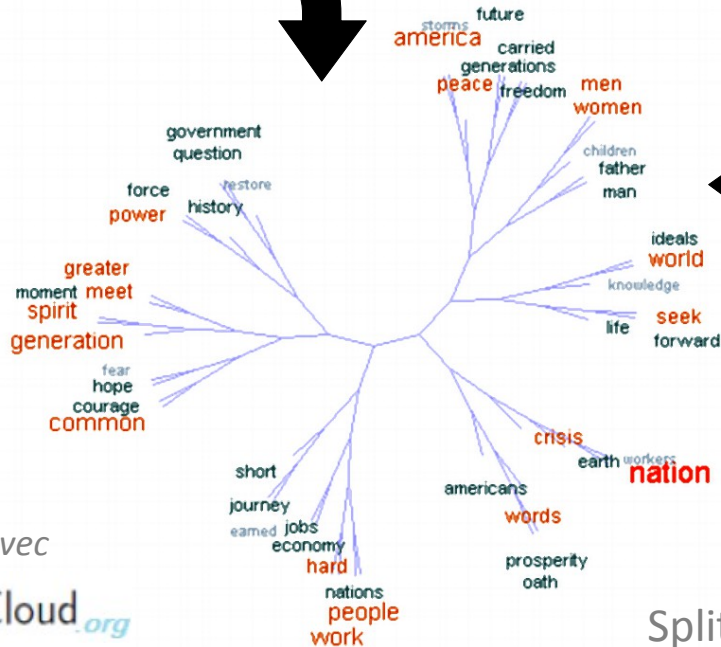
Discours inaugural de Barack Obama en 2008,
Wordle

Le « nuage arboré », une information double

nuage de mots

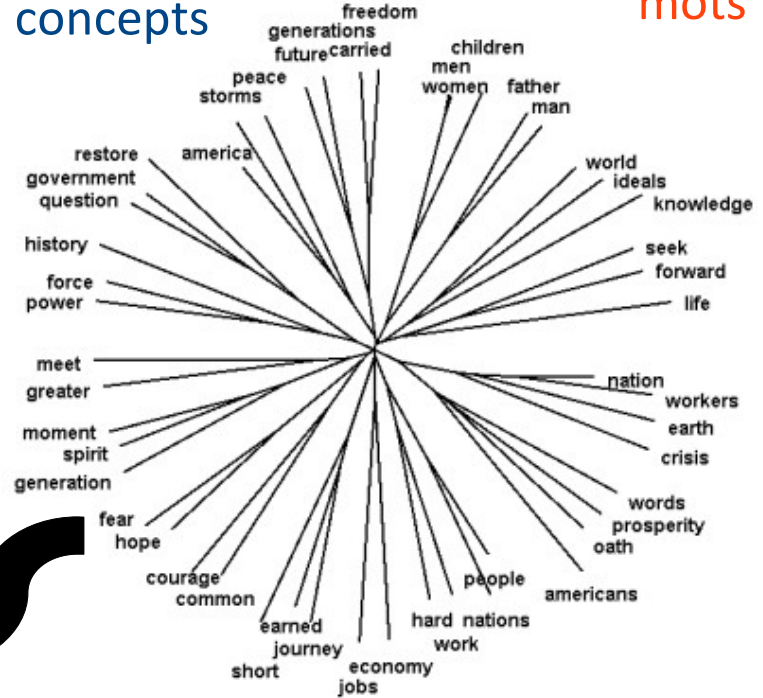


hiérarchie des mots



hiérarchie des concepts

arbre de mots



Discours inaugural de Barack Obama

SplitsTree : Huson & Bryant, *Bioinformatics*, 2006

TreeCloud : Gambette & Véronis, *IFCS'09*

construit avec

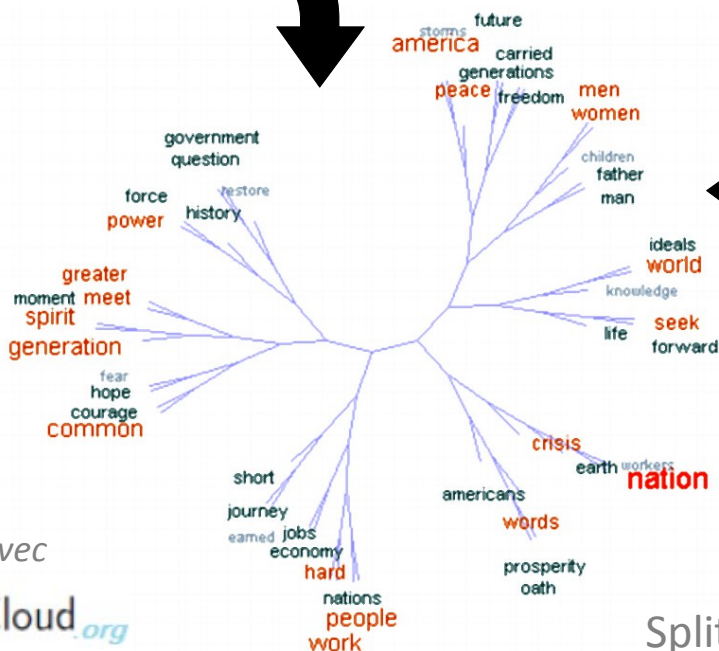
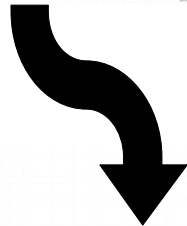


Le « nuage arboré », une information double

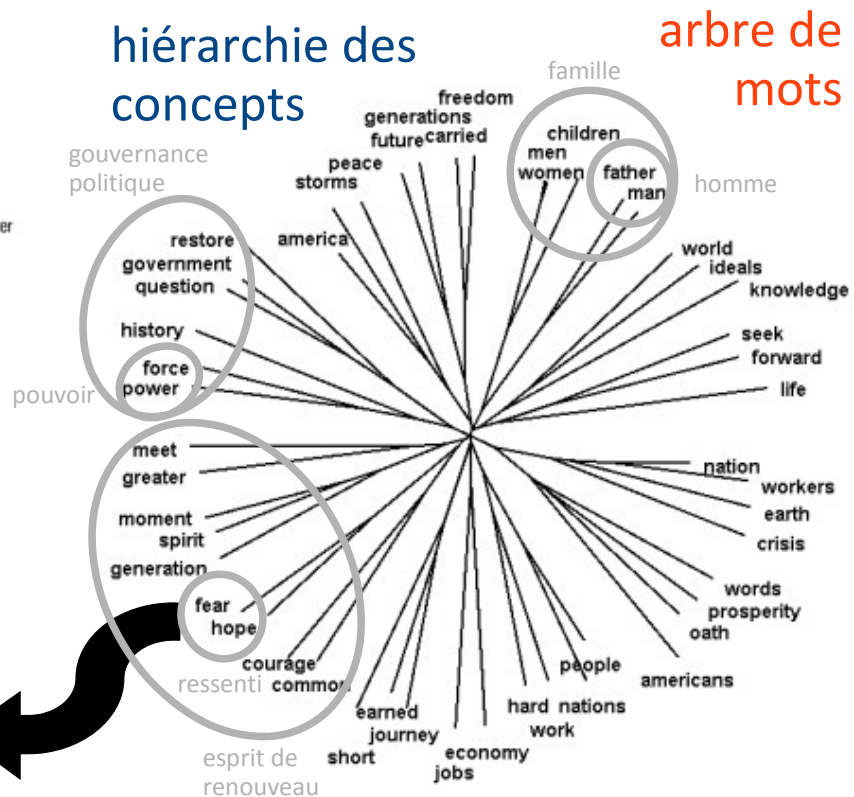
nuage de mots



hiérarchie des mots



hiérarchie des concepts




arbre de mots



Discours inaugural de Barack Obama

construit avec

 TreeCloud.org

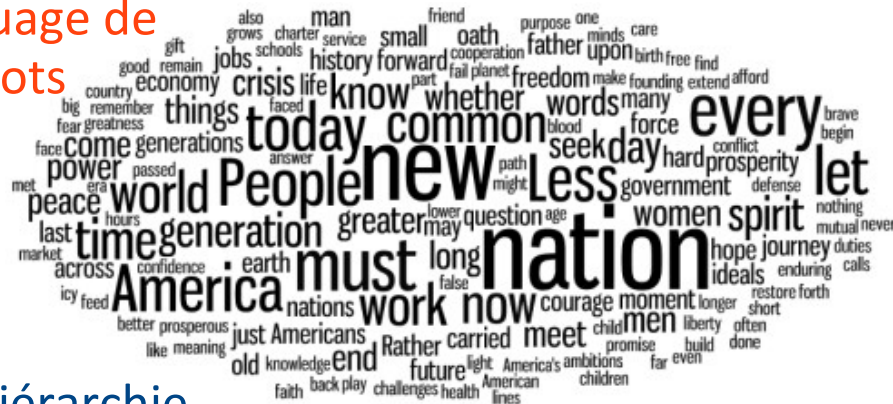
 SplitsTree4

SplitsTree : Huson & Bryant, *Bioinformatics*, 2006

TreeCloud : Gambette & Véronis, *IFCS'09*

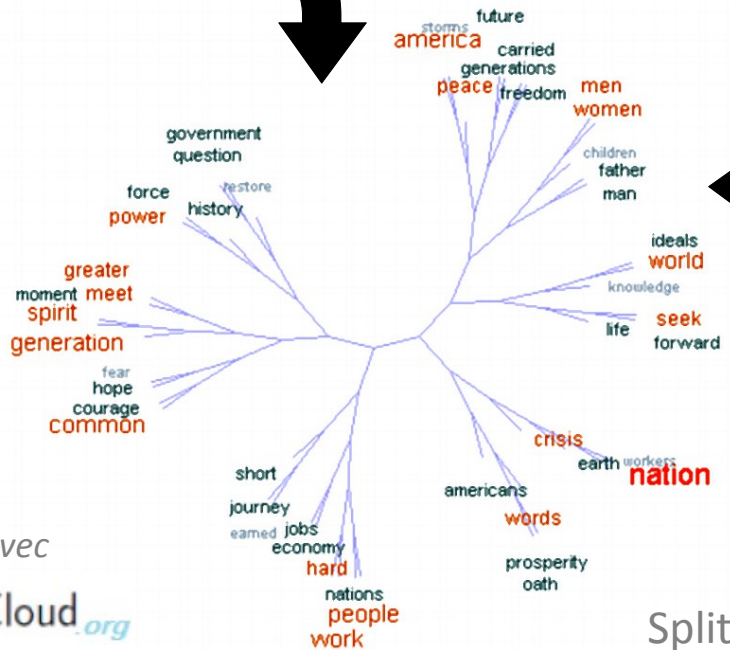
Le « nuage arboré », une information double

nuage de mots



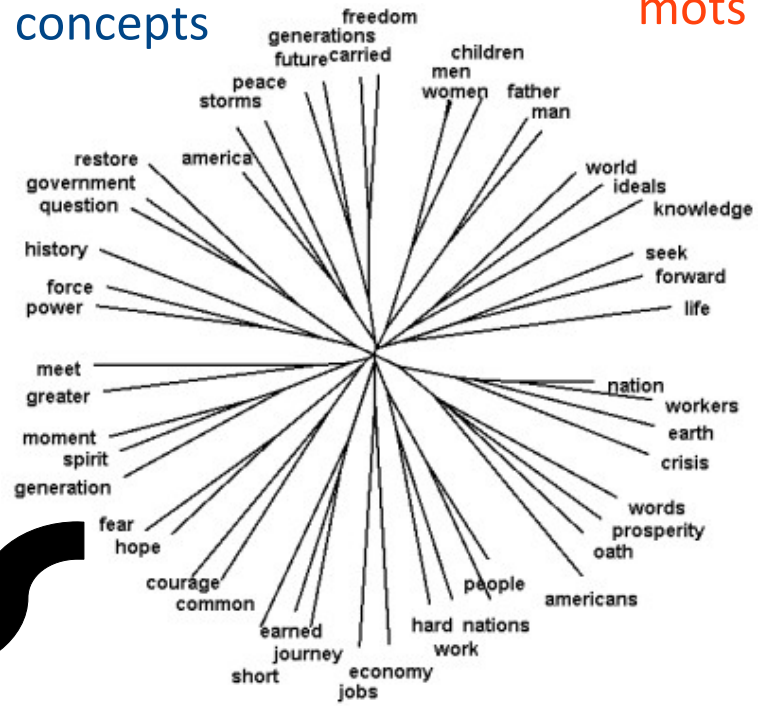
hiérarchie des mots

occurrences



hiérarchie des concepts

arbre de mots



cooccurrences

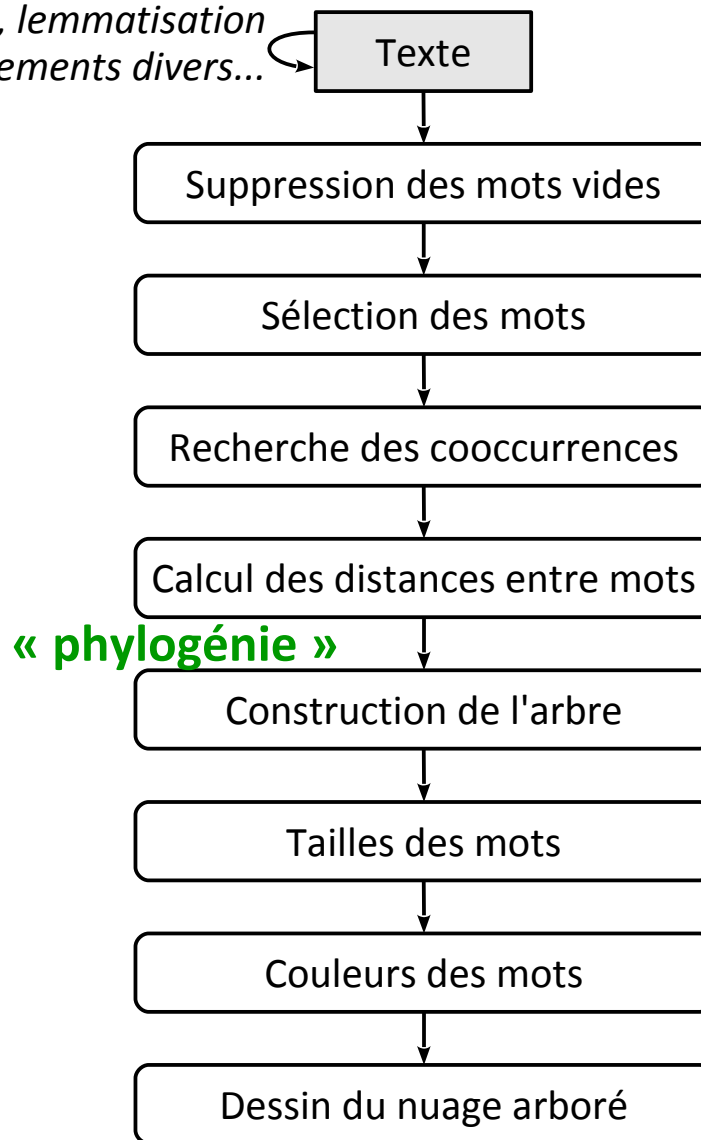
Discours inaugural de Barack Obama

construit avec
TreeCloud.org
SplitsTree4

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

Processus de construction

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

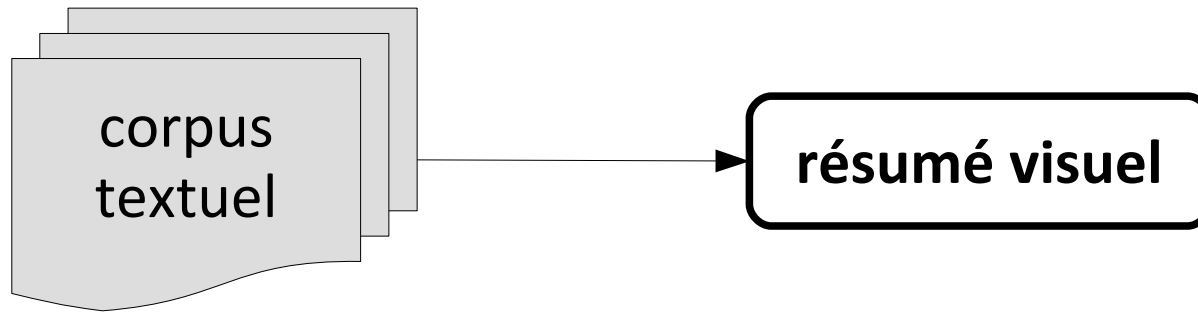
*Appel transparent au logiciel
SplitsTree*

Fréquences ou valeurs personnalisées

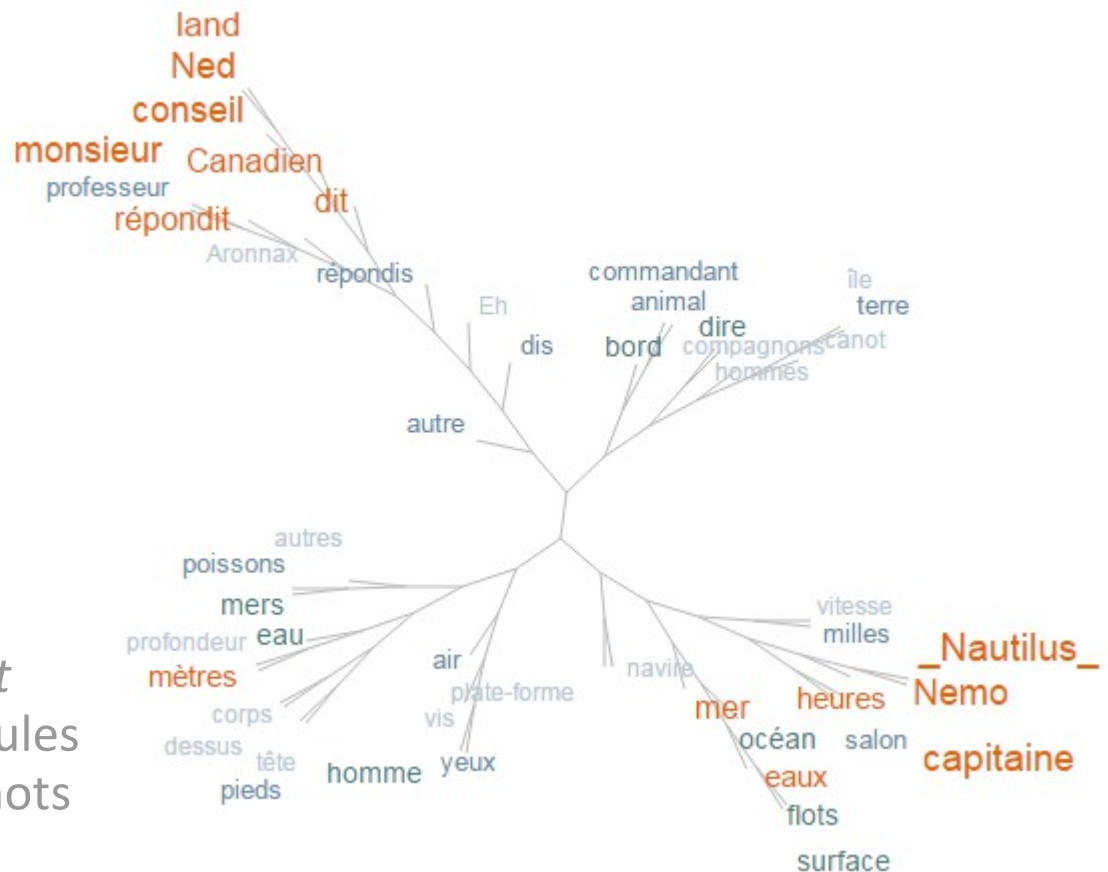
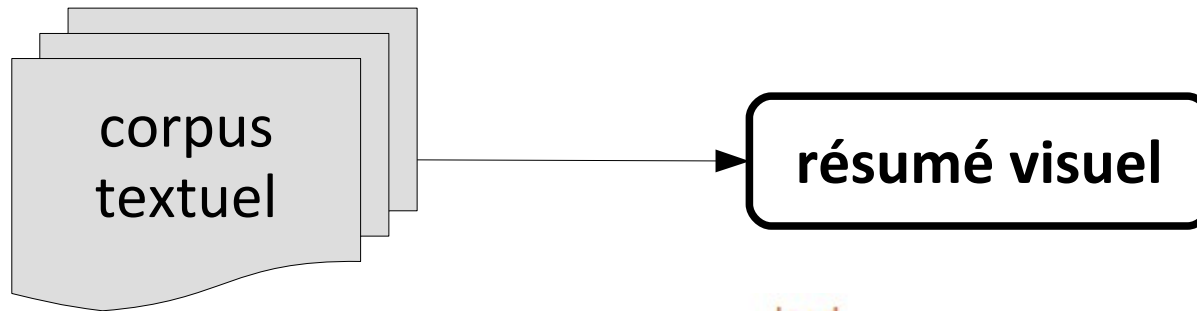
*Fréquences, chronologie, dispersion,
ciblées sur la cooccurrence d'un mot,
ou valeurs personnalisées*

*Appel transparent au logiciel SplitsTree
ou Dendroscope*

Le « nuage arboré », pour quoi faire ?

