

Masters Intelligence stratégique, analyse des risques et territoires &  
Influence, Lobbying et Médias Sociaux (ILMS)  
06/03/2025 – 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

# **Panorama des logiciels de textométrie**

# Quelques logiciels de textométrie

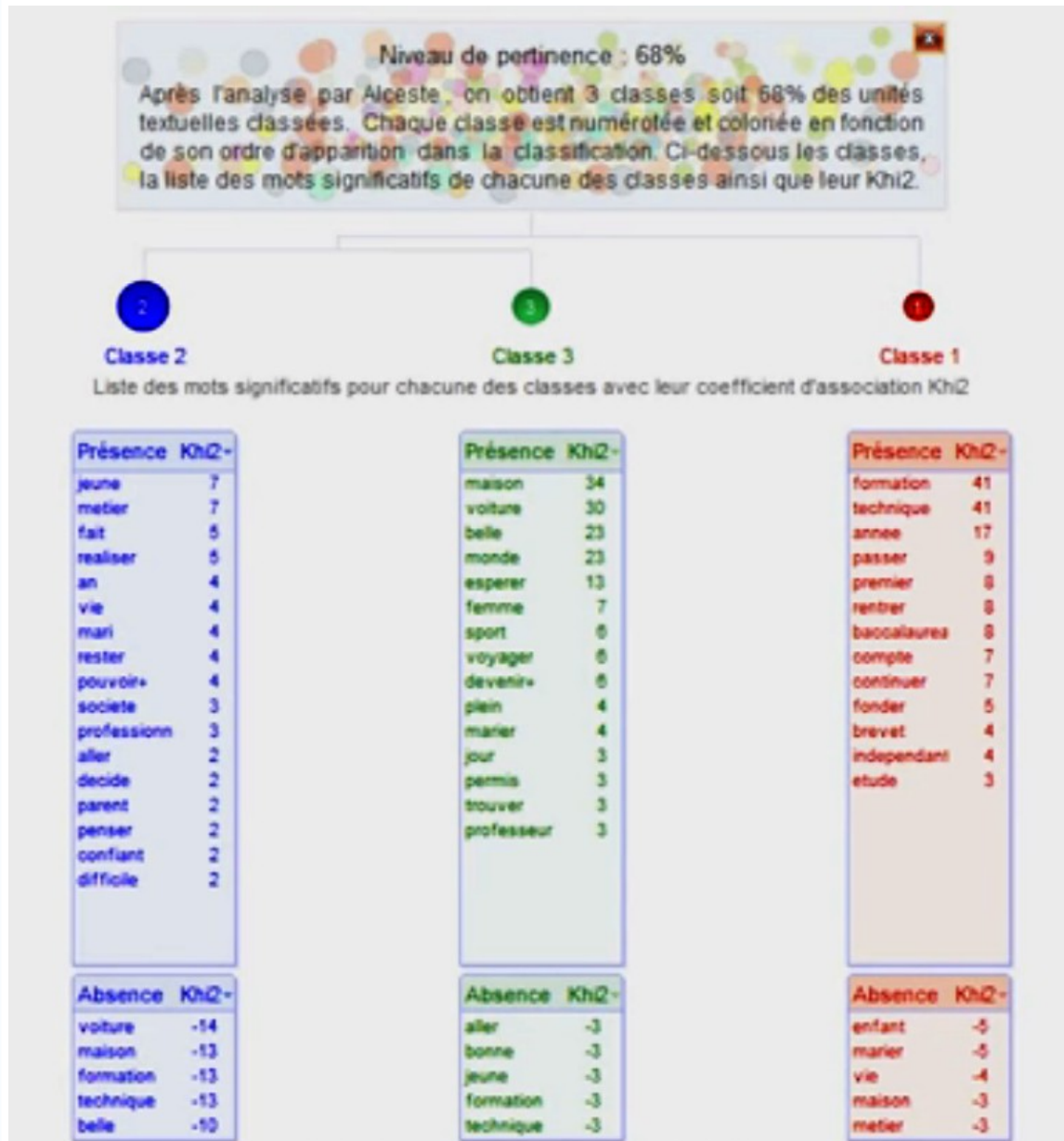
## Alceste (depuis 1983)



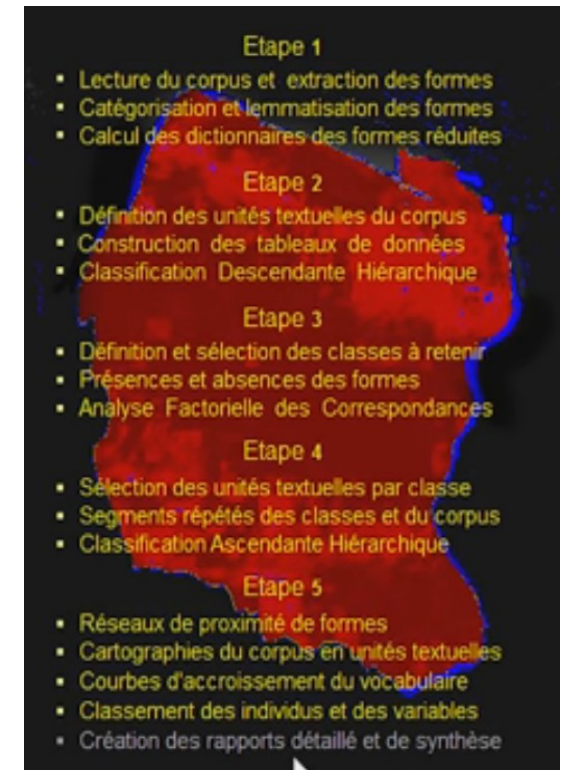
Société IMAGE

<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>

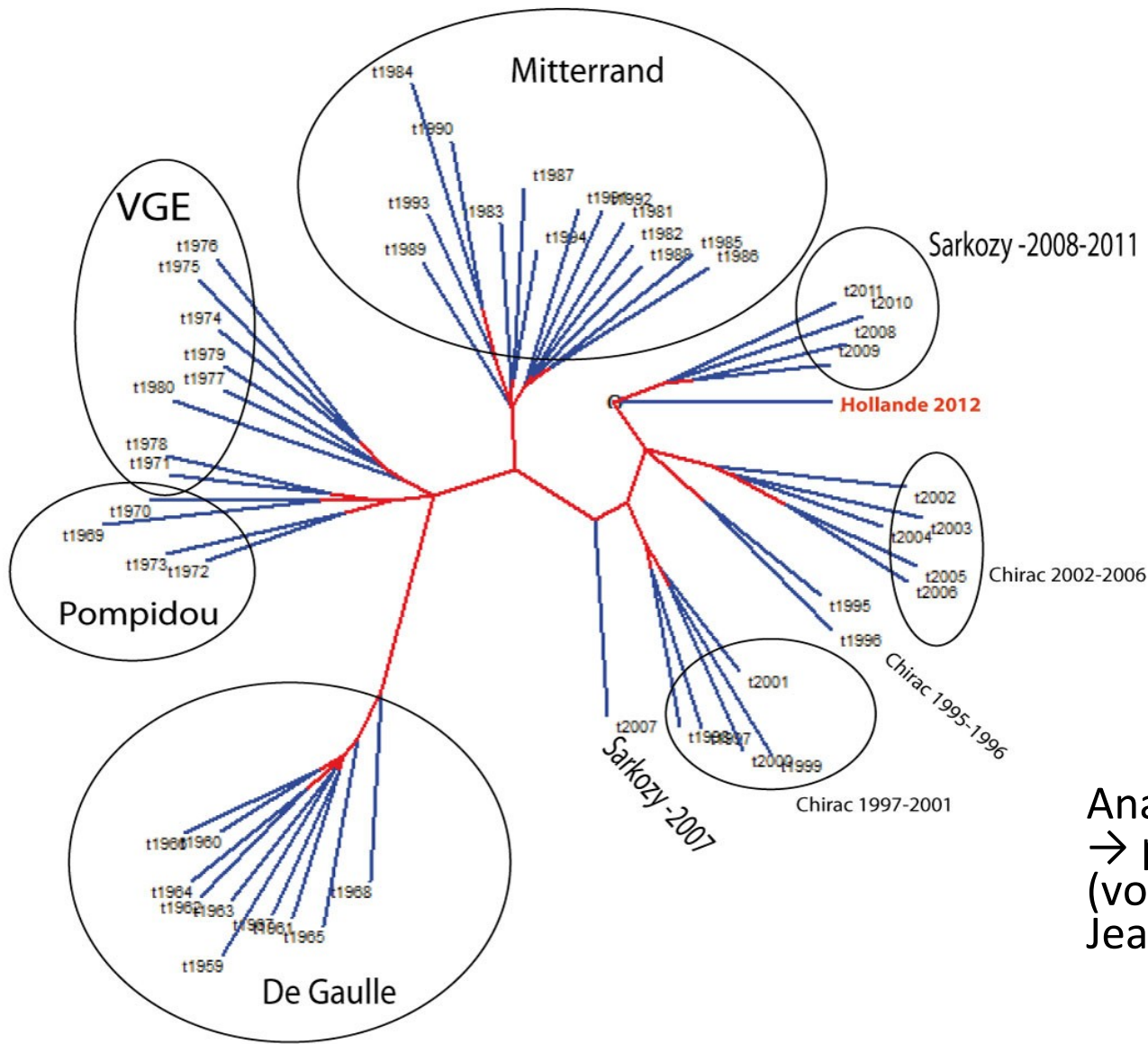
## Alceste (depuis 1983)



Société IMAGE

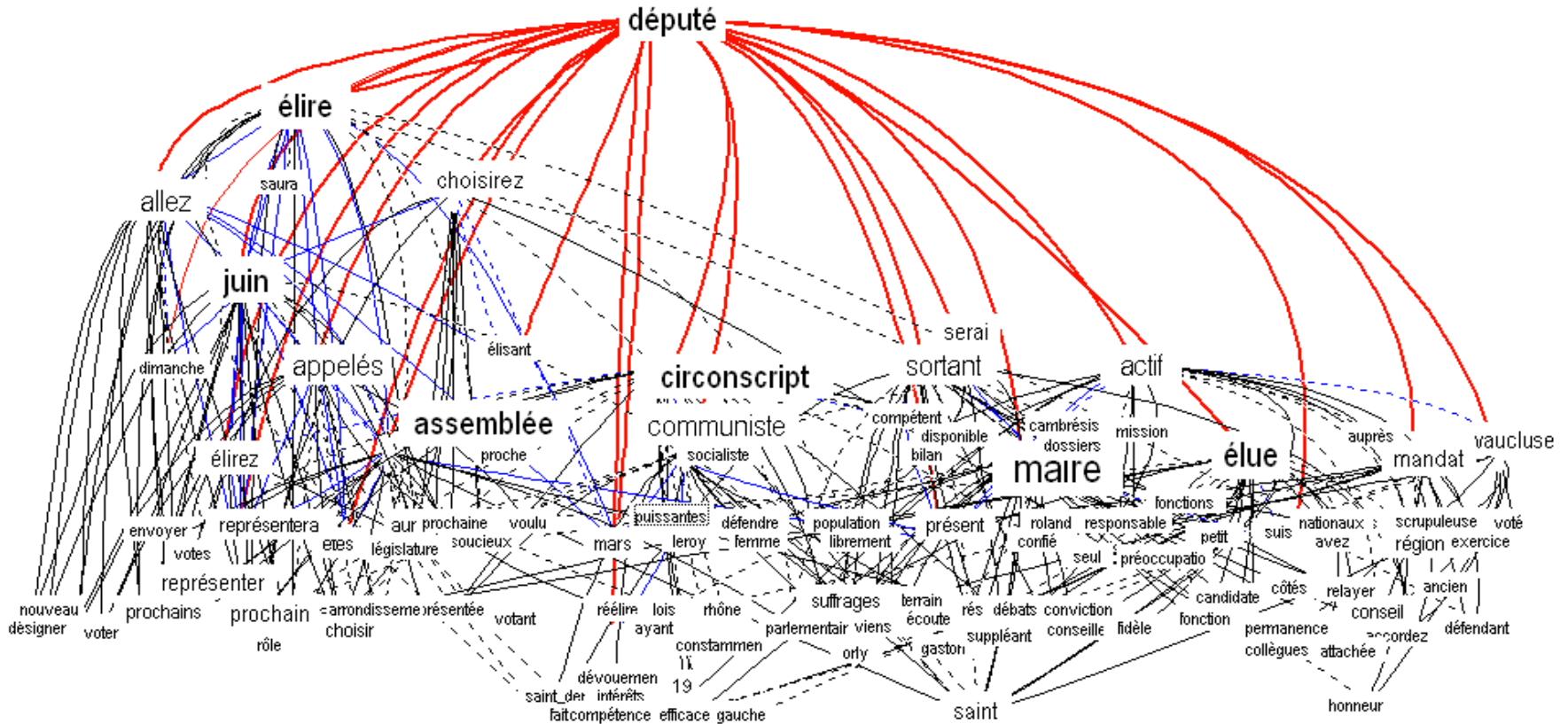
<http://www.image-zafar.com/Logiciel.html>

# Hyperbase



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

# Hyperbase



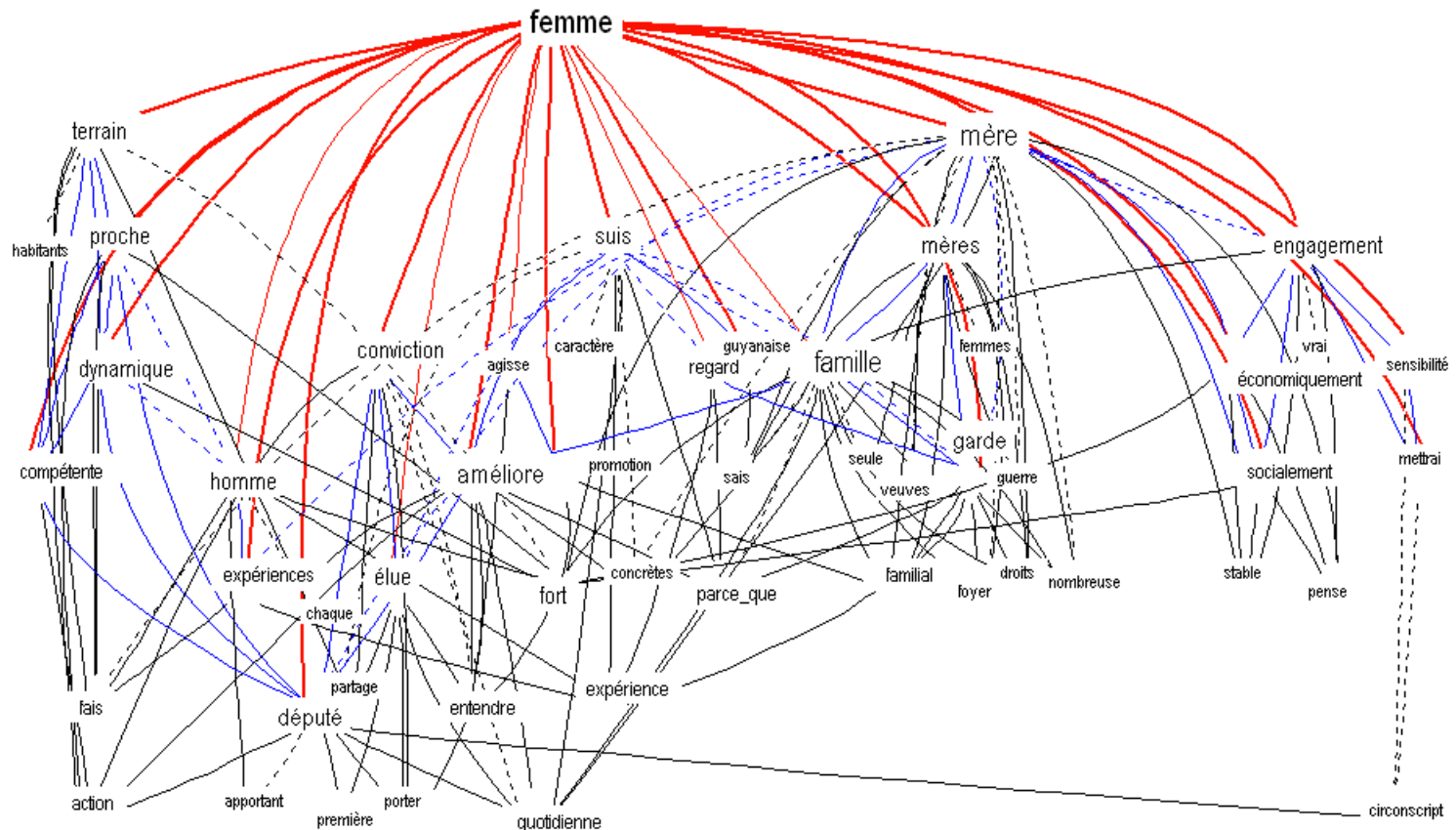
## Réseau de cooccurrents d'un mot

→ graphique des co-occurents directs et indirects du mot-pôle « député » dans l'ensemble du corpus de Professions de foi

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>

# 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

Lexico3 - [Section - Délimiteurs : \$ - vue n°1]

Fichier Traitement Fenêtre

Navigation Rapport Dictionnaire Segments répétés

9 <Aucune>

| 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  |

50  
100  
150  
200  
250  
300  
350  
400  
450  
500  
550  
600

## Segments répétés et carte des sections

Section :

Occurrence :

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 tondus de la "vendée qui ont la patte graissée pour égorger tous les bons citoyens . <edito=1>\$

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

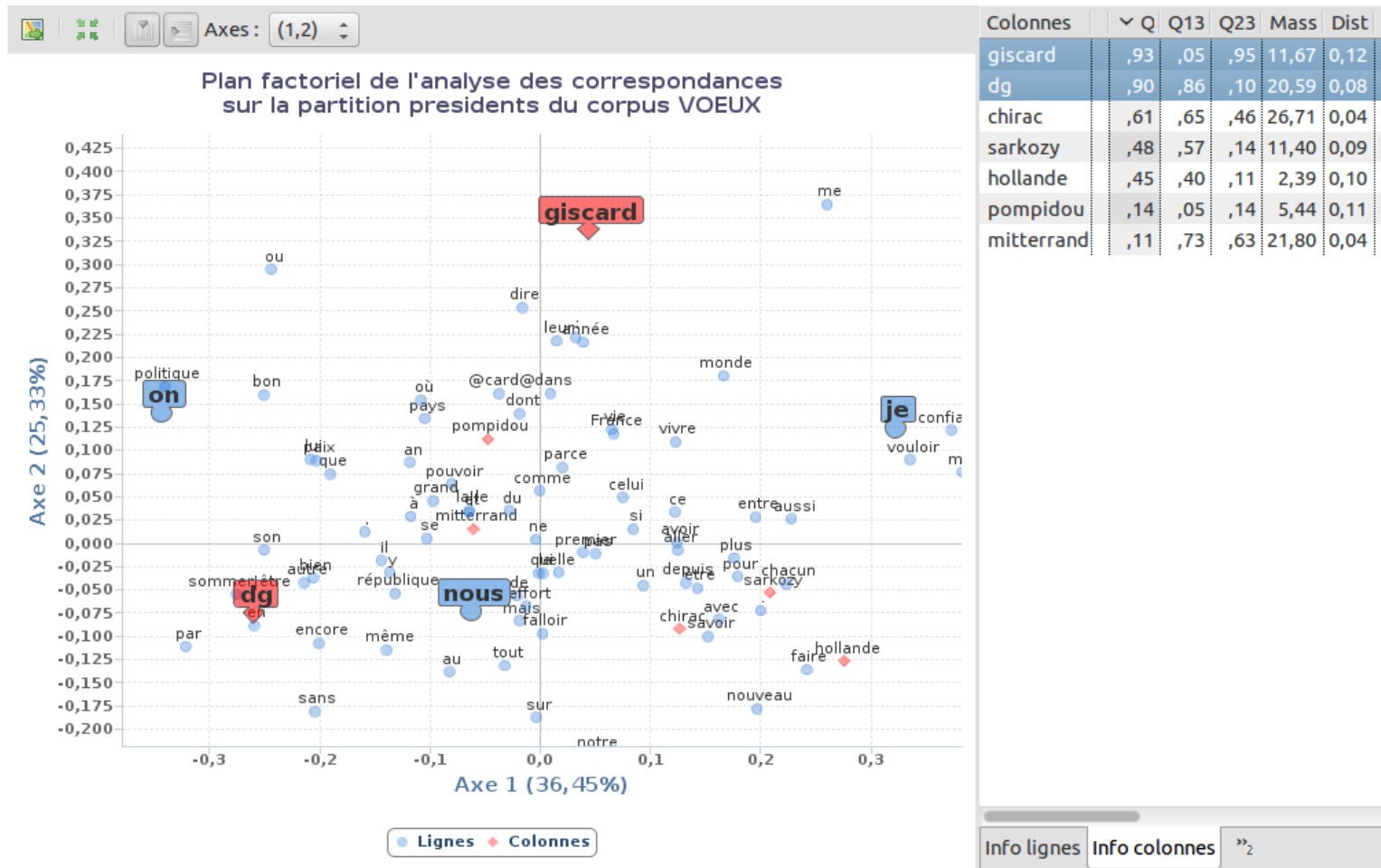
<http://www.image-zafar.com/Logiciel.html>

## 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 voeux présidentiels

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

# Quelques logiciels de textométrie

## Hyperbase (depuis 1989)



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

## Lexico (depuis 1990)



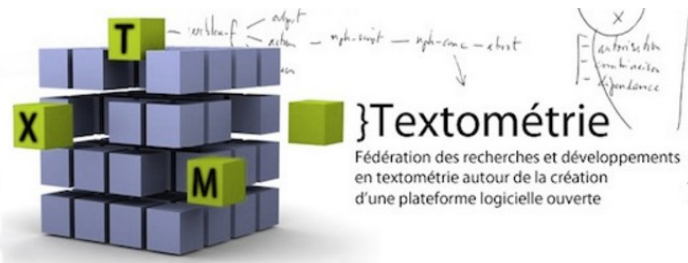
Université Sorbonne nouvelle  
<http://lexi-co.com/>

## Alceste (depuis 1983)



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

## TXM (depuis 2008)



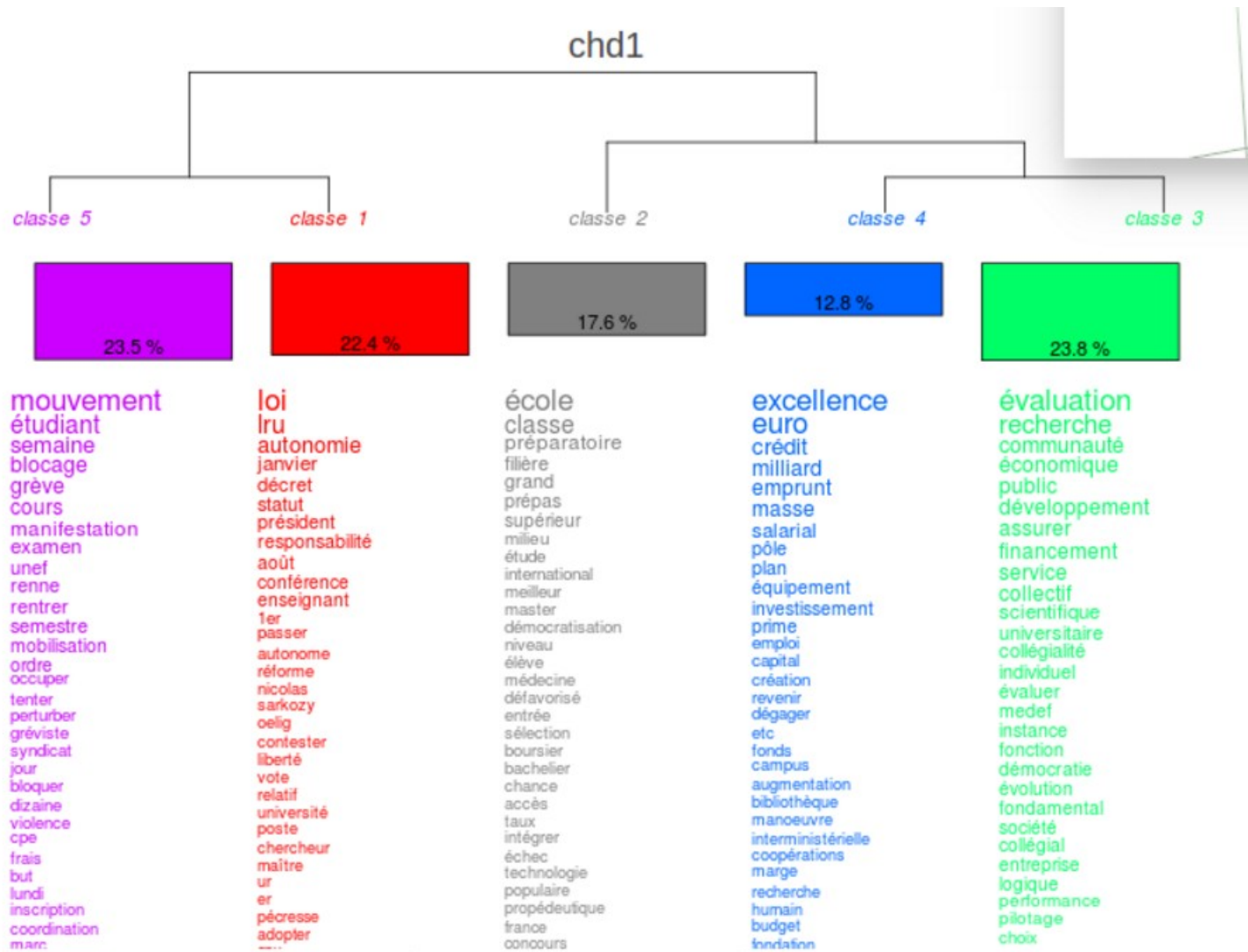
ENS Lyon  
<http://textometrie.ens-lyon.fr/>

