

Symposium on Software and Digital Humanities II  
24/05/2016 – Paris IAS

# *Visualization and analysis of textual corpora with trees of words*

Philippe Gambette

LIGM  
Université Paris-Est  
Marne-la-Vallée

UNIVERSITÉ ———  
— PARIS-EST



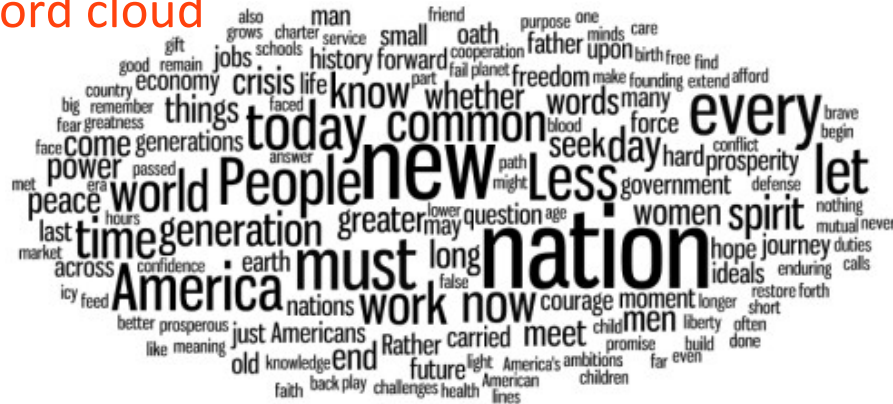
UP  
EM  
UNIVERSITÉ  
PARIS-EST  
MARNE-LA-VALLÉE



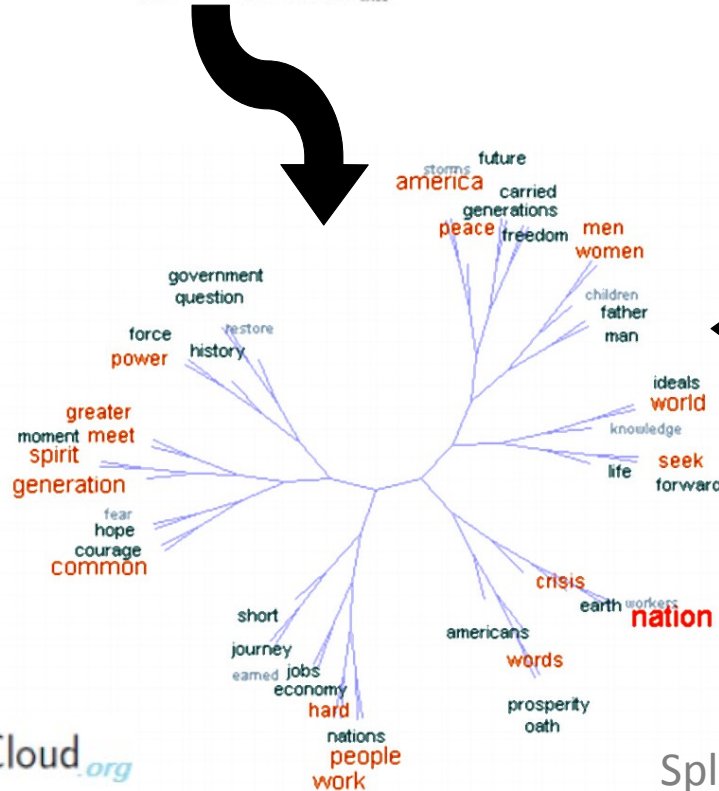
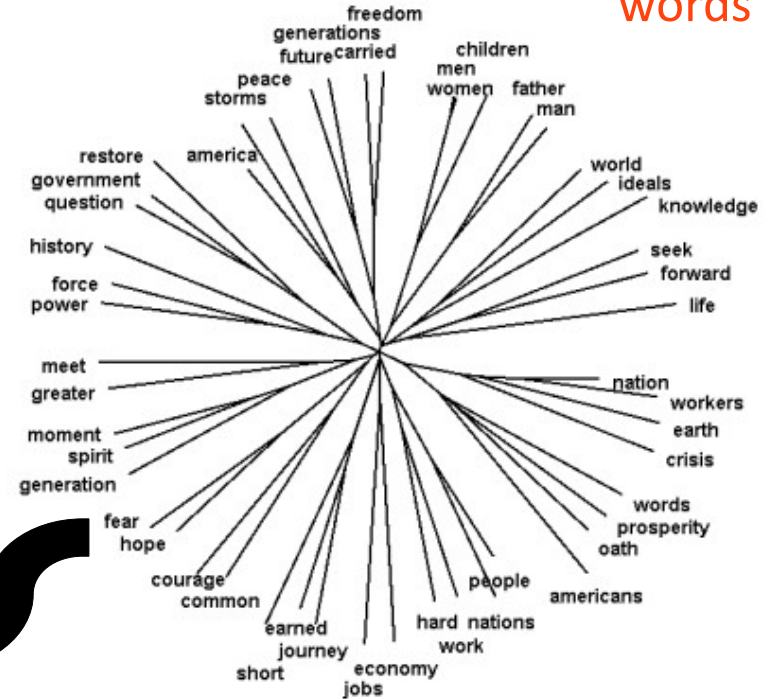


# Combining 2 levels of information with tree clouds

word cloud



tree of words



Inaugural speech of Barack Obama

SplitsTree: Huson & Bryant, *Bioinformatics*, 2006

TreeCloud: Gambette & Véronis, *IFCS'09*

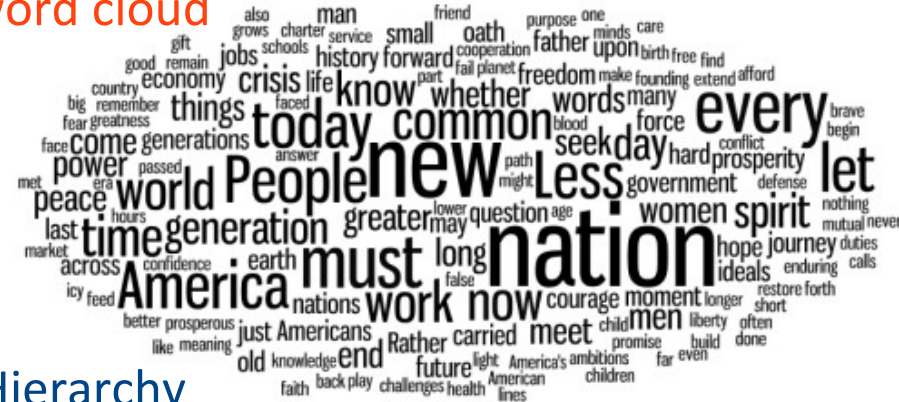
built with

TreeCloud.org

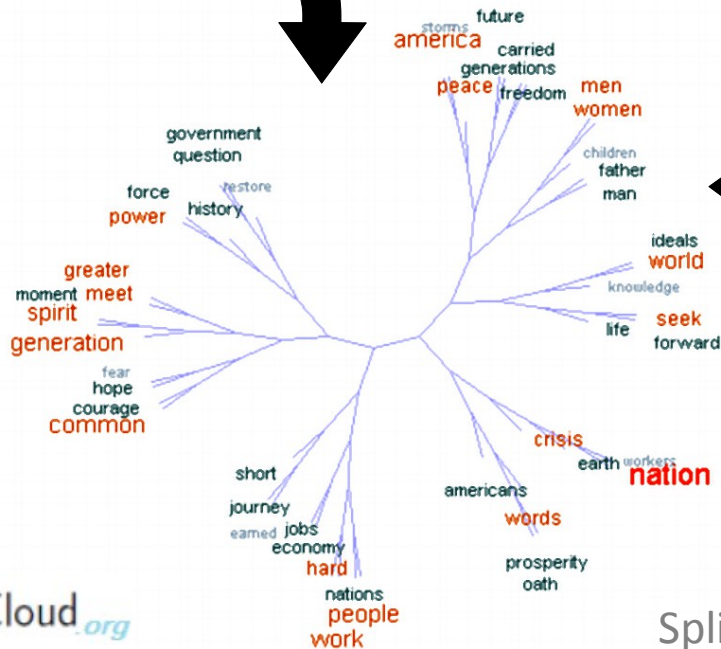
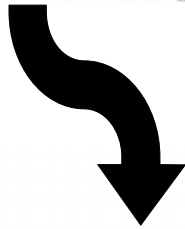
SplitsTree4

# Combining 2 levels of information with tree clouds

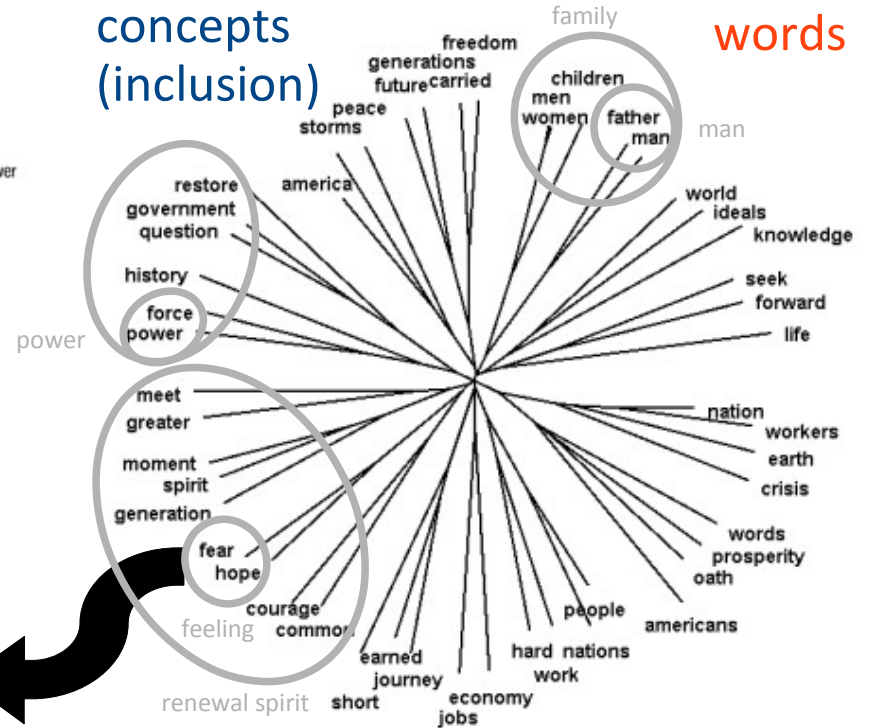
word cloud



Hierarchy of words (frequency)