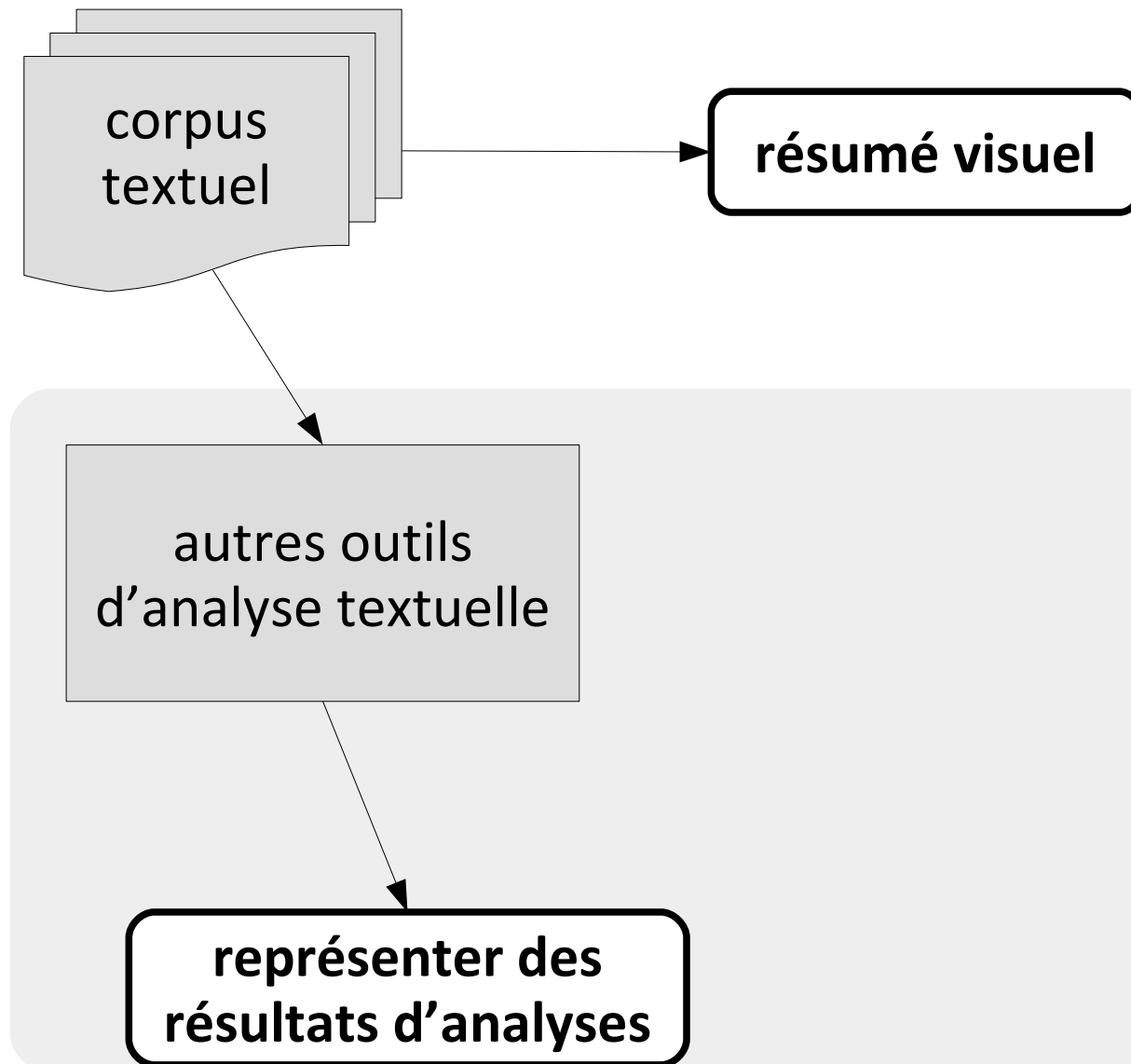


Le « nuage arboré », pour quoi faire ?



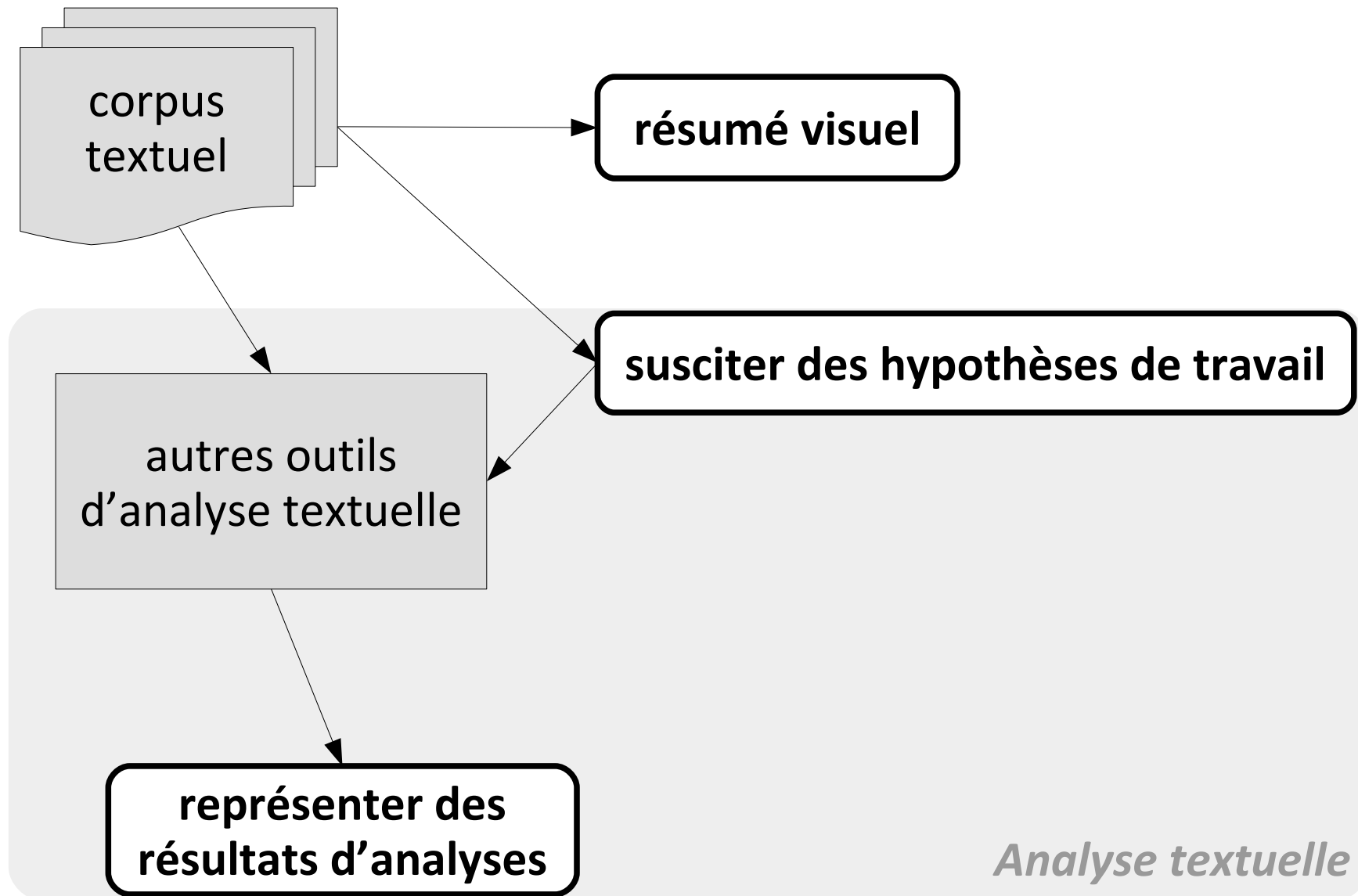
Nuage arboré construit sur Treecloud.org du roman *Vingt mille lieues sous les mers* de Jules Verne (distance Jaccard, 50 mots les plus fréquents)

Le « nuage arboré », pour quoi faire ?



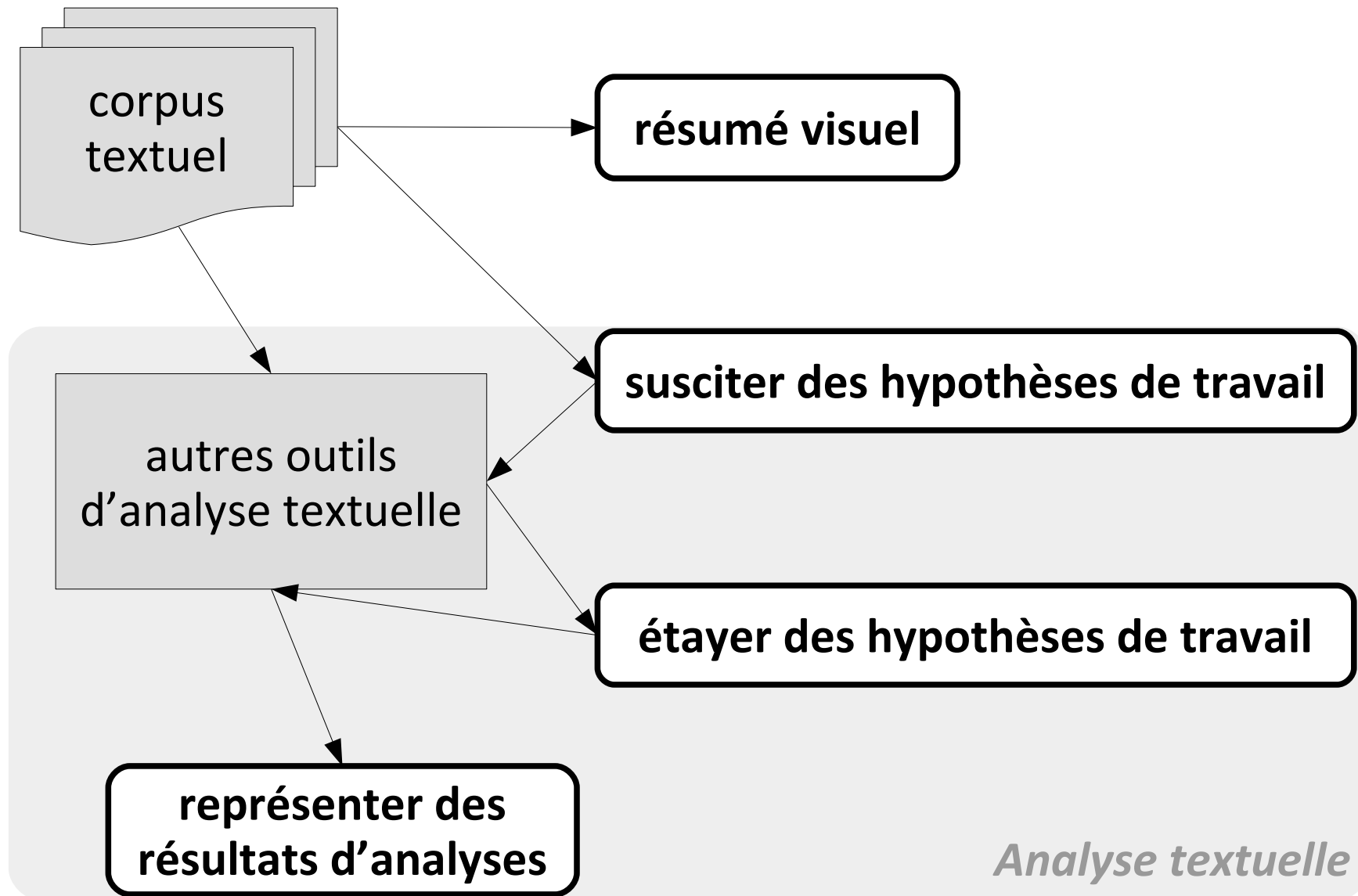
Analyse textuelle

Le « nuage arboré », pour quoi faire ?



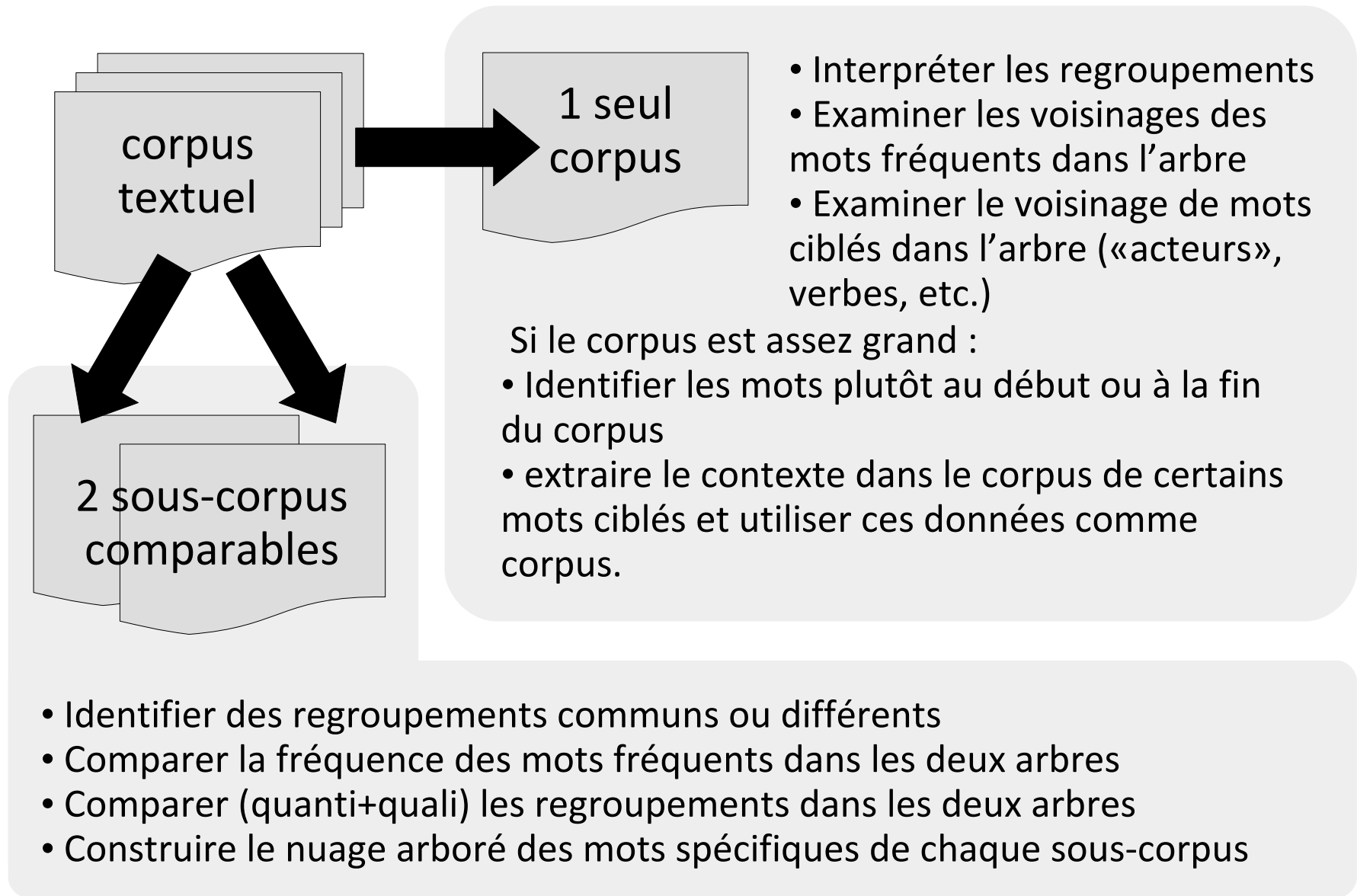
Analyse textuelle

Le « nuage arboré », pour quoi faire ?

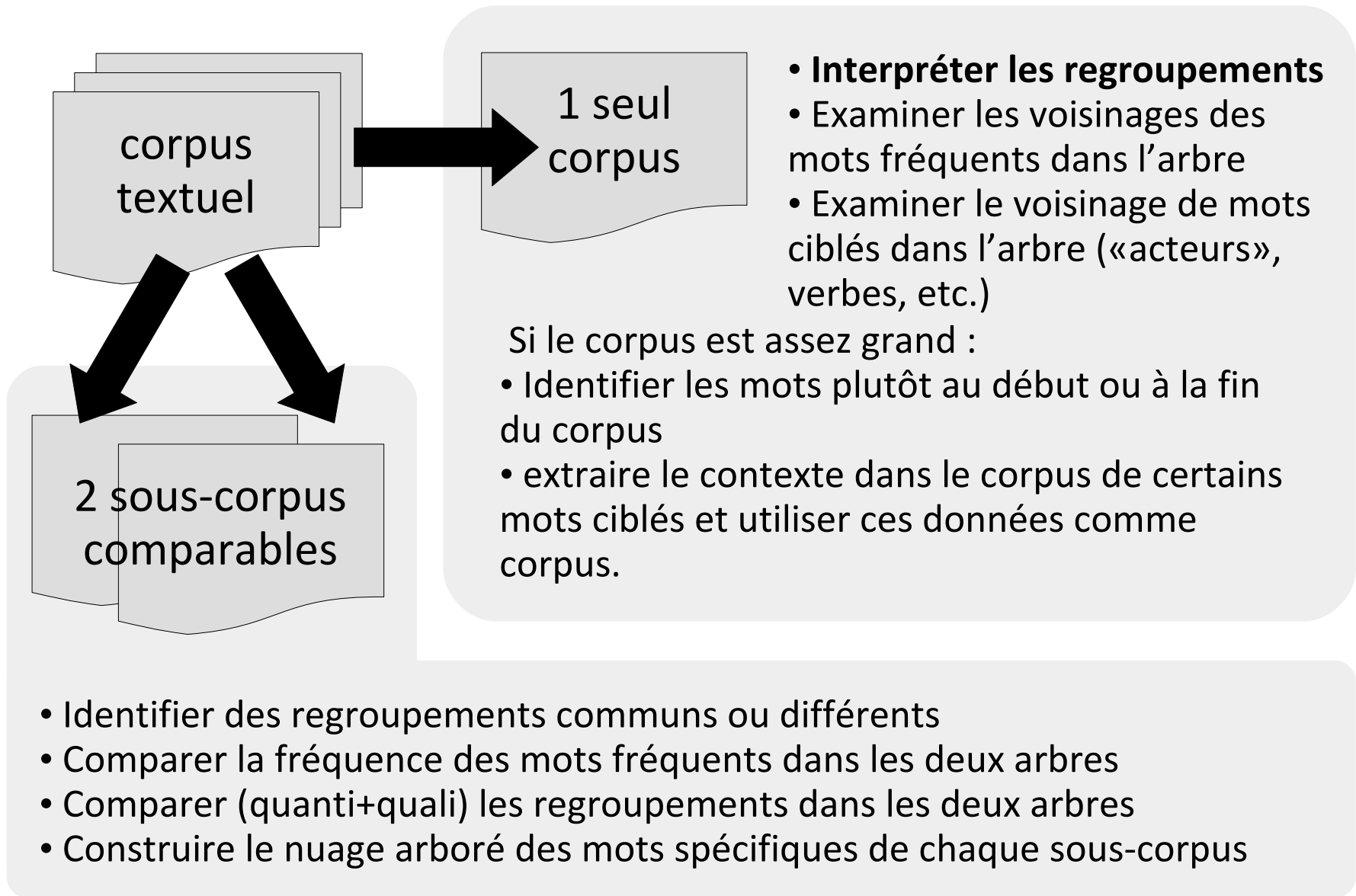


Analyse textuelle

Exploration de corpus avec TreeCloud



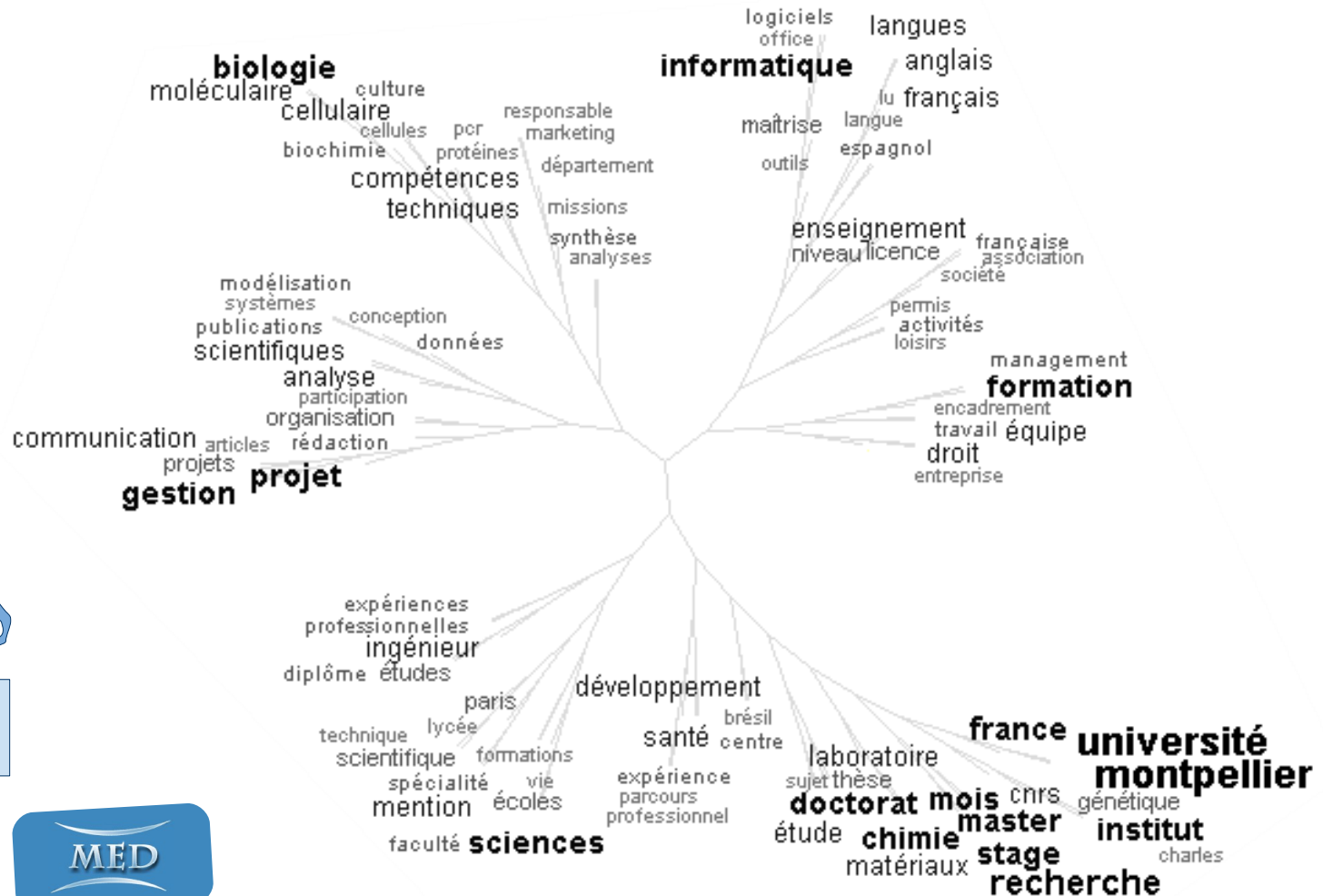
Exploration de corpus avec TreeCloud



Méthode : interpréter les regroupements

Dessiner des « patates »