## Iramuteq (depuis 2009)

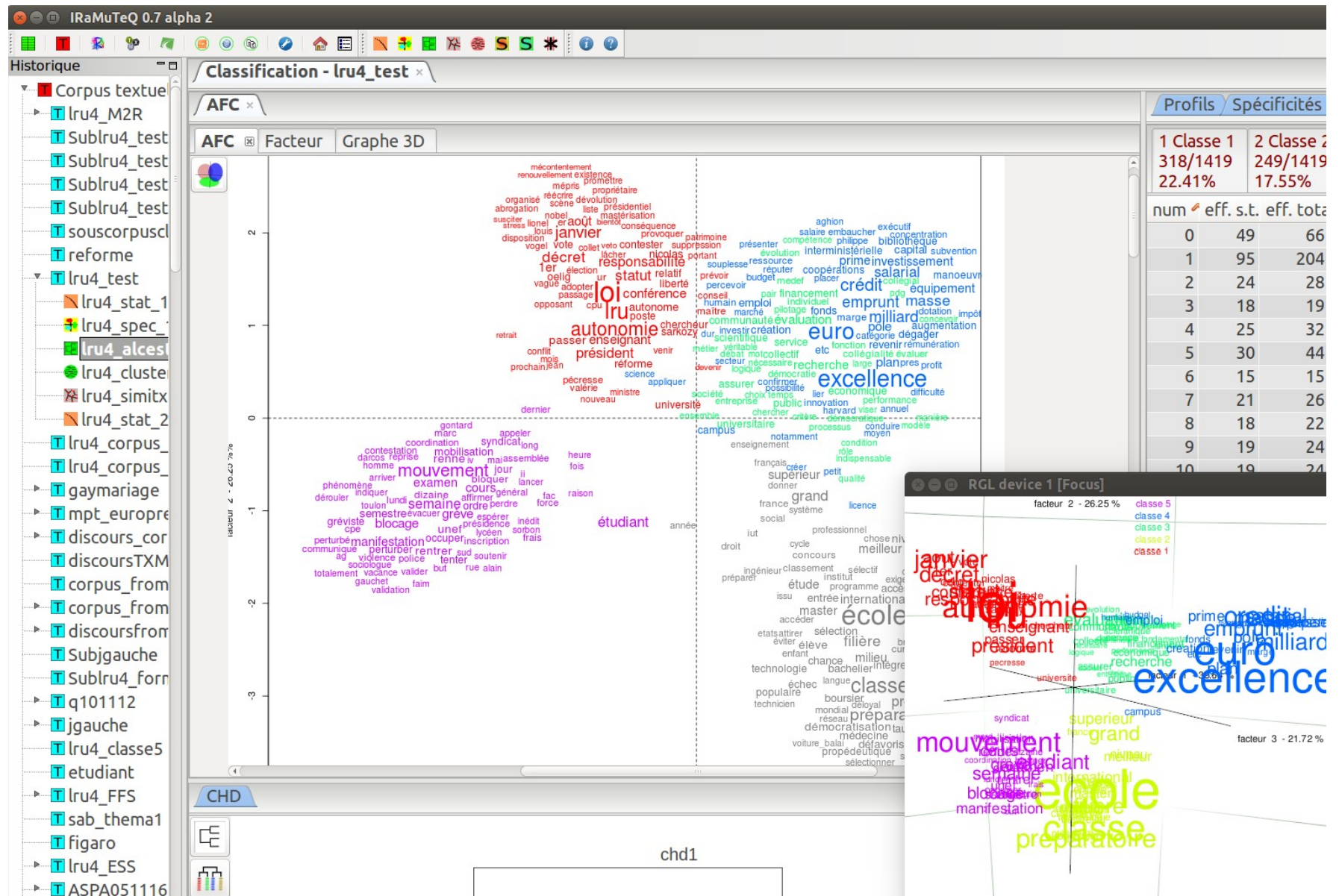


Université de Toulouse  
<http://www.iramuteq.org/>

# Iramuteq



# Iramuteq



# Quelques logiciels de textométrie

## Hyperbase (depuis 1989)



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

## Lexico (depuis 1990)



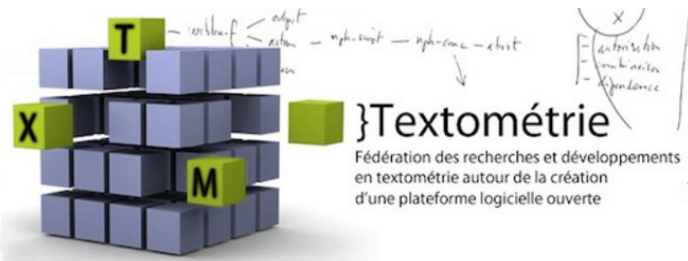
Université Sorbonne nouvelle  
<http://lexi-co.com/>

## Alceste (depuis 1983)



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

## TXM (depuis 2008)



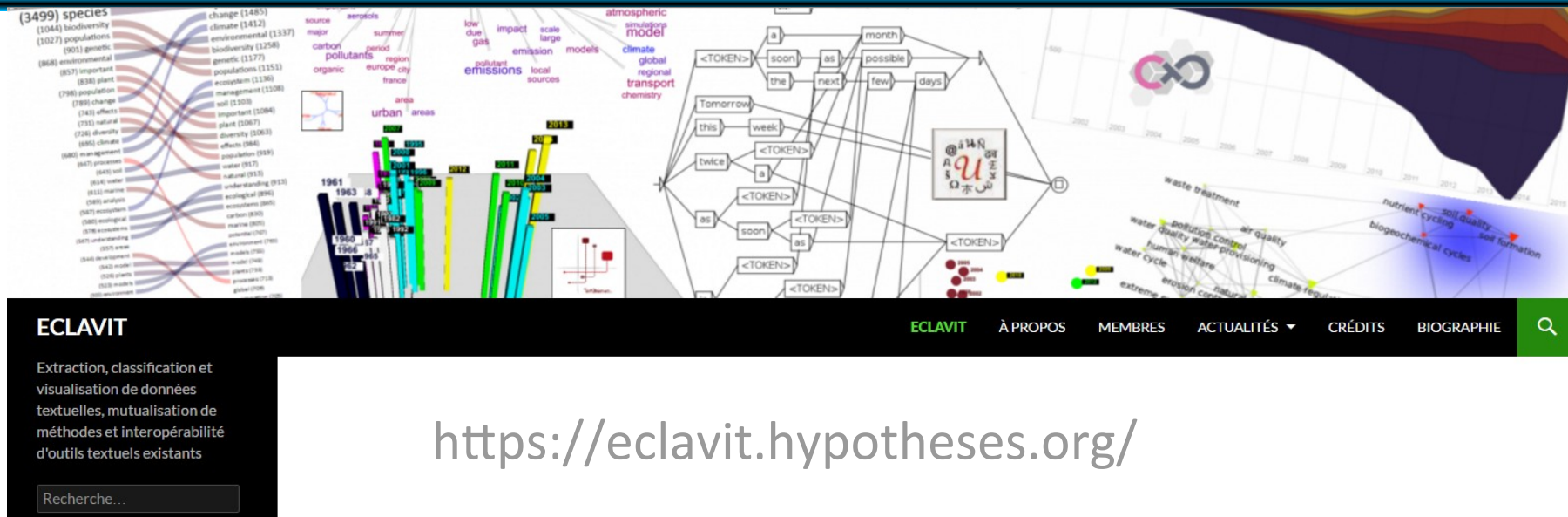
ENS Lyon  
<http://textometrie.ens-lyon.fr/>

## Iramuteq (depuis 2009)



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

# Logiciels d'analyse textuelle à Université Paris-Est



<https://eclavit.hypotheses.org/>

## Logiciels développés à Université Paris-Est :

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

# 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)

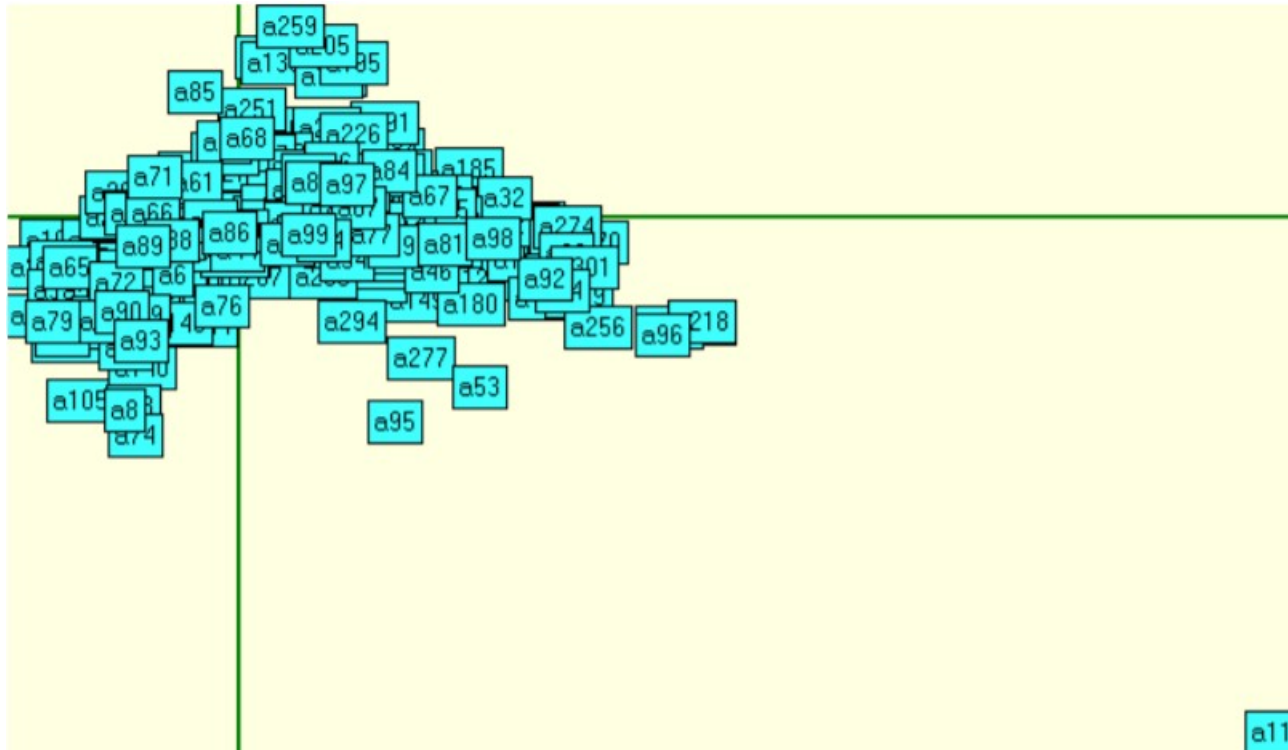
**<annee=2015> <type=article> <doc=a1>** background: lateral, or horizontal, gene transfers are a type of asymmetric evolutionary events where genetic material is transferred from one species to another. in this paper we consider lgt networks, a general model of phylogenetic networks with lateral gene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

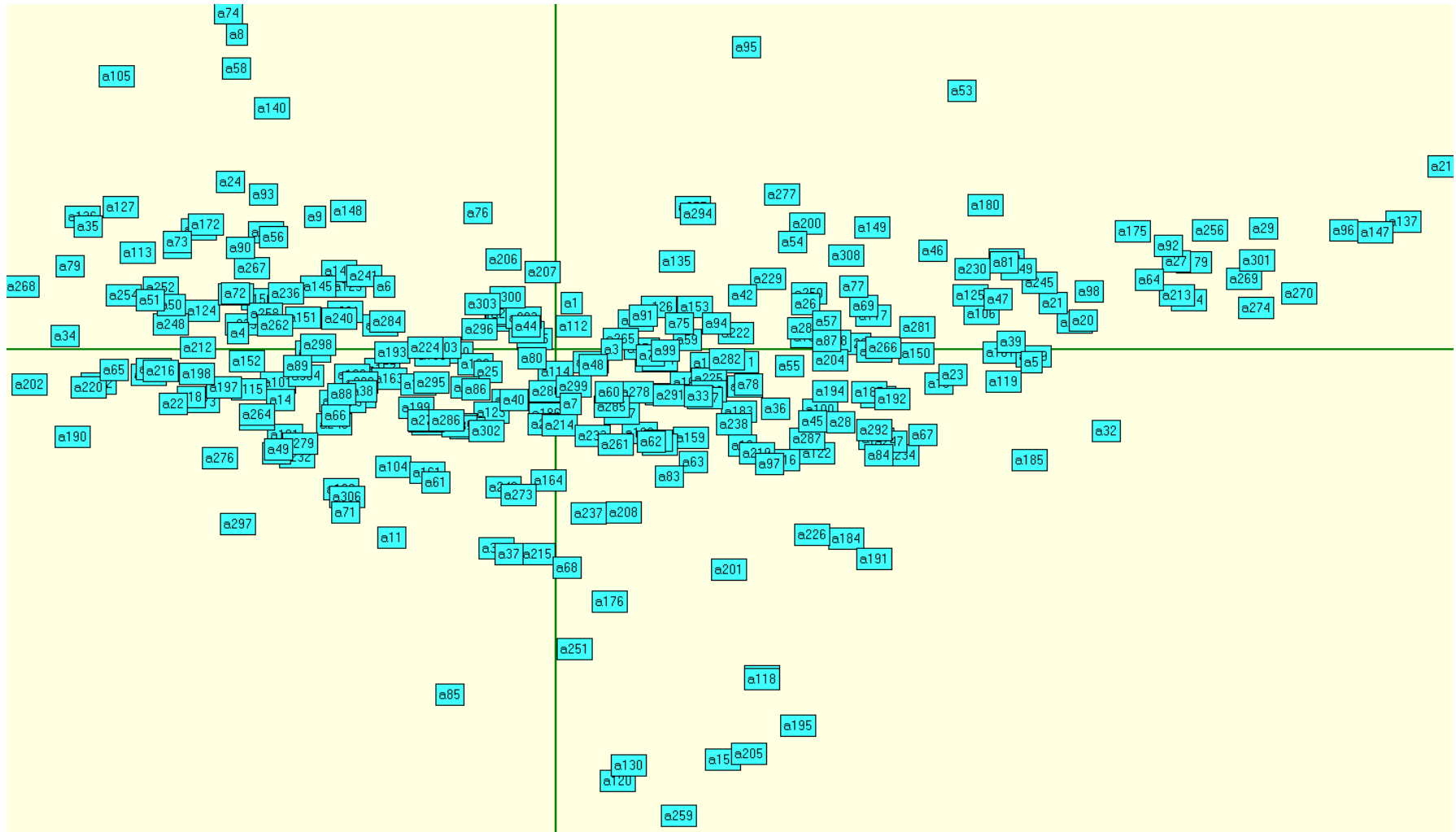
Etape 2) Analyse factorielle des correspondances





# 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    |
| subtree      | 607     | 222       | -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    |
| statistical  | 607      | 222       | -18    |

# Une parenthèse sur la spécificité

---

**Comparer des nombres d'occurrences ?**

Mot présent 10 fois dans le corpus 1 et 5 fois dans le corpus 2

A-t-il plus d'importance dans le corpus 1 que dans le corpus 2 ?

# Une parenthèse sur la spécificité

---

## Comparer des nombres d'occurrences ?

Mot présent 10 fois dans le corpus 1 et 5 fois dans le corpus 2

A-t-il plus d'importance dans le corpus 1 que dans le corpus 2 ?

Le nombre d'occurrences ne suffit pas :  
il peut s'expliquer par différences de tailles de sous-corpus  
(par exemple, corpus 1 deux fois plus grand que le corpus 2)

# Une parenthèse sur la spécificité

---

## Comparer des fréquences ?

Mot présent à une fréquence de :

- 2% dans le corpus 1
- 1% dans le corpus 2

Deux fois plus présent dans le corpus 1

A-t-il plus d'importance dans le corpus 1 que dans le corpus 2 ?

# Une parenthèse sur la spécificité

## Comparer des fréquences ?

Mot présent à une fréquence de :

- 2% dans le corpus 1
- 1% dans le corpus 2

Deux fois plus présent dans le corpus 1

A-t-il plus d'importance dans le corpus 1 que dans le corpus 2 ?

La différence de fréquences ne suffit pas,  
elle pourrait être due au hasard !

Exemple : si le corpus 1 et le corpus 2 contiennent chacun 100 mots ,  
seulement une occurrence d'écart !

# Une parenthèse sur la spécificité

---

Modèle **hypergéométrique** :

on imagine que les mots sont jetés au hasard dans les sous-corpus, avec des probabilités proportionnelles à la taille de chacun.

Si corpus 1 et corpus 2 de même taille (100 mots chacun), un mot présent 3 fois dans la totalité du corpus :

# Une parenthèse sur la spécificité

Modèle **hypergéométrique** :

on imagine que les mots sont jetés au hasard dans les sous-corpus, avec des probabilités proportionnelles à la taille de chacun.

Si corpus 1 et corpus 2 de même taille (100 mots chacun), un mot présent 3 fois dans la totalité du corpus :

- 1 chance sur 2 que la première occurrence soit dans le corpus 1
- 1 chance sur 4 que les 2 premières occurrences soient dans le corpus 1
- 1 chance sur 8 que les 3 occurrences soient dans le corpus 1

# Une parenthèse sur la spécificité

Modèle **hypergéométrique** :

on imagine que les mots sont jetés au hasard dans les sous-corpus, avec des probabilités proportionnelles à la taille de chacun.

Si corpus 1 et corpus 2 de même taille (100 mots chacun), un mot présent 3 fois dans la totalité du corpus :

- 1 chance sur 2 que la première occurrence soit dans le corpus 1
- 1 chance sur 4 que les 2 premières occurrences soient dans le corpus 1
- 1 chance sur 8 que les 3 occurrences soient dans le corpus 1
- 2 chances sur 8 (= 1 chance sur 4) que les 3 occurrences soient dans le corpus 1, ou les trois dans le corpus 2

# Une parenthèse sur la spécificité

Modèle **hypergéométrique** :

on imagine que les mots sont jetés au hasard dans les sous-corpus, avec des probabilités proportionnelles à la taille de chacun.

Si corpus 1 et corpus 2 de même taille (100 mots chacun), un mot présent 3 fois dans la totalité du corpus :

- 1 chance sur 2 que la première occurrence soit dans le corpus 1
- 1 chance sur 4 que les 2 premières occurrences soient dans le corpus 1
- 1 chance sur 8 que les 3 occurrences soient dans le corpus 1
- 2 chances sur 8 (= 1 chance sur 4) que les 3 occurrences soient dans le corpus 1, ou les trois dans le corpus 2
- 3 chances sur 4 d'avoir une répartition avec deux occurrences dans un corpus et une occurrence dans l'autre

# Une parenthèse sur la spécificité

Modèle **hypergéométrique** :

on imagine que les mots sont jetés au hasard dans les sous-corpus, avec des probabilités proportionnelles à la taille de chacun.

Si corpus 1 et corpus 2 de même taille (100 mots chacun), un mot présent 3 fois dans la totalité du corpus :

- 1 chance sur 2 que la première occurrence soit dans le corpus 1
- 1 chance sur 4 que les 2 premières occurrences soient dans le corpus 1
- 1 chance sur 8 que les 3 occurrences soient dans le corpus 1
- 2 chances sur 8 (= 1 chance sur 4) que les 3 occurrences soient dans le corpus 1, ou les trois dans le corpus 2
- 3 chances sur 4 d'avoir une répartition avec deux occurrences dans un corpus et une occurrence dans l'autre
- **la situation observée (2% de fréquence dans le corpus 1, 1% dans le corpus 2) avait 75% de chances d'arriver avec cette répartition "au hasard" !**

# Une parenthèse sur la spécificité

**La spécificité** d'un mot dans un sous-corpus est un score :

- supérieur à 2 si le mot est sur-représenté dans ce sous-corpus, et qu'il est peu probable que ce soit dû au hasard ;
- compris entre -2 et 2 s'il est impossible de déterminer si le mot est sur-représenté ou sous-représenté dans le corpus ; le mot est considéré comme "banal" dans ce sous-corpus
- Inférieur à -2 si le mot est sous-représenté dans le corpus, et qu'il est peu probable que ce soit dû au hasard.

Lebart, L. & Salem, A. (1994). *Statistique Textuelle*.

Chapitre 6 : Éléments caractéristiques, réponses ou textes modaux

# Une parenthèse sur la spécificité

La **spécificité** d'un mot dans un sous-corpus est un score :

- supérieur à 2 si le mot est sur-représenté dans ce sous-corpus, et qu'il est peu probable que ce soit dû au hasard ;
- compris entre -2 et 2 s'il est impossible de déterminer si le mot est sur-représenté ou sous-représenté dans le corpus ; le mot est considéré comme "banal" dans ce sous-corpus
- Inférieur à -2 si le mot est sous-représenté dans le corpus, et qu'il est peu probable que ce soit dû au hasard.

La "valeur absolue" de la spécificité est d'autant plus **grande** qu'il est **peu probable** que le phénomène soit dû au hasard :

- spécificité = 1,96 : 5% de chances
- spécificité = 2,33 : 1% de chances
- spécificité = 3,09 : 1 chance sur 1000

# Une parenthèse sur la spécificité

La **spécificité** d'un mot dans un sous-corpus est un score :

- supérieur à 2 si le mot est sur-représenté dans ce sous-corpus, et qu'il est peu probable que ce soit dû au hasard ;
- compris entre -2 et 2 s'il est impossible de déterminer si le mot est sur-représenté ou sous-représenté dans le corpus ; le mot est considéré comme "banal" dans ce sous-corpus
- Inférieur à -2 si le mot est sous-représenté dans le corpus, et qu'il est peu probable que ce soit dû au hasard.

La "valeur absolue" de la spécificité est d'autant plus **grande** qu'il est **peu probable** que le phénomène soit dû au hasard :

- spécificité = 1,96 : 5% de chances
- spécificité = 2,33 : 1% de chances
- spécificité = 3,09 : 1 chance sur 1000

Scores similaires :  
écart réduit, TF-IDF

# 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    |
| statistical  | 687      | 222       | -18    |

Attention aux  
conclusions  
tirées : tests  
multiples...

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

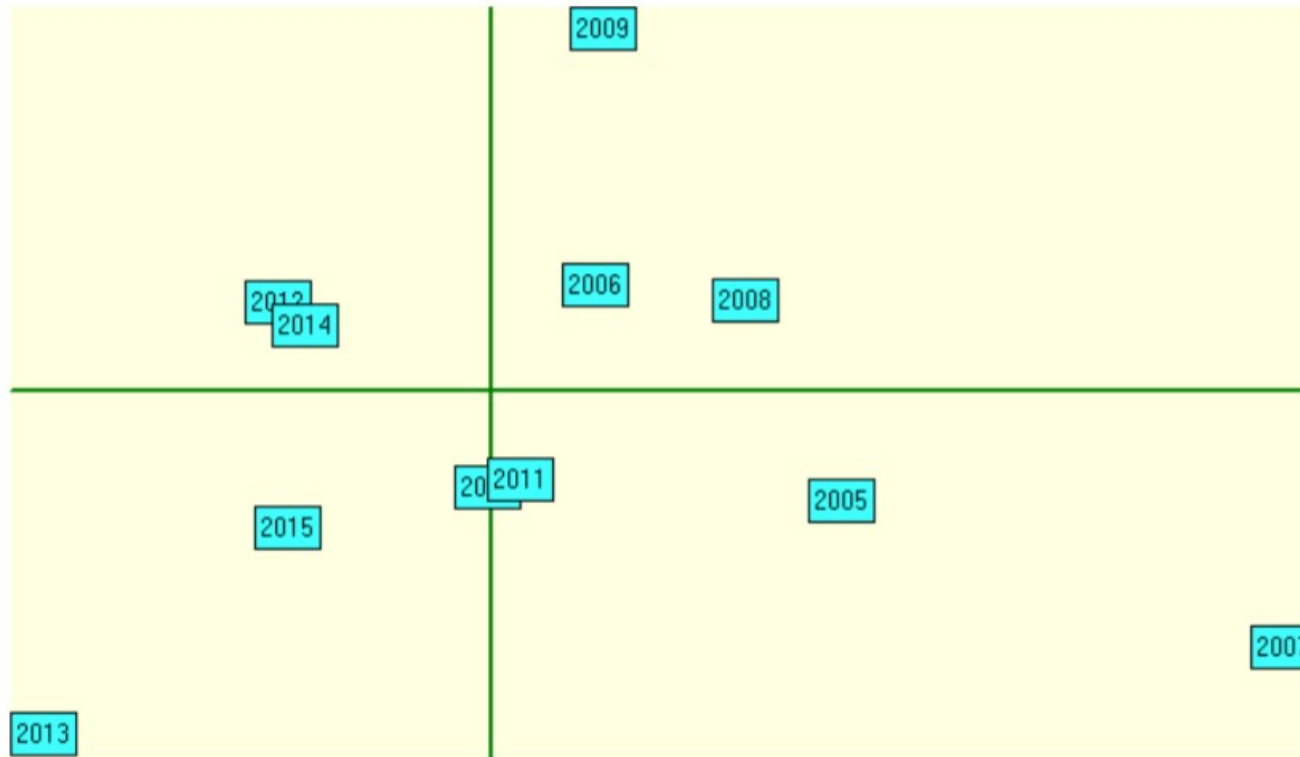
# Exemple d'exploration : articles scientifiques

## 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  $n$  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  
two networks 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

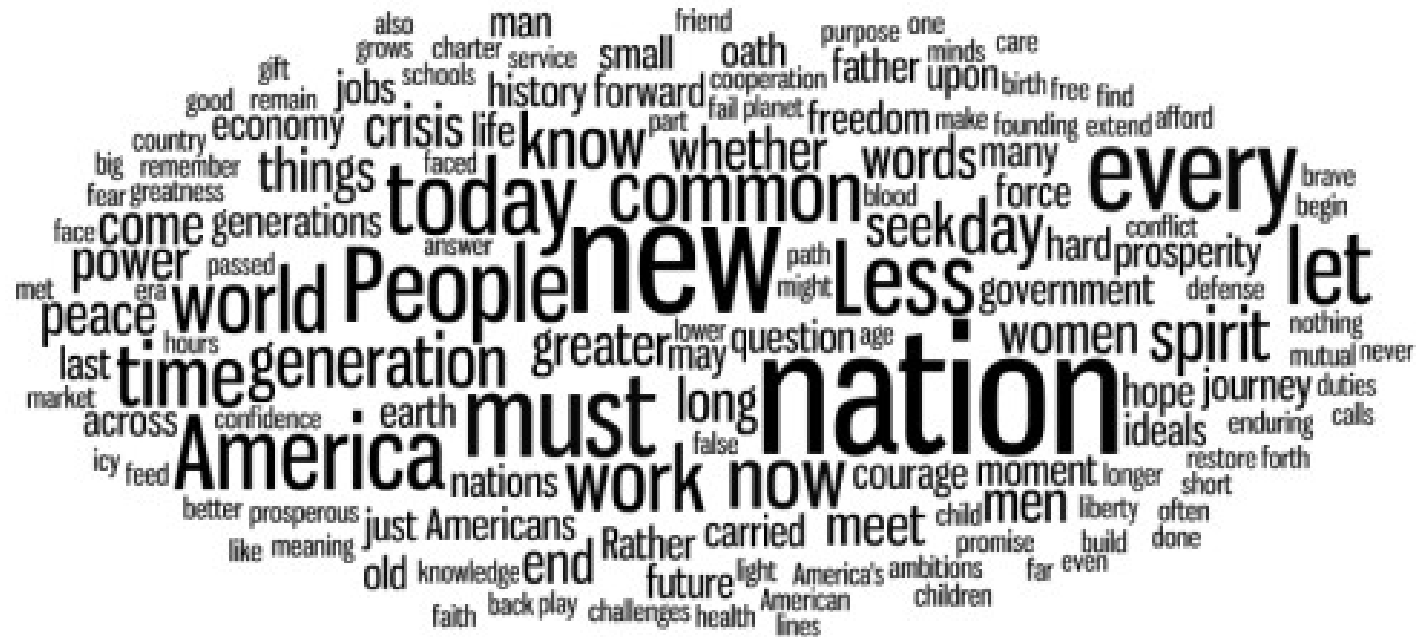
Test de l'effet d'un paramètre : l'année de publication