Hierarchy of concepts (inclusion)

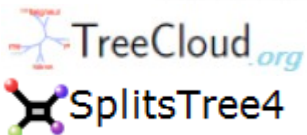


tree of words

Inaugural speech of Barack Obama

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

built with



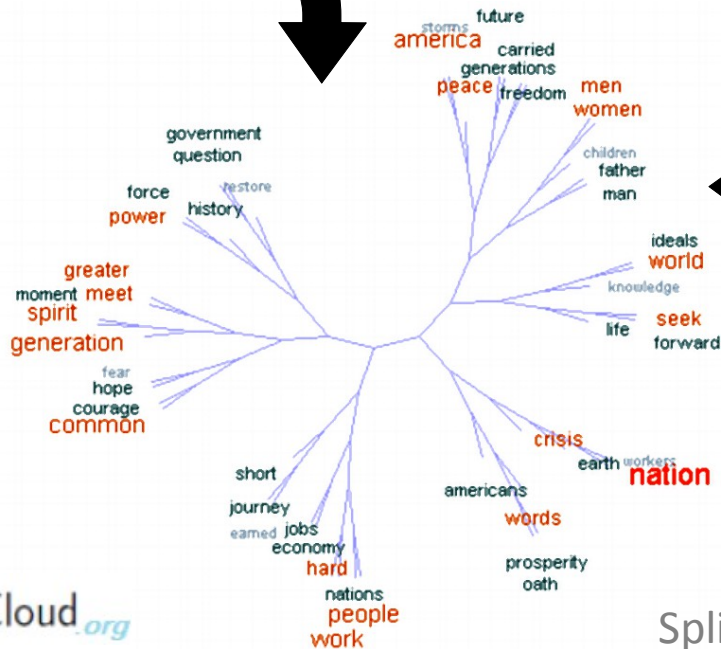
# Combining 2 levels of information with tree clouds

word cloud

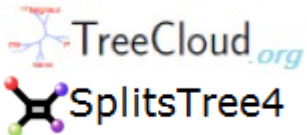


Hierarchy of words (frequency)

occurrences

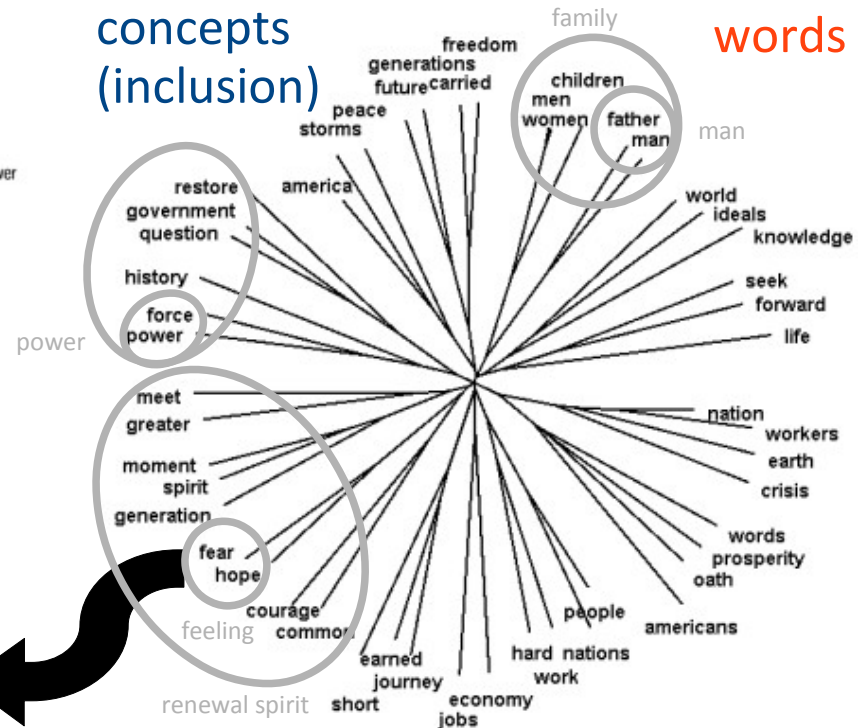


built with



Hierarchy of concepts (inclusion)

tree of words



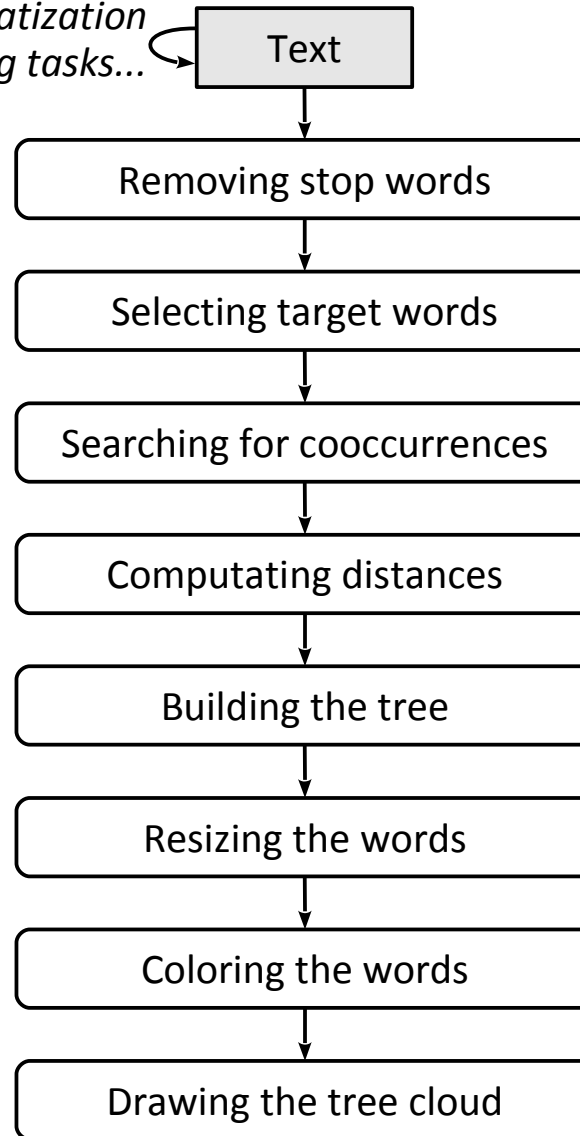
cooccurrences

Inaugural speech of Barack Obama

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

# Building a tree cloud

*Concordance of a word, lemmatization  
or other preprocessing tasks...*



## **Available in the TreeCloud standalone application**

*Stoplists (English, French, Spanish, etc.)*

*n most frequent words, or word with at least k occurrences, or user list*

*Cooccurrence window (parameters: size + sliding step), or separating character*

*12 formulas evaluating the cooccurrence distance between two words*

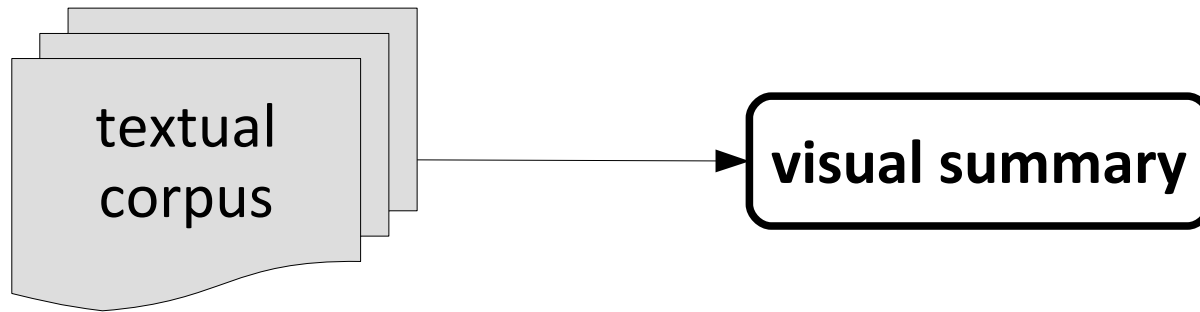
*Automatic call to SplitsTree (Neighbor-Joining)*

*Frequencies or user values*

*Frequency, average position in the text, dispersion, reflecting cooccurrence with a target word or user values*