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



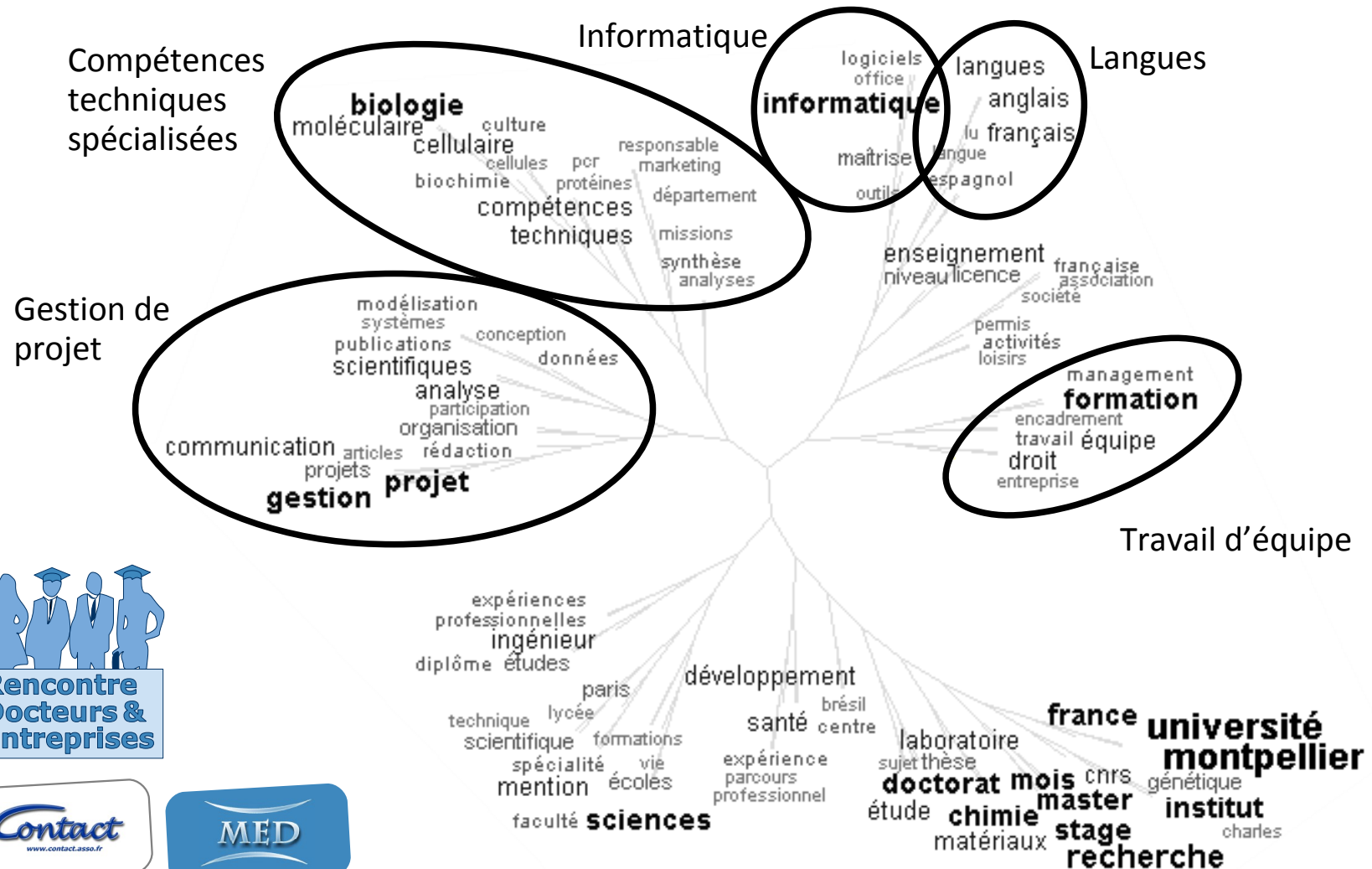
**Rencontre
Docteurs &
Entreprises**



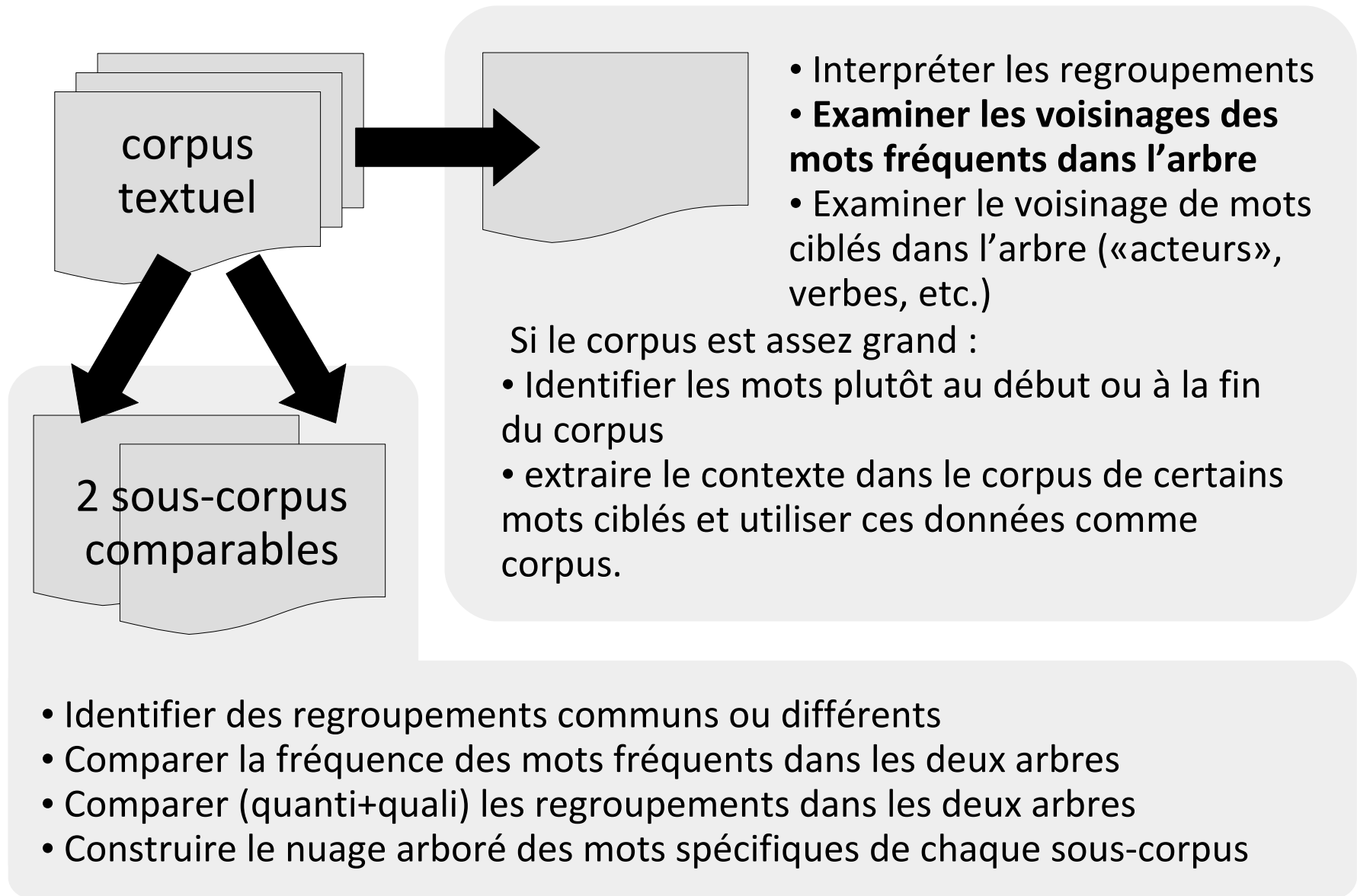
Méthode : interpréter les regroupements

Dessiner des « patates »

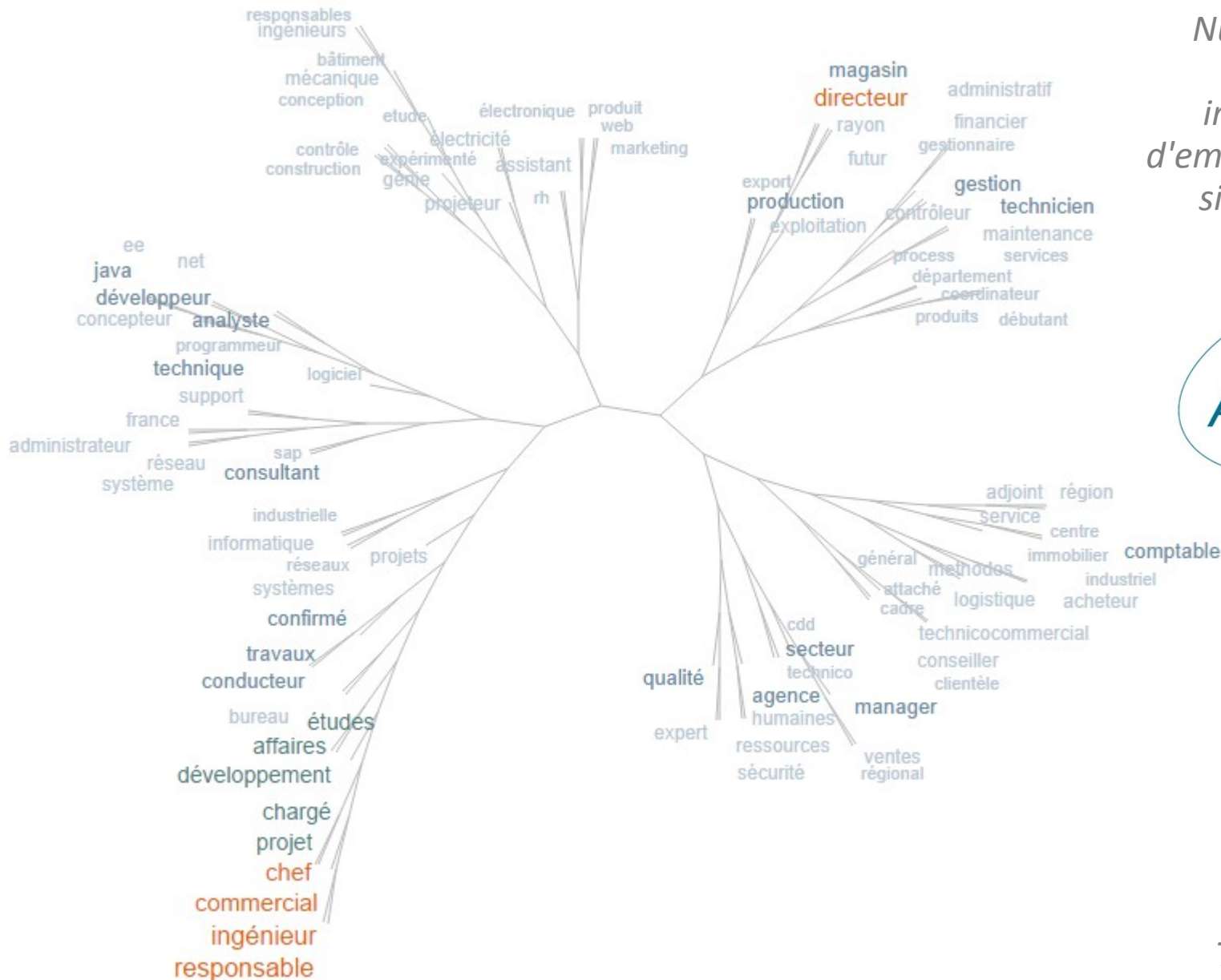
Corpus : une centaine de CV soumis à une rencontre docteurs-entreprises



Exploration de corpus avec TreeCloud



Méthode : voisinage des mots fréquents

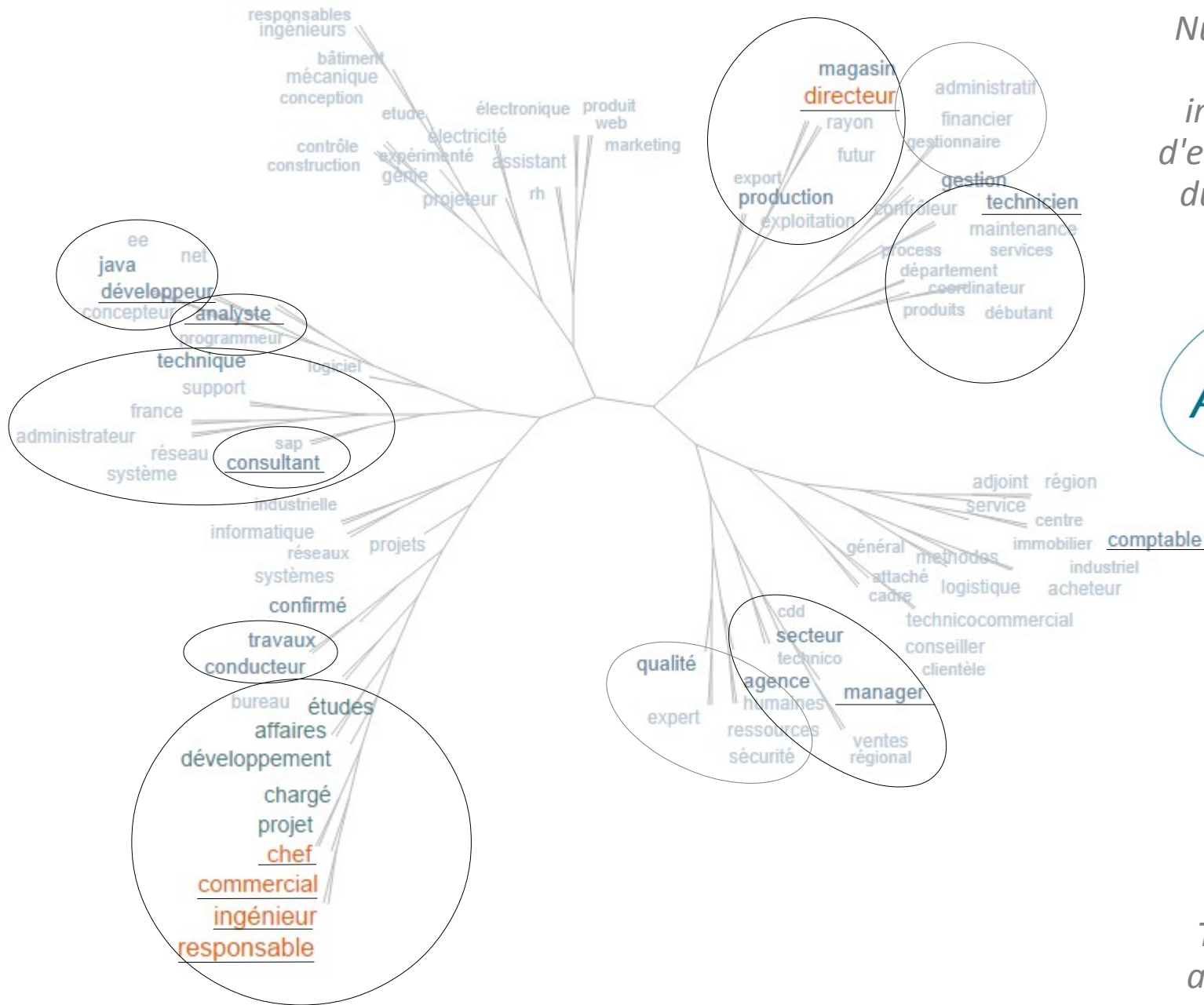


Nuage arboré de plus de 4800 intitulés d'offres d'emploi extraits du site de l'APEC en avril 2011.



Travail de 2011 avec Paola Salle

Méthode : voisinage des mots fréquents

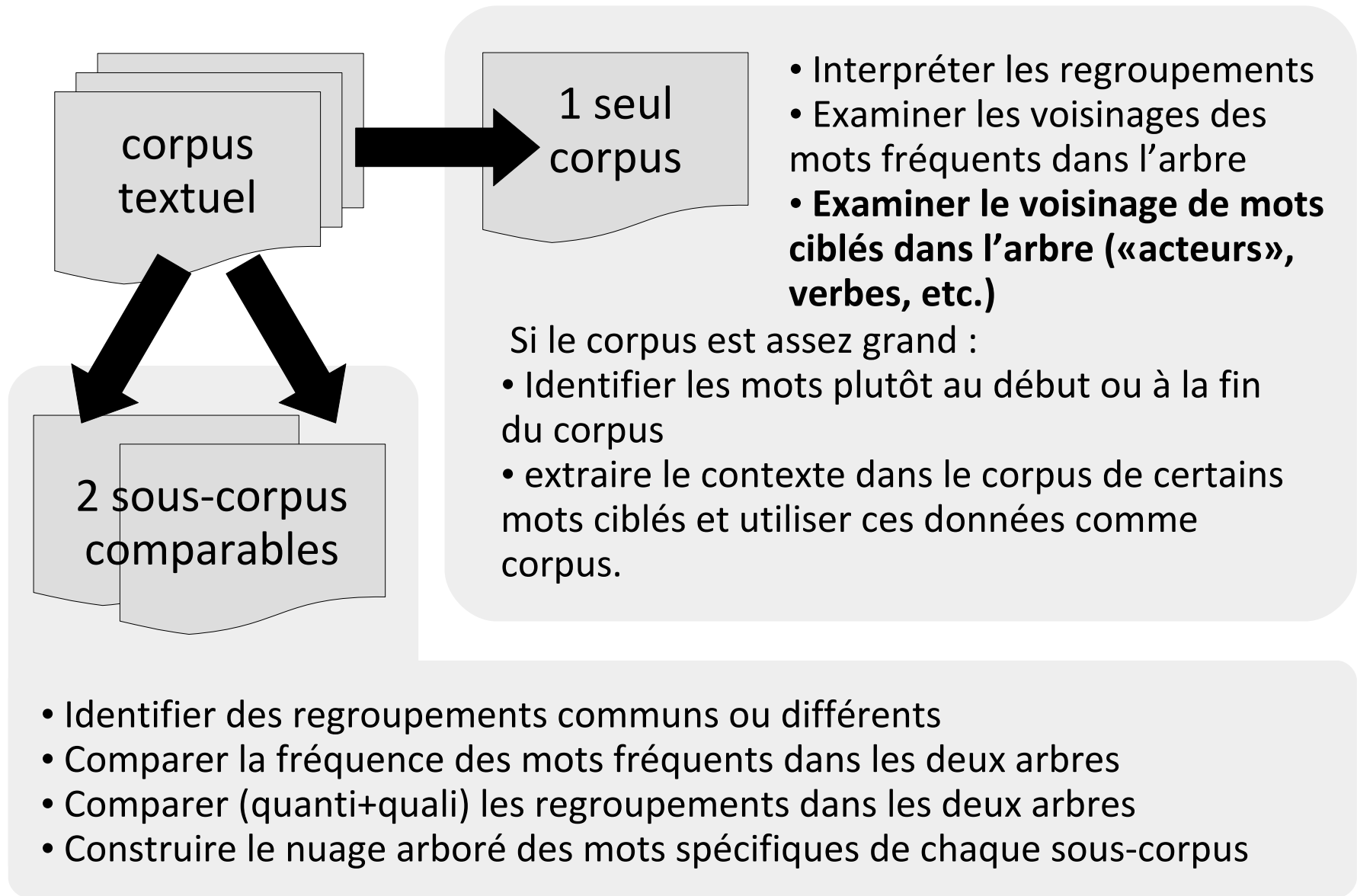


Nuage arboré de plus de 4800 intitulés d'offres d'emploi extraites du site de l'APEC en avril 2011.