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>

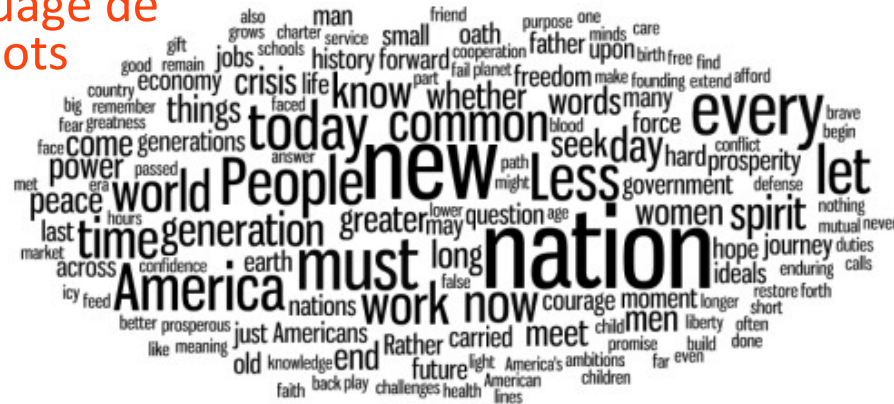
# **Arbres de mots**

## Le « nuage arboré », une information double

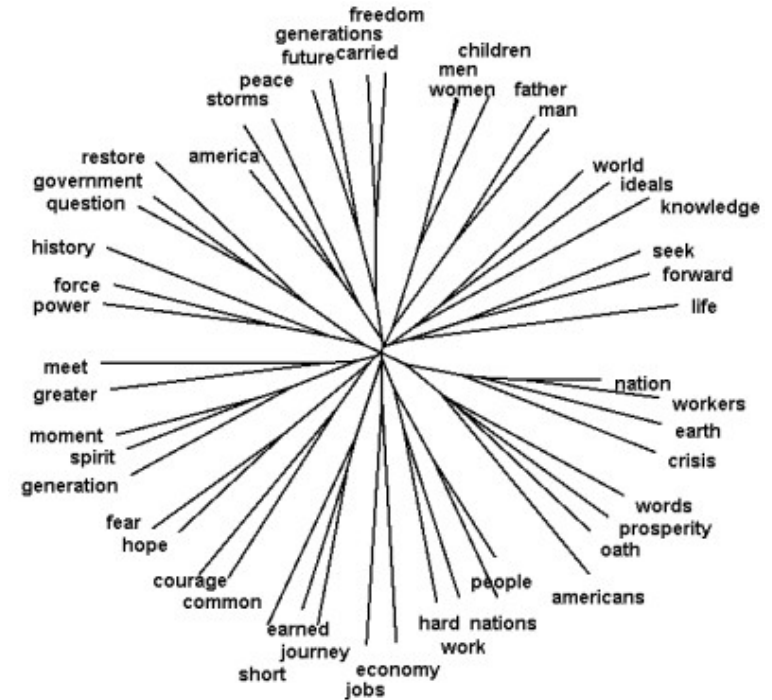


Discours inaugural de Barack Obama en 2008,  
Wordle

nuage de mots



arbre de mots



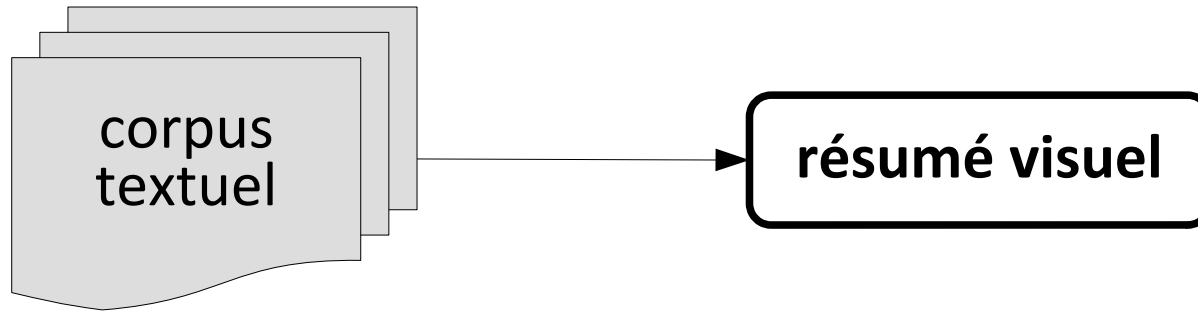




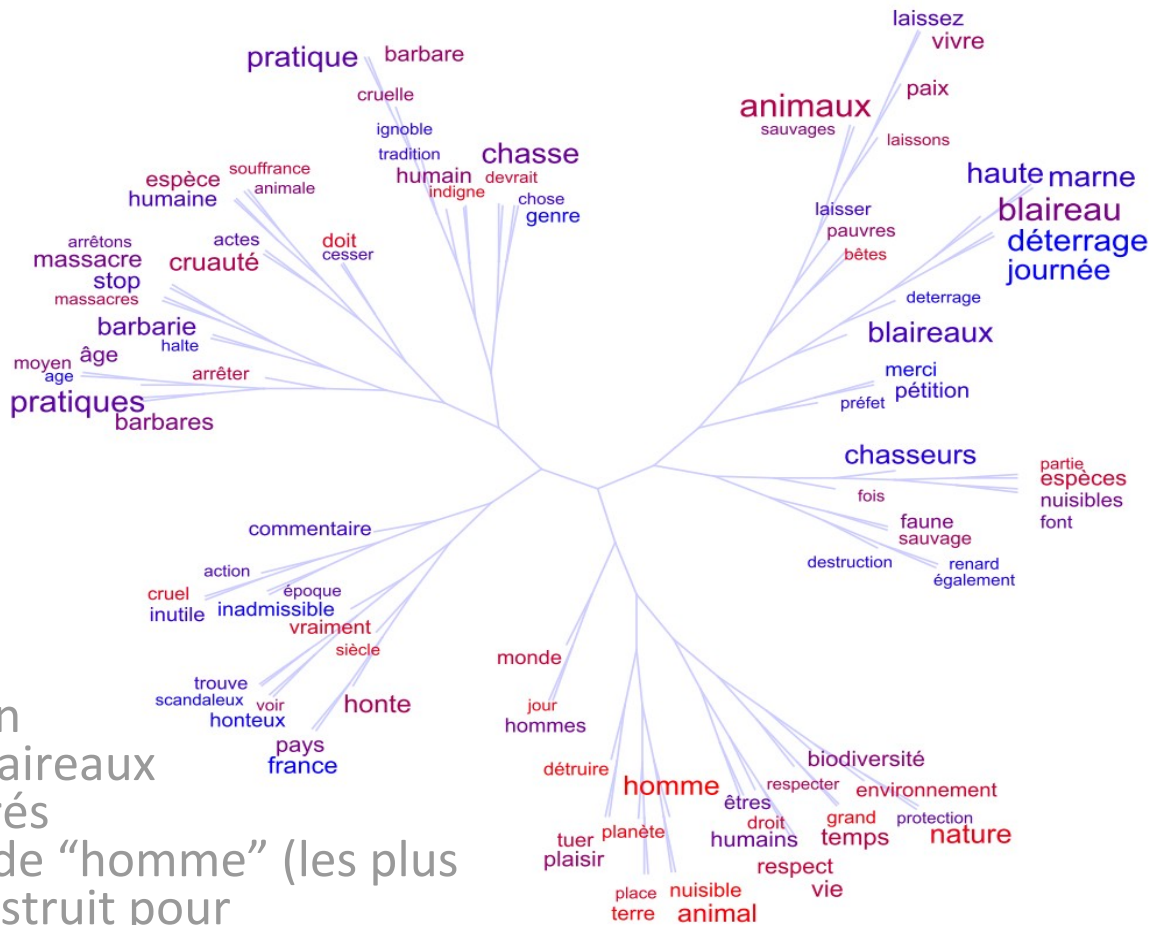
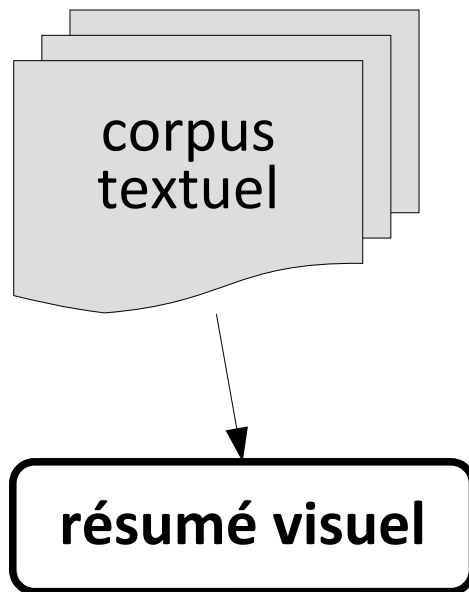




# Le « nuage arboré », pour quoi faire ?

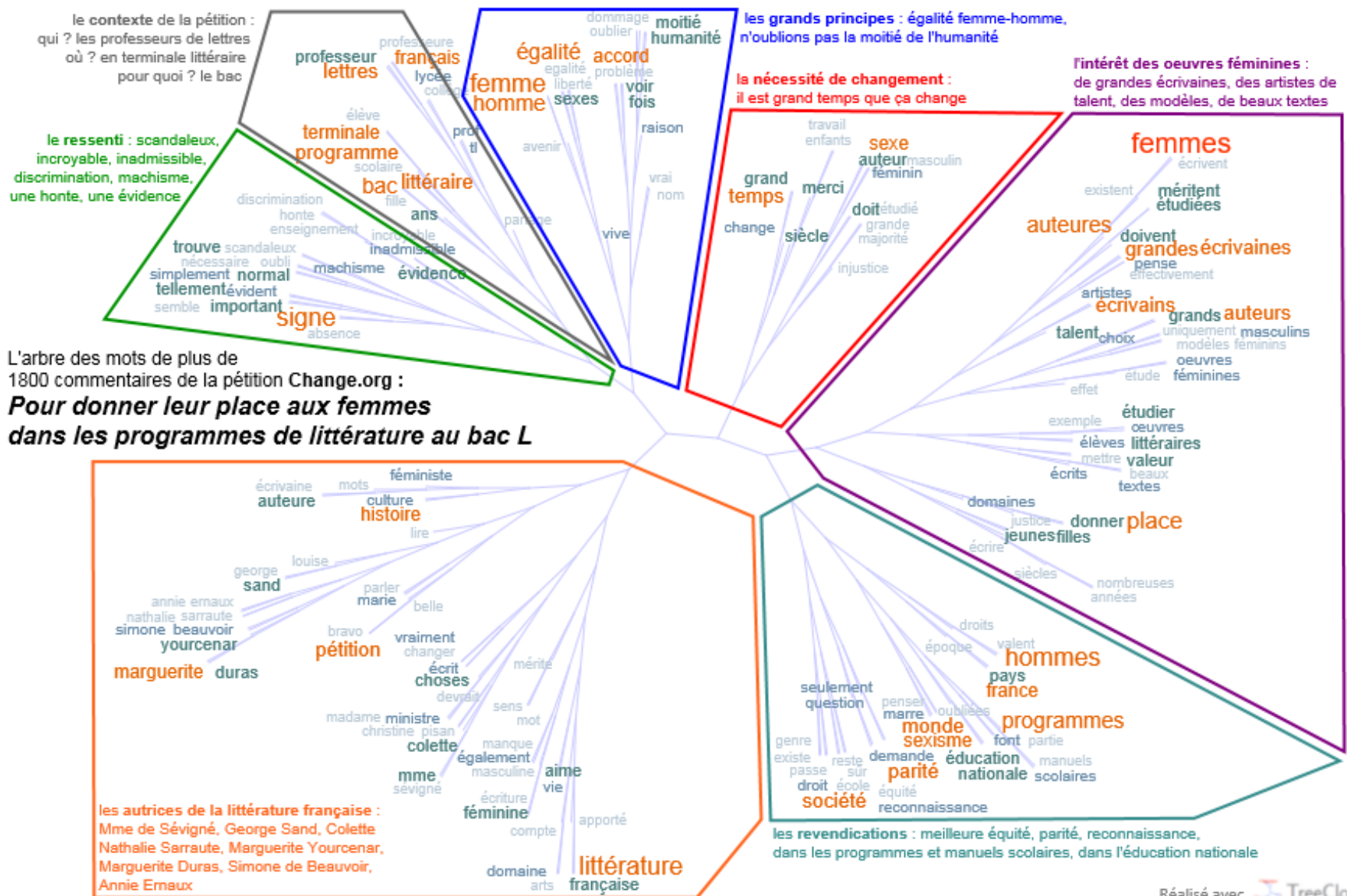


# Le « nuage arboré », pour quoi faire ?

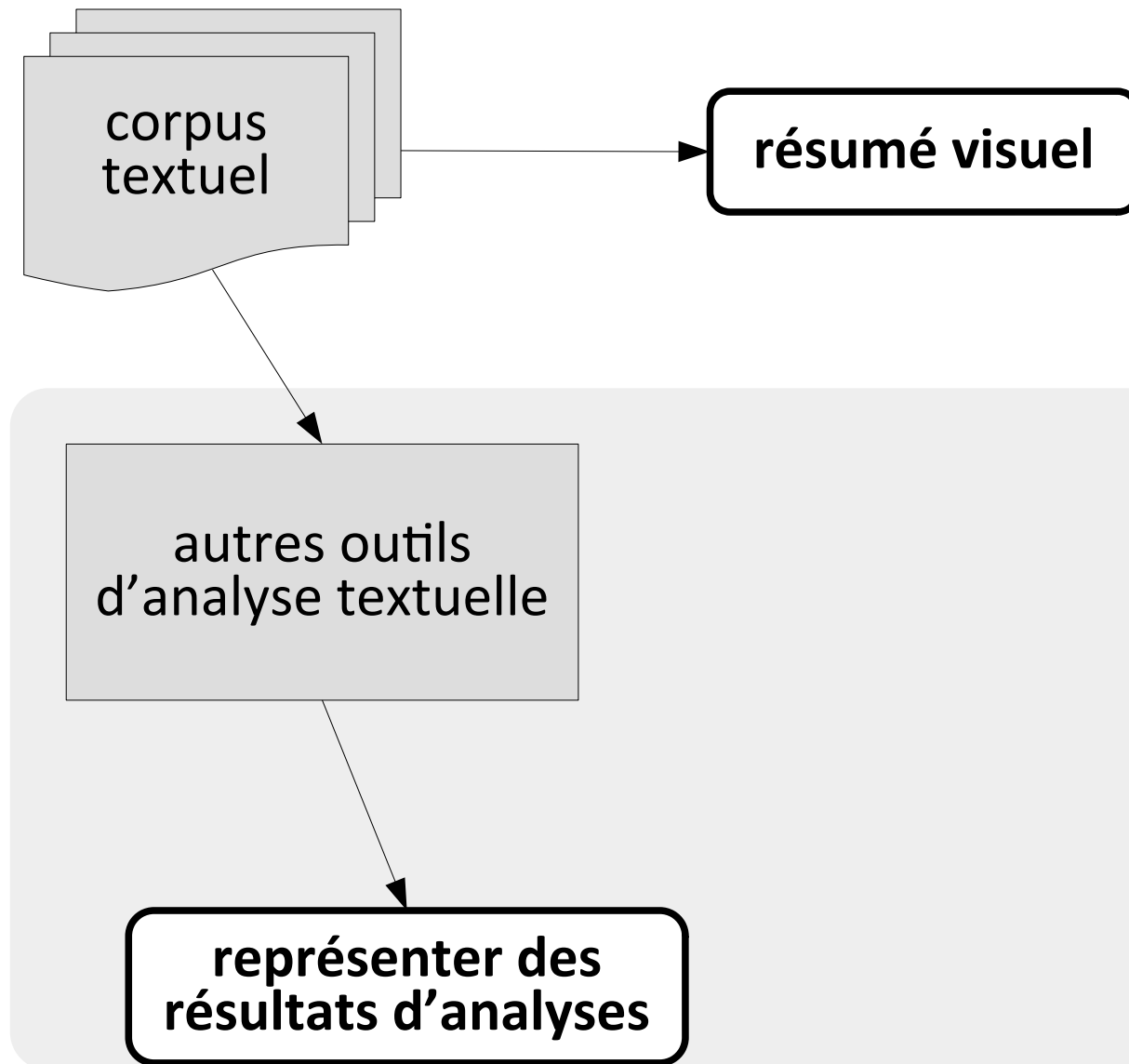


Nuage arboré des 100 mots les plus fréquents des commentaires d'une pétition contre l'extermination de blaireaux sur *lapetition.be*, mots colorés en fonction de la proximité de "homme" (les plus cooccurrents en rouge), construit pour R. Matuszewicz et M. Legris-Revel (projet ANR APPEL)