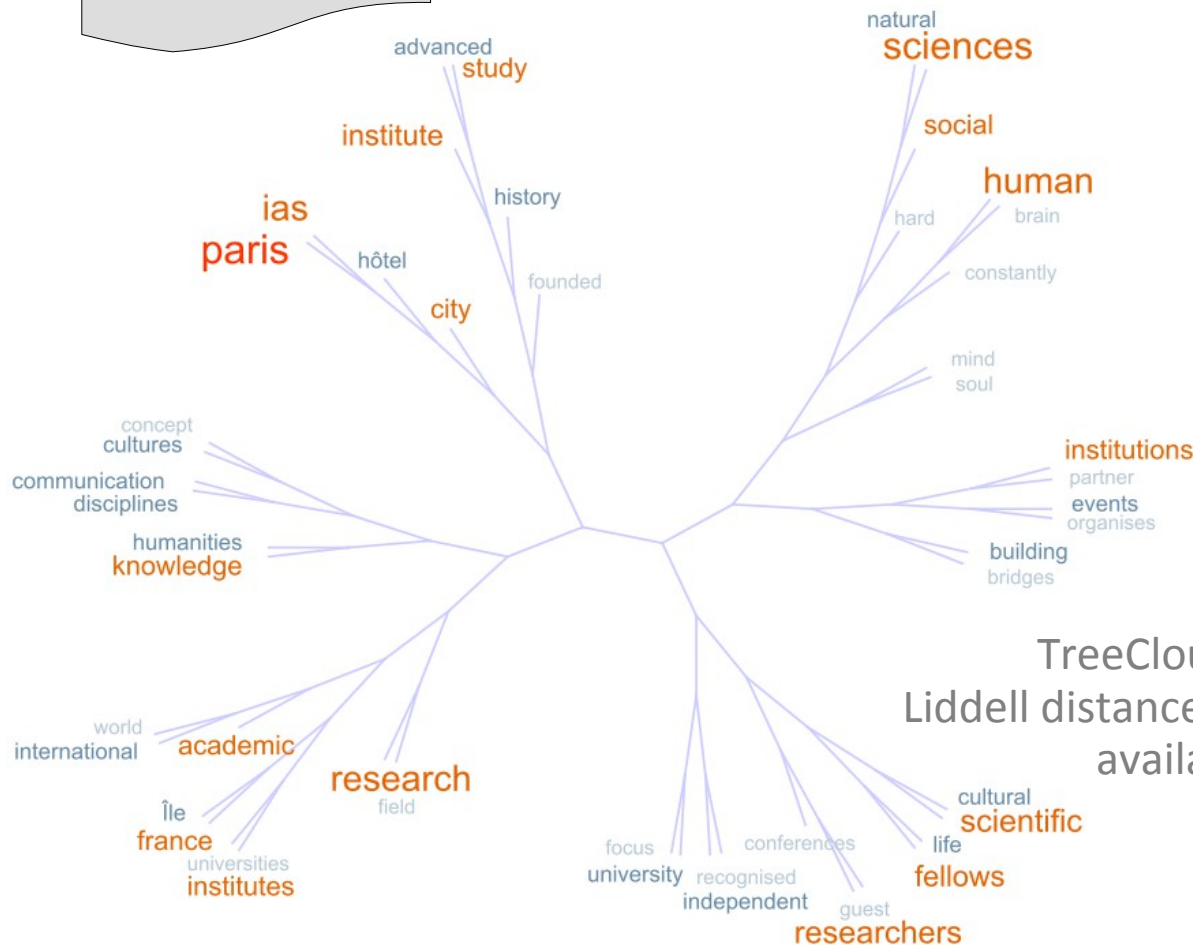
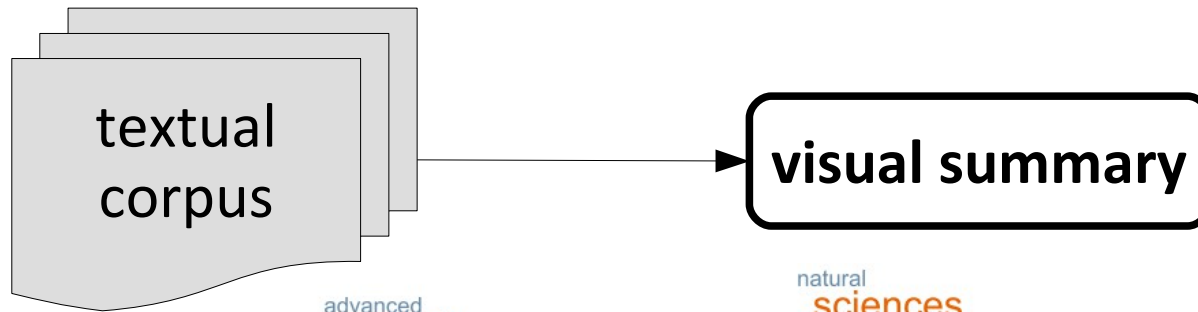
*Automatic call to SplitsTree ou Dendroscope*