Travail de 2011 avec Paola Salle

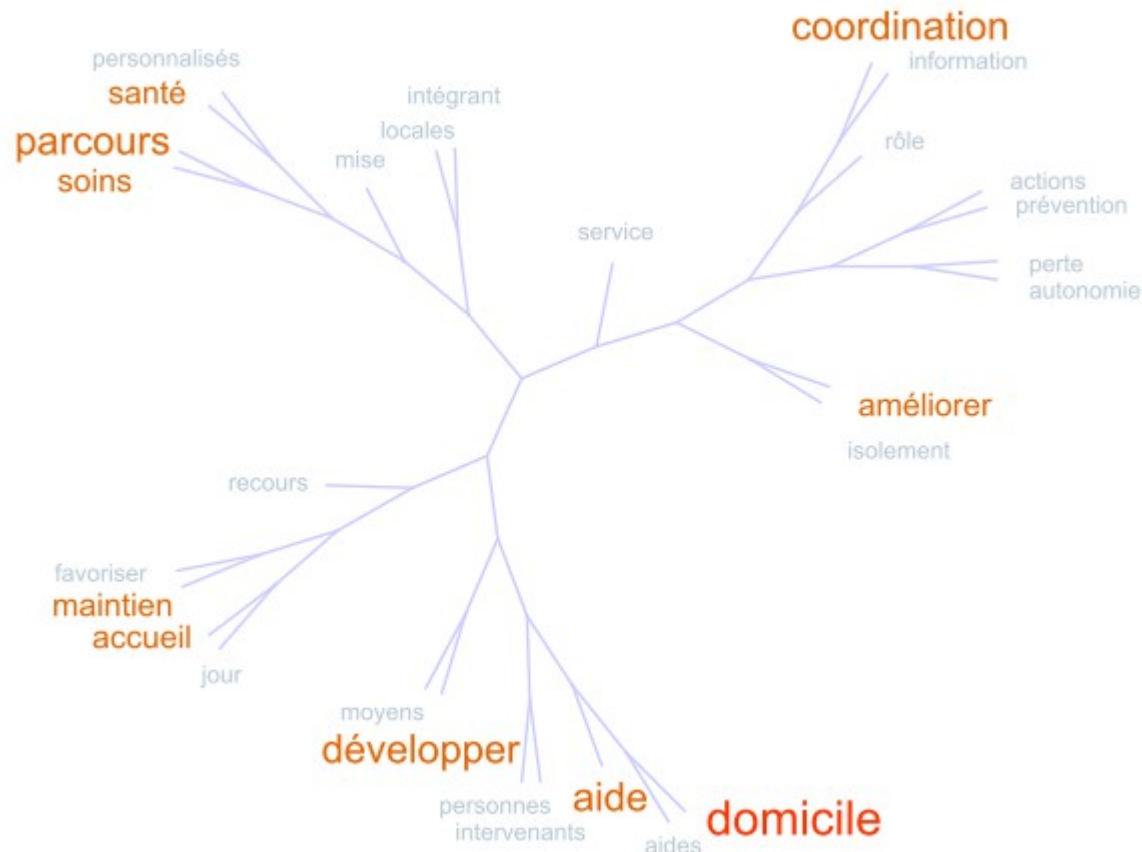
Exploration de corpus avec TreeCloud



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

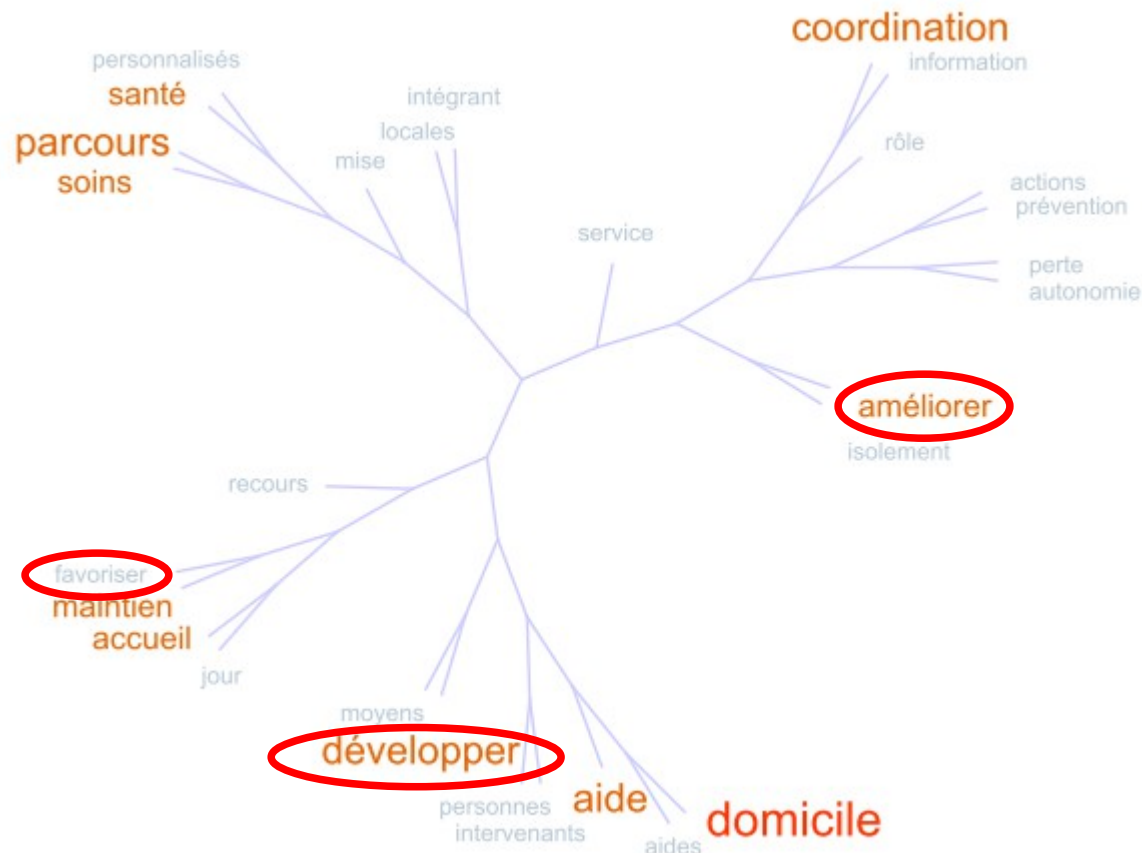
Suggestions d'améliorations :



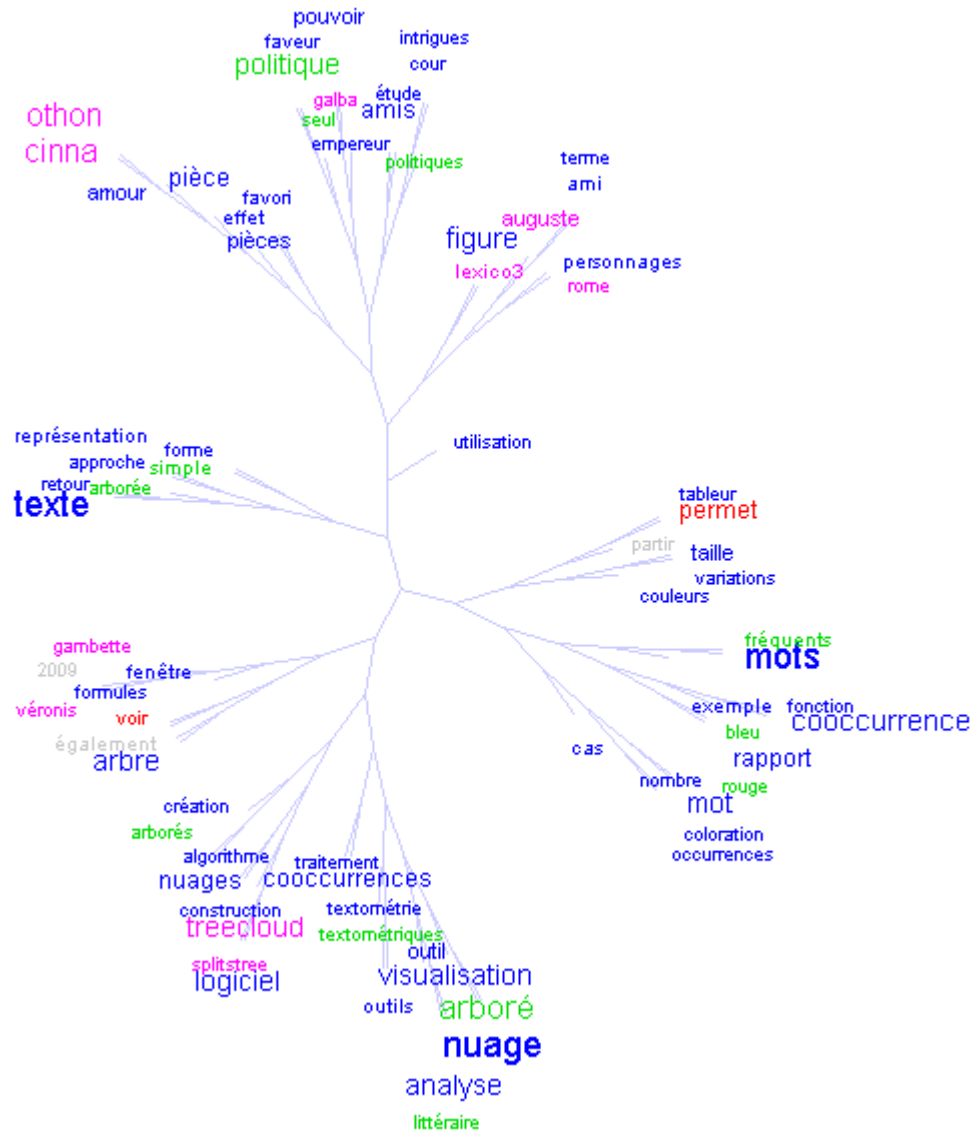
Méthode : voisinage des verbes

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 :



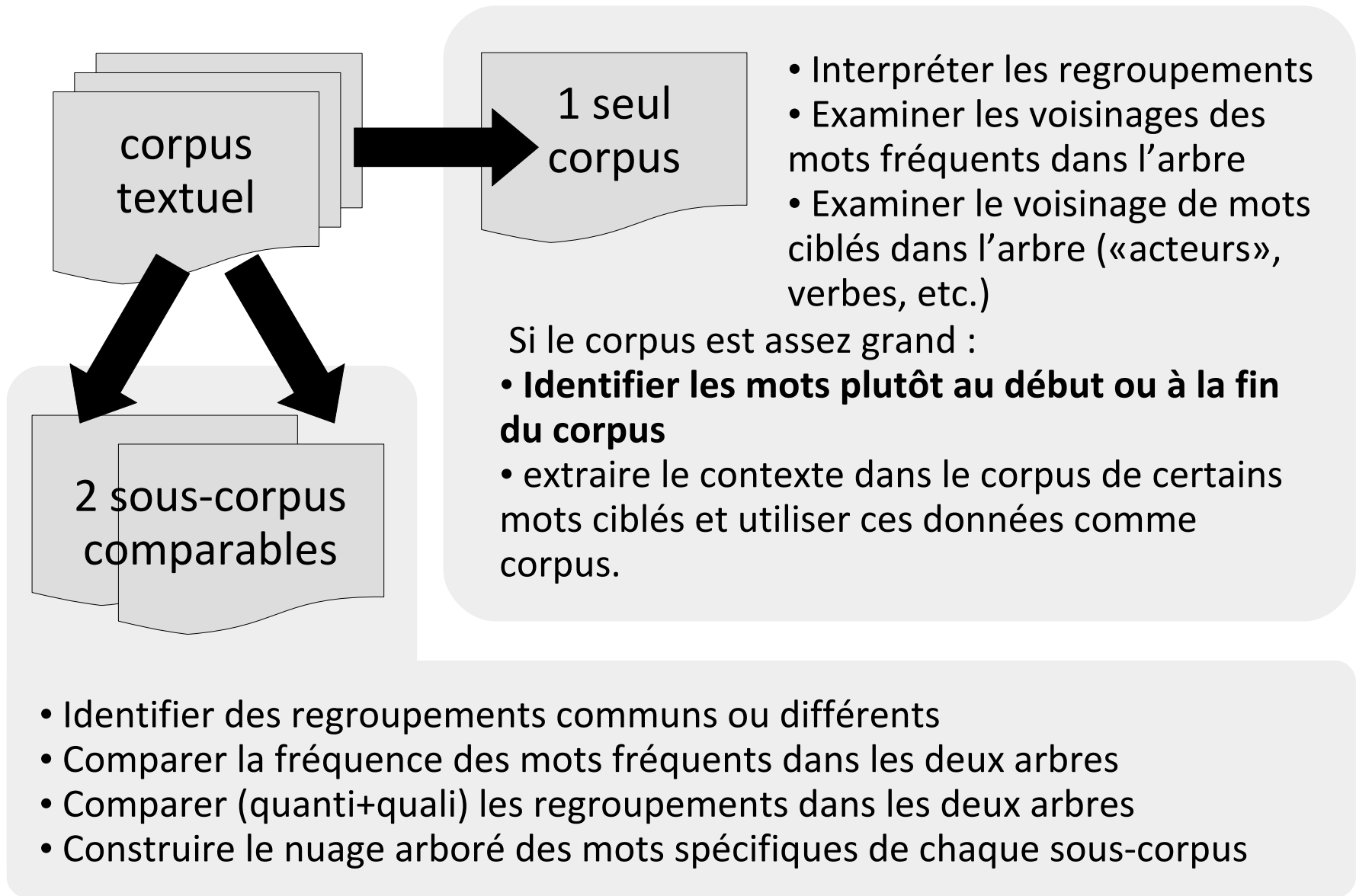
Perspective : coloration grammaticale



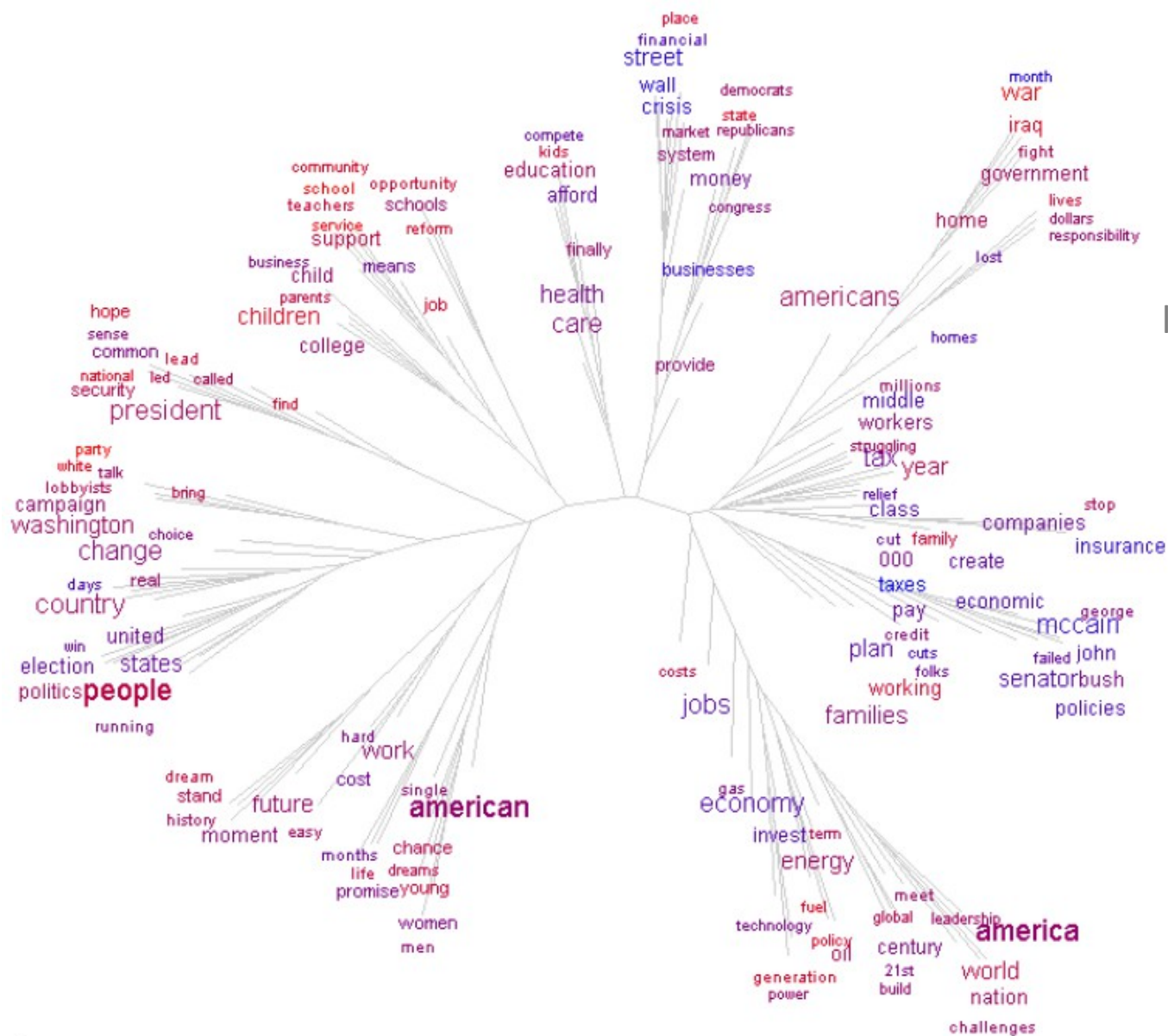
noms
adjectifs
verbes
noms propres

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



Méthode : mots au début ou à la fin

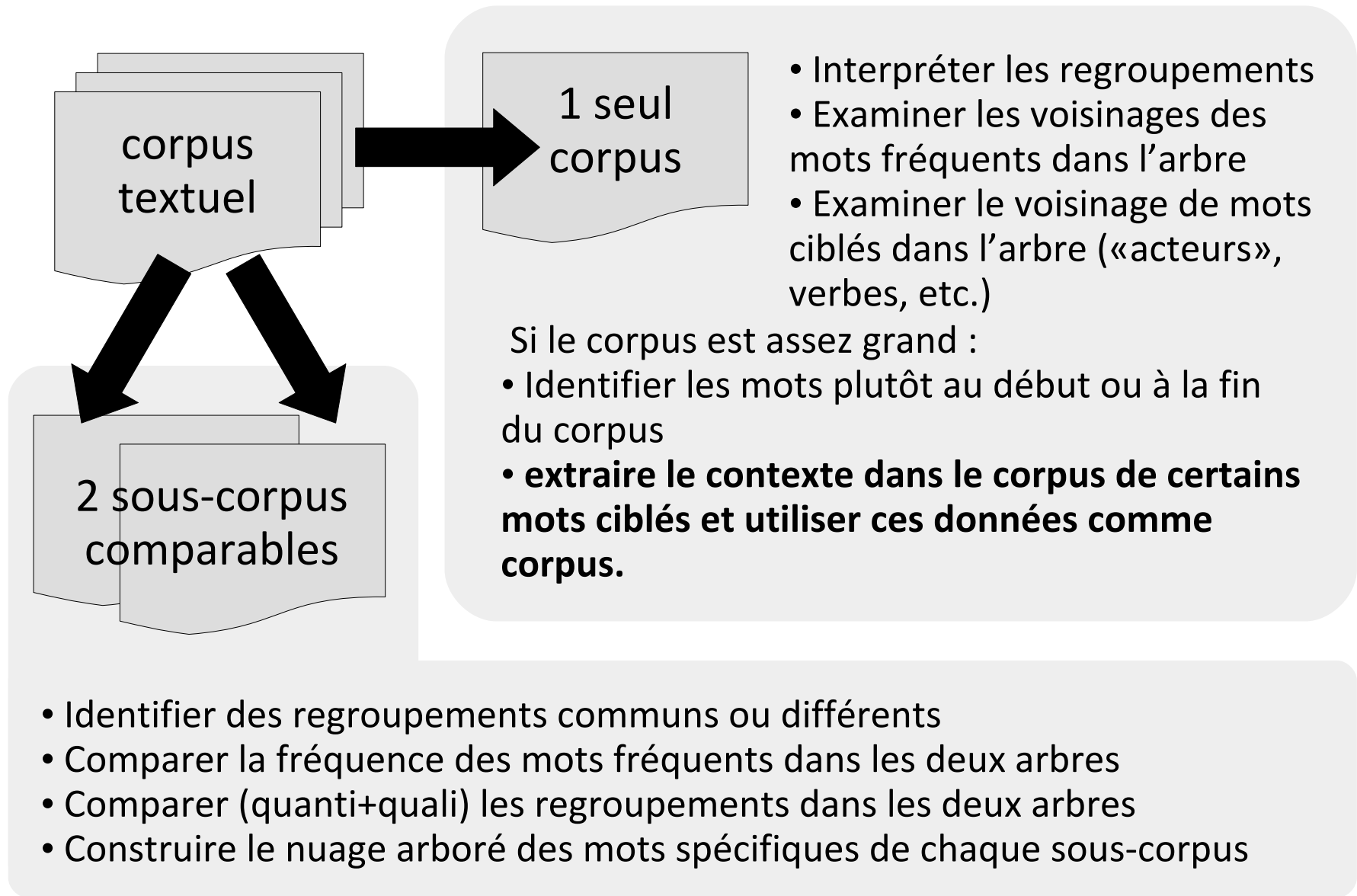


Nuage arboré de l'ensemble des discours de campagne de 2008 de Barack Obama, coloration chronologique