# Le « nuage arboré », pour quoi faire ?

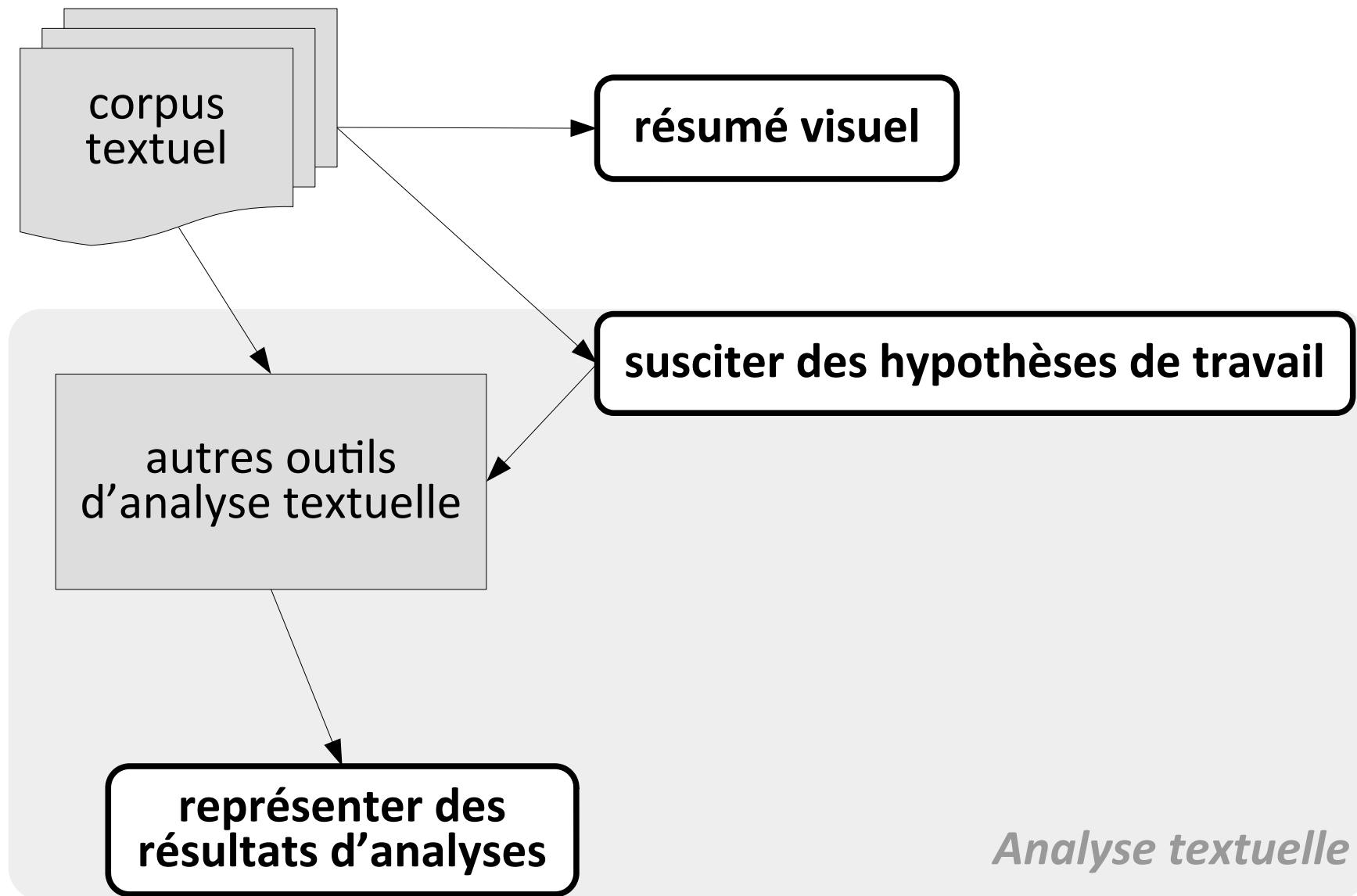


# Le « nuage arboré », pour quoi faire ?



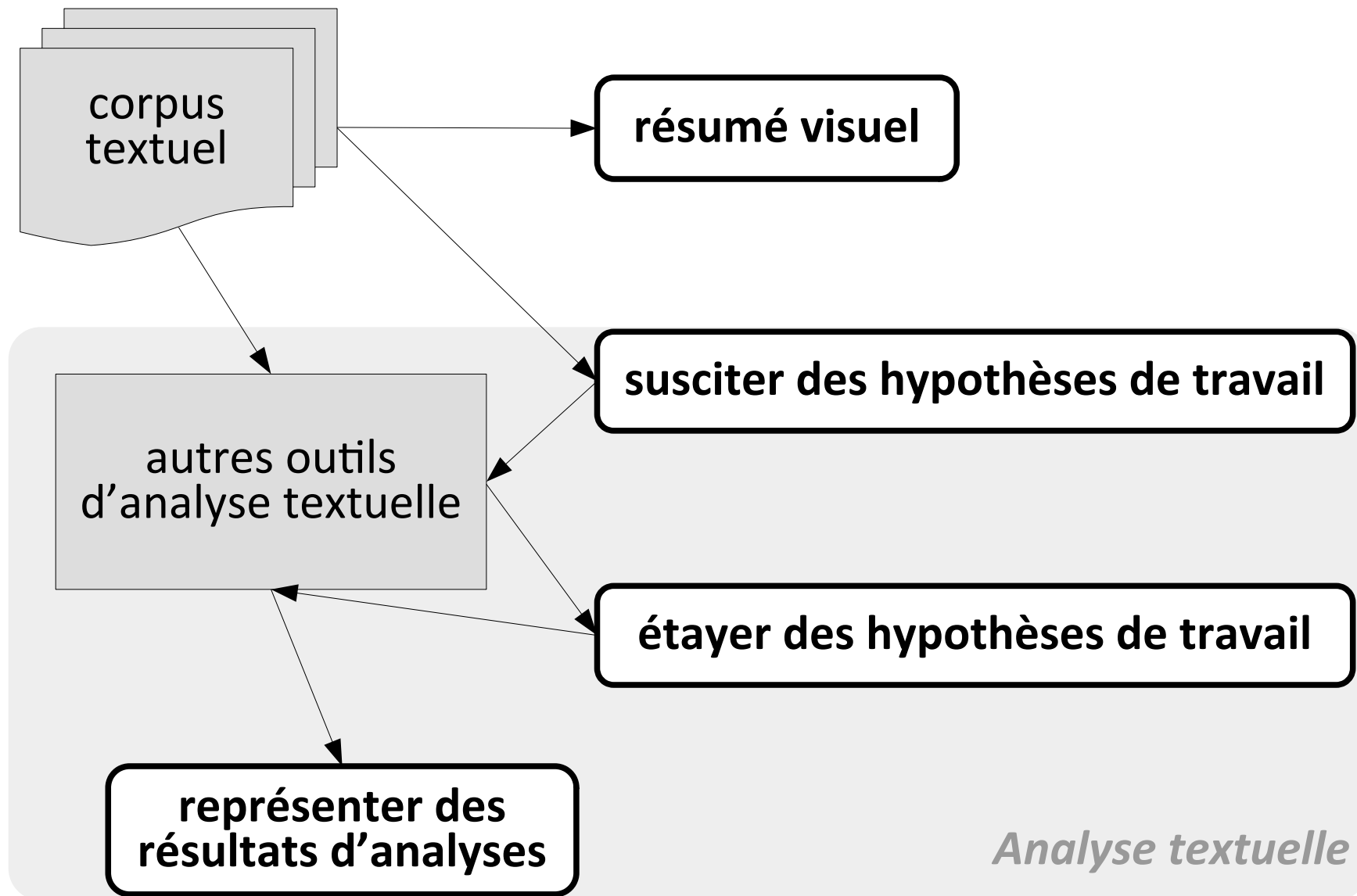
*Analyse textuelle*

# Le « nuage arboré », pour quoi faire ?



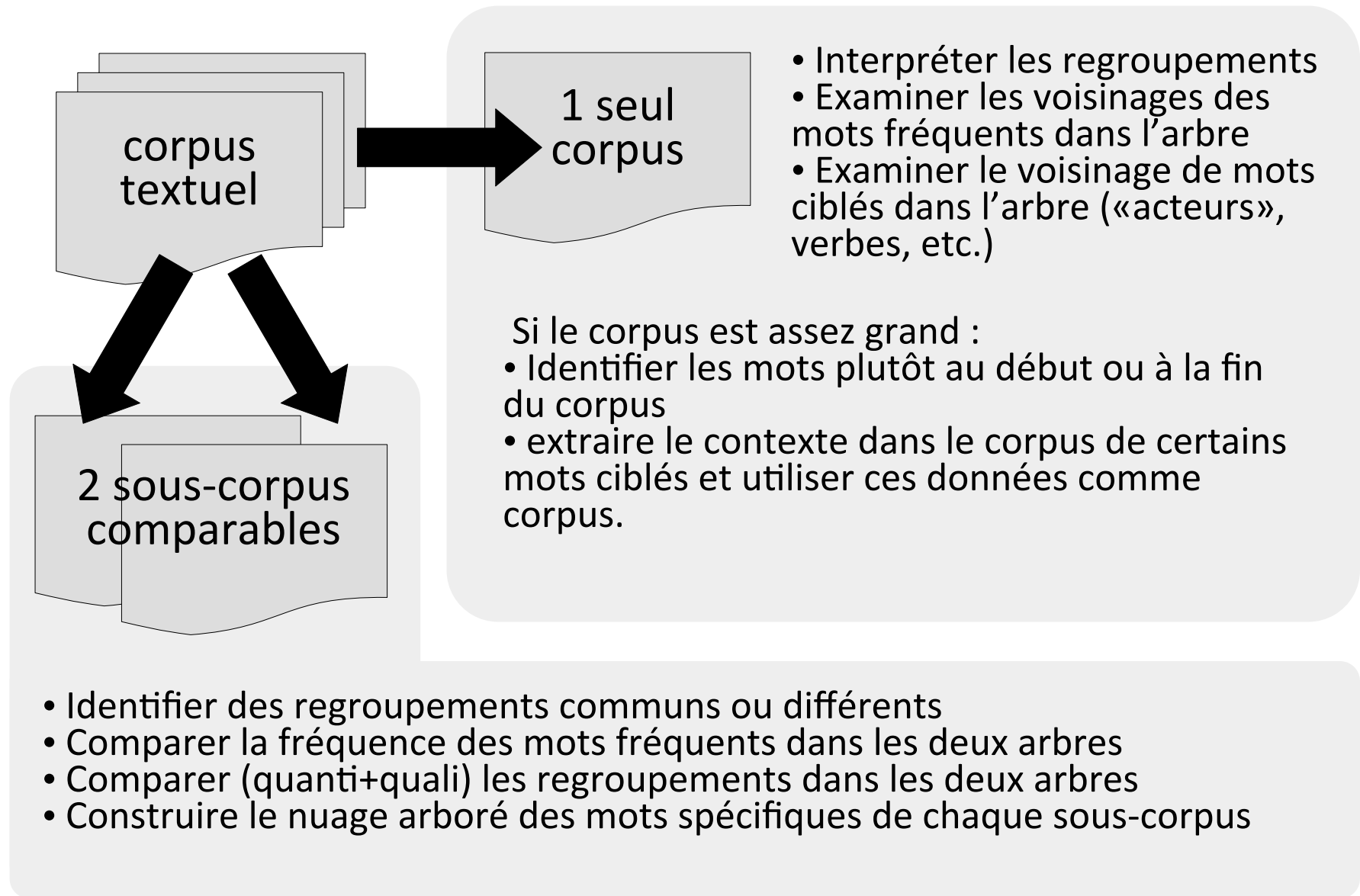
*Analyse textuelle*

# Le « nuage arboré », pour quoi faire ?

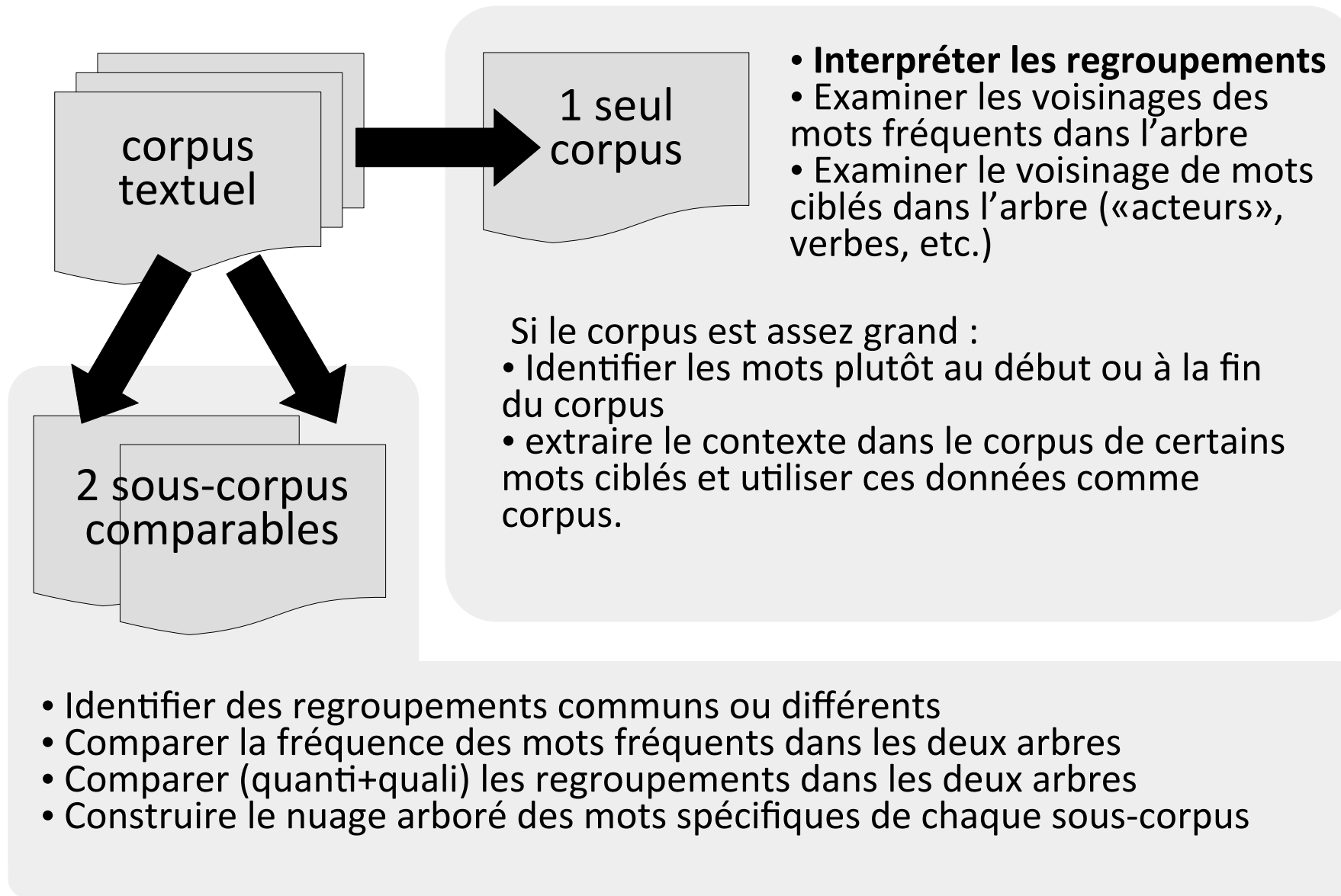


*Analyse textuelle*

# Exploration de corpus avec TreeCloud



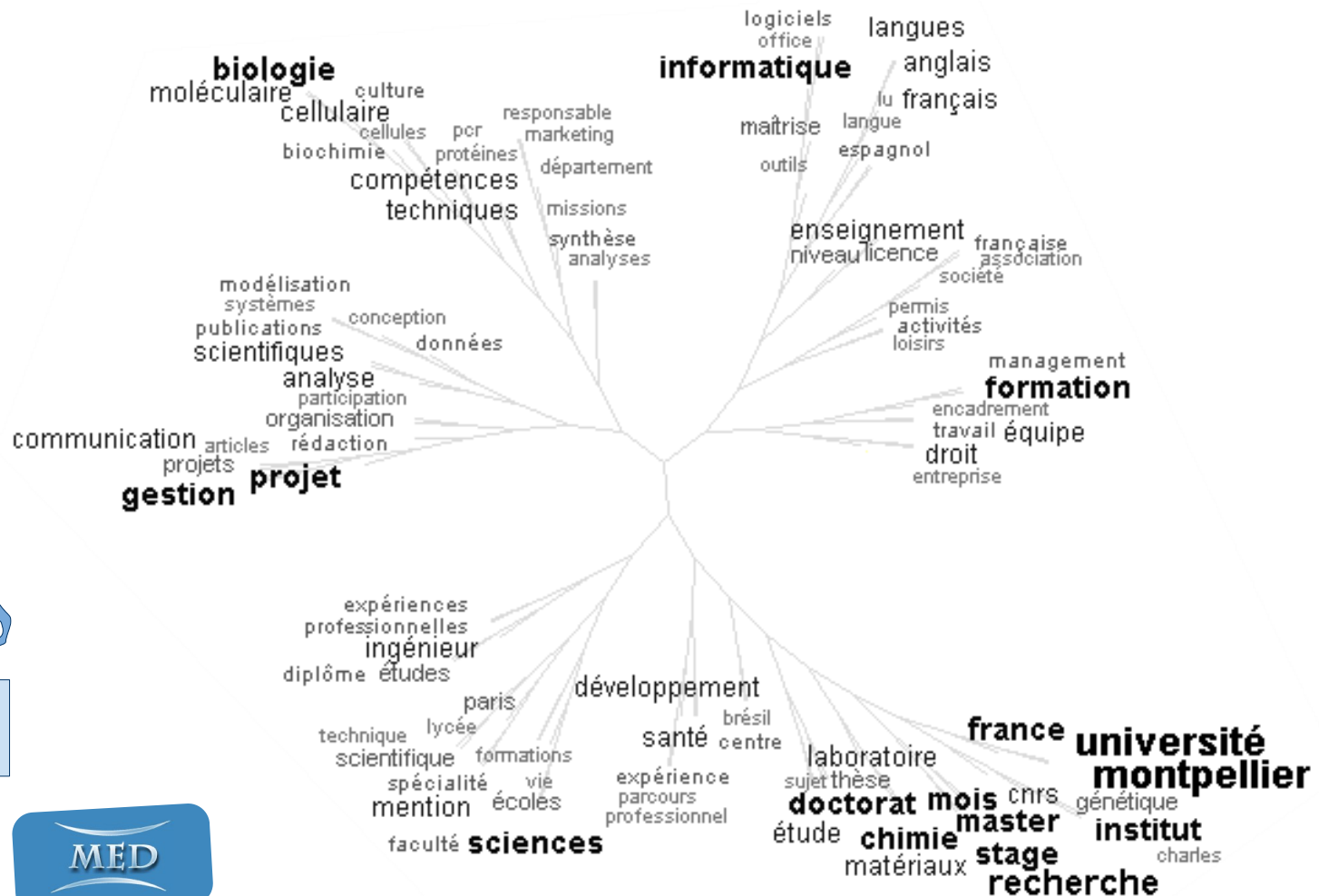
# Exploration de corpus avec TreeCloud



# Méthode : interpréter les regroupements

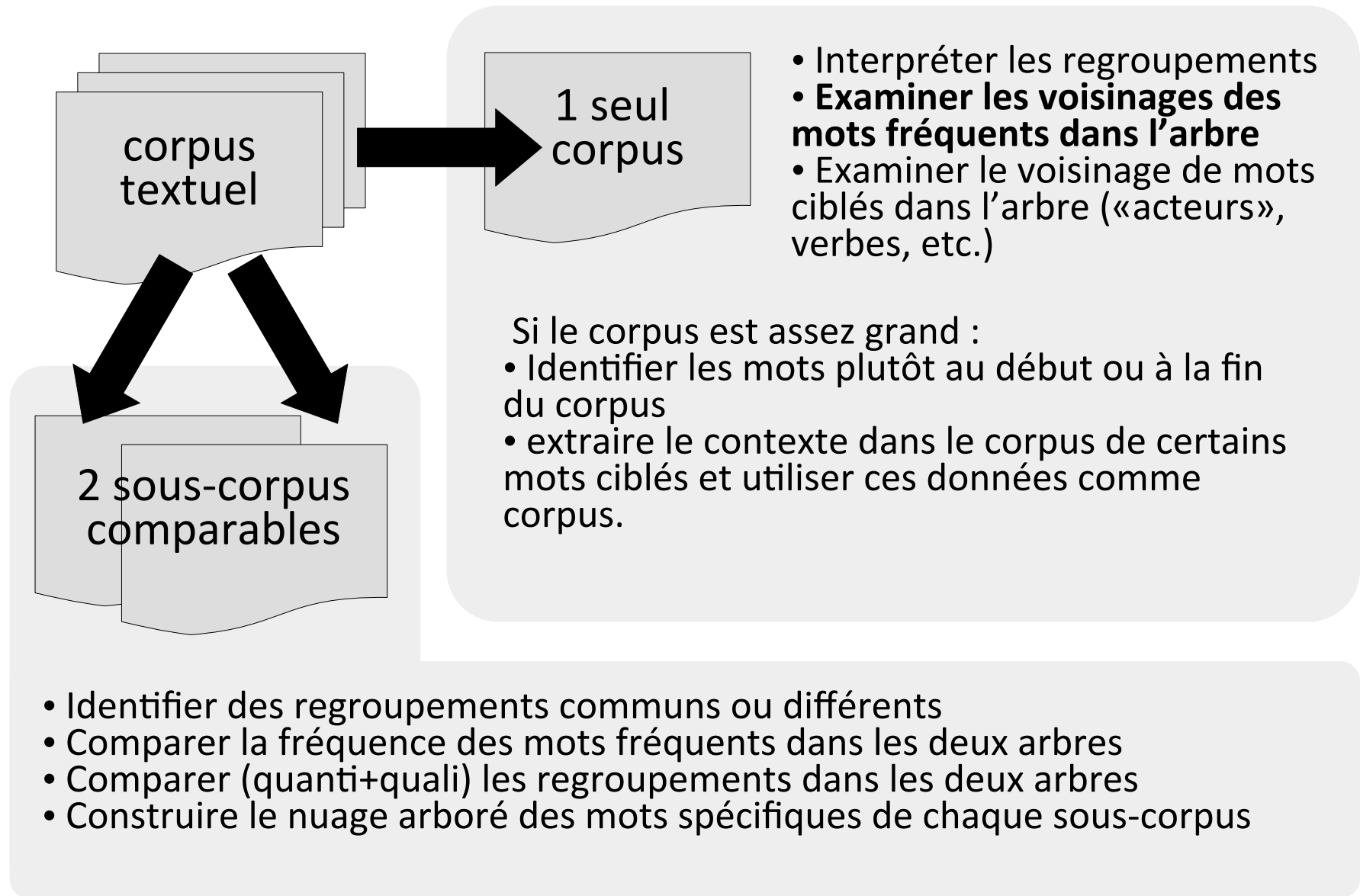
## Dessiner des « patates »

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



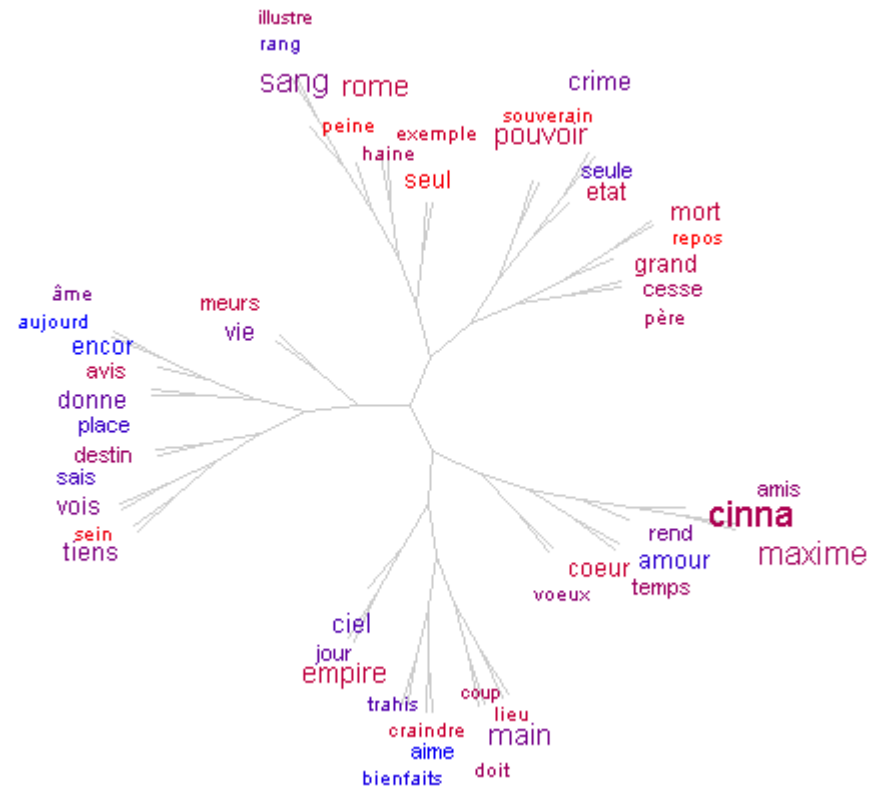


# Exploration de corpus avec TreeCloud



# Méthode : voisinage des mots fréquents

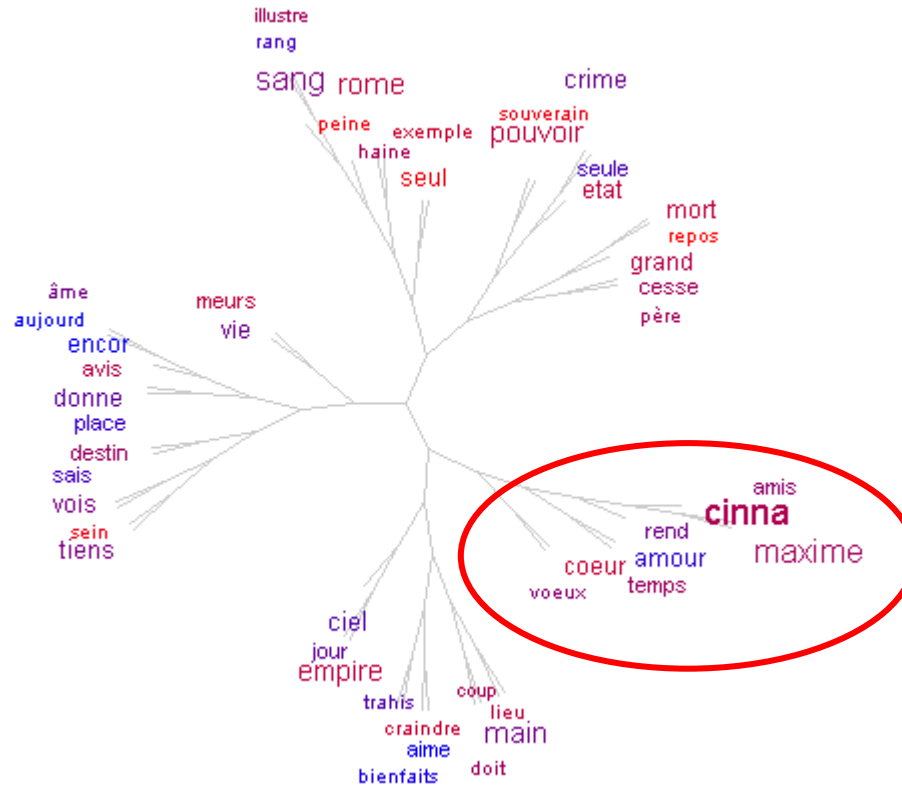
Amstutz & Gambette,  
JADT 2010



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

## Méthode : voisinage des mots fréquents

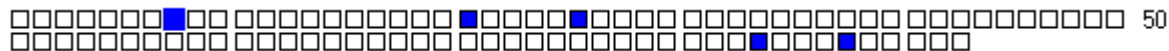
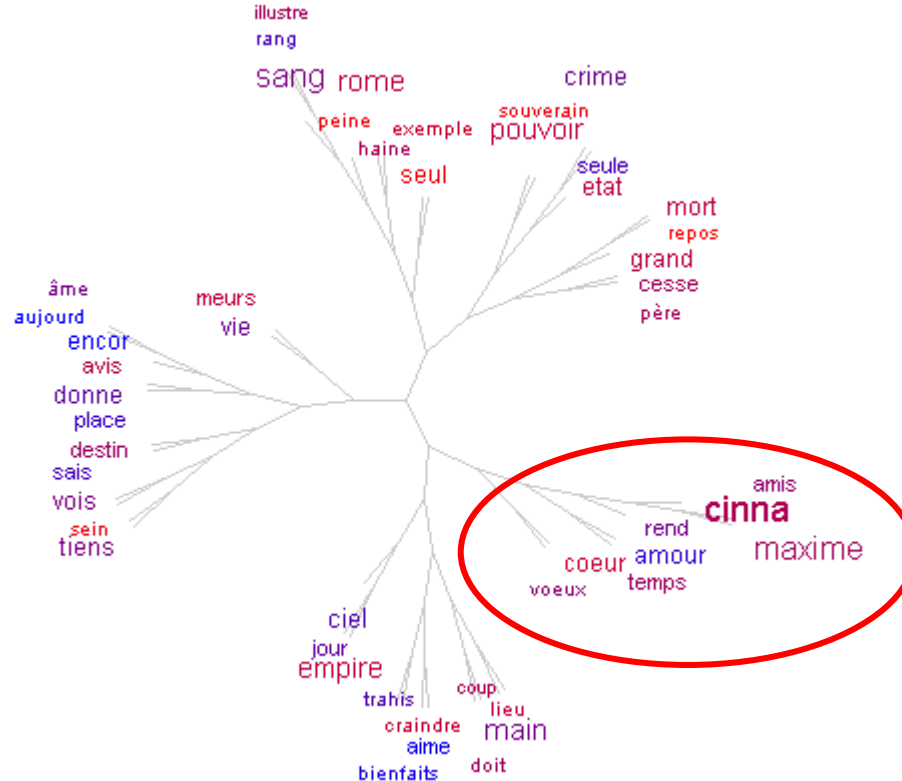
Amstutz & Gambette,  
JADT 2010



*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.

Amstutz & Gambette,  
JADT 2010

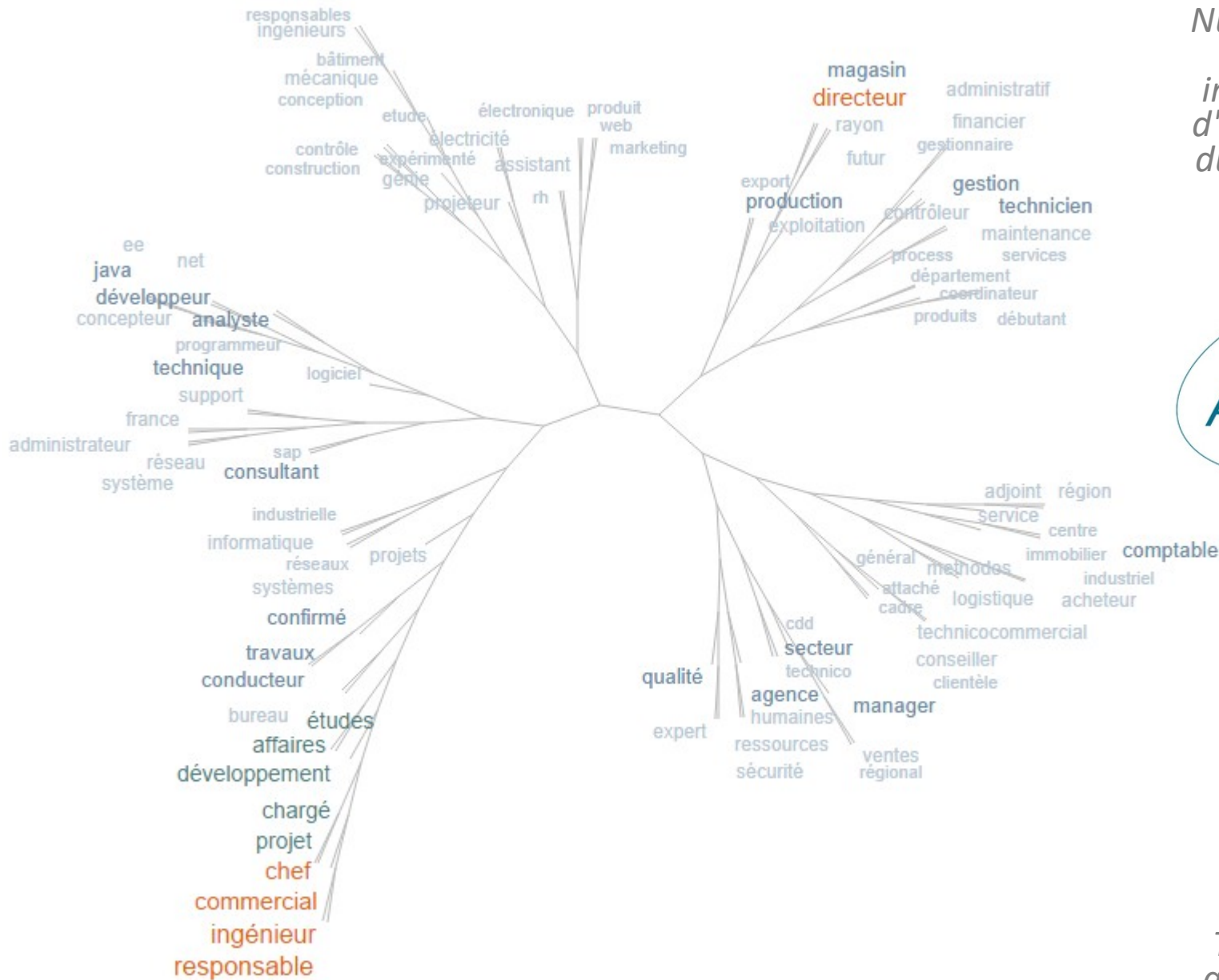


Carte des sections Lexico3 et contextes de « amis » dans les paroles d'Auguste dans Cinna.