# Why use tree clouds?



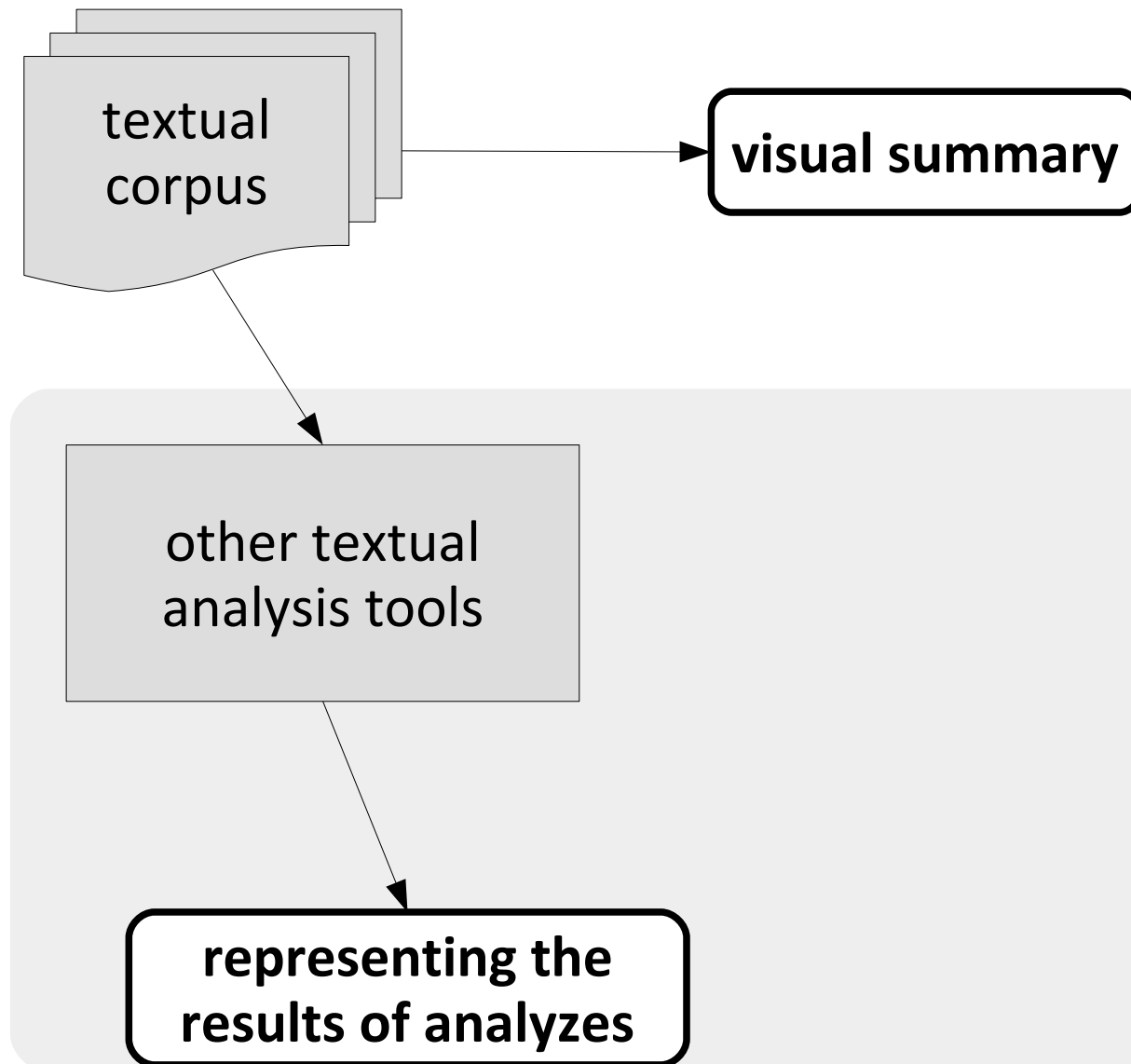


# Why use tree clouds?



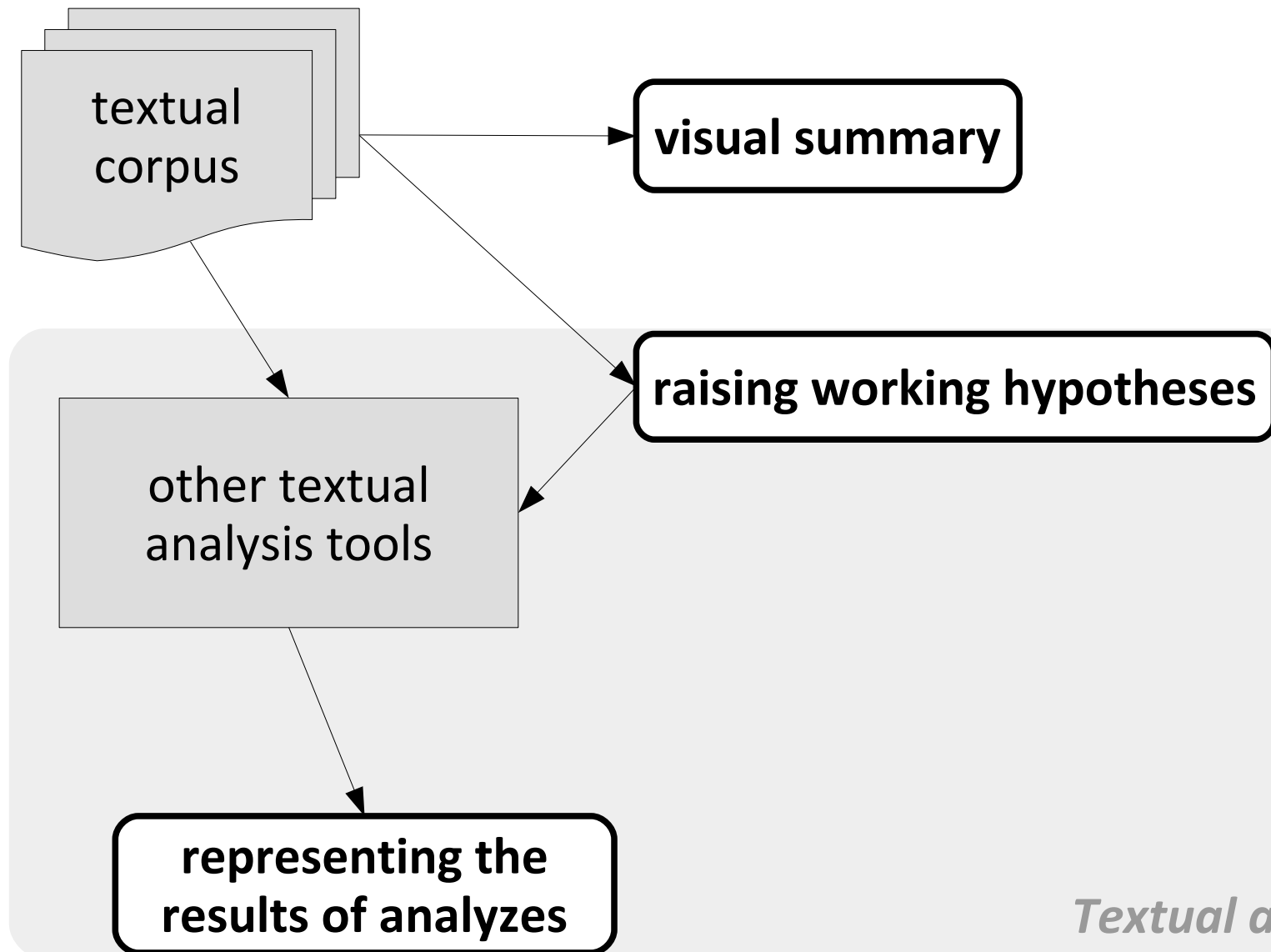
TreeCloud (50 words, 10 word window, Liddell distance) of the description of Paris IAS available on page <http://www.paris-iaea.fr/en/presentation-of-the-institute/mission-and-history>