début de la campagne
fin de la campagne

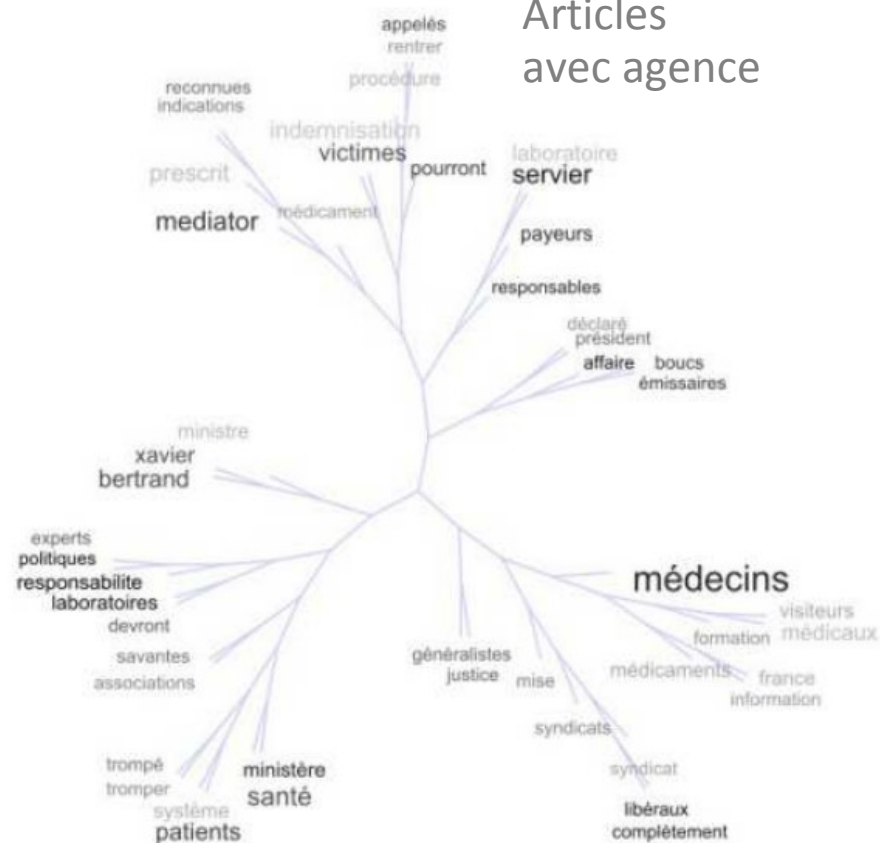
Gambette & Véronis, IFCS 2009

Exploration de corpus avec TreeCloud

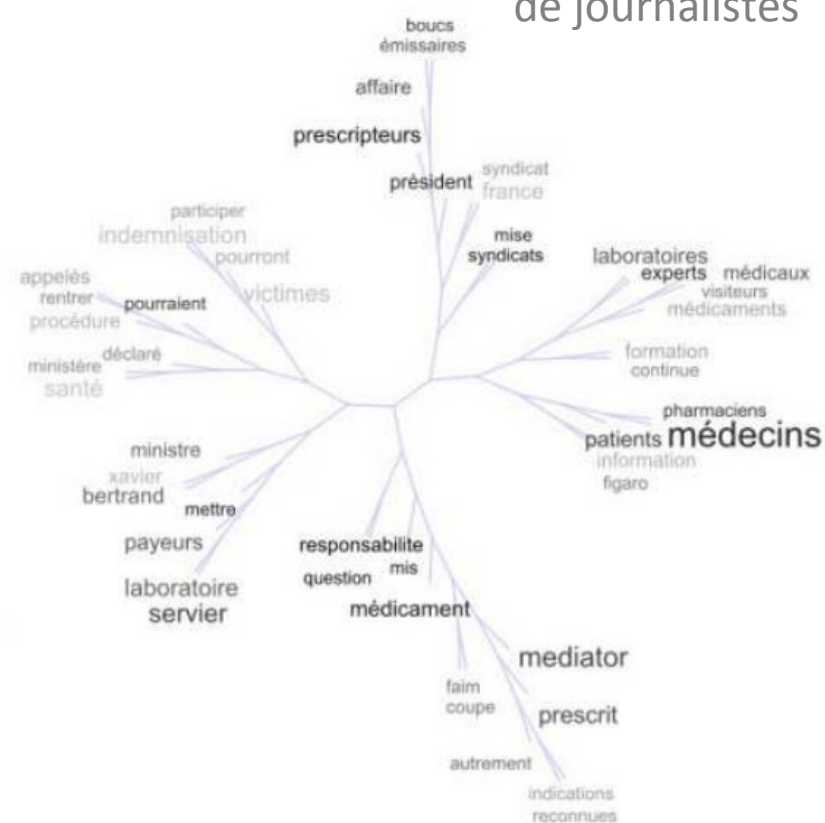


Nuages arborés des contextes de « médecins »

Articles avec agence



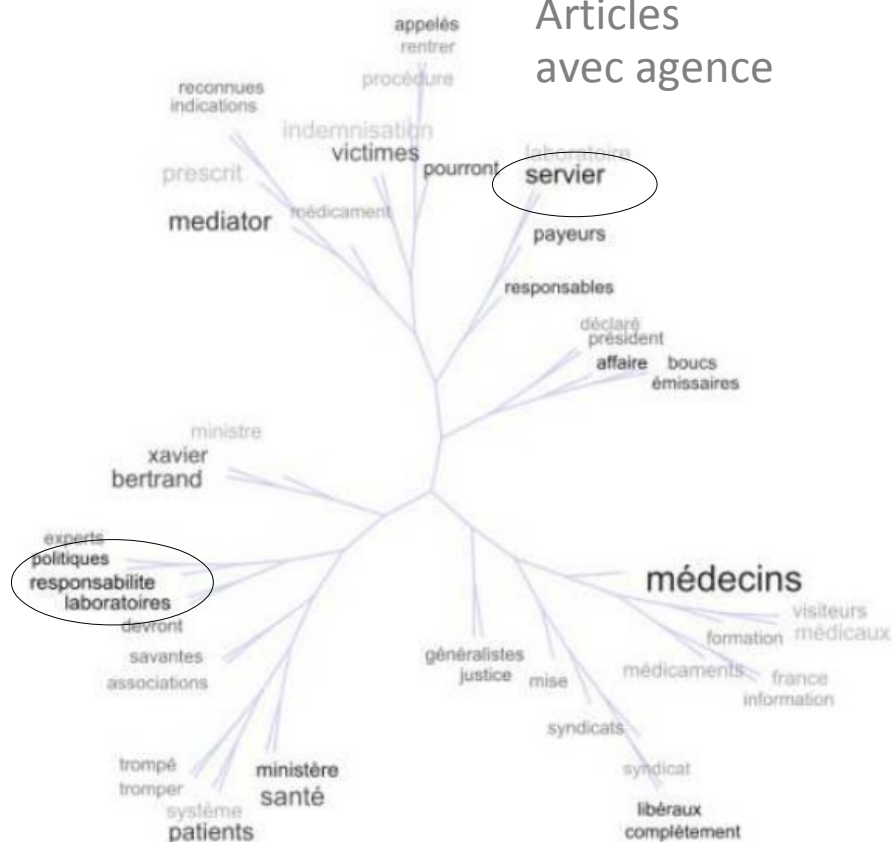
Articles de journalistes



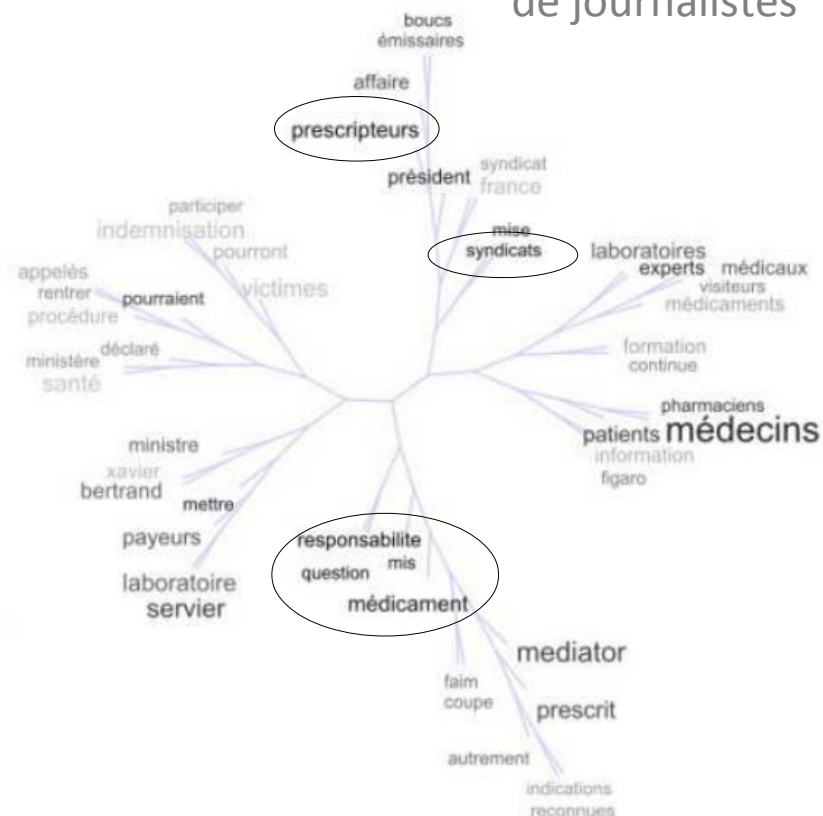
Nuage arboré des 50 mots les plus fréquents des contextes (10 mots avant et 10 mots après) du mot médecins dans le sous-corpus des articles sur le Mediator, colorés par le degré de cooccurrence avec le mot responsabilités (en noir pour les mots les plus cooccurents), construit par TreeCloud avec la formule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

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

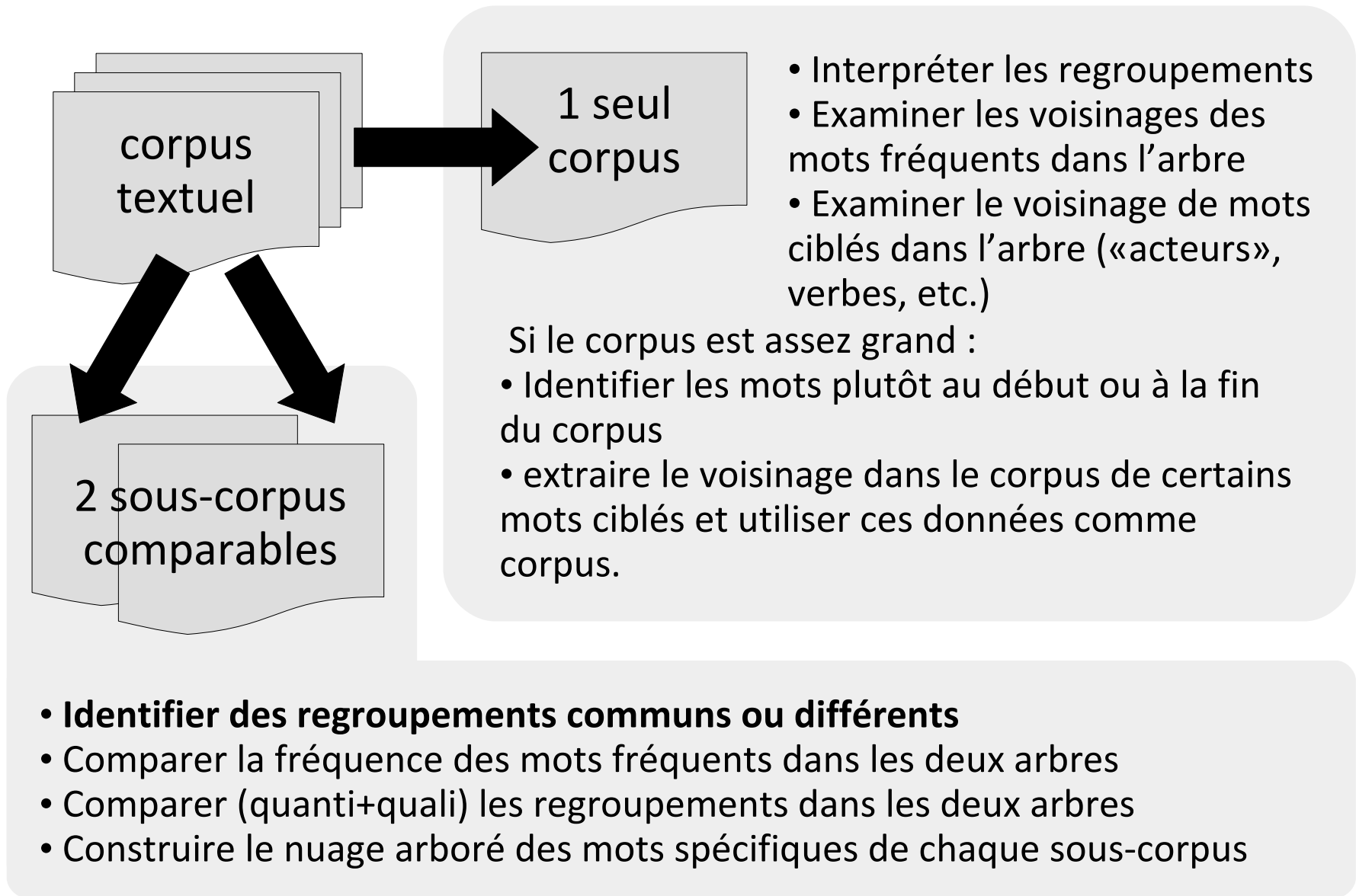


Illustration sur le corpus Mediator

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

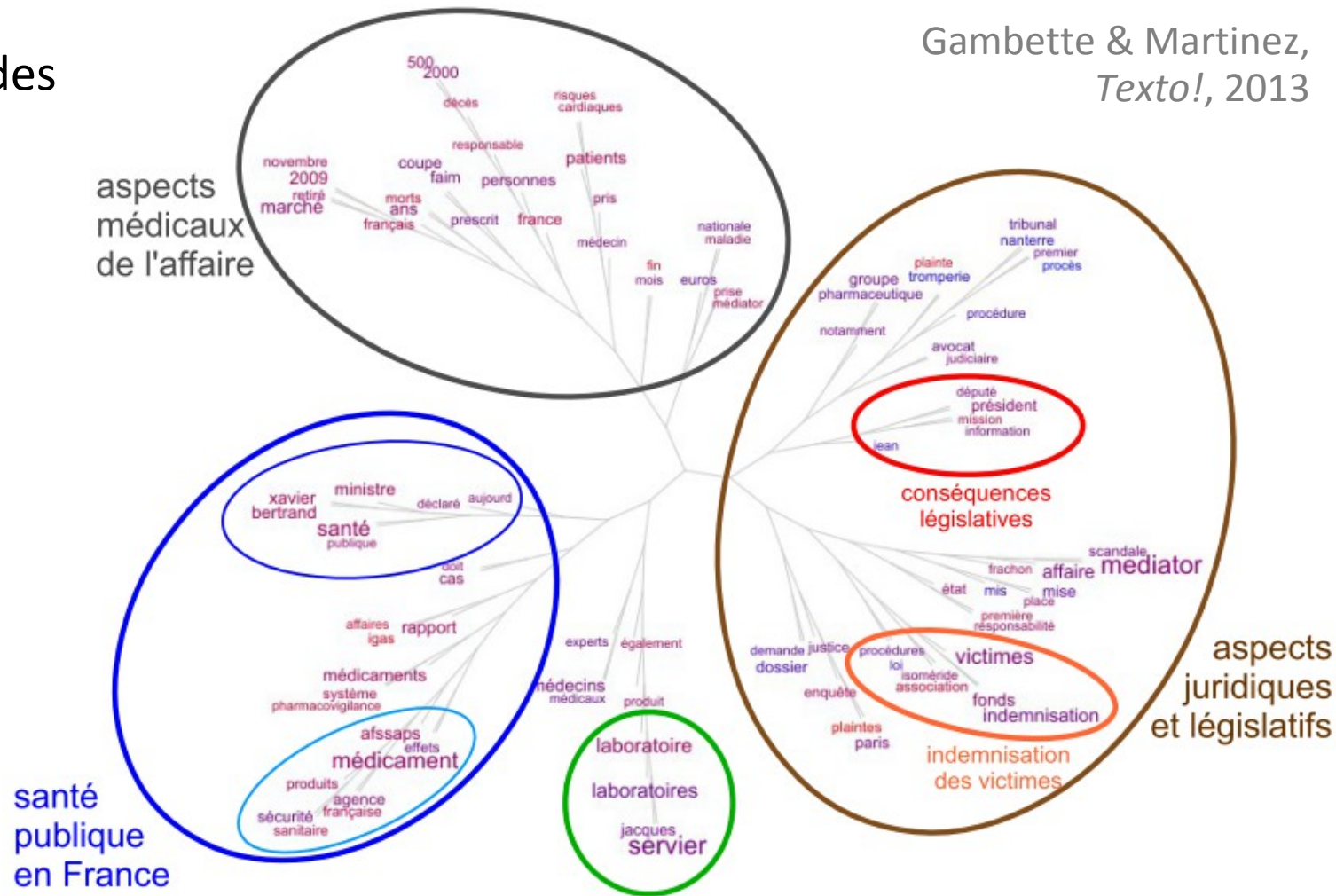
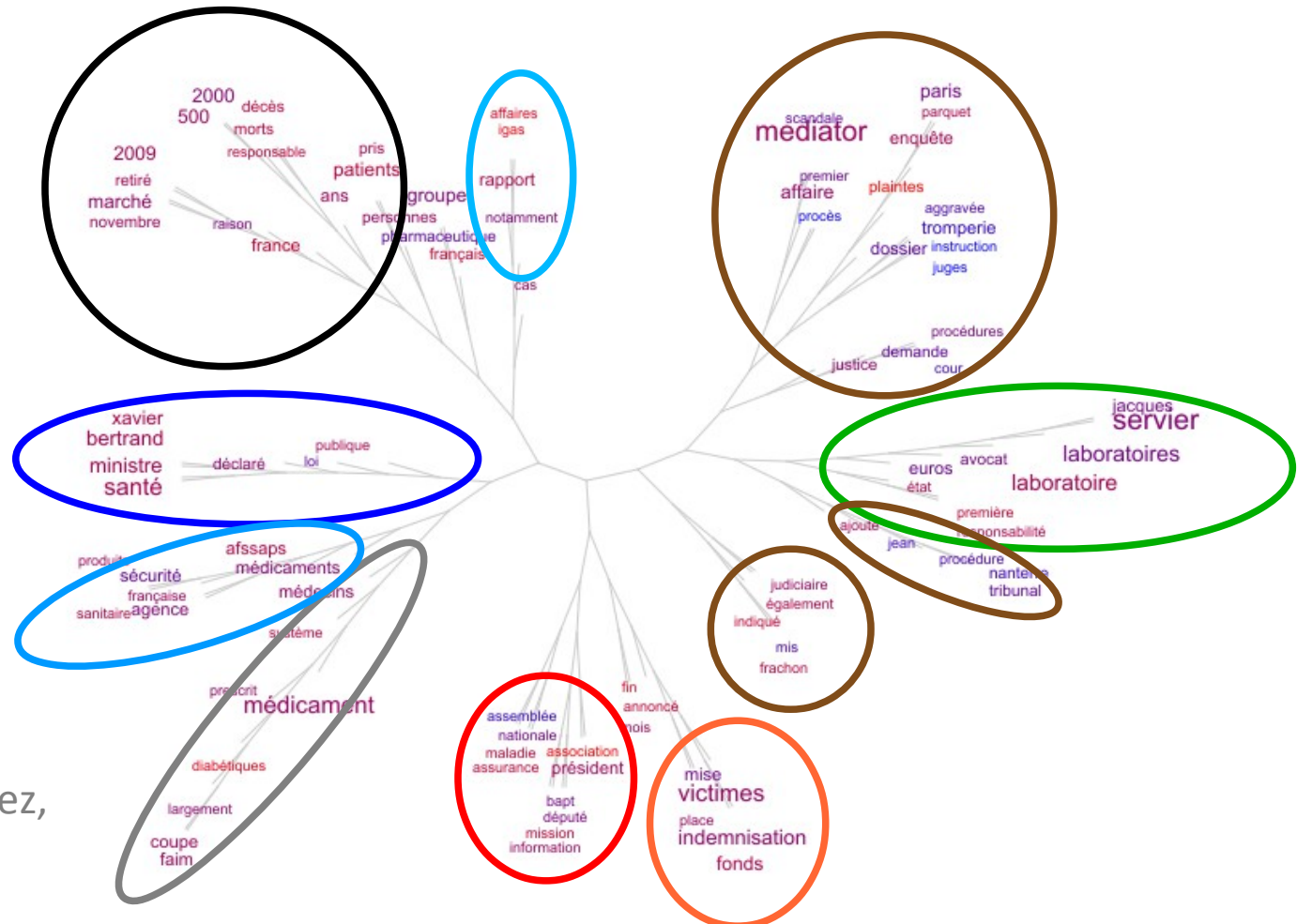


Illustration sur le corpus Mediator

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



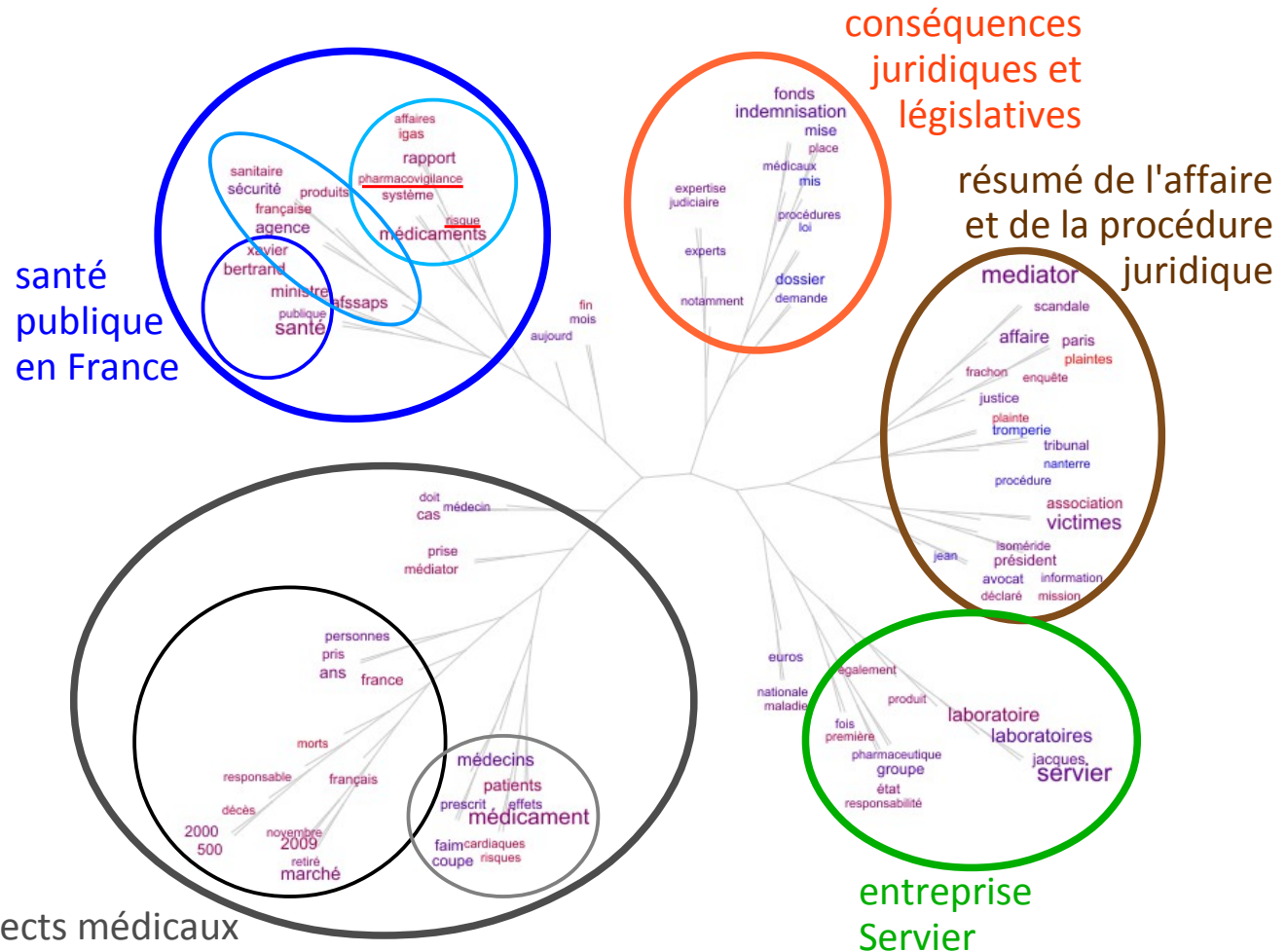
Gambette & Martinez,
Texto!, 2013

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.