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

## Méthode : voisinage des mots fréquents

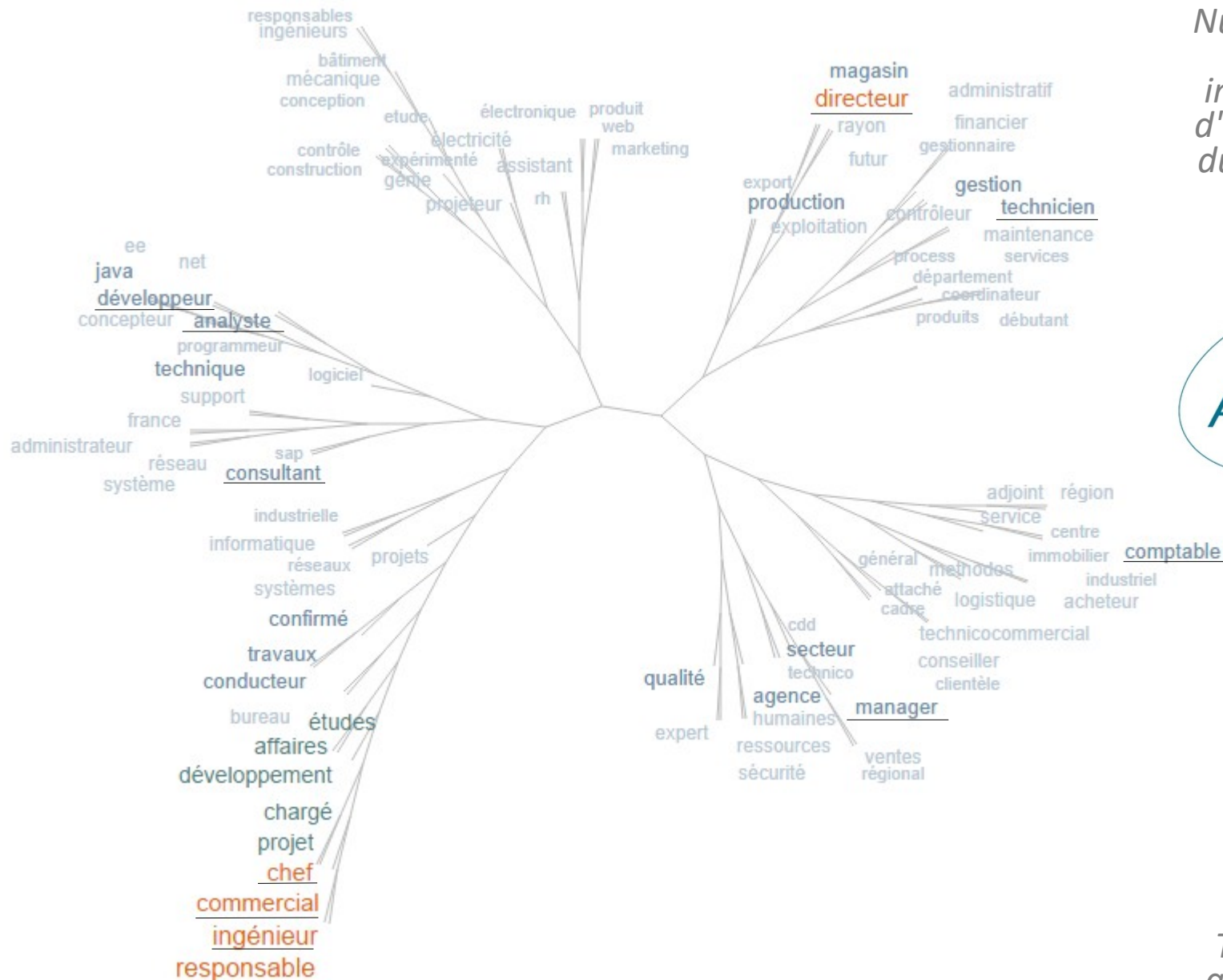
*Nuage arboré de  
plus de 4800  
intitulés d'offres  
d'emploi extraits  
du site de l'APEC  
en avril 2011*



*Travail de 2011  
avec Paola Salle*

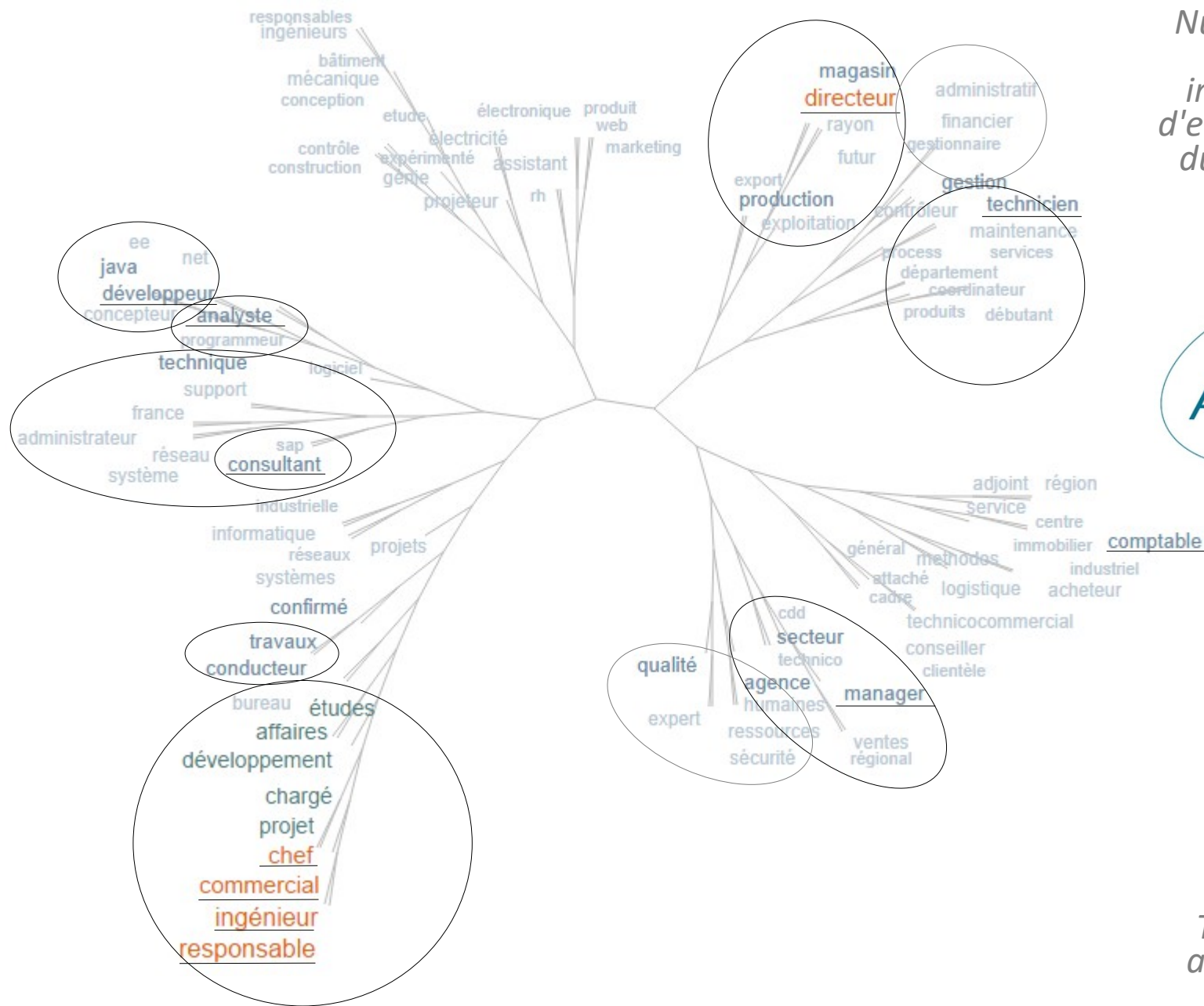
## Méthode : voisinage des mots fréquents

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



*Travail de 2011  
avec Paola Salle*

# Méthode : voisinage des mots fréquents

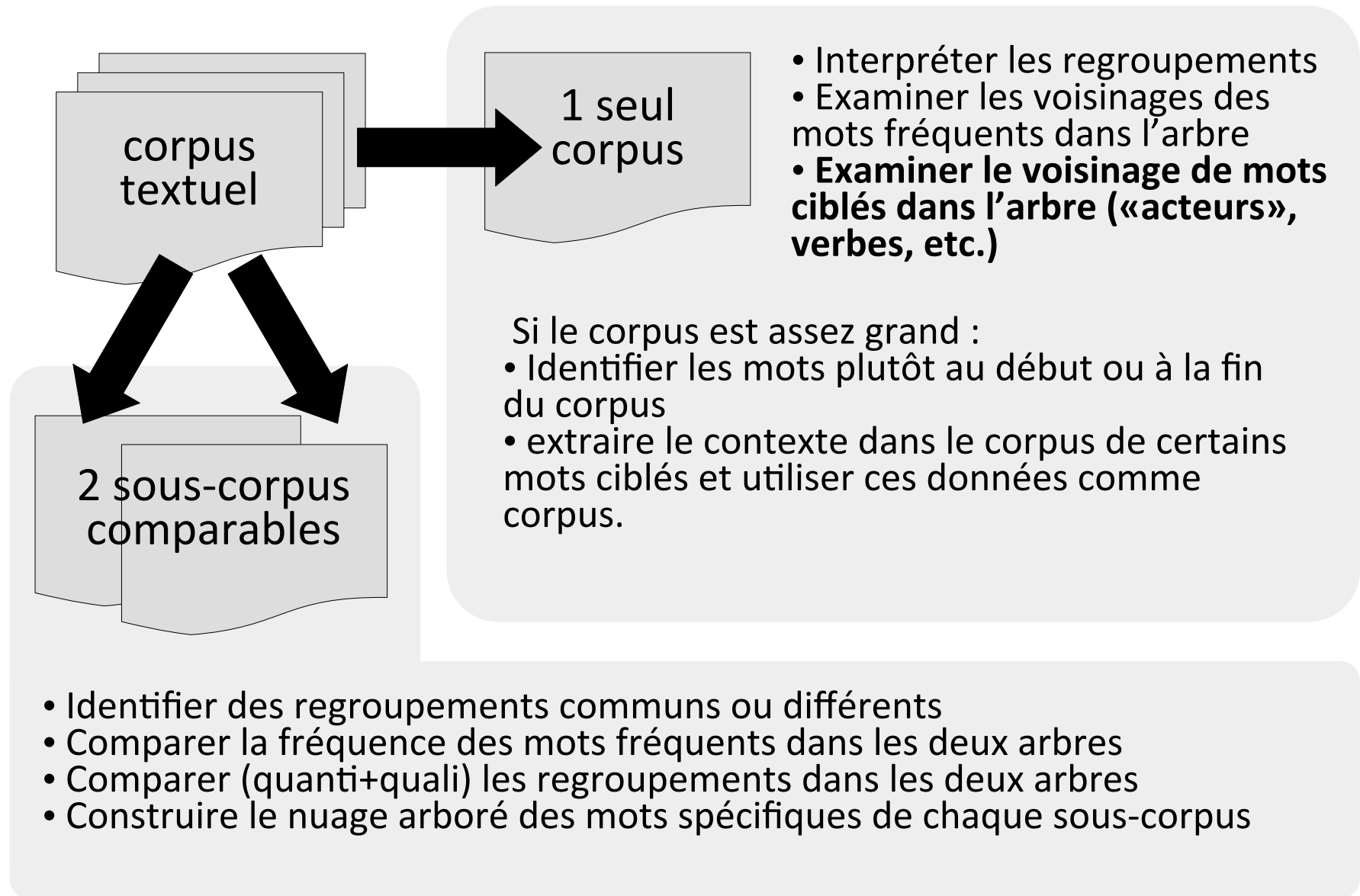


*Nuage arboré de plus de 4800 intitulés d'offres d'emploi extraites du site de l'APEC en avril 2011.*



*Travail de 2011 avec Paola Salle*

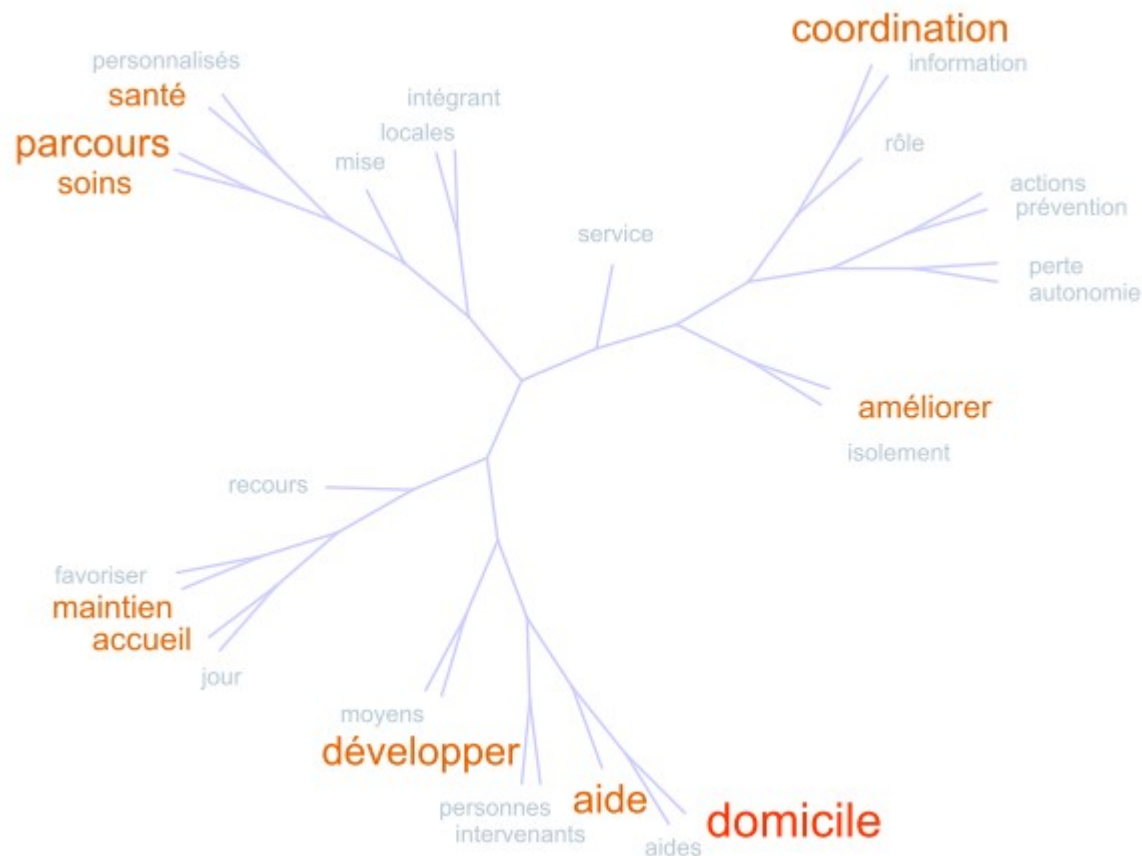
# Exploration de corpus avec TreeCloud



# 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

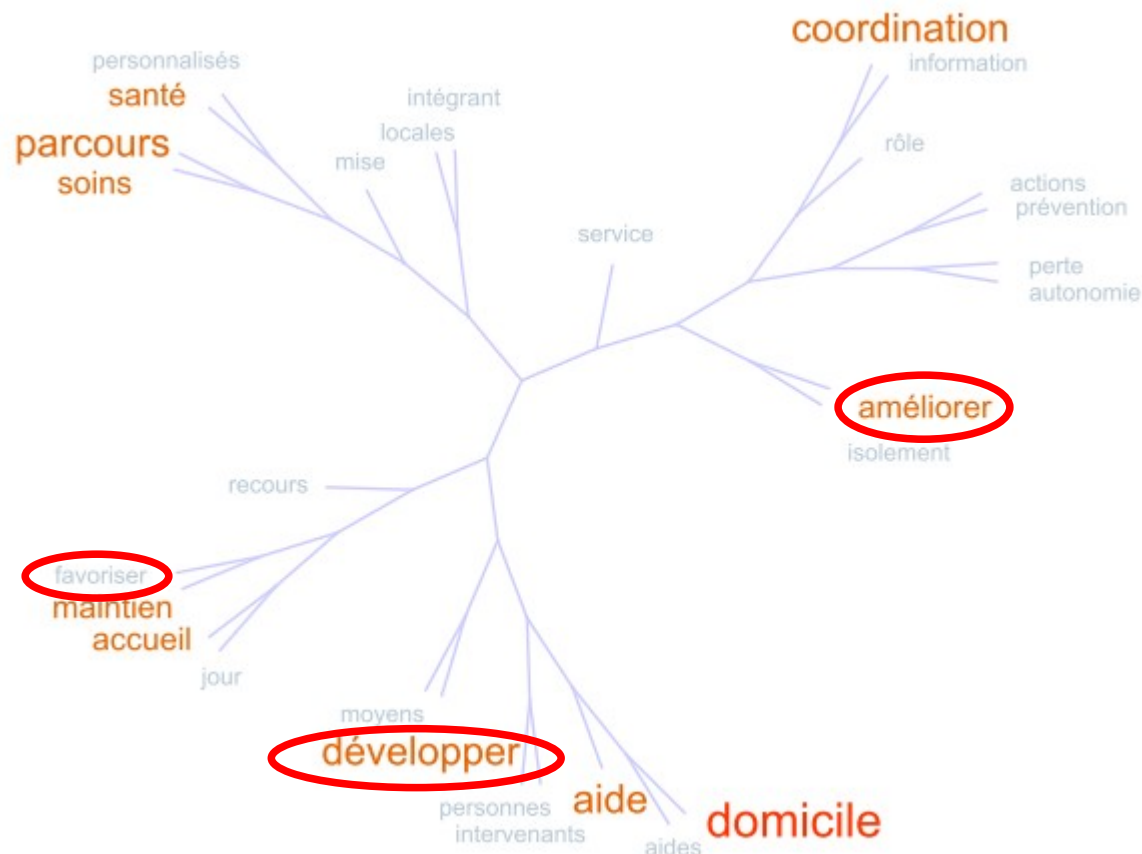
Suggestions d'améliorations :



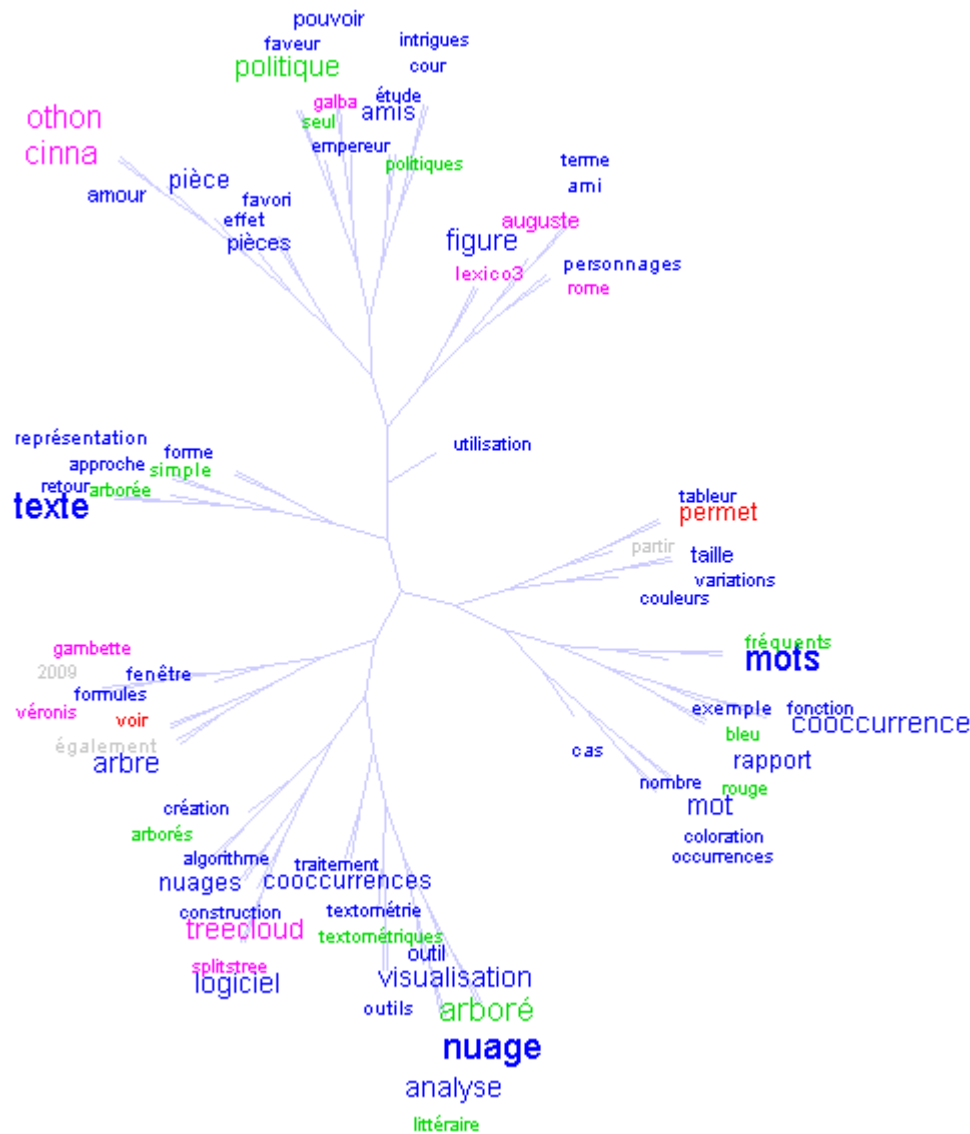
# Méthode : voisinage des verbes

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 :



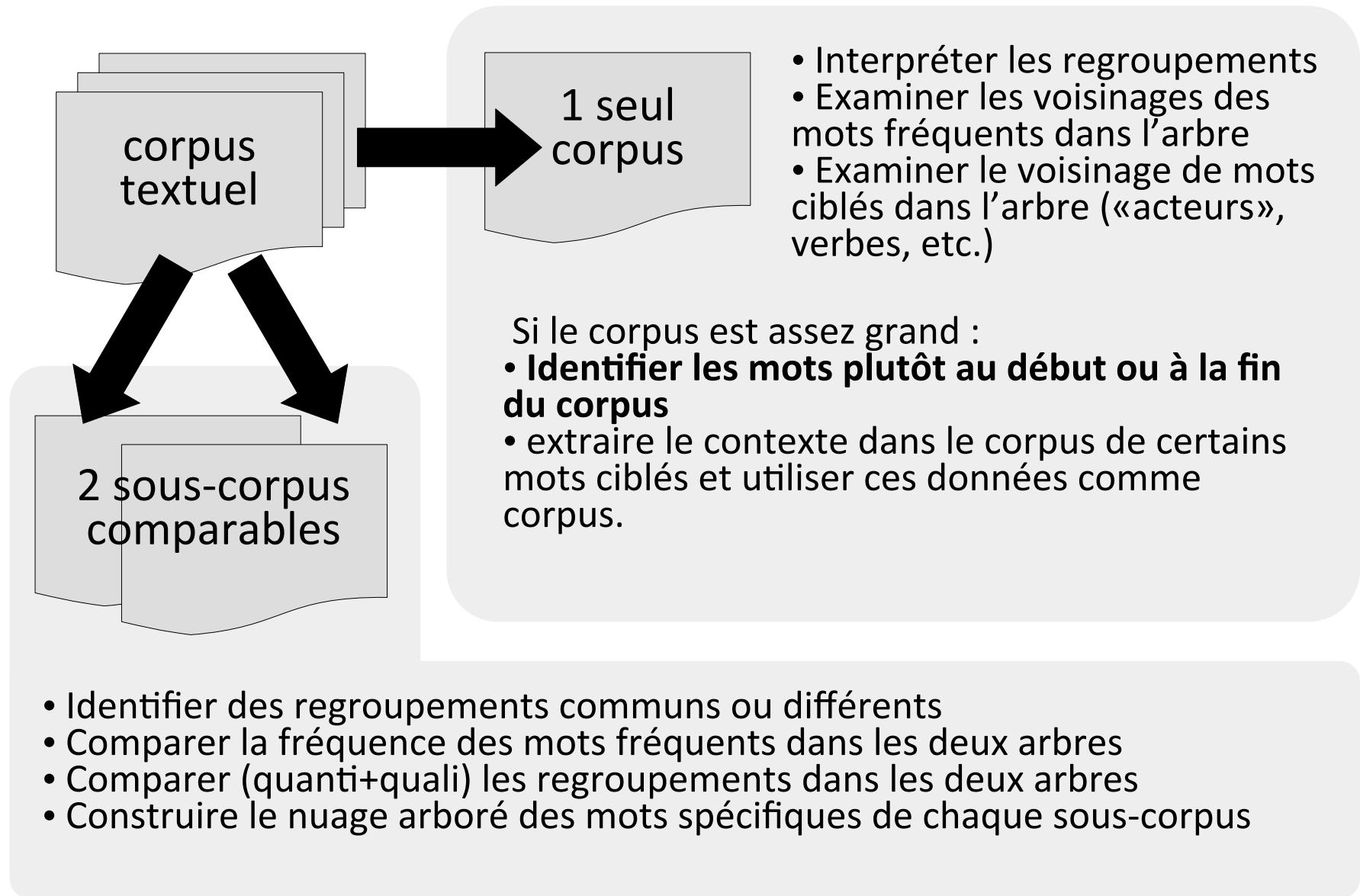
# Perspective : coloration grammaticale



noms  
adjectifs  
verbes  
noms propres

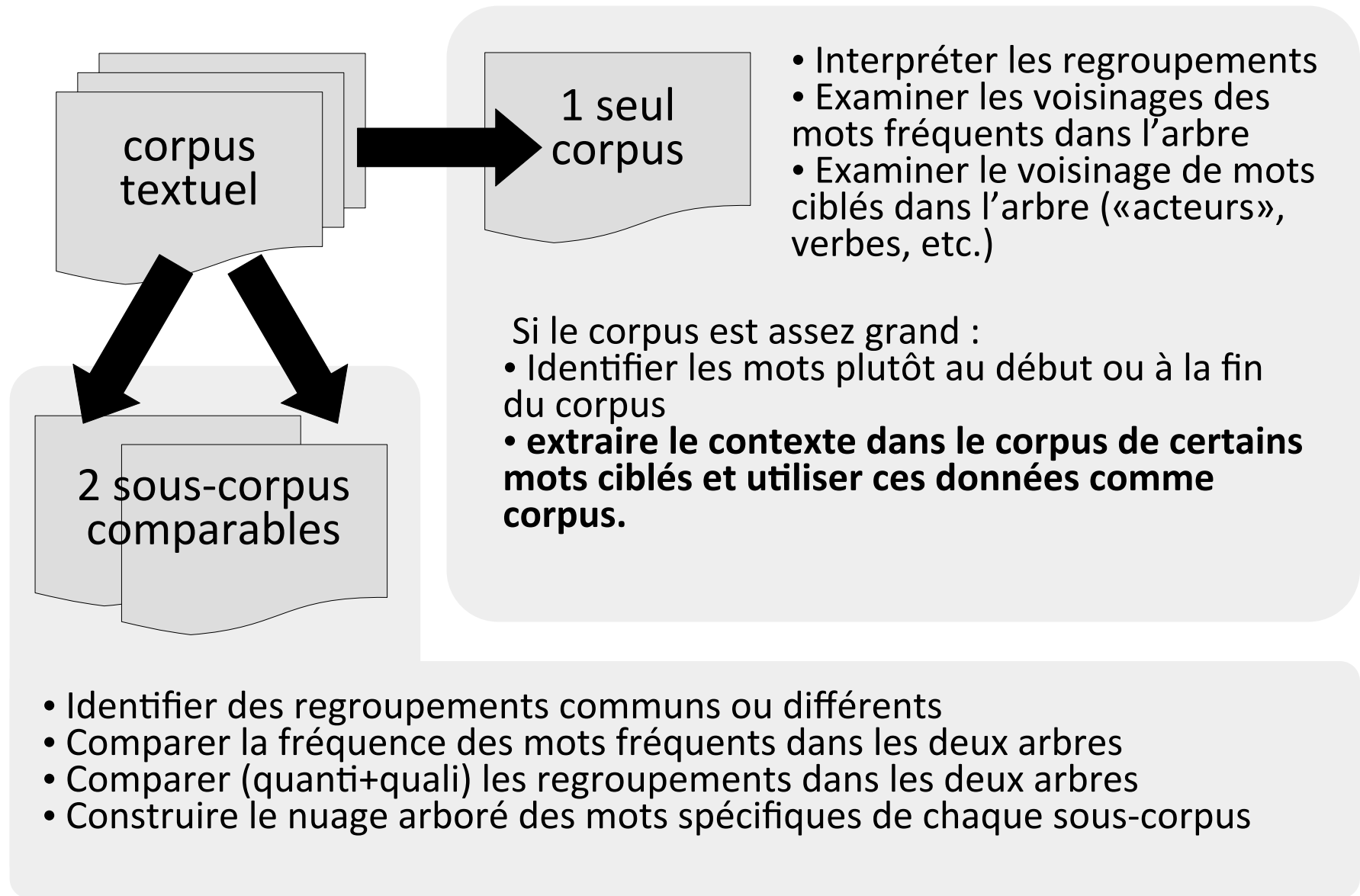
Nuage arboré des mots  
apparaissant 5 fois ou plus dans  
l'article d'Amstutz & Gambette,  
JADT 2010, distance Liddell,  
fenêtre de 20 mots, coloration  
personnalisée à partir d'un  
étiquetage TreeTagger