# Why use tree clouds?



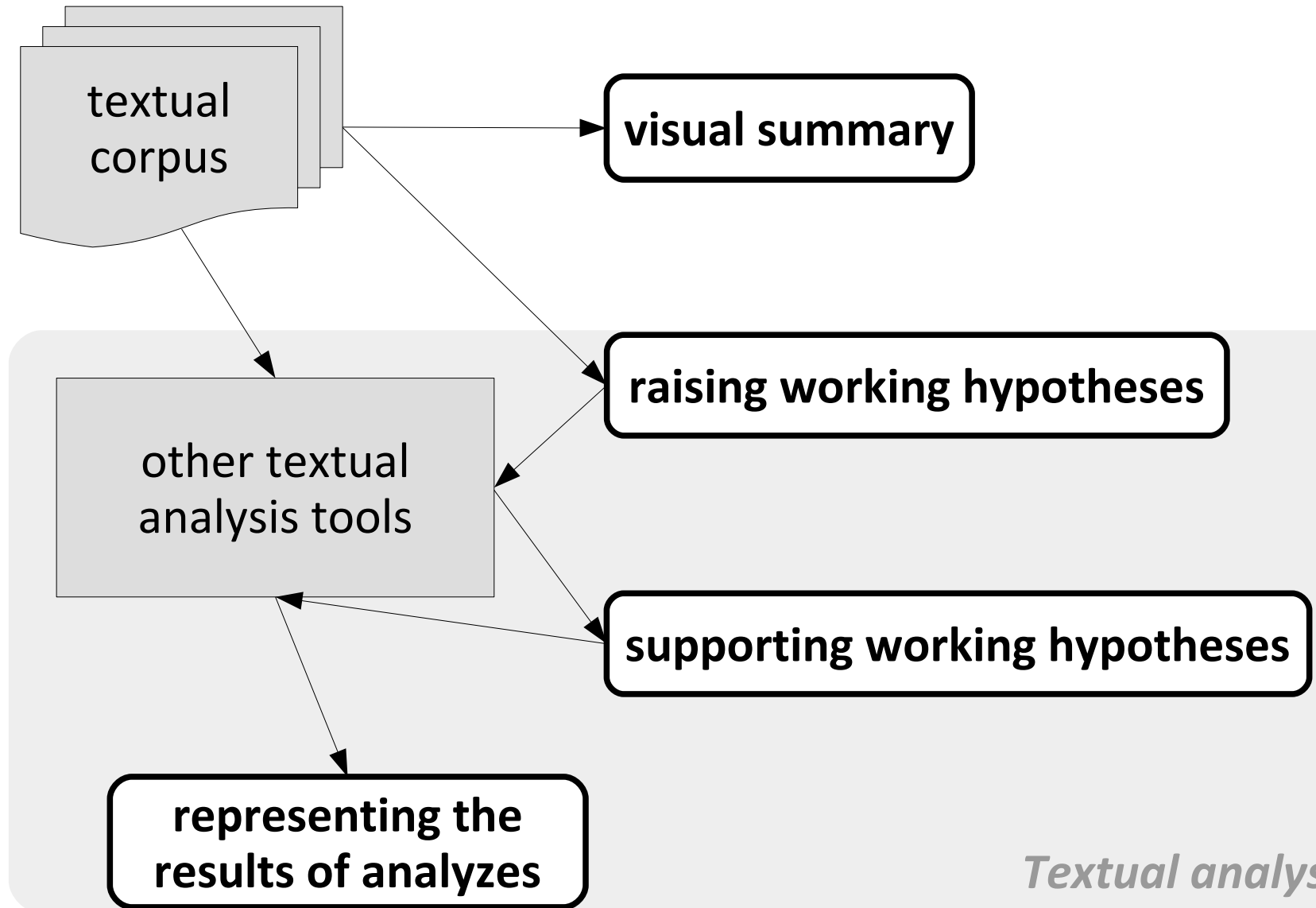
*Textual analysis*

# Why use tree clouds?



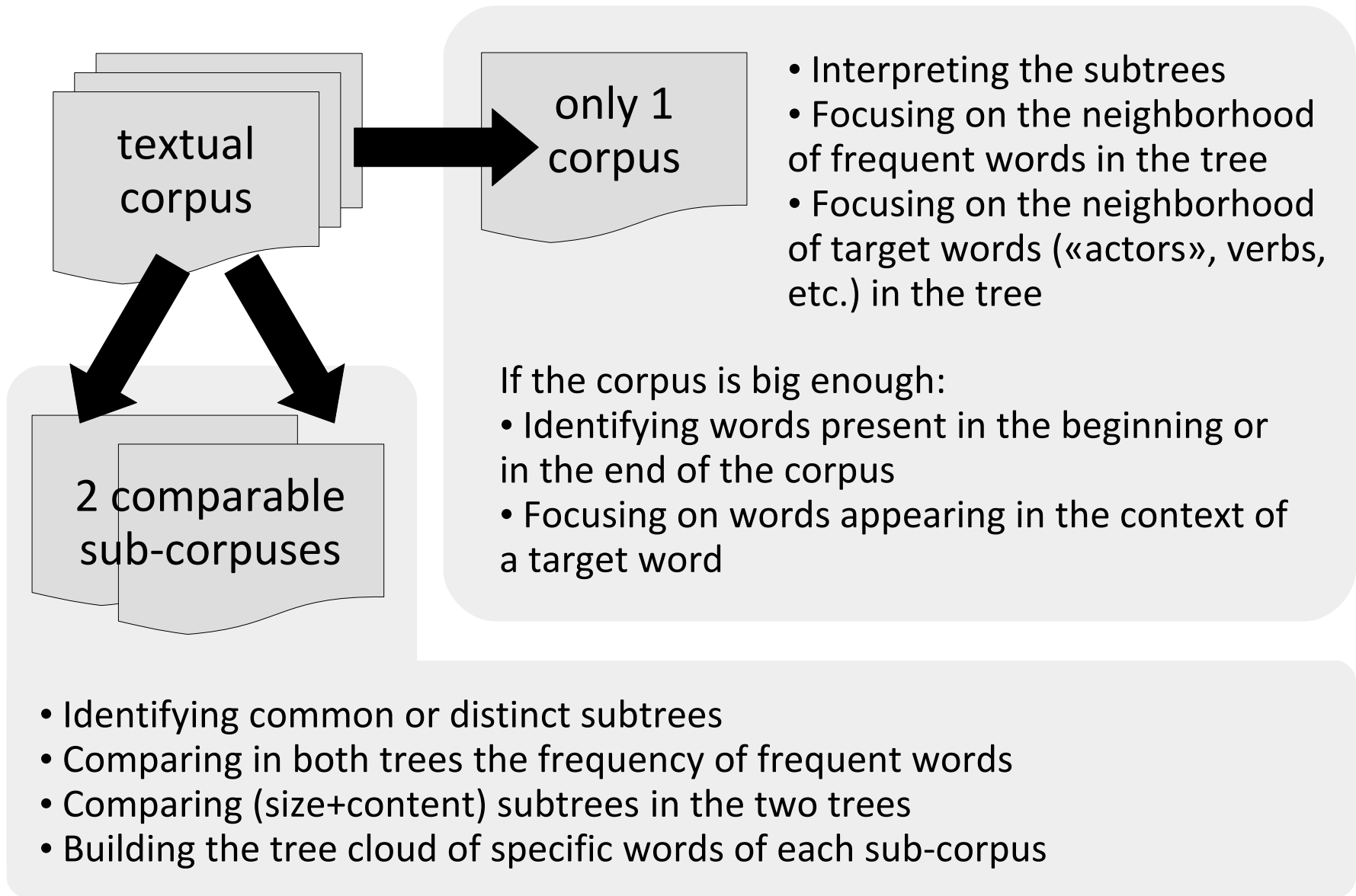
*Textual analysis*

# Why use tree clouds?

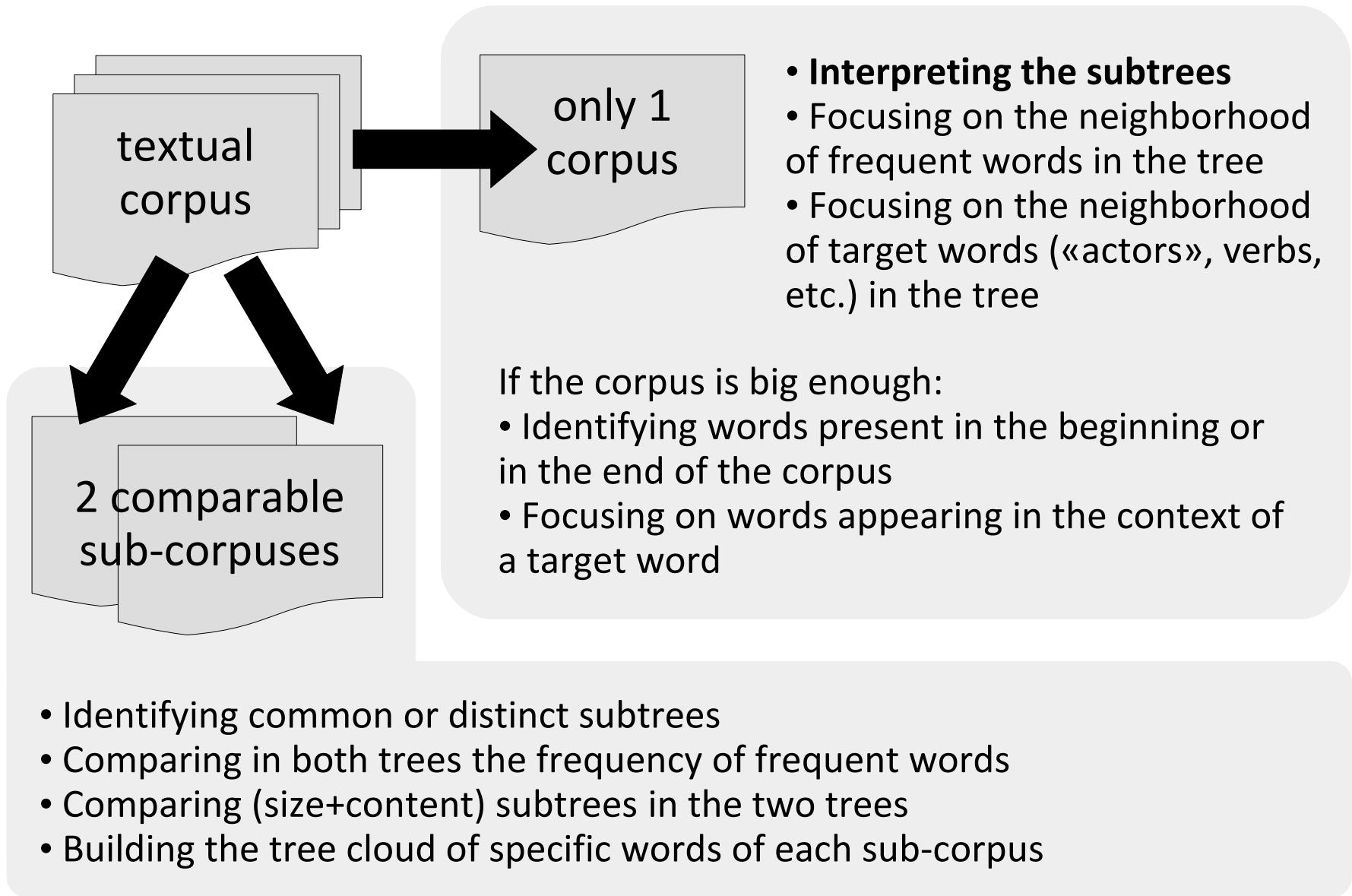


*Textual analysis*

# Corpus exploration with TreeCloud



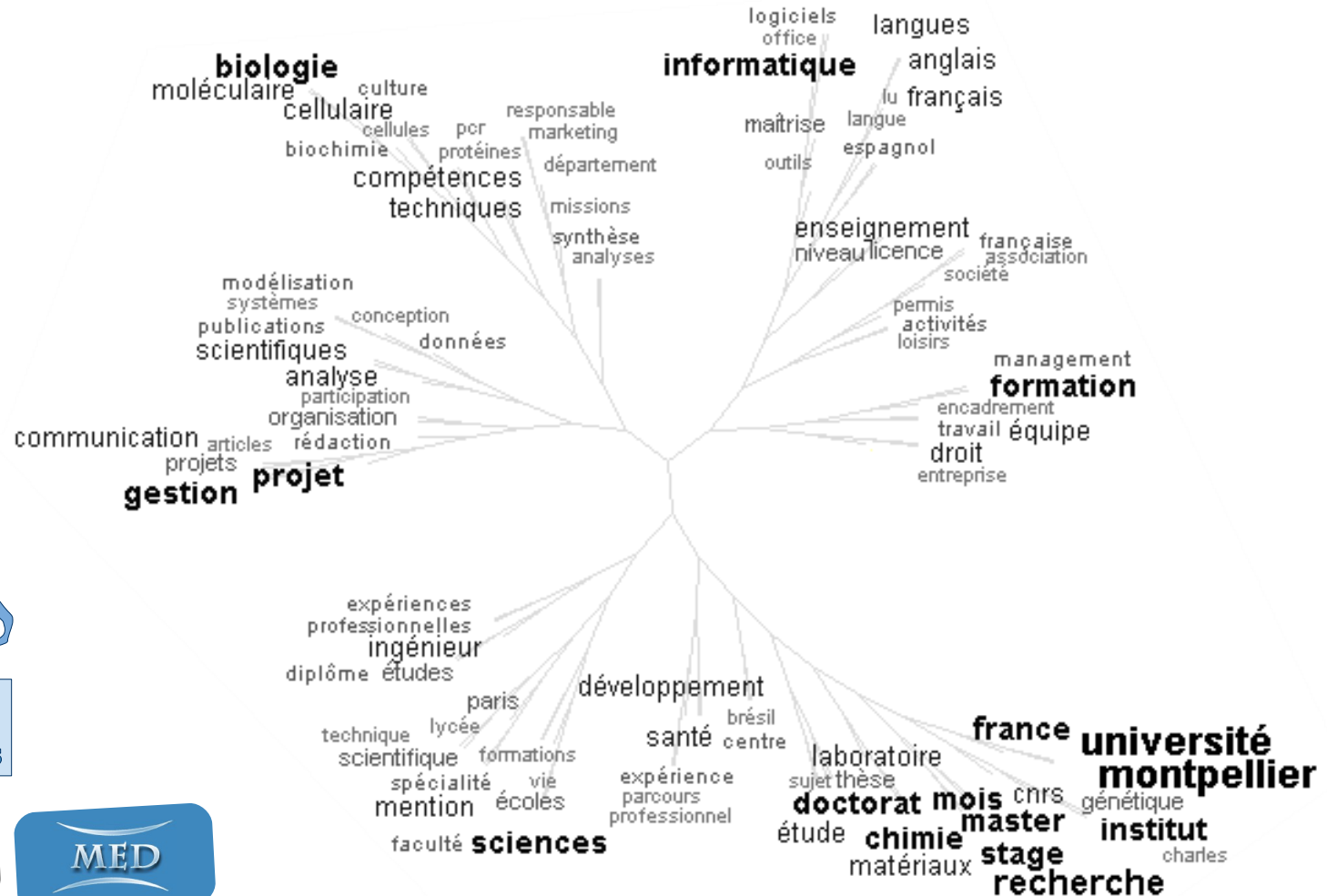
# Corpus exploration with TreeCloud



# Interpreting the subtrees

## Drawing « potatos »