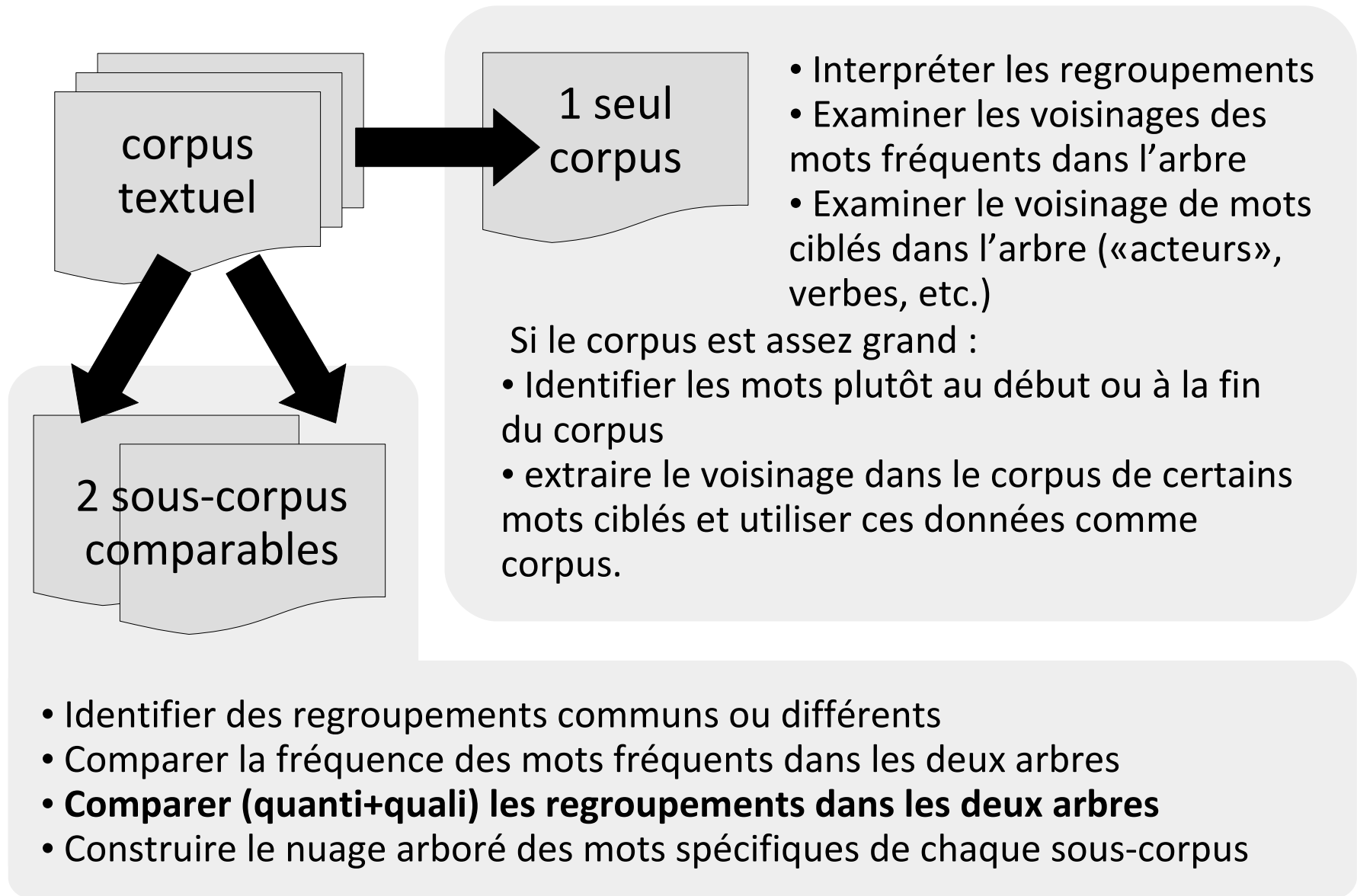
Articles
de journalistes



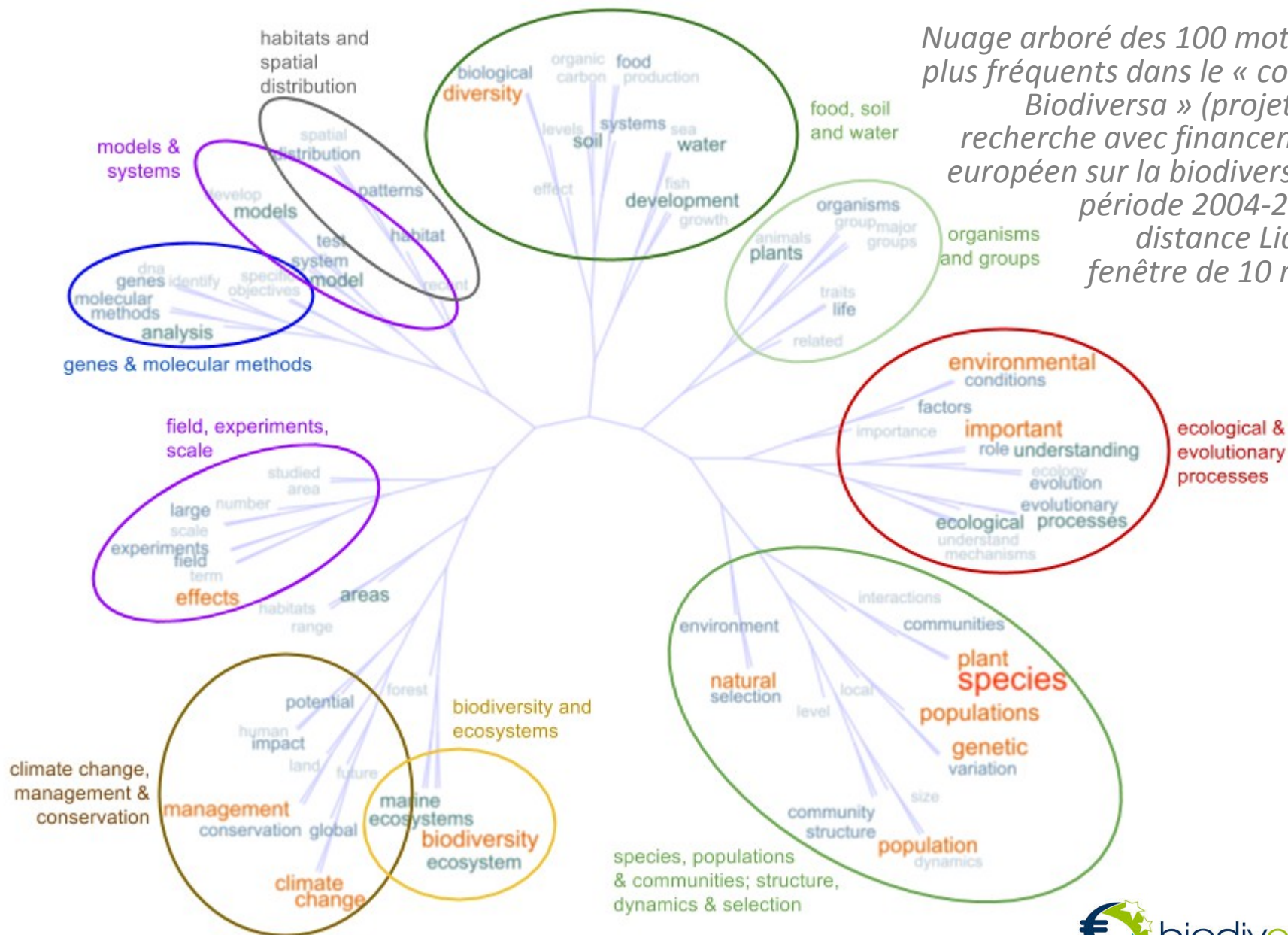
Gambette & Martinez,
Texto!, 2013

aspects médicaux

Exploration de corpus avec TreeCloud

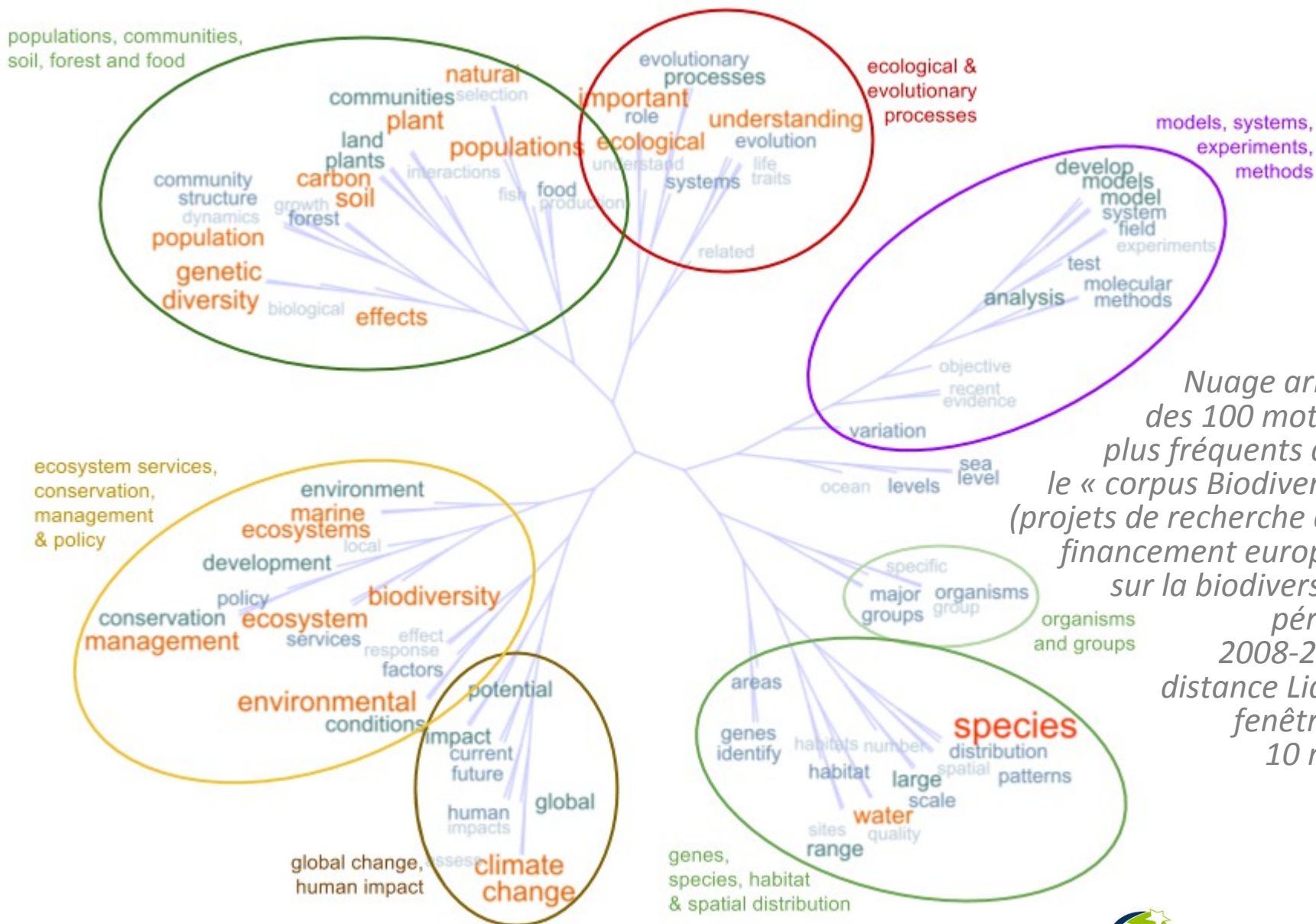


Méthode : comparaison de voisinages dans l'arbre



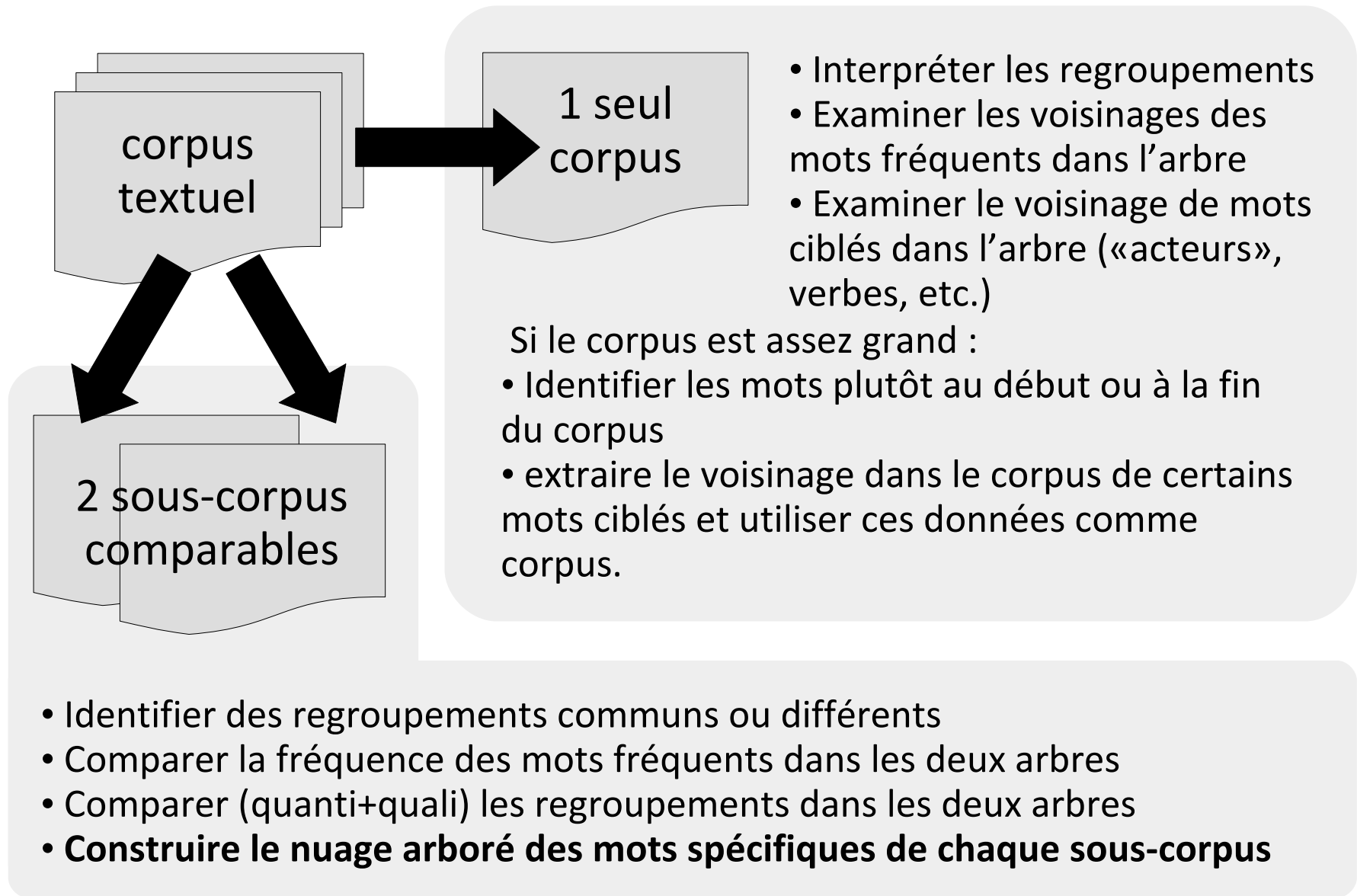
Nuage arboré des 100 mots les plus fréquents dans le « corpus Biodiversa » (projets de recherche avec financement européen sur la biodiversité), période 2004-2007, distance Liddel, fenêtre de 10 mots

Méthode : comparaison de voisinages dans l'arbre



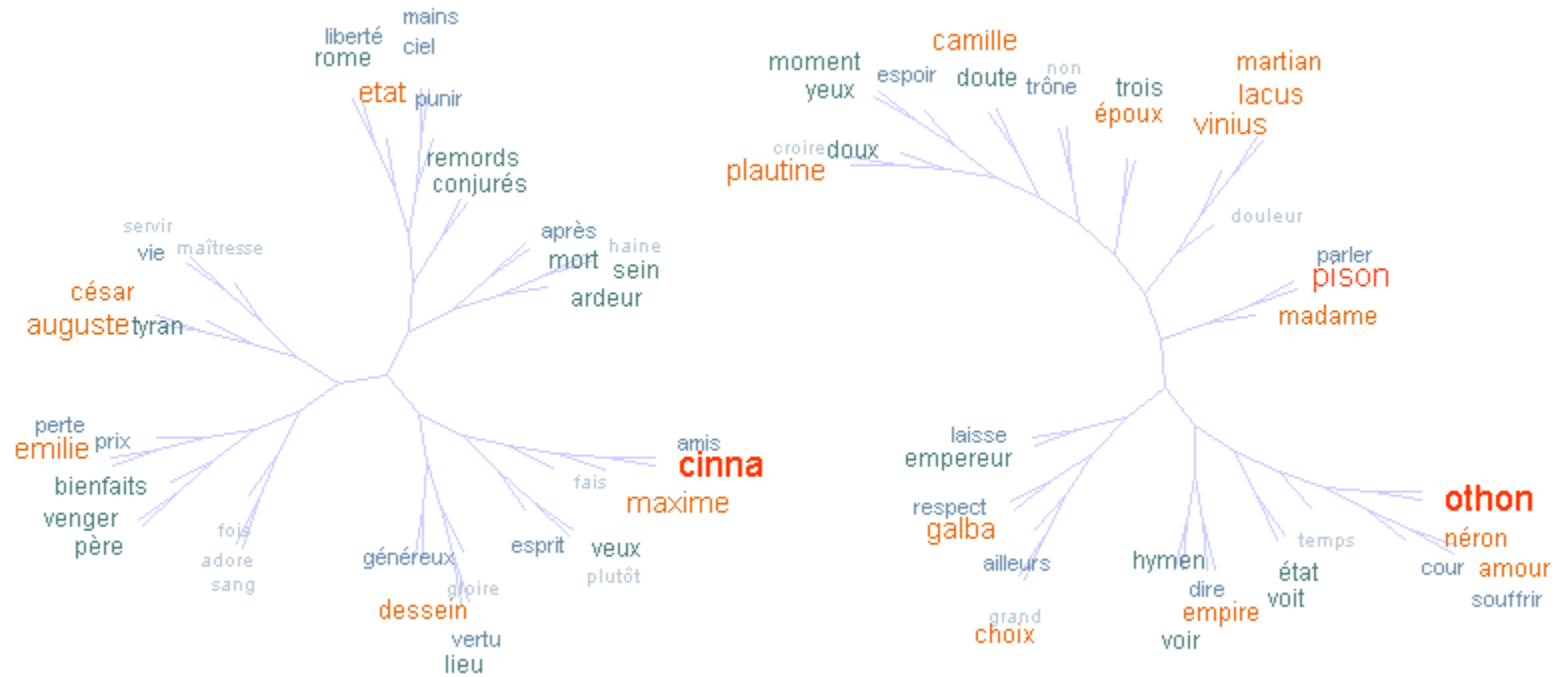
Nuage arboré des 100 mots les plus fréquents dans le « corpus Biodiversa » (projets de recherche avec financement européen sur la biodiversité), période 2008-2011, distance Liddel, fenêtre de 10 mots