# Exploration de corpus avec TreeCloud



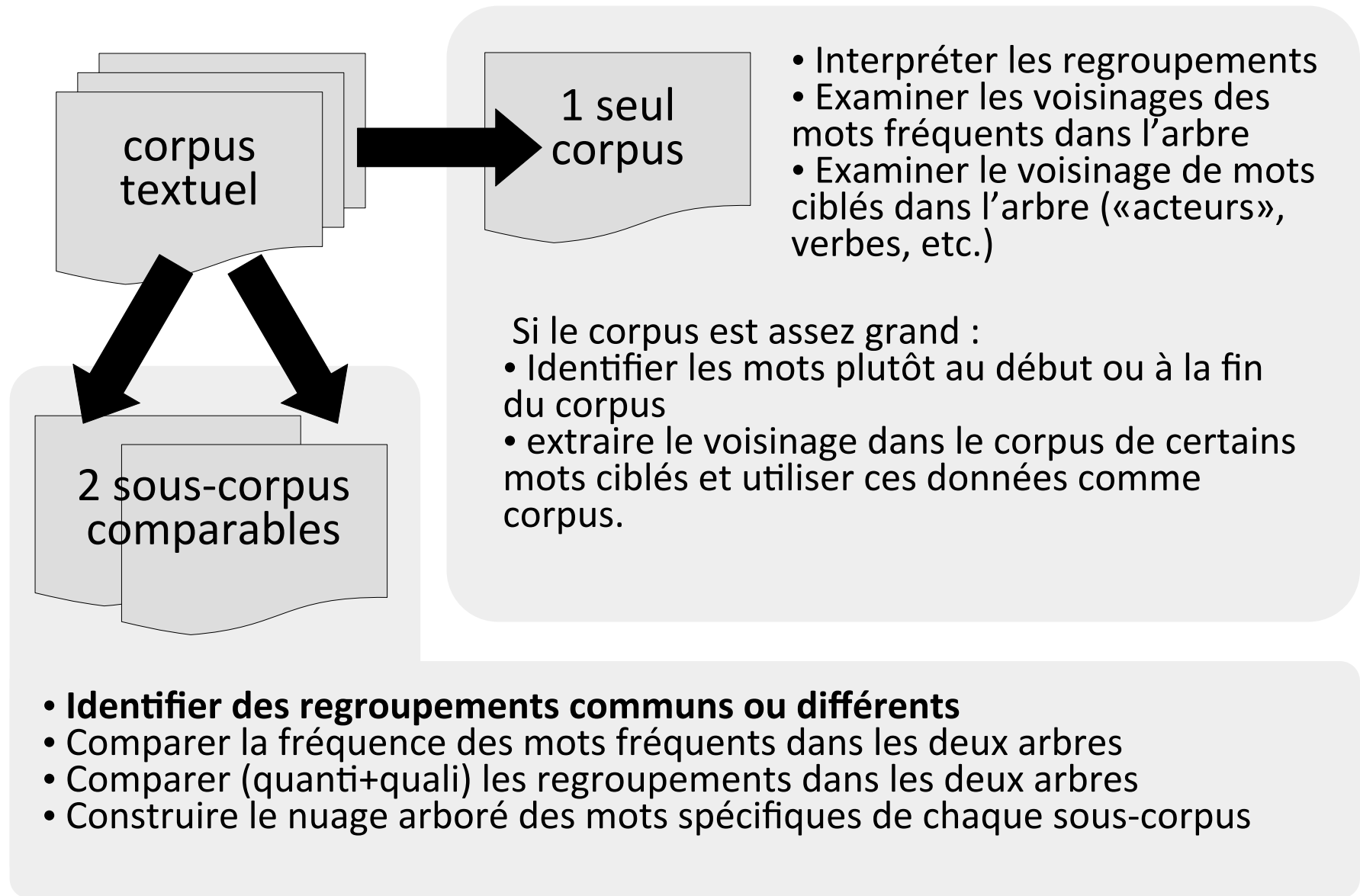


# Exploration de corpus avec TreeCloud





# Exploration de corpus avec TreeCloud



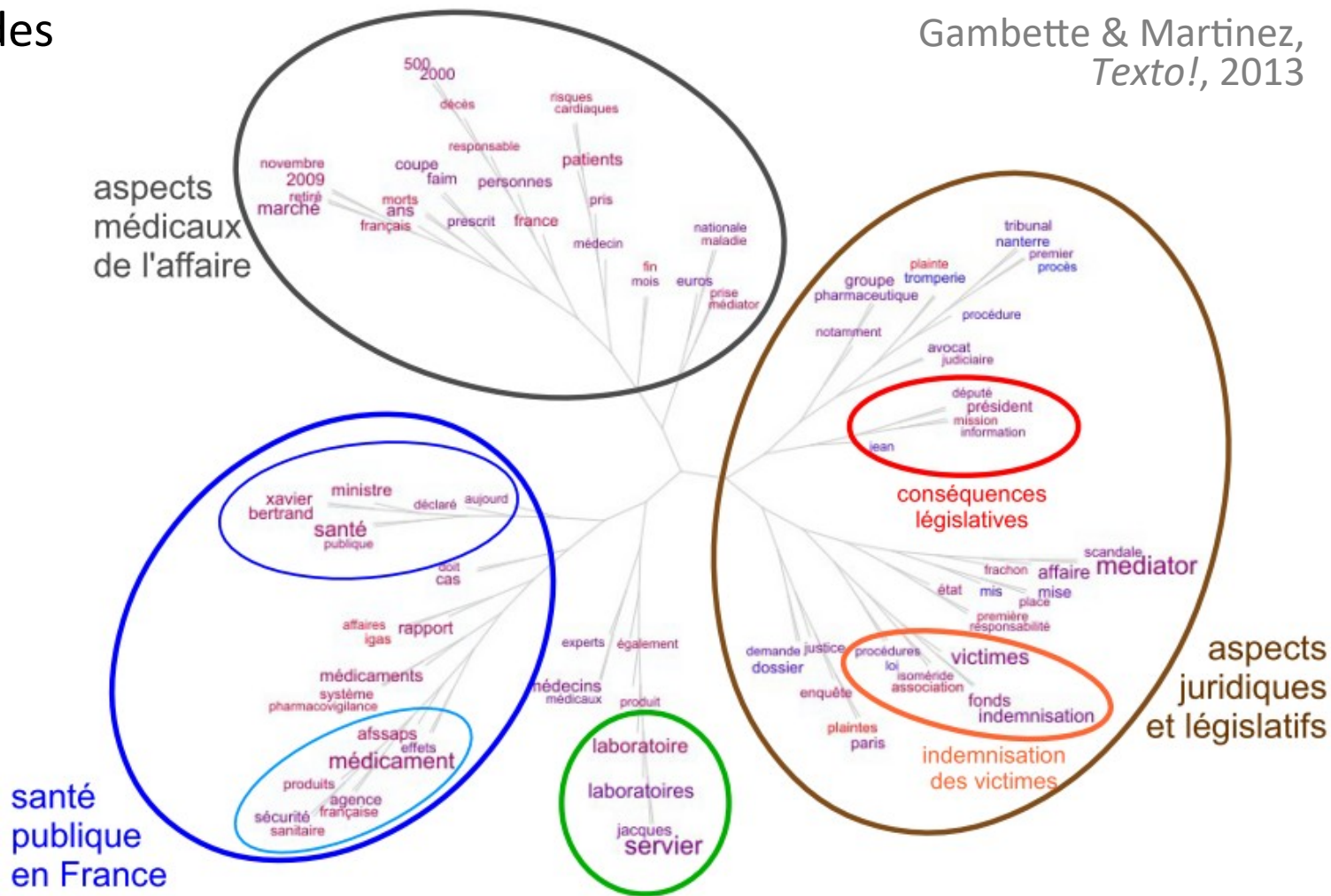
# Illustration sur le corpus Mediator

## *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

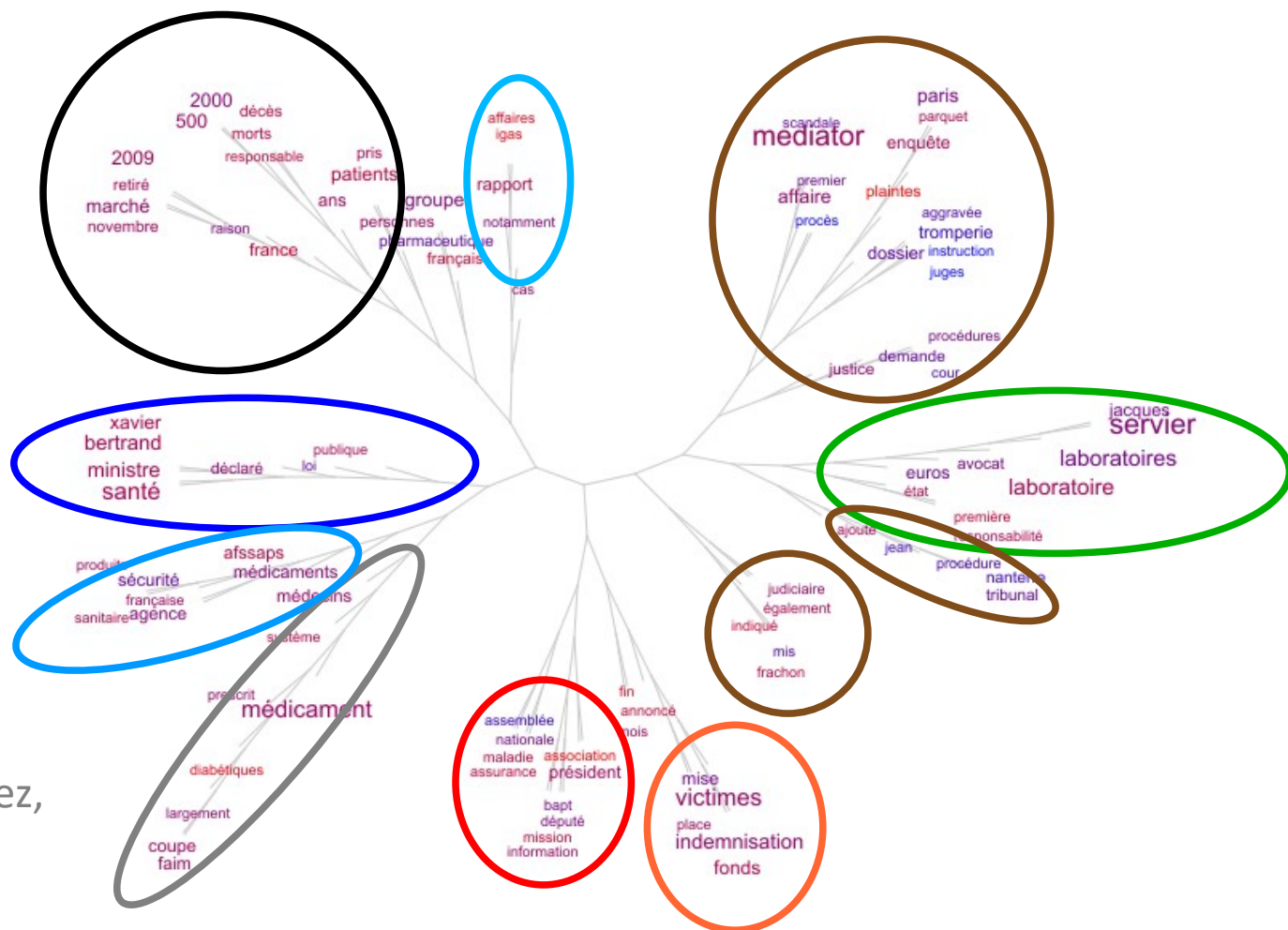


# Illustration sur le corpus Mediator

## *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



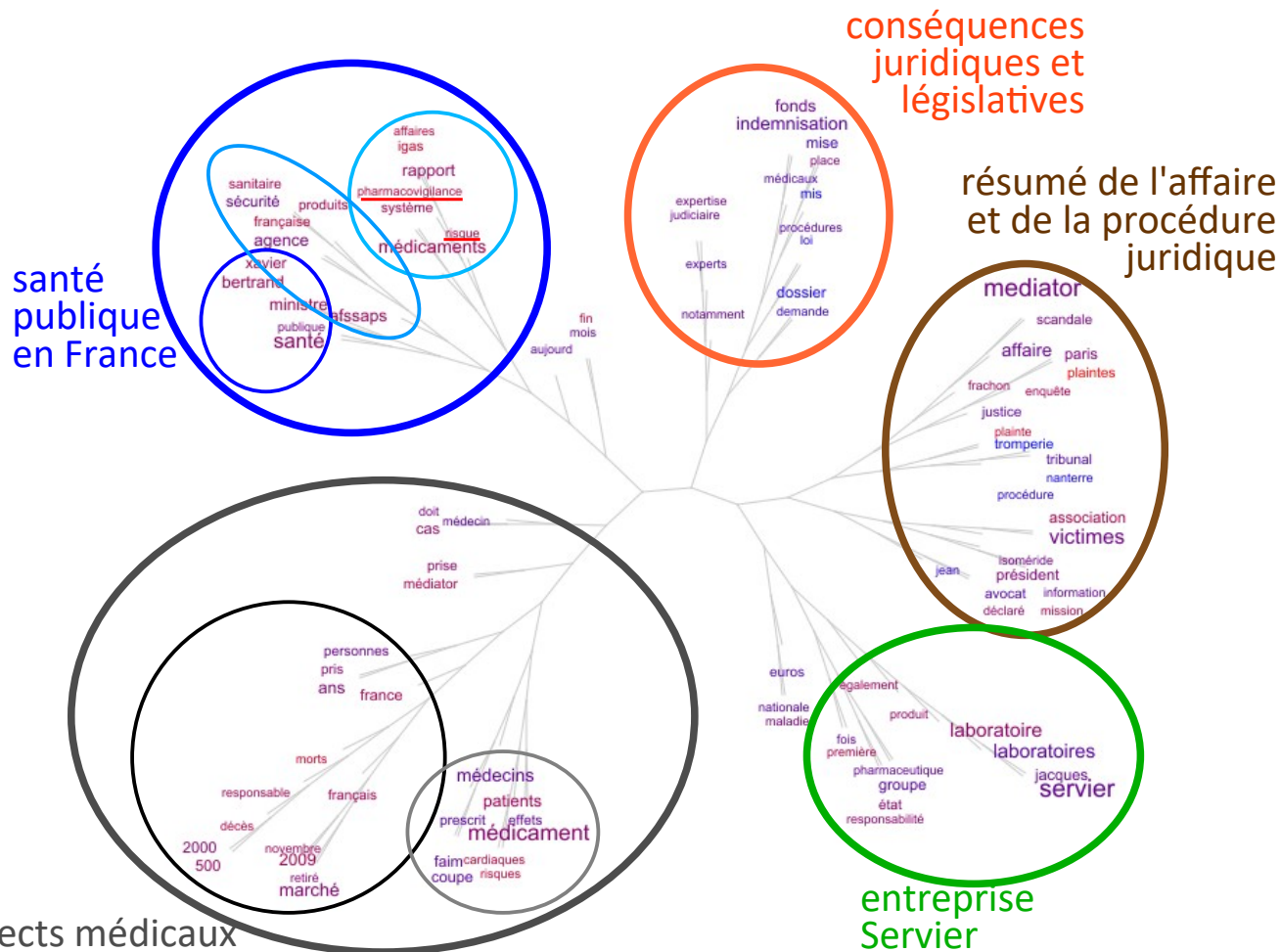
Gambette & Martinez,  
*Texto!*, 2013

# Illustration sur le corpus Mediator

## *Comparer les articles d'agences et articles de journalistes*

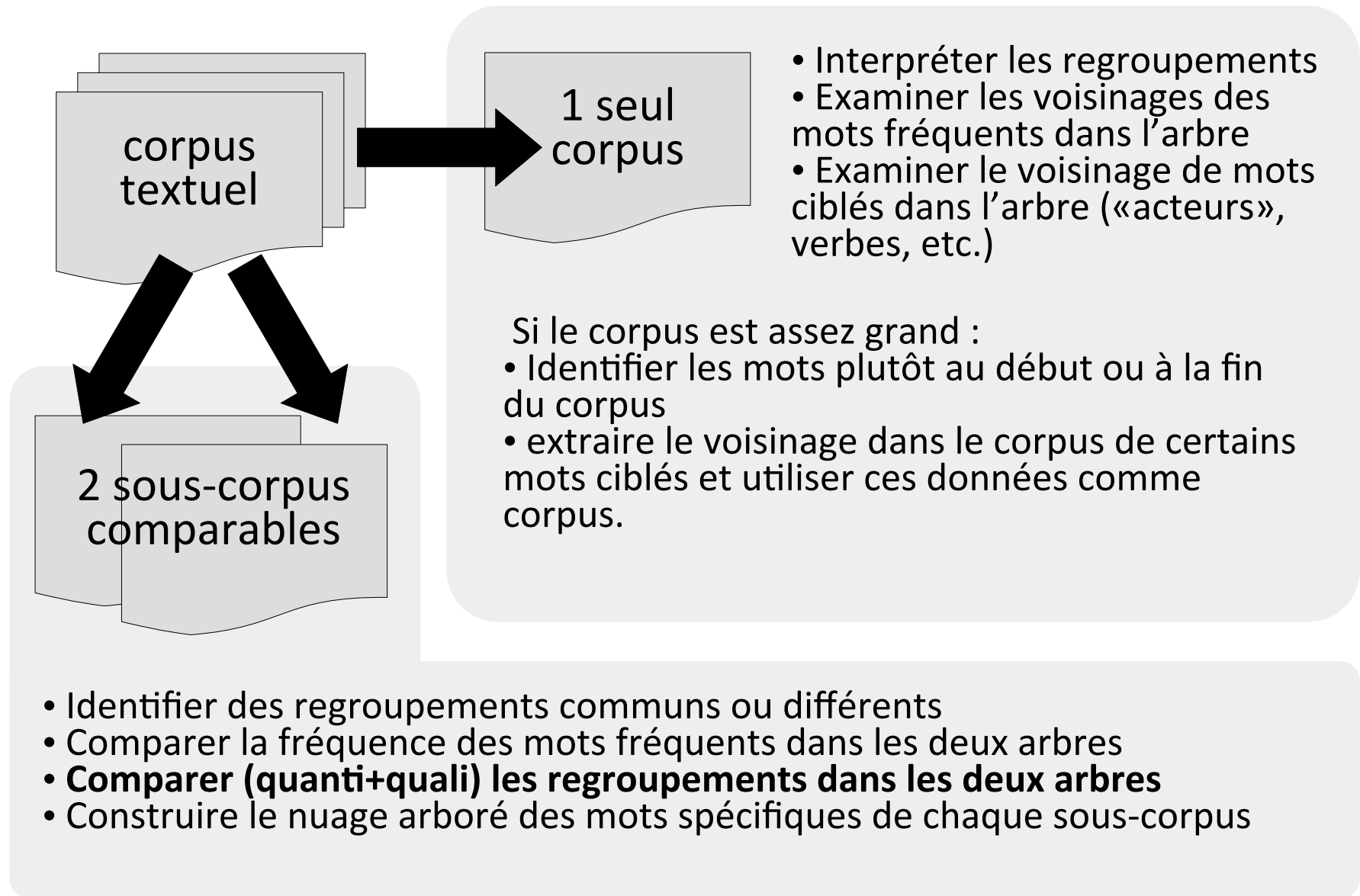
Corpus : 595 articles d'agences contre 1496 articles de journalistes de 2011 évoquant l'affaire du Mediator dans la presse française.