Corpus: 100 CVs of PhD candidates and PhDs attending a meeting with companies



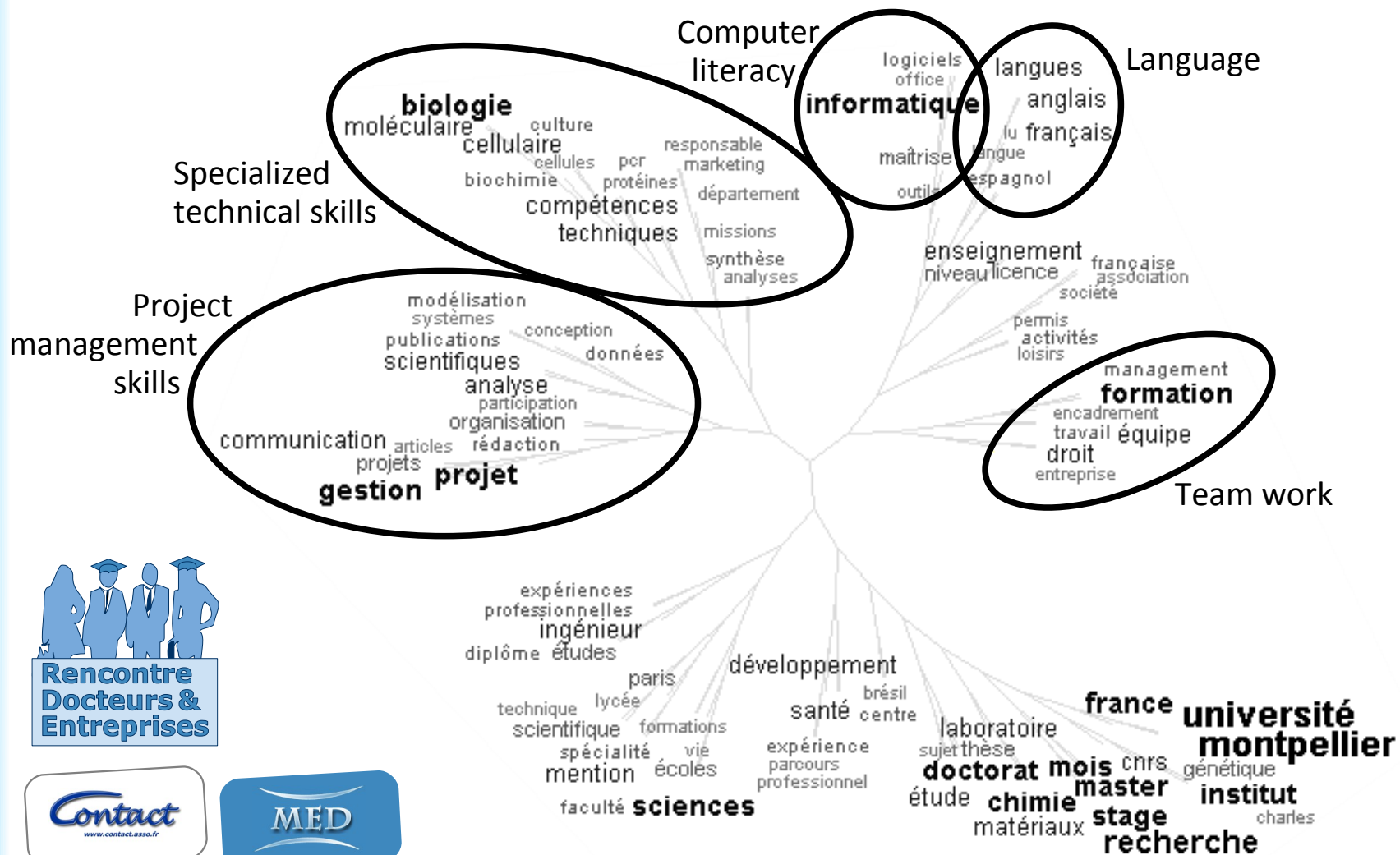
Rencontre  
Docteurs &  
Entreprises



# Interpreting the subtrees

## Drawing « potatoes »

Corpus: 100 CVs of PhD candidates and PhDs attending a meeting with companies



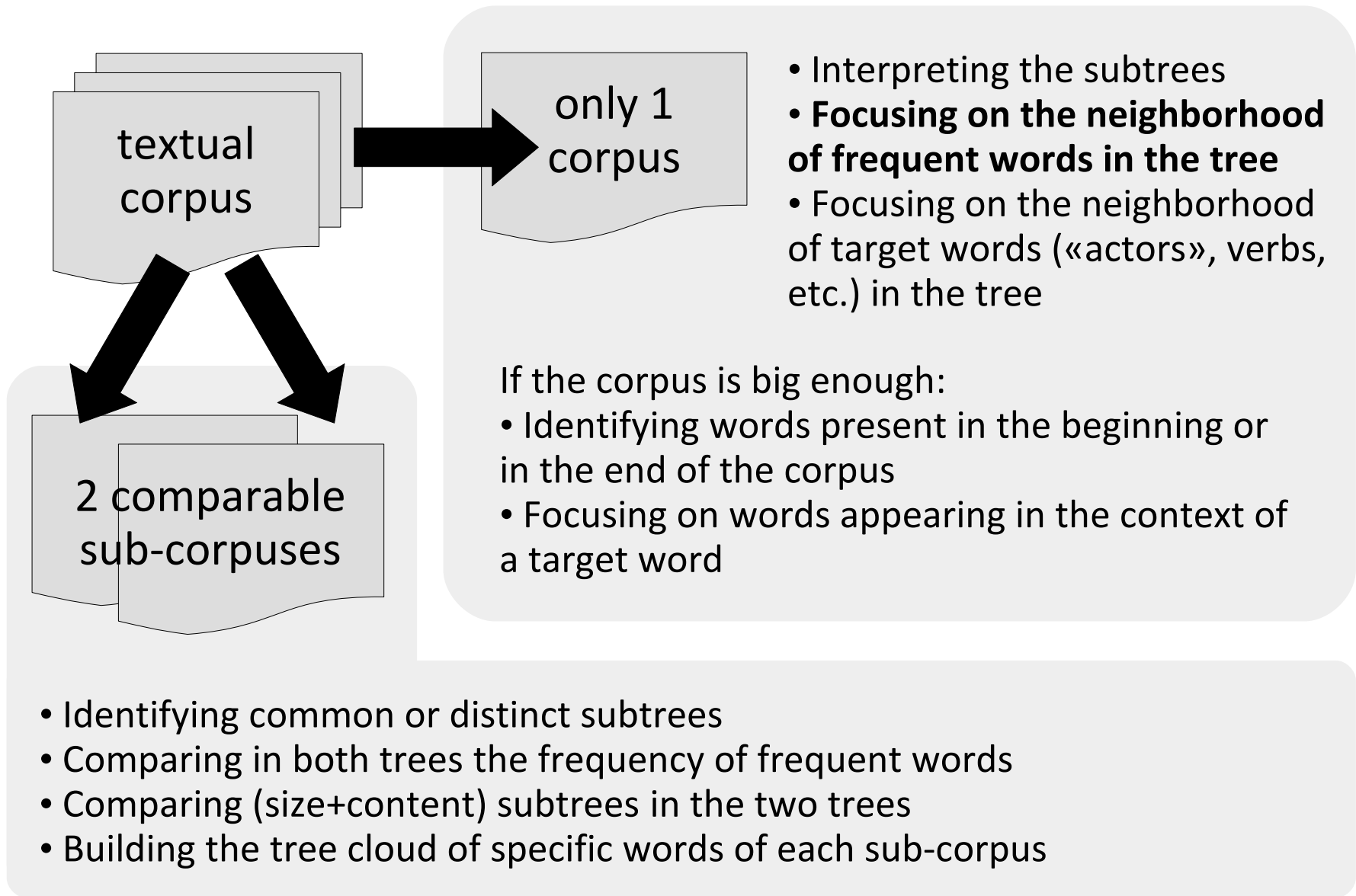
Rencontre  
Docteurs &  
Entreprises

Contact  
[www.contact.asso.fr](http://www.contact.asso.fr)