Exploration de corpus avec TreeCloud



Méthode : comparaison des spécifiques

Amstutz & Gambette,
JADT 2010



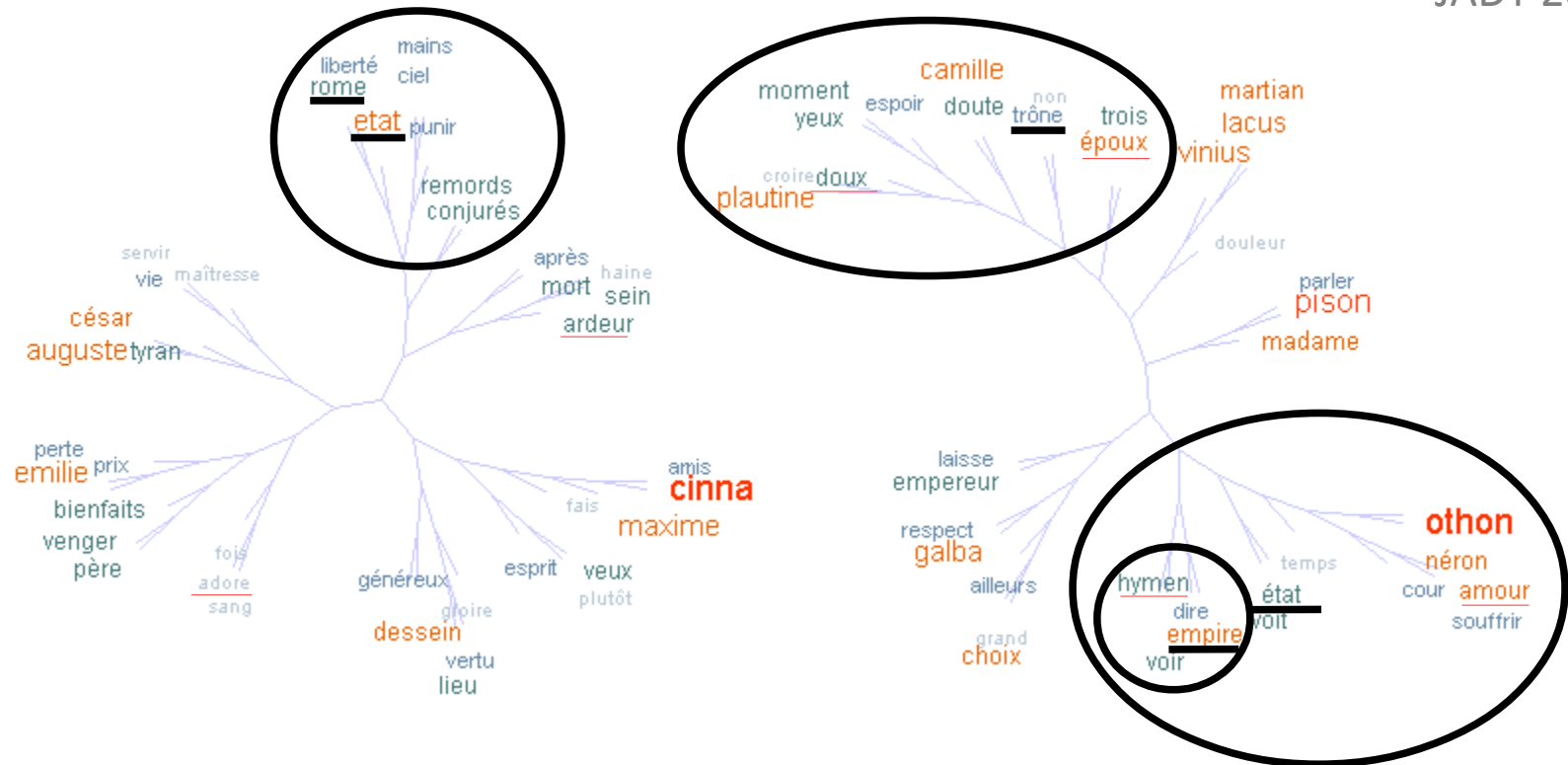
*Nuages arborés des **mots spécifiques** de Cinna et Othon, dimensionnés et colorés d'après leur spécificité calculée dans Lexico3.*



Quels moyens au service de la cause politique ?

Méthode : comparaison des spécifiques

Amstutz & Gambette,
JADT 2010

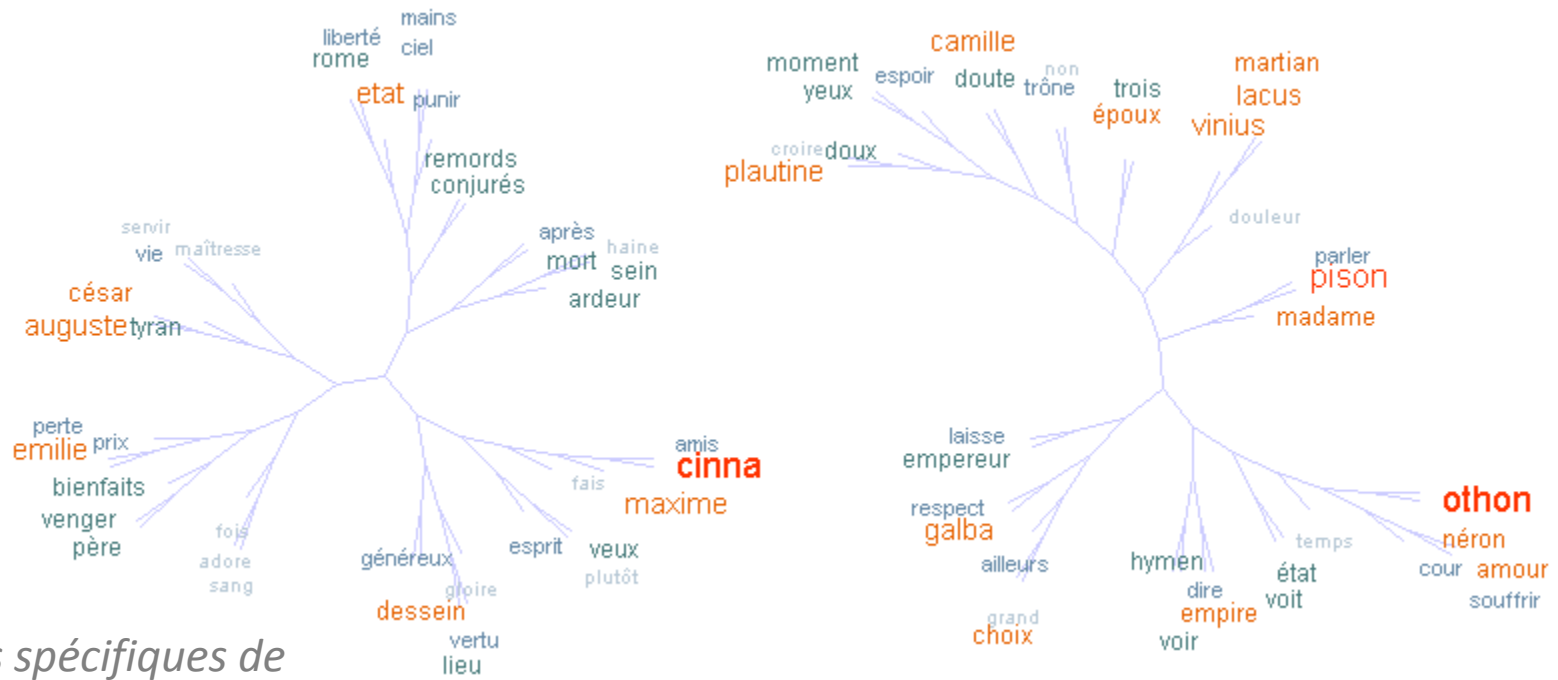


*Nuages arborés des **mots spécifiques** de Cinna et Othon, dimensionnés et colorés d'après leur spécificité calculée dans Lexico3.*



Quels moyens au service de la cause politique ?

Méthode : comparaison des spécifiques



mots spécifiques de Cinna et Othon d'après Lexico3

| | <i>Cinna</i> | <i>Othon</i> |
|--|--------------------|---|
| 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 |

Comparaison avec d'autres visualisations

nuage arboré
(TreeCloud)



réseau de mots
(PhraseNet d'IBM ManyEyes)

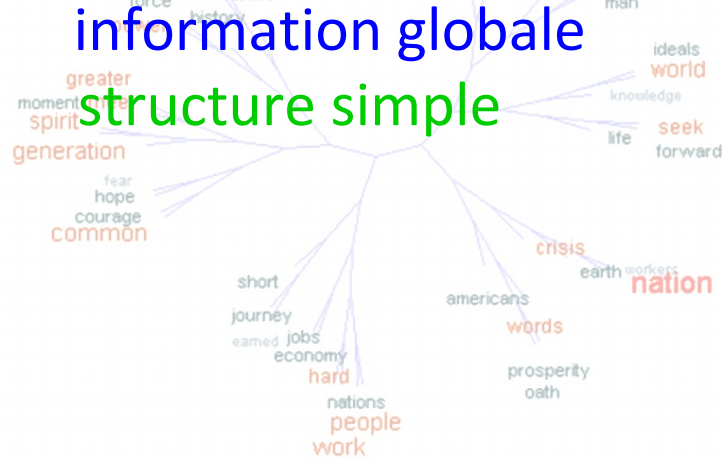


projection des mots (Astartex)



Comparaison avec d'autres visualisations

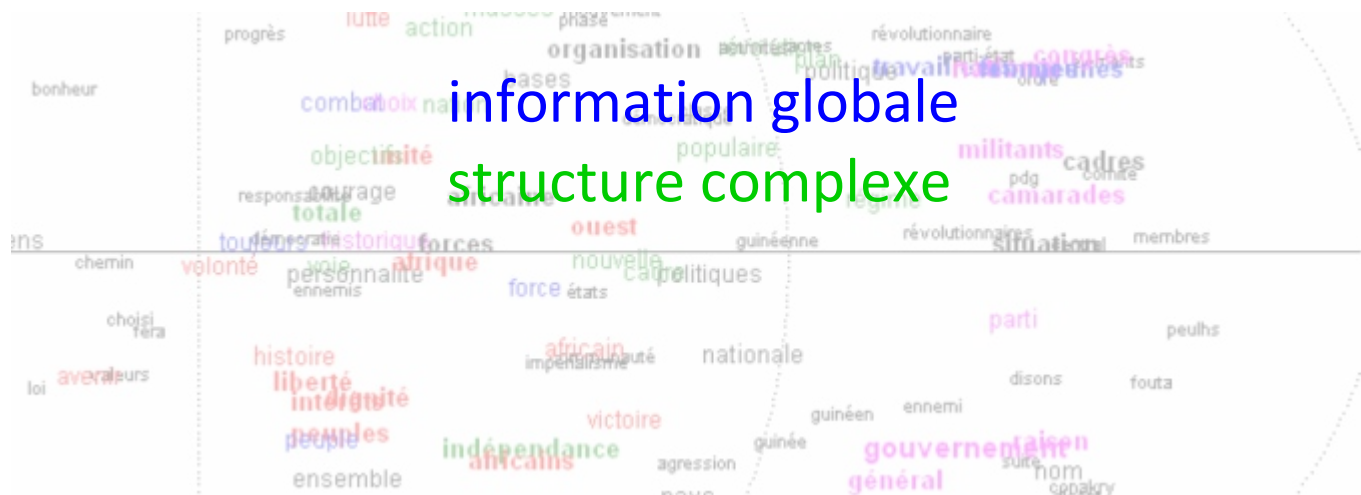
nuage arboré (TreeCloud)



réseau de mots (PhraseNet d'IBM ManyEyes)



projection des mots (Astartex)



Comparaison avec d'autres visualisations

nuage arboré (TreeCloud)



réseau de mots (PhraseNet d'IBM ManyEyes)

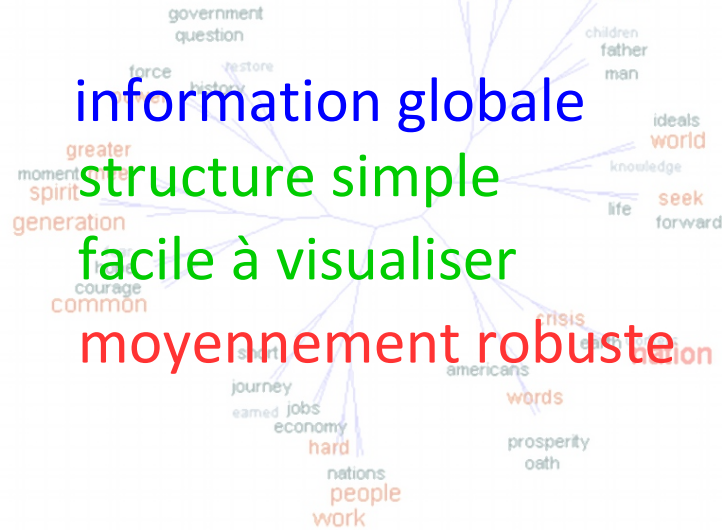


projection des mots (Astartex)



Comparaison avec d'autres visualisations

nuage arboré (TreeCloud)



information globale

structure simple

facile à visualiser

moyennement robuste

réseau de mots (PhraseNet d'IBM ManyEyes)



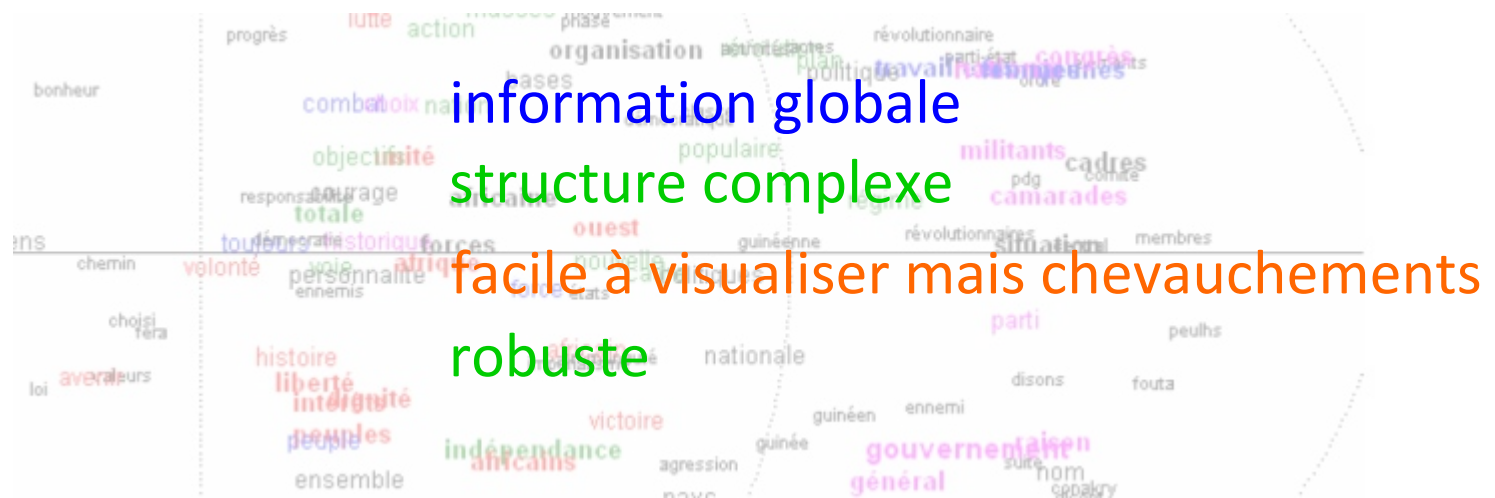
information locale

structure complexe

difficile à visualiser

robuste

projection des mots (Astartex)



information globale

structure complexe

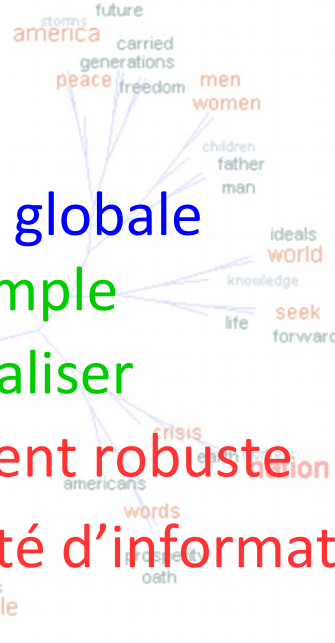
facile à visualiser mais chevauchements

robuste

Comparaison avec d'autres visualisations

nuage arboré (TreeCloud)

information globale
structure simple
facile à visualiser
moyennement robuste
faible densité d'information



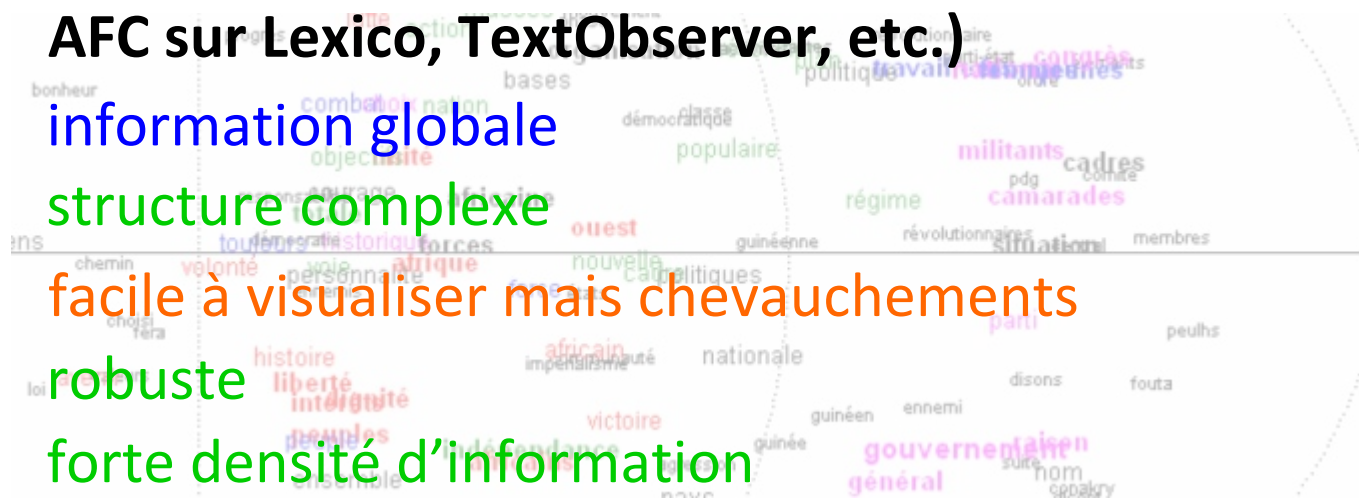
réseau de mots (PhraseNet d'IBM ManyEyes, Tropes)

information locale
structure complexe
difficile à visualiser
robuste
forte densité d'information



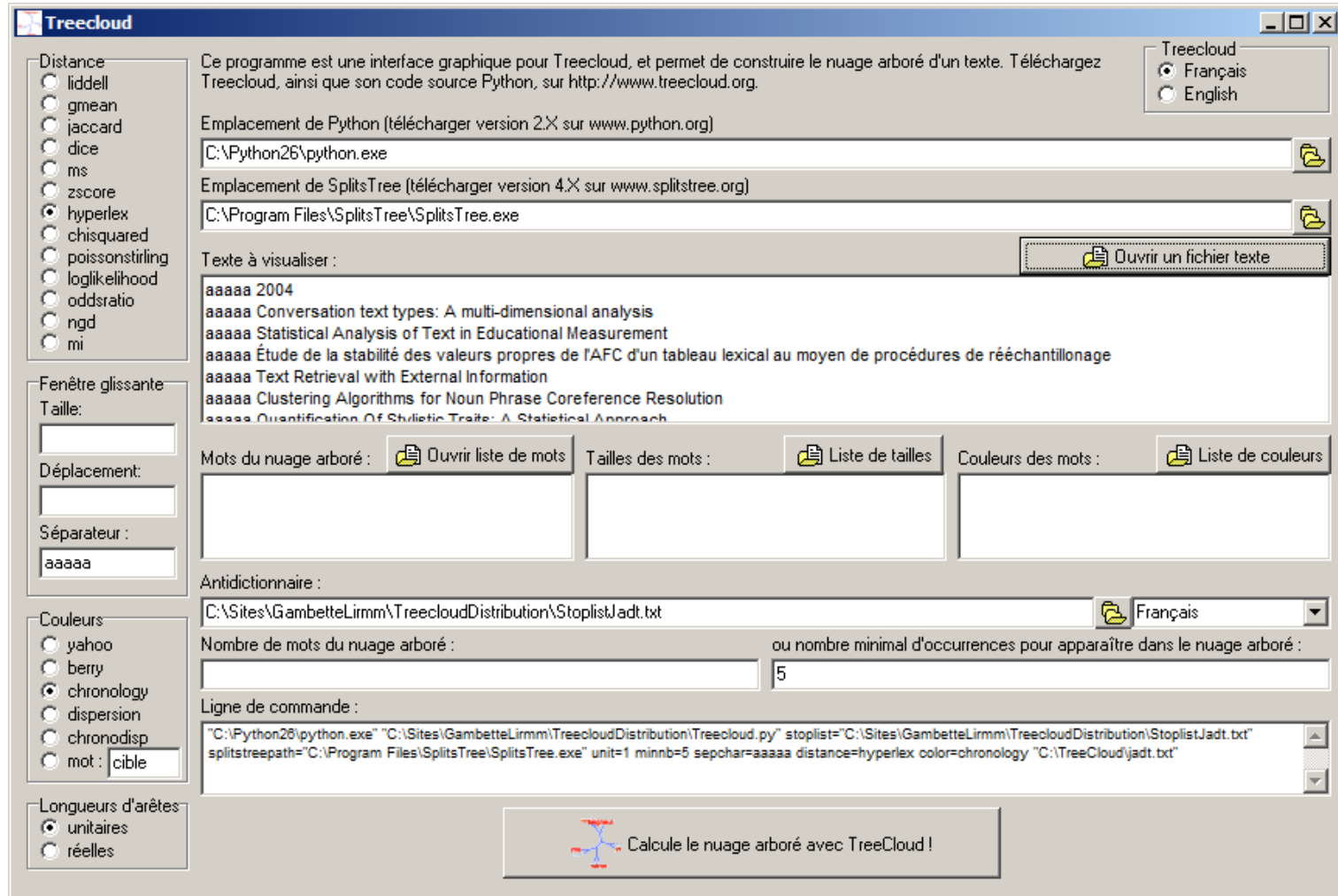
projection des mots (Astartex, AFC sur Lexico, TextObserver, etc.)