Articles  
de journalistes

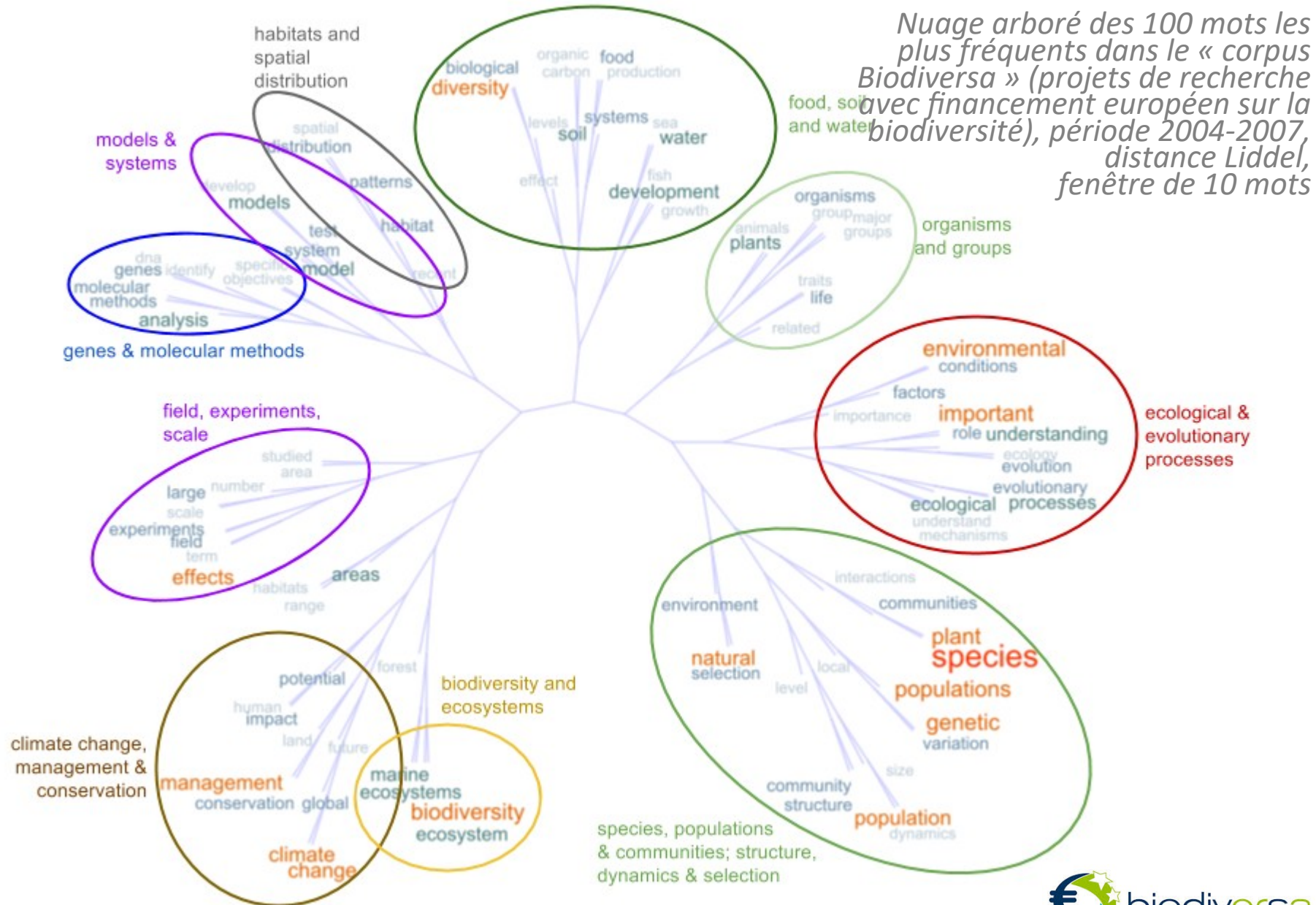


Gambette & Martinez,  
*Texto!*, 2013

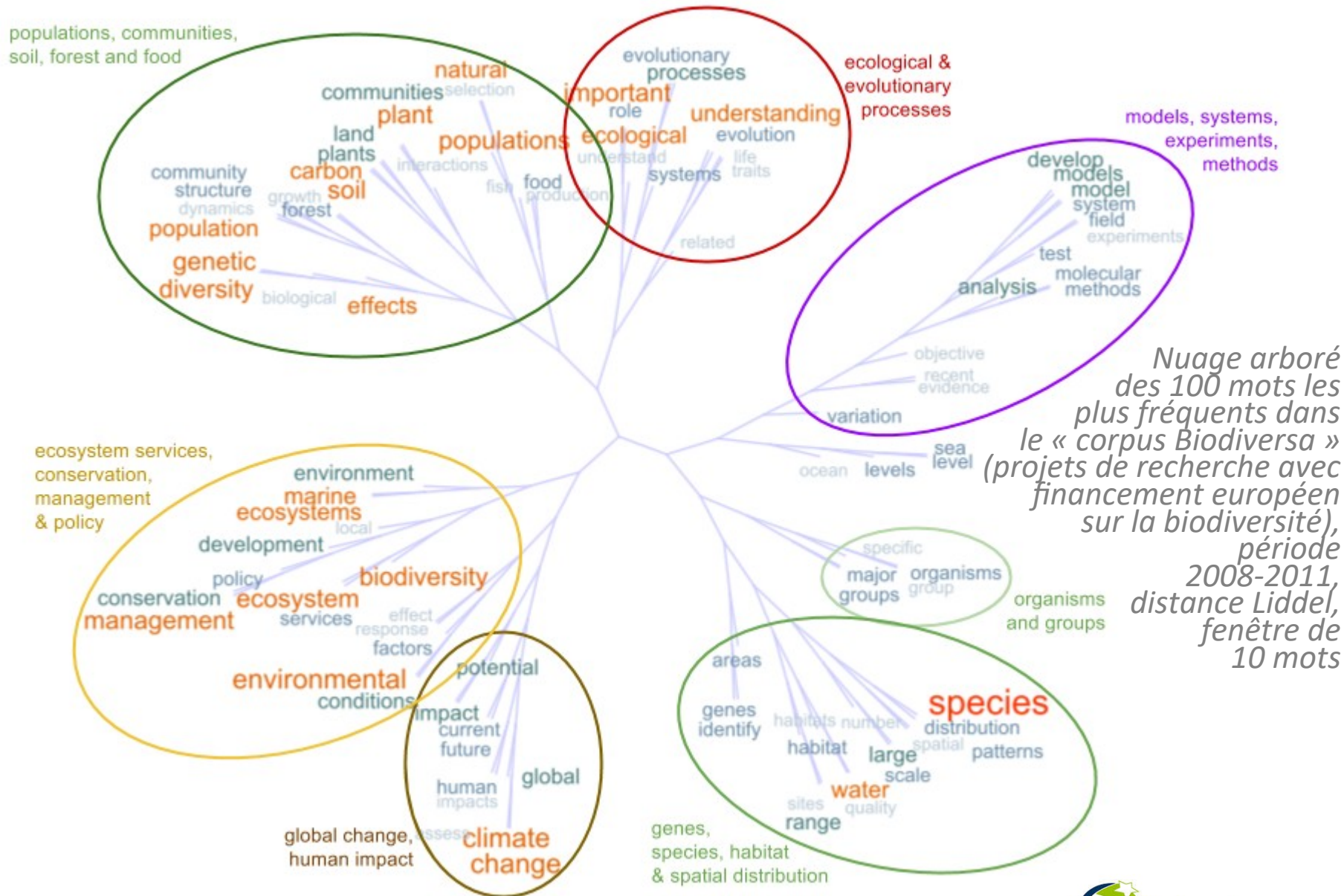
# Exploration de corpus avec TreeCloud



# Méthode : comparaison de voisinages dans l'arbre



# Méthode : comparaison de voisinages dans l'arbre

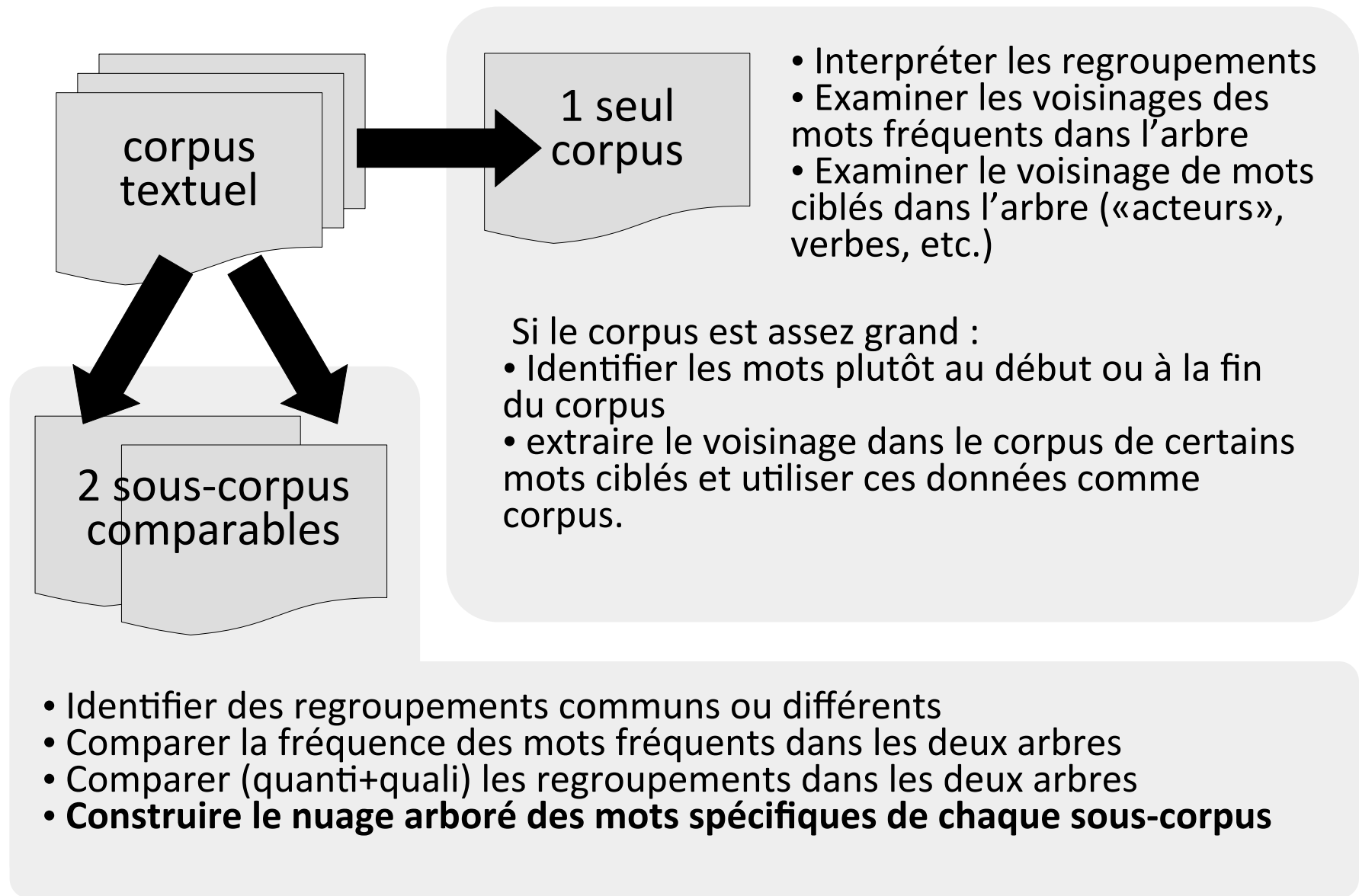


*Nuage arboré  
des 100 mots les  
plus fréquents dans  
le « corpus Biodiversa »  
(projets de recherche avec  
financement européen  
sur la biodiversité),  
période  
2008-2011,  
distance Liddel,  
fenêtre de  
10 mots*



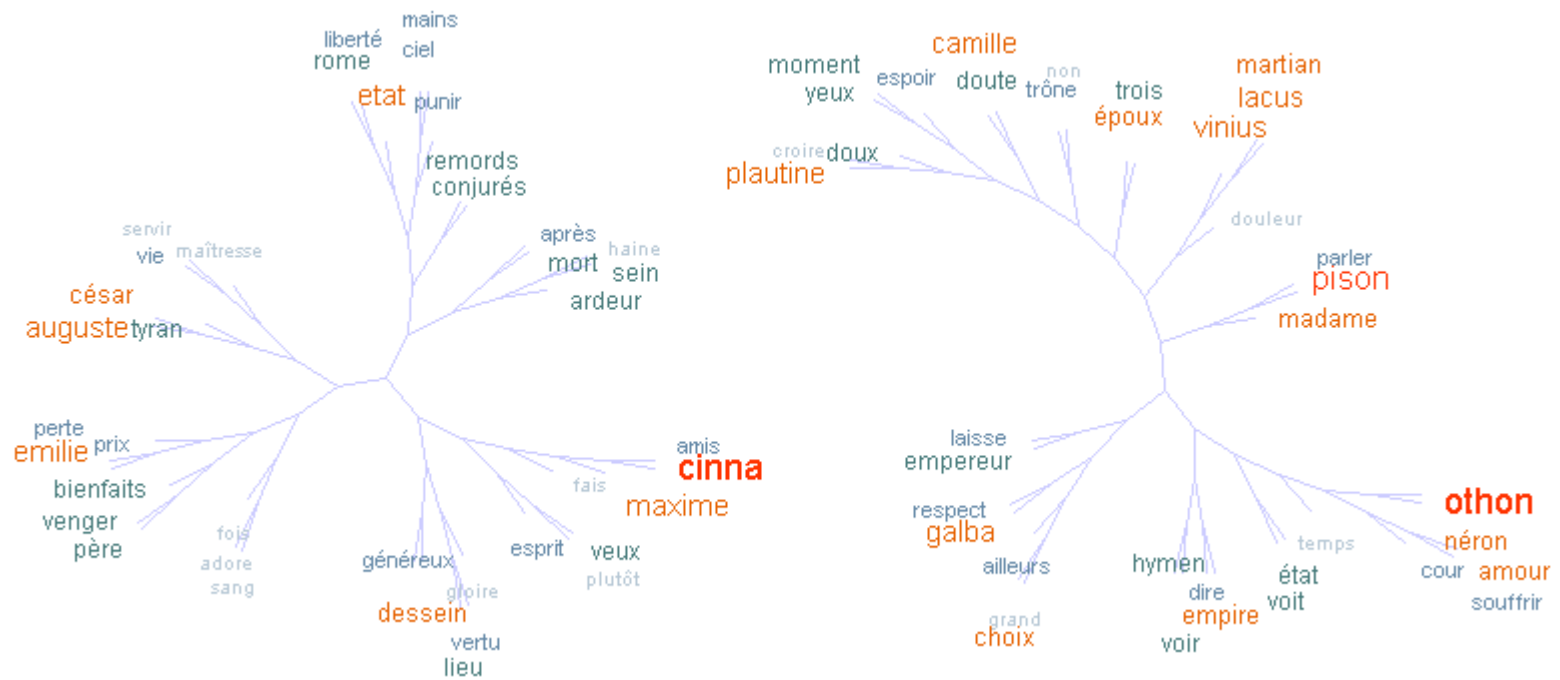


# Exploration de corpus avec TreeCloud



# Méthode : comparaison des spécifiques

Amstutz & Gambette,  
JADT 2010



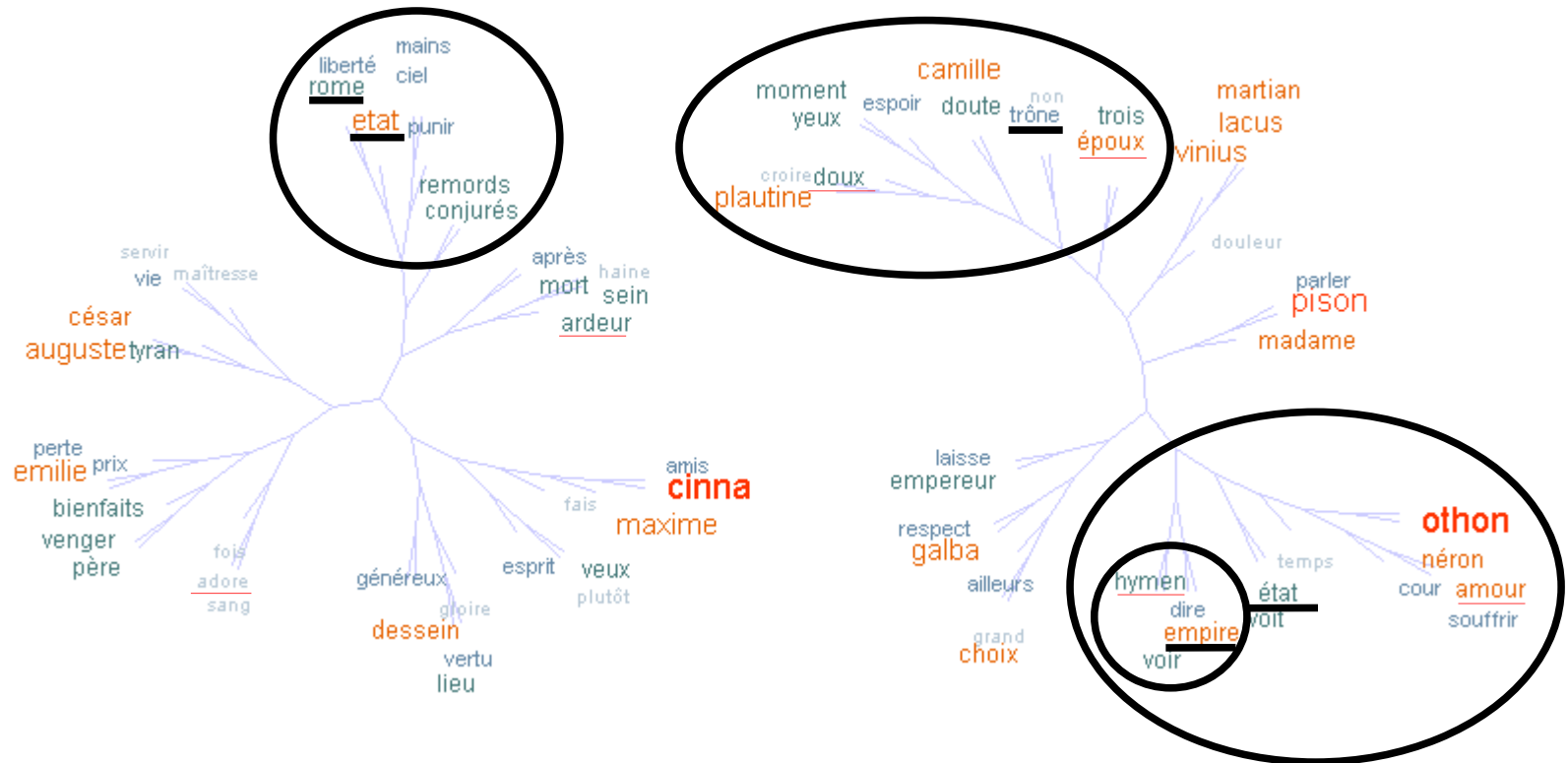
*Nuages arborés des mots spécifiques de Cinna et Othon, dimensionnés et colorés d'après leur spécificité calculée dans Lexico3.*

**Quels moyens au service de la cause politique ?**



# Méthode : comparaison des spécifiques

Amstutz & Gambette,  
JADT 2010

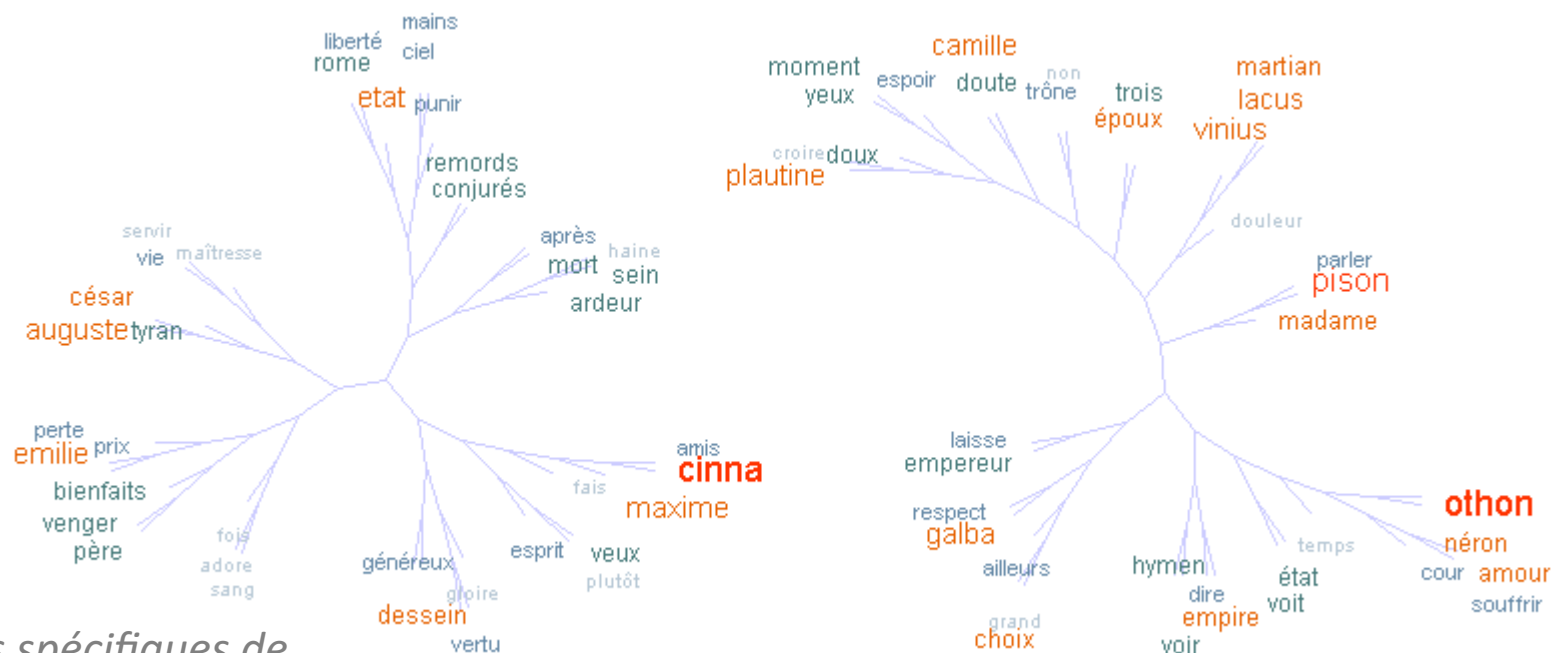


*Nuages arborés des **mots spécifiques** de Cinna et Othon, dimensionnés et colorés d'après leur spécificité calculée dans Lexico3.*

**Quels moyens au service de la cause politique ?**



# Méthode : comparaison des spécifiques

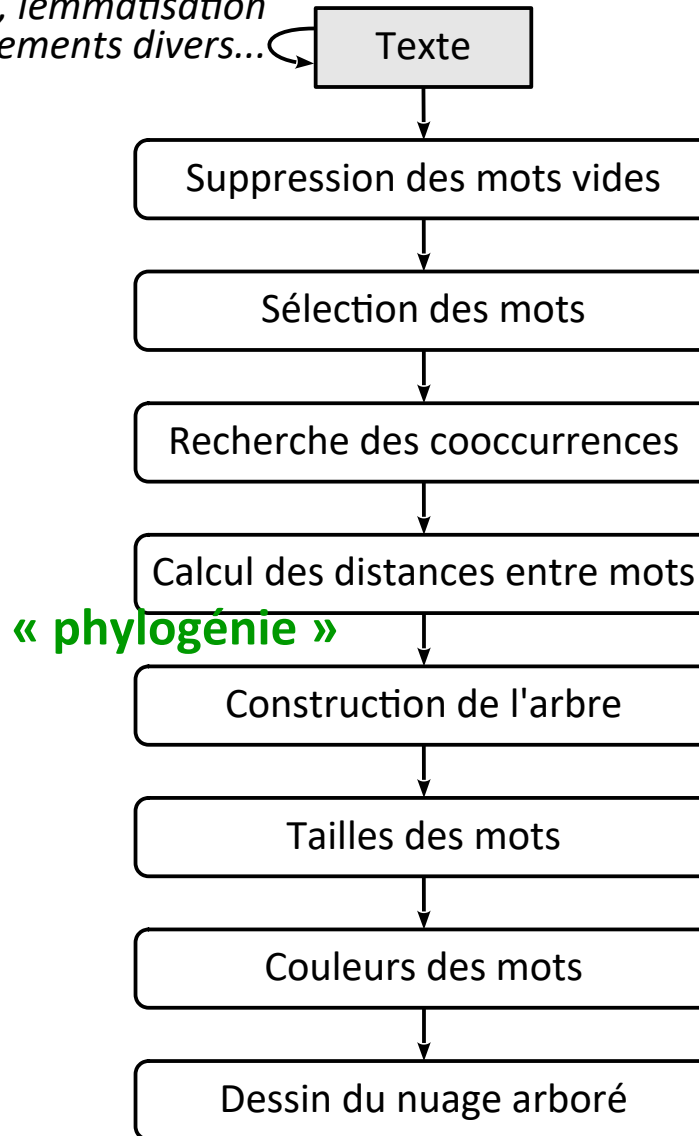


*mots spécifiques de  
Cinna et Othon d'après Lexico3*

|                                                                    | <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                      |

# 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*

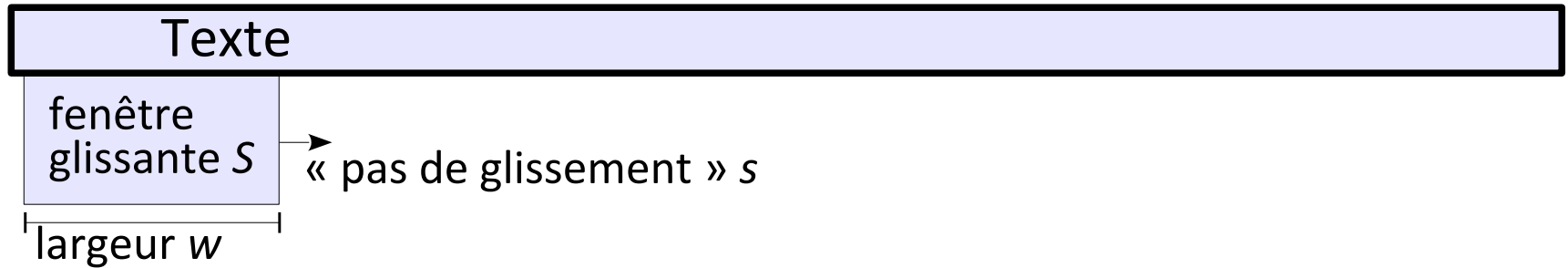
*Fréquences ou valeurs personnalisées*

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

*Appel transparent au logiciel SplitsTree  
ou Dendroscope*

# Calcul des proximités entre mots

Déplacement d'une « fenêtre glissante » tout au long du texte pour compter, pour chaque paire de mots  $u$  et  $v$ , leurs cooccurrences :



Nombre de cooccurrences  
entre  $u$  et  $v$

| nombre de<br>fenêtres   | mot $u$ dans<br>$S$ | mot $u$ pas<br>dans $S$ |
|-------------------------|---------------------|-------------------------|
| mot $v$ dans<br>$S$     | 5                   | 10                      |
| mot $v$ pas<br>dans $S$ | 50                  | 935                     |

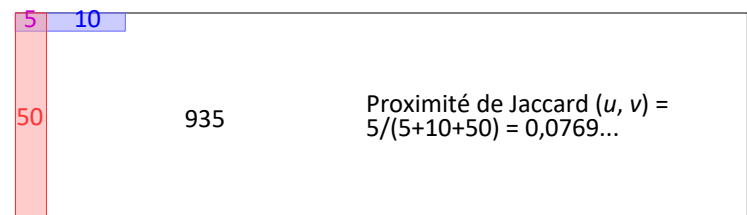
Exemple : texte de 991 mots avec 55  
occurrences de  $u$  et 15 de  $v$ ,  $w=10$ ,  $s=1$



Score de cooccurrence  
entre  $u$  et  $v$

*chi squared, mutual information,  
liddel, dice, jaccard, gmean,  
hyperlex, minimum sensitivity,  
odds ratio, zscore, log likelihood,  
poisson-stirling...*

Evert,  
*Statistics of words cooccurrences*, Thèse, 2005



# Arbres phylogénétiques et arbres de mots

## ESPÈCES

Séquences ADN

Données sur  
les feuilles

## MOTS

Position des mots

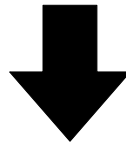
# Arbres phylogénétiques et arbres de mots

## 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

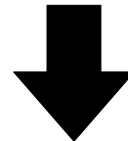
Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

## 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

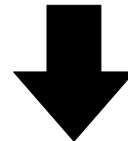
Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

## MOTS

Position des mots

Distances fondées sur la cooccurrence entre les deux mots

Données sur les feuilles



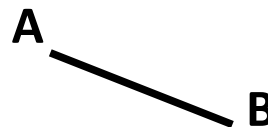
Distances entre les feuilles

|     | 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

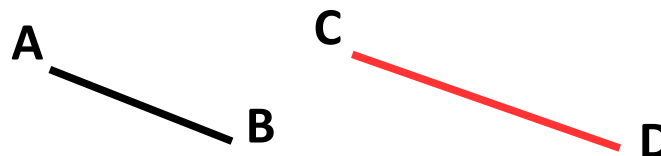
Distances entre les feuilles

|     | 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)

## MOTS

Position des mots

Distances fondées sur la cooccurrence entre les deux mots

Données sur les feuilles

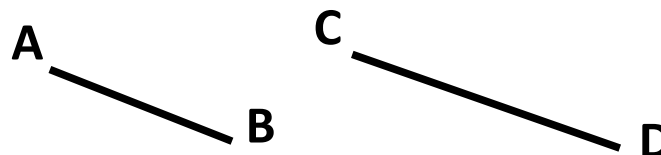
Distances entre les feuilles

|     | 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

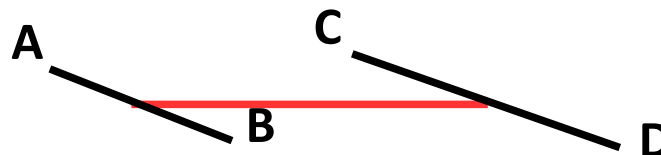
Distances entre les feuilles

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



*classification hiérarchique ascendante*  
*algorithme UPGMA*

Arbre



# Arbres phylogénétiques et arbres de mots

## ESPÈCES

Séquences ADN

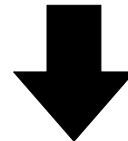
Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