MED

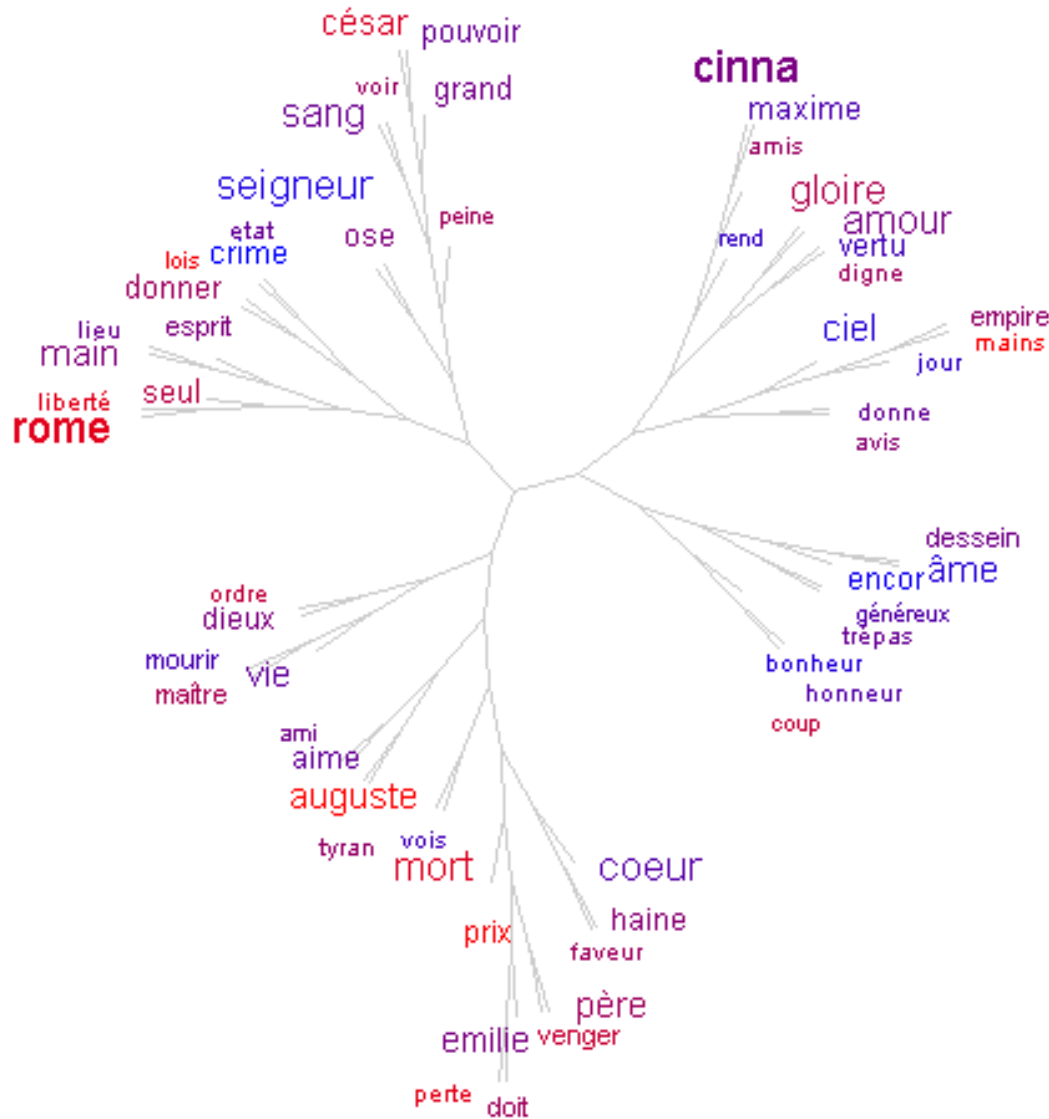


# Corpus exploration with TreeCloud



# Neighborhood of frequent words

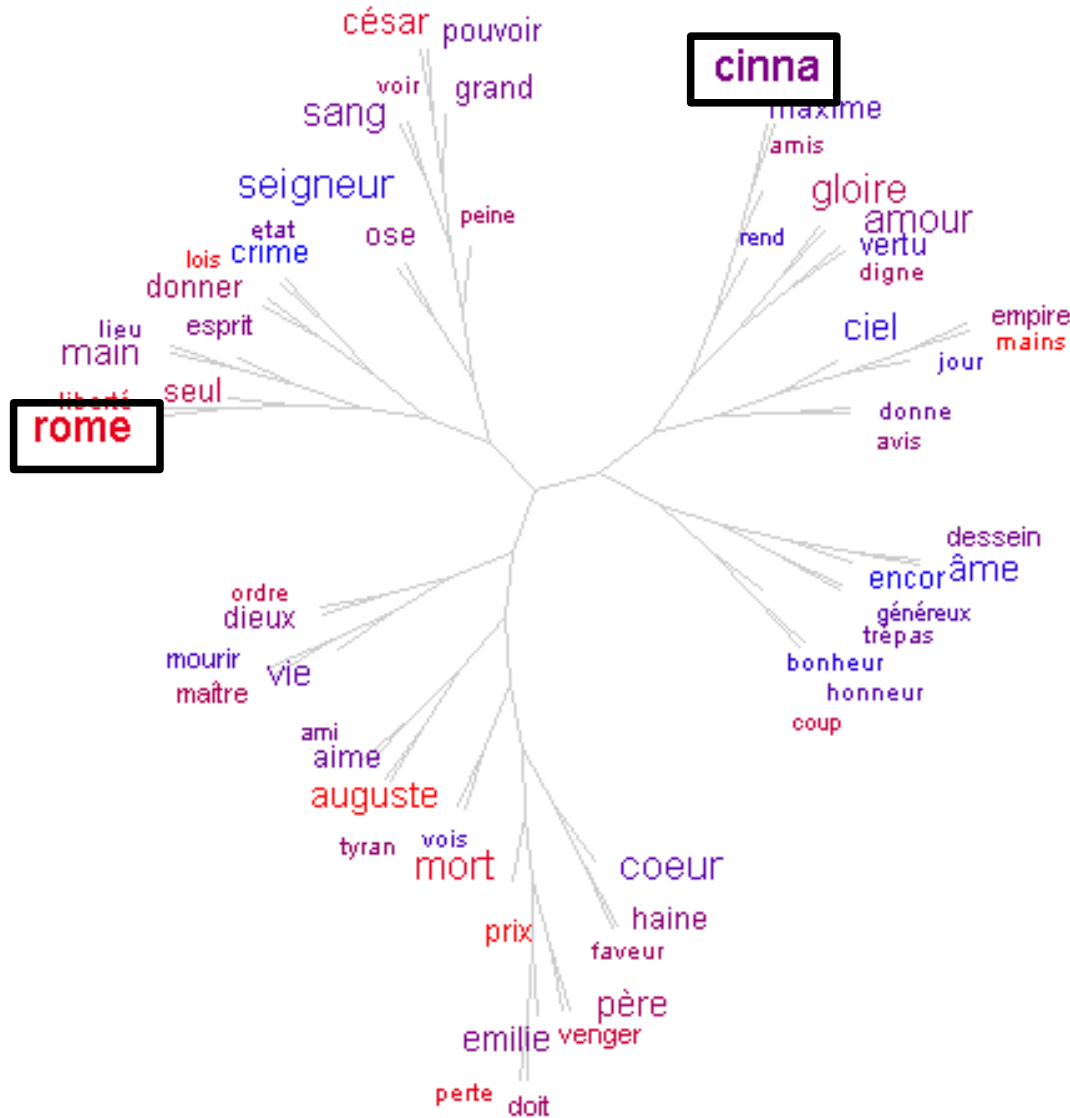
Amstutz & Gambette, JADT 2010



Tree cloud of the 60 most frequent words in *Cinna* by Corneille (Liddell distance, 20 word window), colored chronologically (red in the beginning, blue in the end)

# Neighborhood of frequent words

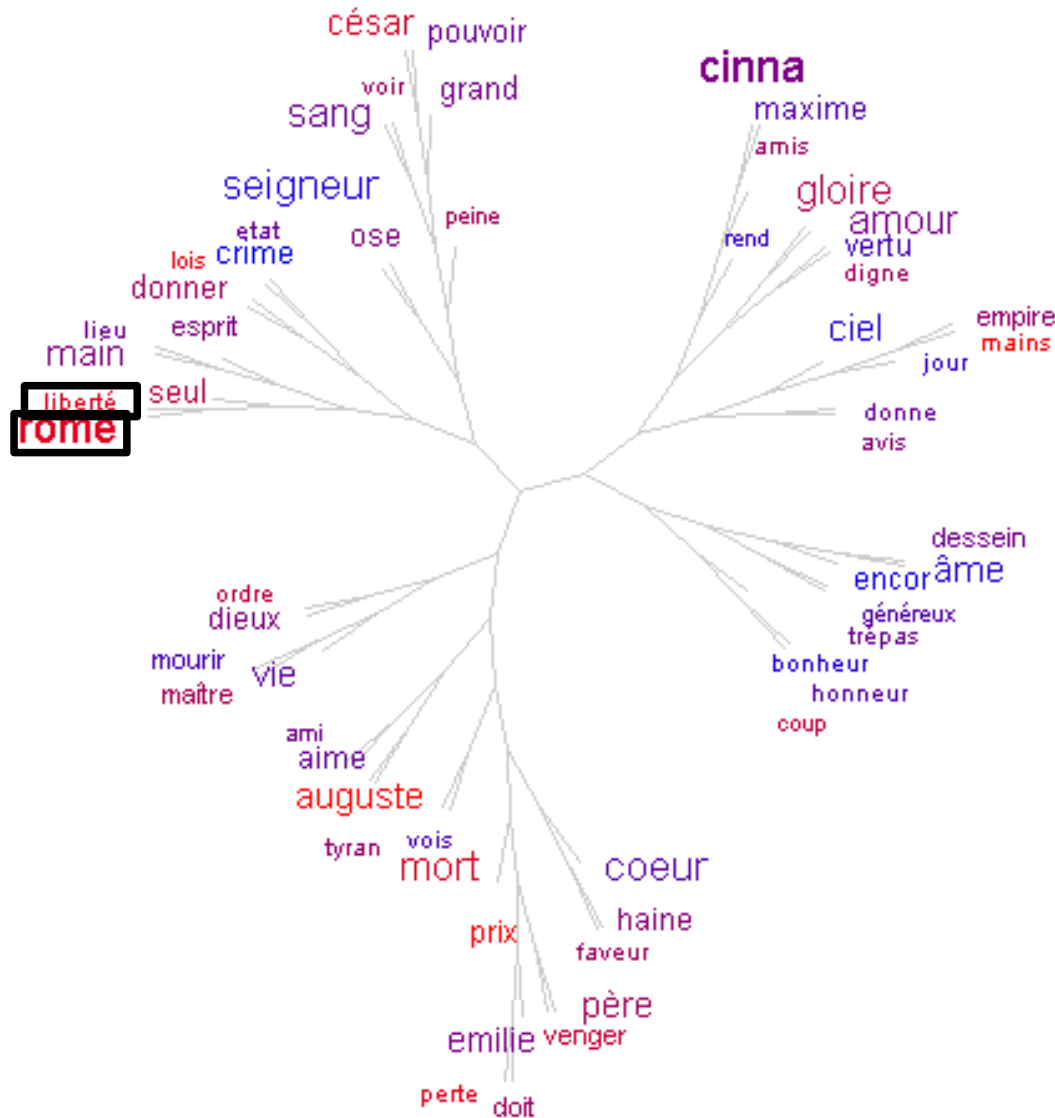
Amstutz & Gambette, JADT 2010



Tree cloud of the 60 most frequent words in *Cinna* by Corneille (Liddell distance, 20 word window), colored chronologically (red in the beginning, blue in the end)