information globale
structure complexe
facile à visualiser mais chevauchements
robuste
forte densité d'information



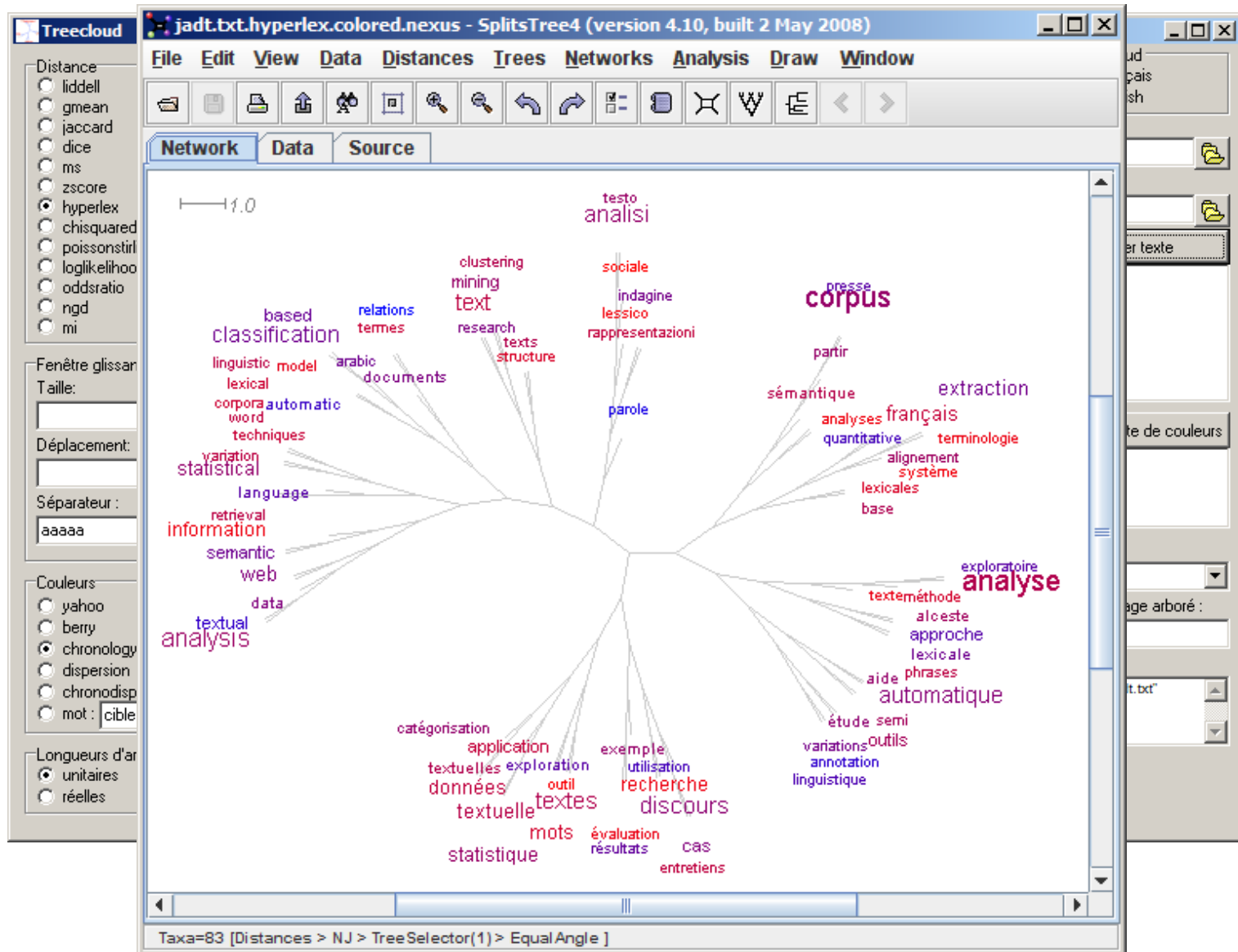
Implémentations

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



Implémentations

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



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 proximity inside the text. The first tree cloud appeared on [Jean Véronis's blog](#) in December 2007, you can now [create your own with this website](#), or [with the TreeCloud software](#).

Create your own tree cloud online!

Ce site web vous permet de générer des **nuages arborés** à partir d'un texte, c'est à dire des nuages de mots disposés autour d'un arbre qui indique leur proximité dans le texte. Le premier nuage arboré est apparu sur le [blog de Jean Véronis](#) en décembre 2007, vous pouvez maintenant [créer les vôtres avec ce site web](#), ou [avec le logiciel TreeCloud](#).

Créez vos propres nuages arborés en ligne !

Documents :



If you use TreeCloud or this website, please cite www.treecloud.org or: Philippe Gambette et Jean Véronis: *Visualising a Text with a Tree Cloud*, In Locarek-Junge H. and Weihs C., editors, *Classification as a Tool of Research, Proc. of IFC'S'09 (11th Conference of the International Federation of Classification Societies)*, to appear, 2010 ([supplementary material](#)).

Pour des exemples d'utilisation de la visualisation en nuage arboré, vous pouvez lire :

Delphine Amstutz et Philippe Gambette: *Utilisation de la visualisation en nuage arboré pour l'analyse littéraire*, *Proc. of JADT'10 (10th International Conference on statistical analysis of textual data)*, à paraître, 2010 ([matériel supplémentaire](#)).



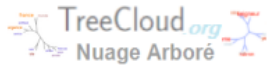
www.treecloud.org

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



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

www.treecloud.org

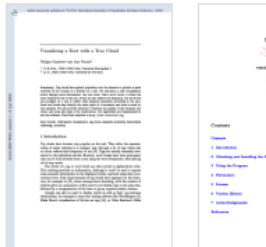
This website helps you to generate tree cloud words are arranged on a tree which reflects The first tree cloud appeared on [Jean Véronis](#) create your own with this website, or with t

Create your own tree cloud online

Ce site web vous permet de générer des 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. of Societies](#)), to appear, 2010 ([supplémentaire](#))

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



Créer! Téléchargements Galerie A propos FAQ

Créez vos propres nuages arborés !

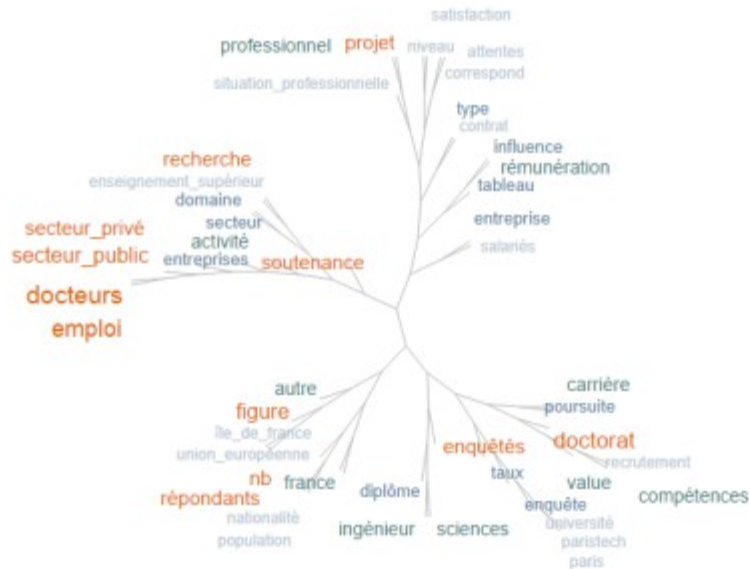
Collez votre texte dans le cadre ci-dessous, puis cliquez sur *Envoyer* ! Attention, l'utilisateur suivant verra votre texte quand il se connectera au site, si vous ne voulez pas faire apparaître vos textes, installez plutôt [TreeCloud](#) sur votre machine.

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



TreeCloud.org
Nuage Arboré

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

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 [this website](#)

Create your own tree cloud online

Le site web vous permet de générer des nuages de mots disposés autour d'un arbre. Le premier nuage arboré est apparu sur le site [http://www.adoc-tm.com/2013rapport.pdf](#) vous pouvez maintenant [créer les vôtres avec ce site](#)

Créez vos propres nuages arborés

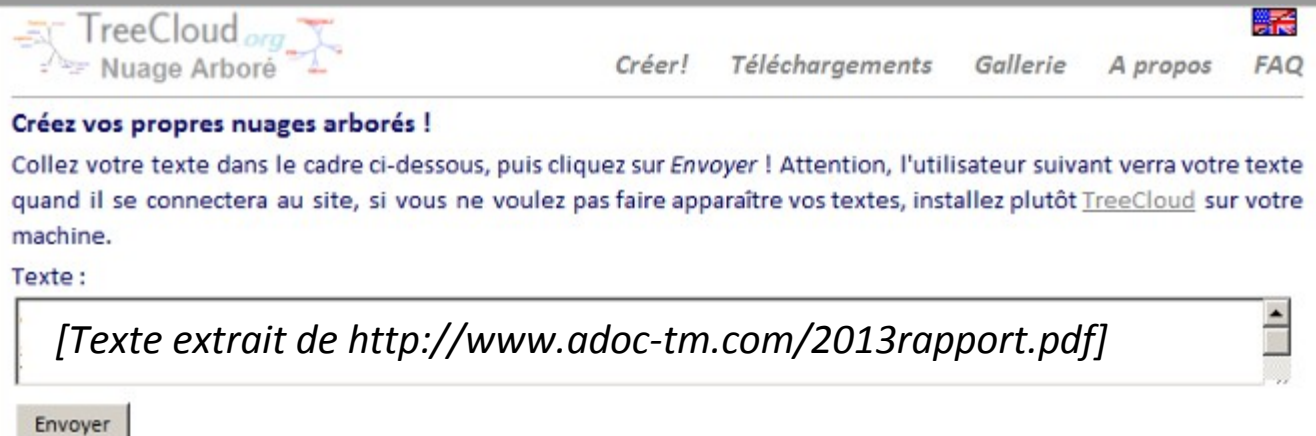
Documents :



If you use TreeCloud or this website, please contact Philippe Gambette et Jean Véronis: [Visual Classification as a Tool of Research, Proc. of the 10th International Conference on Visual Information Systems \(VISIS 2010\)](#), to appear, 2010 ([supplémentaire](#))

Pour des exemples d'utilisation de la visualisation de mots, voir Delphine Amstutz et Philippe Gambette: [Utah at the JADT'10 \(10th International Conference on Visual Information Systems\)](#) ([supplémentaire](#)).

www.treecloud.org



TreeCloud.org
Nuage Arboré

Créer! Téléchargements Galerie A propos FAQ

Créez vos propres nuages arborés !

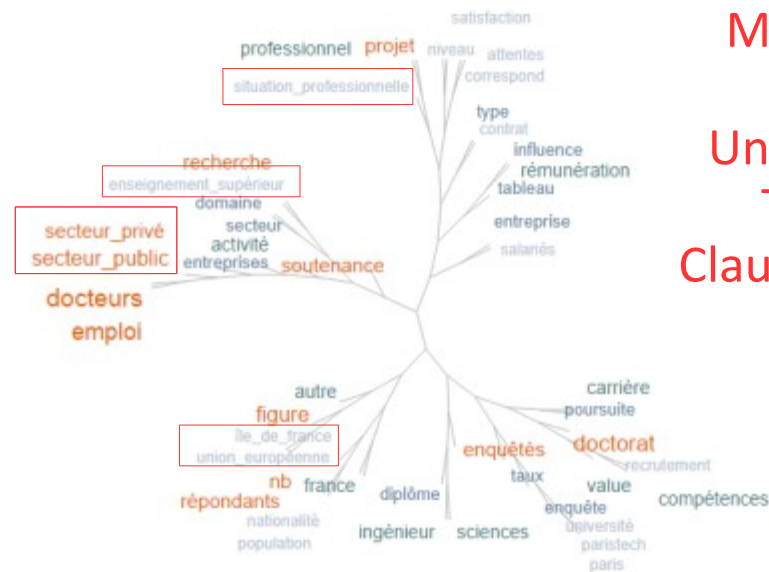
Collez votre texte dans le cadre ci-dessous, puis cliquez sur *Envoyer* ! Attention, l'utilisateur suivant verra votre texte quand il se connectera au site, si vous ne voulez pas faire apparaître vos textes, installez plutôt [TreeCloud](#) sur votre machine.

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.



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

Interface web



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

www.treecloud.org

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 structure. 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 arborés** des nuages de mots disposés autour d'un arbre qui indique leur structure sémantique. 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

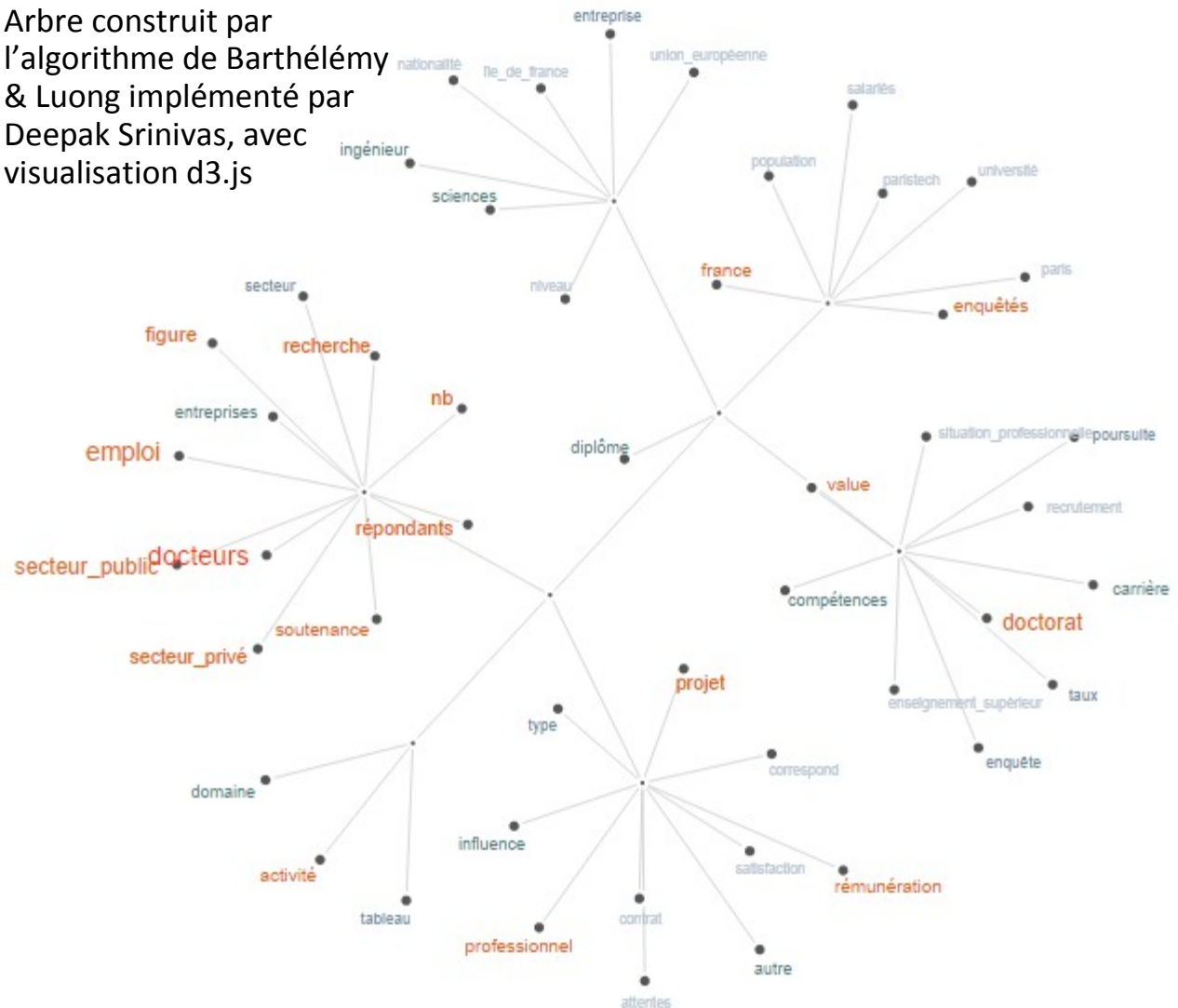
Documents :



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 Informatics in Society), to appear, 2010 ([supplementary material](#)).

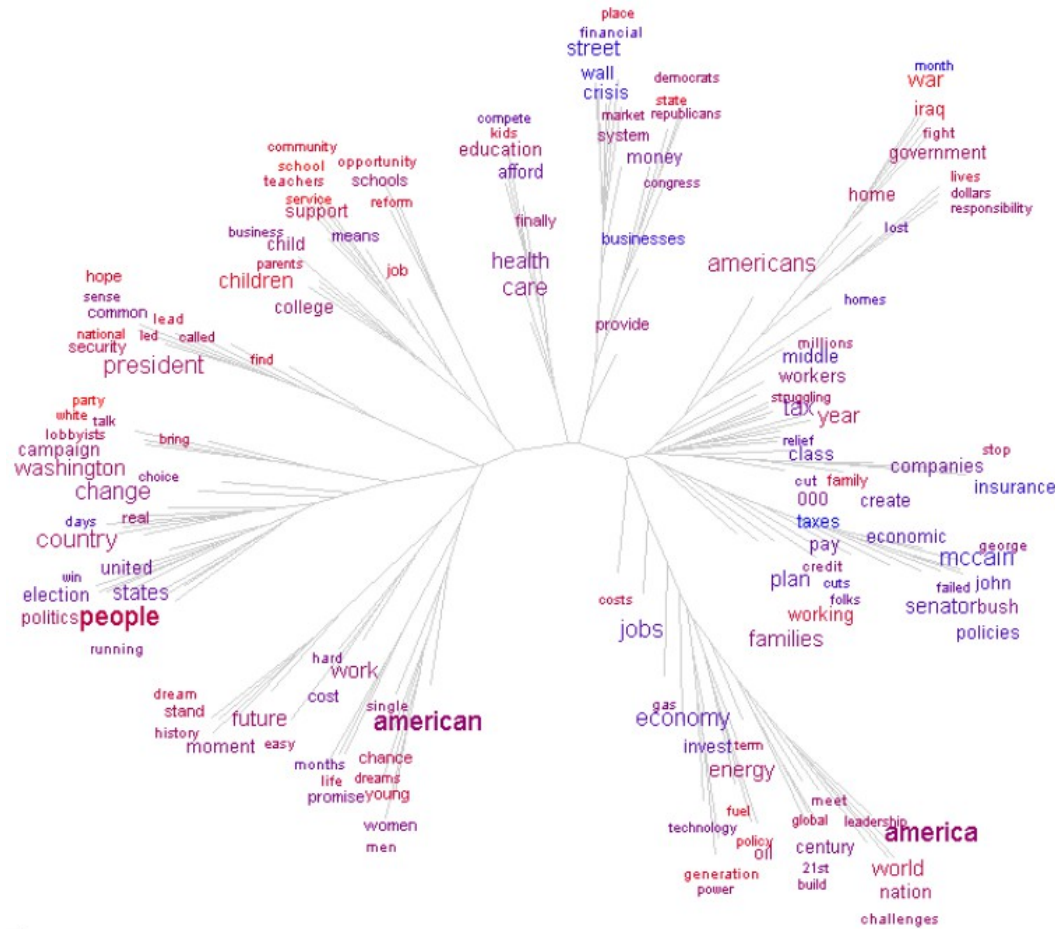
Pour des exemples d'utilisation de la visualisation en nuage arboré, voir Delphine Amstutz et Philippe Gambette: *Utilisation de la visualisation en nuage arboré*, Proc. of *JADT'10* (10th International Conference on Statistical Data Analysis in Text and Information Systems), to appear, 2010 ([supplémentaire](#)).

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



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)

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>

Co-auteurs des travaux en cours :

- Edna Hernandez : méthodologie d'utilisation de TreeCloud pour les analyses exploratoires
- Claude Martineau : intégration de prétraitements Unitex dans TreeCloud
- Deepak Srinivas : visualisation avec bibliothèque d3.js