## MOTS

Position des mots

Distances fondées sur la cooccurrence entre les deux mots

Données sur les feuilles



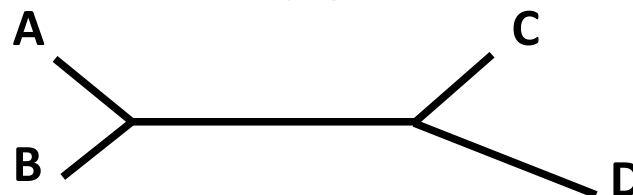
Distances entre les feuilles

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



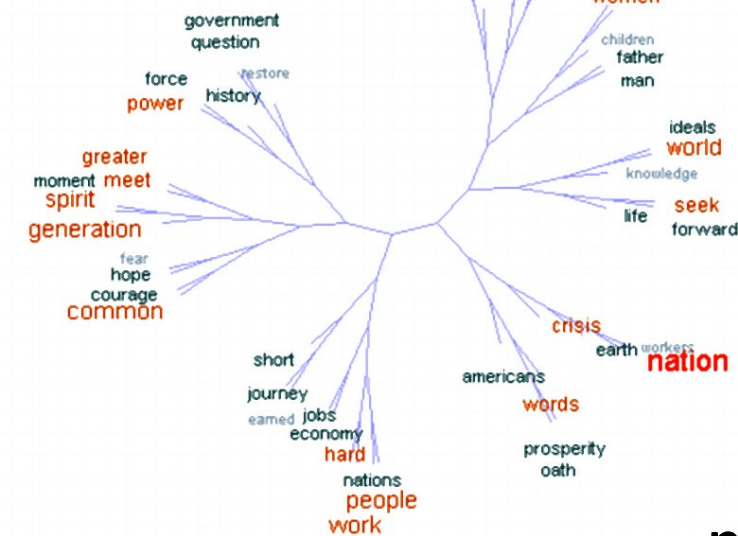
*classification hiérarchique ascendante  
algorithme UPGMA*

Arbre



# Comparaison avec d'autres visualisations

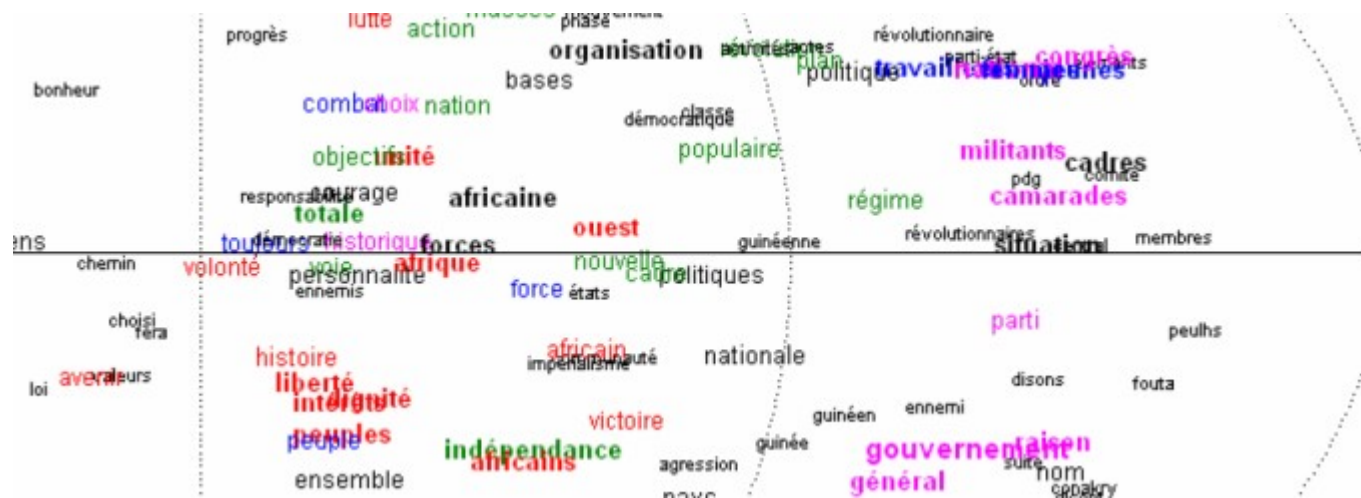
## nuage arboré (TreeCloud)



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



## projection des mots (Astartex)



# Comparaison avec d'autres visualisations

## nuage arboré (TreeCloud)



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



## projection des mots (Astartex)





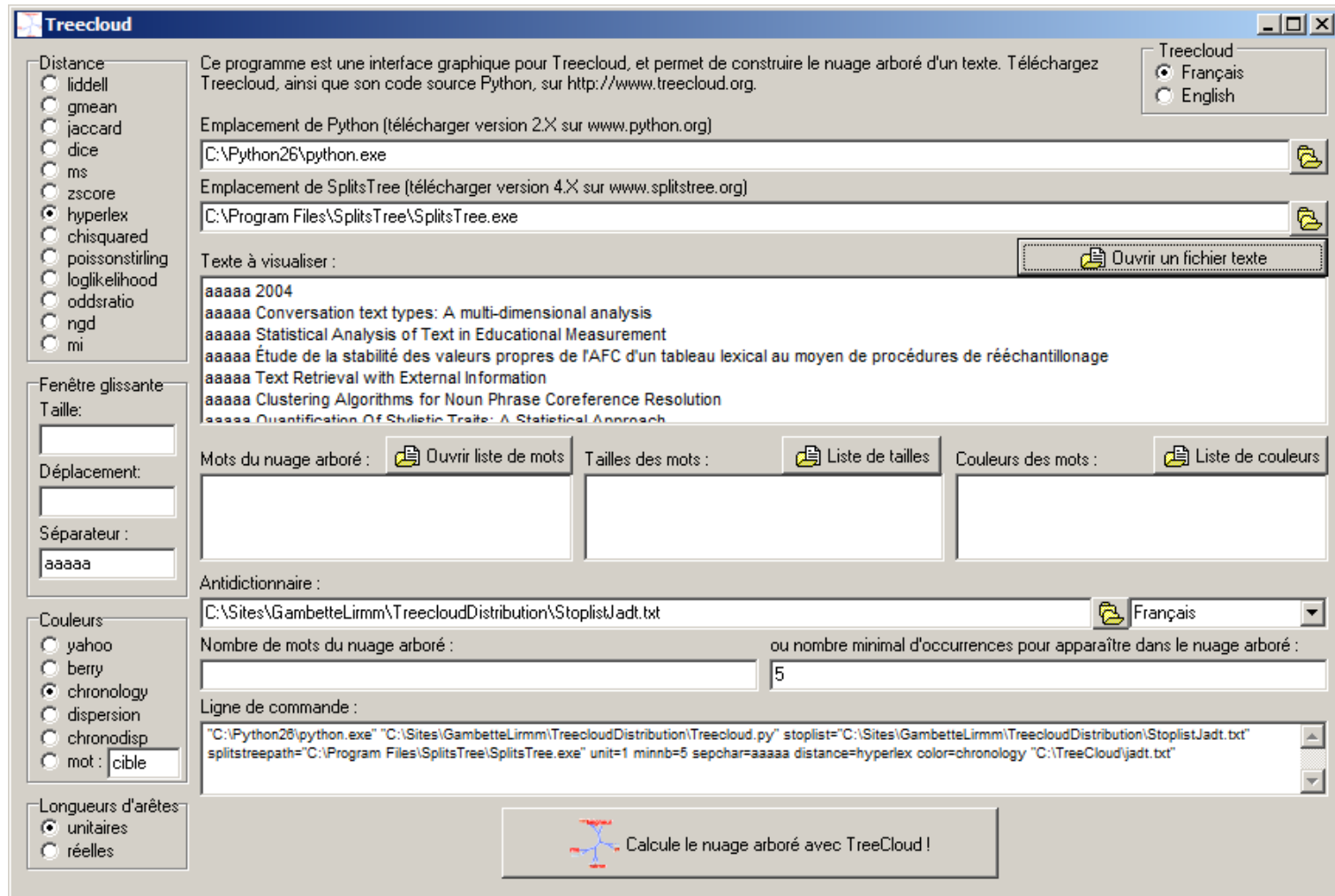






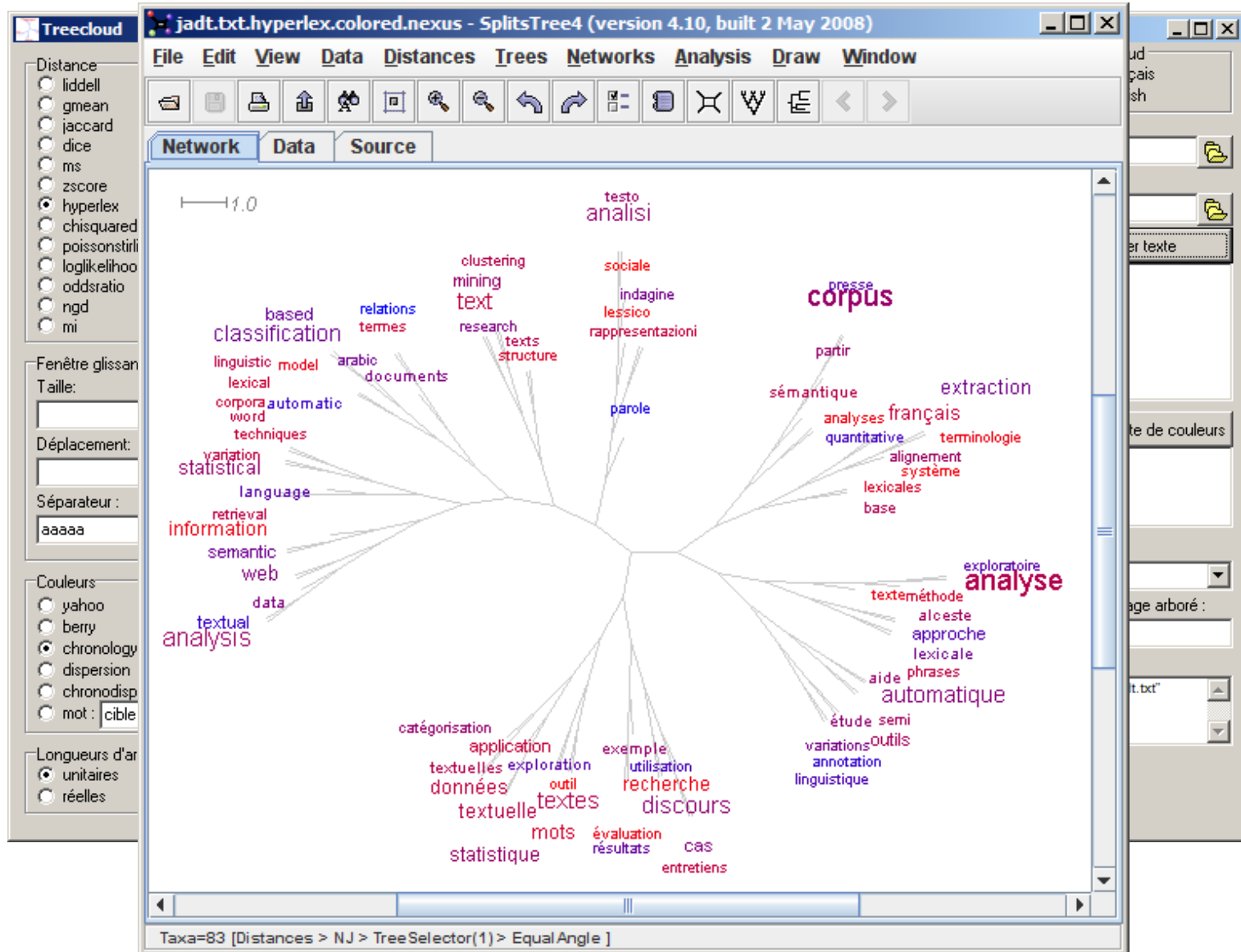
# Implémentations

## 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

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](http://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 IFCV'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](#)).

© 2007-2010 - Jean Véronis, Philippe Gambette, Jean-Charles Bontemps

[www.treecloud.org](http://www.treecloud.org)

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

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

Développements supplémentaires avec d3.js par Deepak Srinivas



# Interface web

[www.treecloud.org](http://www.treecloud.org)

[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 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: [Visualizing a Word Cloud as a Tree Cloud](#) 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: [Visualizing a Word Cloud as a Tree Cloud](#) [JADT'10](#) (10th International Conference supplémentaire).

© 2007-2010 - Je

[Créer!](#) [Téléchargements](#) [Galerie](#) [A propos](#) [FAQ](#)

**Créez vos propres nuages arborés !**

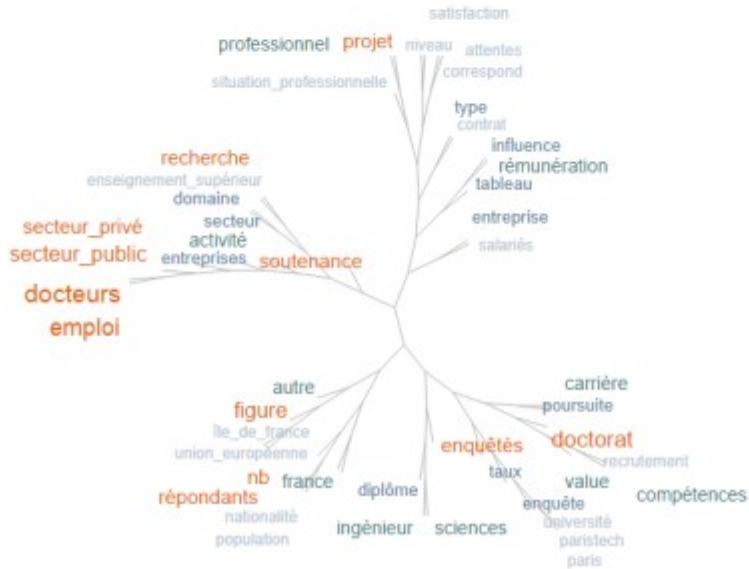
Collez votre texte dans le cadre ci-dessous, puis cliquez sur *Envoyer* ! Attention, l'utilisateur suivant verra votre texte quand il se connectera au site, si vous ne voulez pas faire apparaître vos textes, installez plutôt [TreeCloud](#) sur votre machine.

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

[www.treecloud.org](http://www.treecloud.org)

[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 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 ([supplementary r](#)

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

© 2007-2010 - Je

[Créer!](#) [Téléchargements](#) [Galerie](#) [A propos](#) [FAQ](#)

**Créez vos propres nuages arborés !**

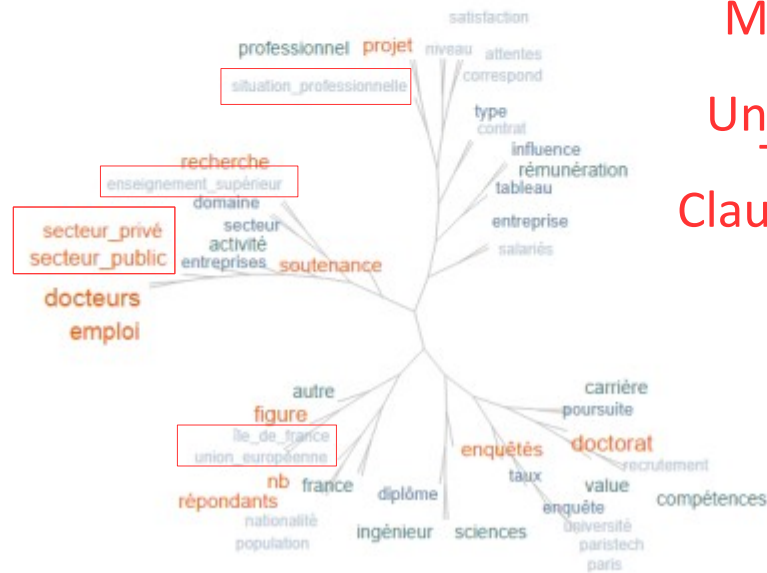
Collez votre texte dans le cadre ci-dessous, puis cliquez sur *Envoyer* ! Attention, l'utilisateur suivant verra votre texte quand il se connectera au site, si vous ne voulez pas faire apparaître vos textes, installez plutôt [TreeCloud](#) sur votre machine.

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

[www.treecloud.org](http://www.treecloud.org)

 **TreeCloud**  
Nuage Arboré

[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 structure.

The first tree cloud appeared on [Jean Véronis's blog](#). You can [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 de textes. Les mots sont disposés autour d'un arbre qui reflète 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 le logiciel TreeCloud](#).

### Créez vos propres nuages arborés en ligne

#### Documents :

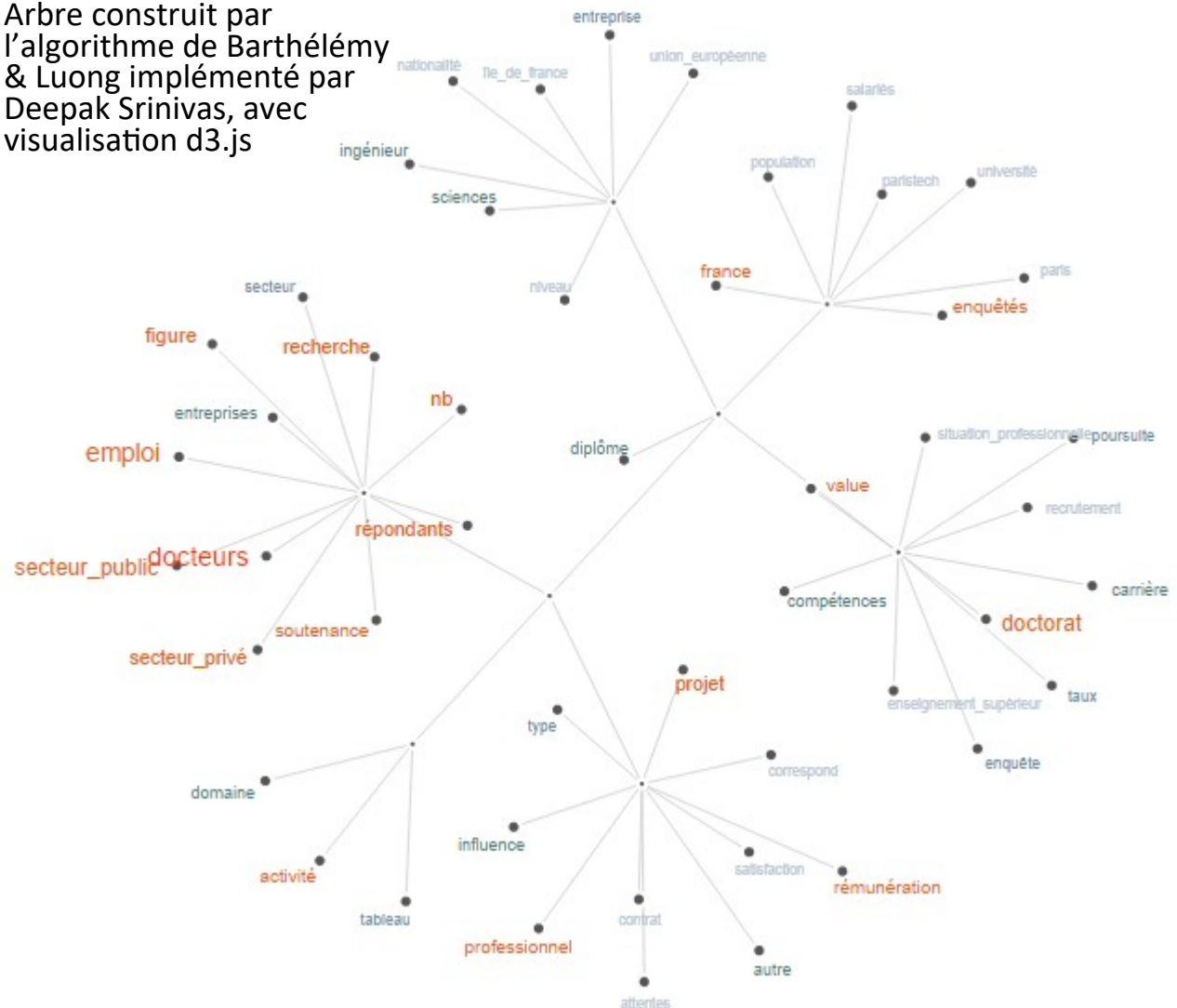


If you use TreeCloud or this website, please cite [www.treecloud.org](#), [Philippe Gambette](#) et [Jean Véronis](#): *Visualising a Text: Classification as a Tool of Research*, *Proc. of IJCS'09 (10th International Conference on Intelligent and Intelligent Societies)*, to appear, 2010 ([supplementary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

Pour des exemples d'utilisation de la visualisation en ligne, voir [Delphine Amstutz et Philippe Gambette](#): *Utilisation de TreeCloud*, *IJADT'10 (10th International Conference on Intelligent and Intelligent Societies)*, to appear, 2010 ([supplémentaire](#)).

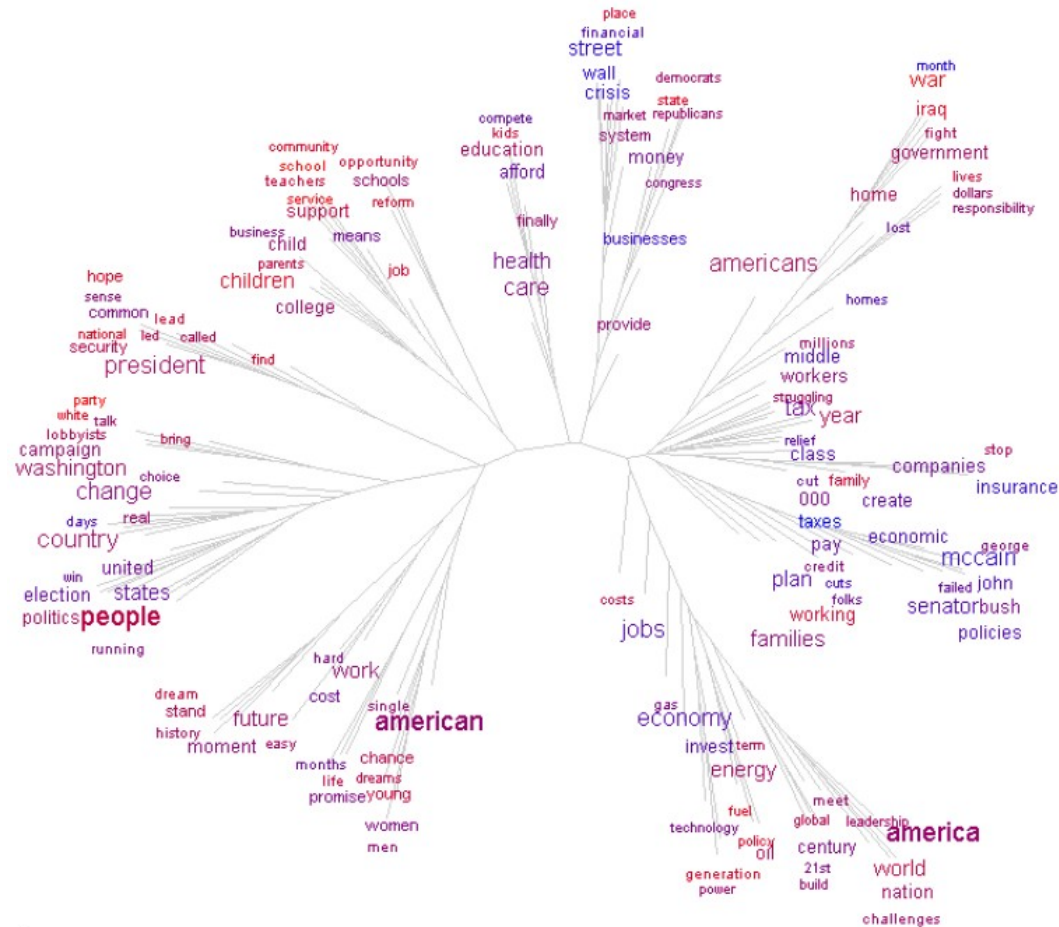
© 2007-2010 - Jean Véronis, Philippe Gambette

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)

# 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

- 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-litteraire>

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 divers

Utilisation de formules de tableur pour pré-traiter des corpus :

- **Données d'enquête :**

- ajouter “ a a a a a a a a ” à la fin de chaque réponse à une question ouverte, pour éviter que les mots d'une réponse soient considérés proches des mots de la réponse suivante (si *fenêtre glissante* paramétrée à 10 mots).
- possibilité de filtrer les lignes pour sélectionner uniquement les réponses d'un échantillon donné.

- **Données d'entretien :**

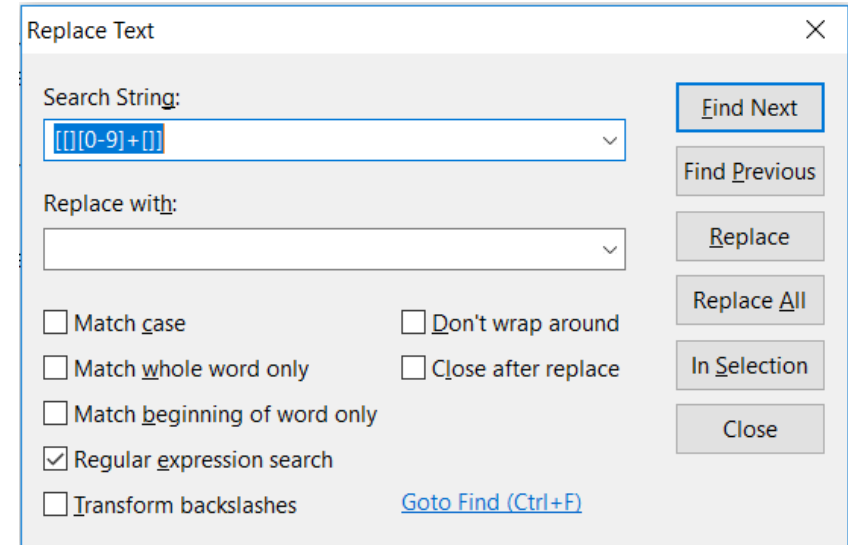
- mettre le nom du locuteur sur la ligne précédant ses paroles
- possibilité d'utiliser une formule pour créer une colonne avec le nom du locuteur sur chaque ligne
- Filtre pour sélectionner seulement les paroles d'un locuteur

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

# 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. Ce 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 :

Contextes

Corrélations

|   | 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)

 Export

 Cancel