# Neighborhood of frequent words

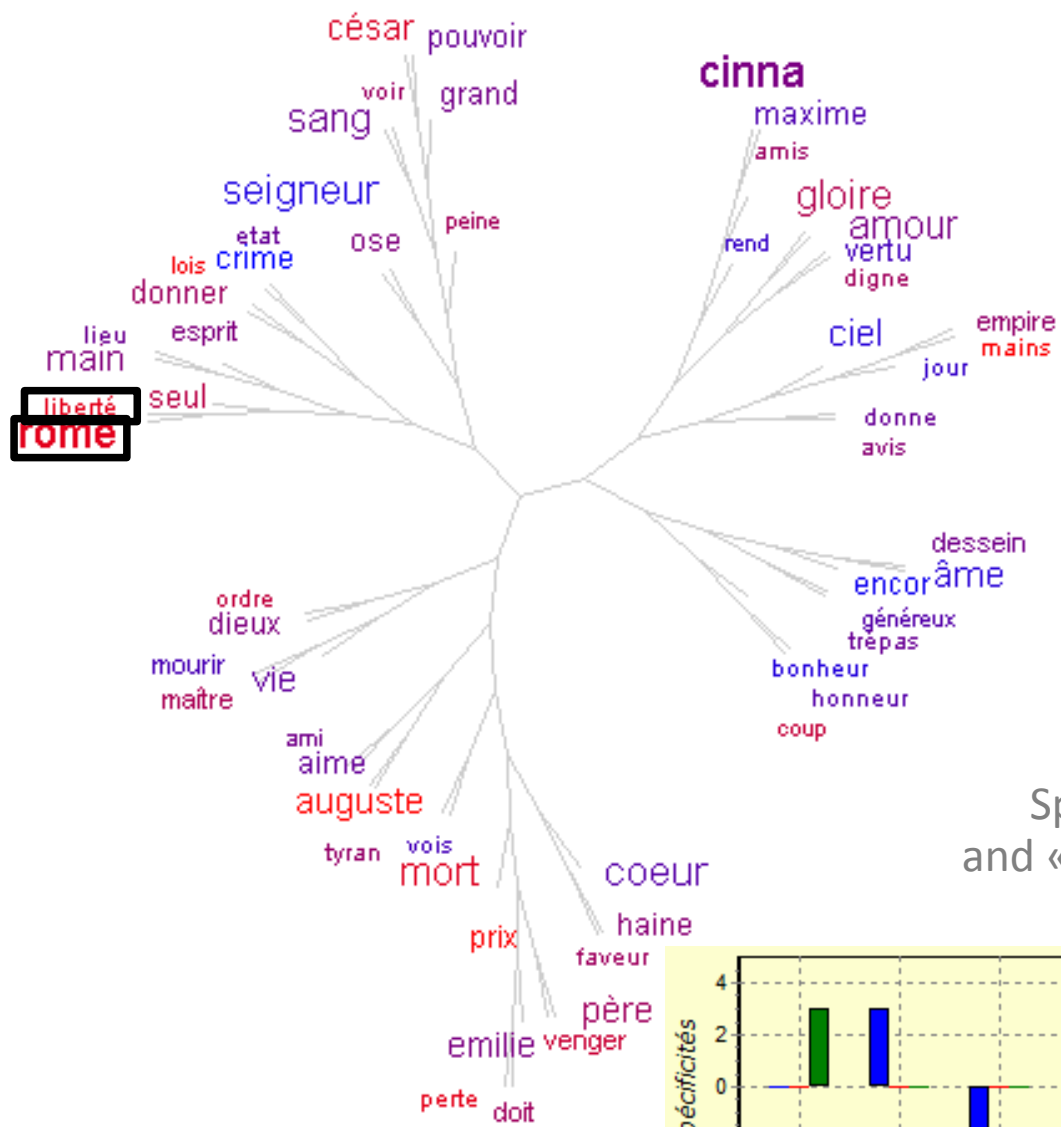
Amstutz & Gambette, JADT 2010



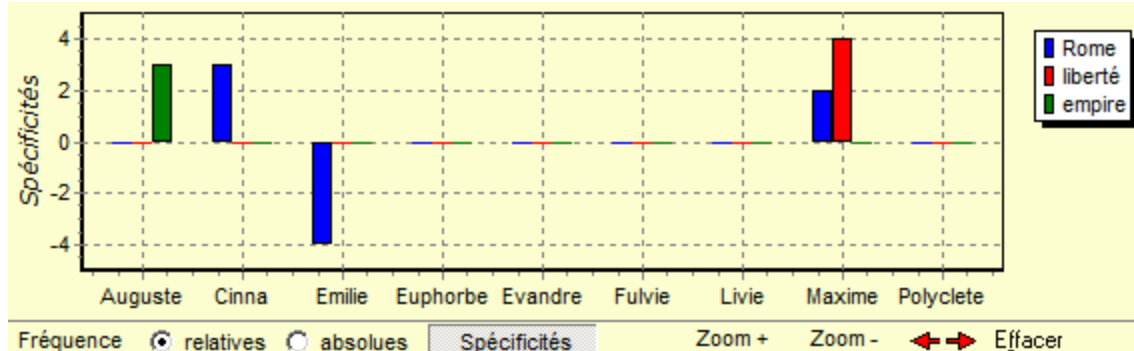
Tree cloud of the 60 most frequent words in *Cinna* by Corneille (Liddell distance, 20 word window), colored chronologically (red in the beginning, blue in the end)

# Neighborhood of frequent words

Amstutz & Gambette, JADT 2010

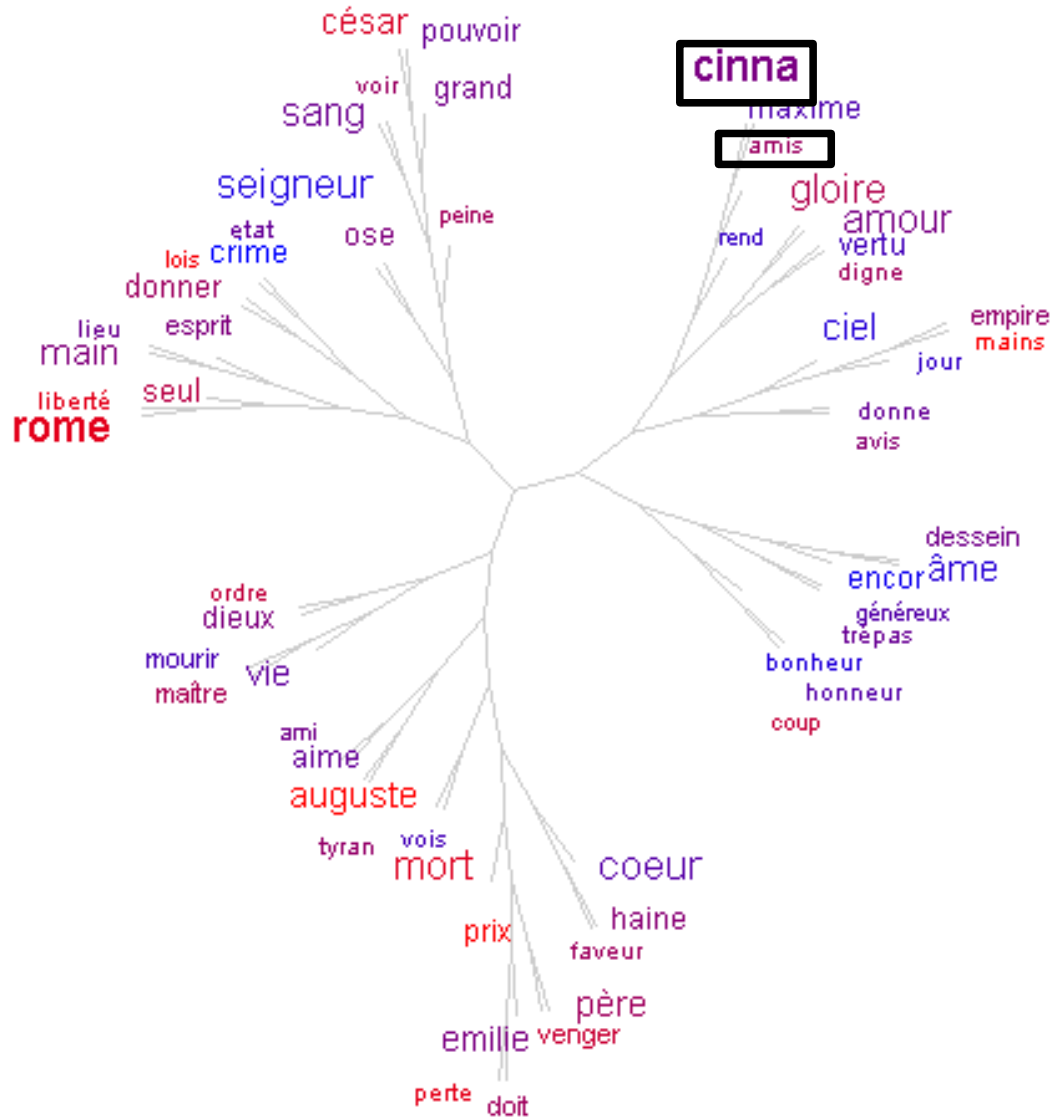


Specificity score of « Rome », « liberté » and « empire », for the characters of *Cinna*, according to Lexico3



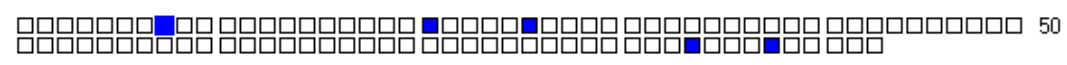
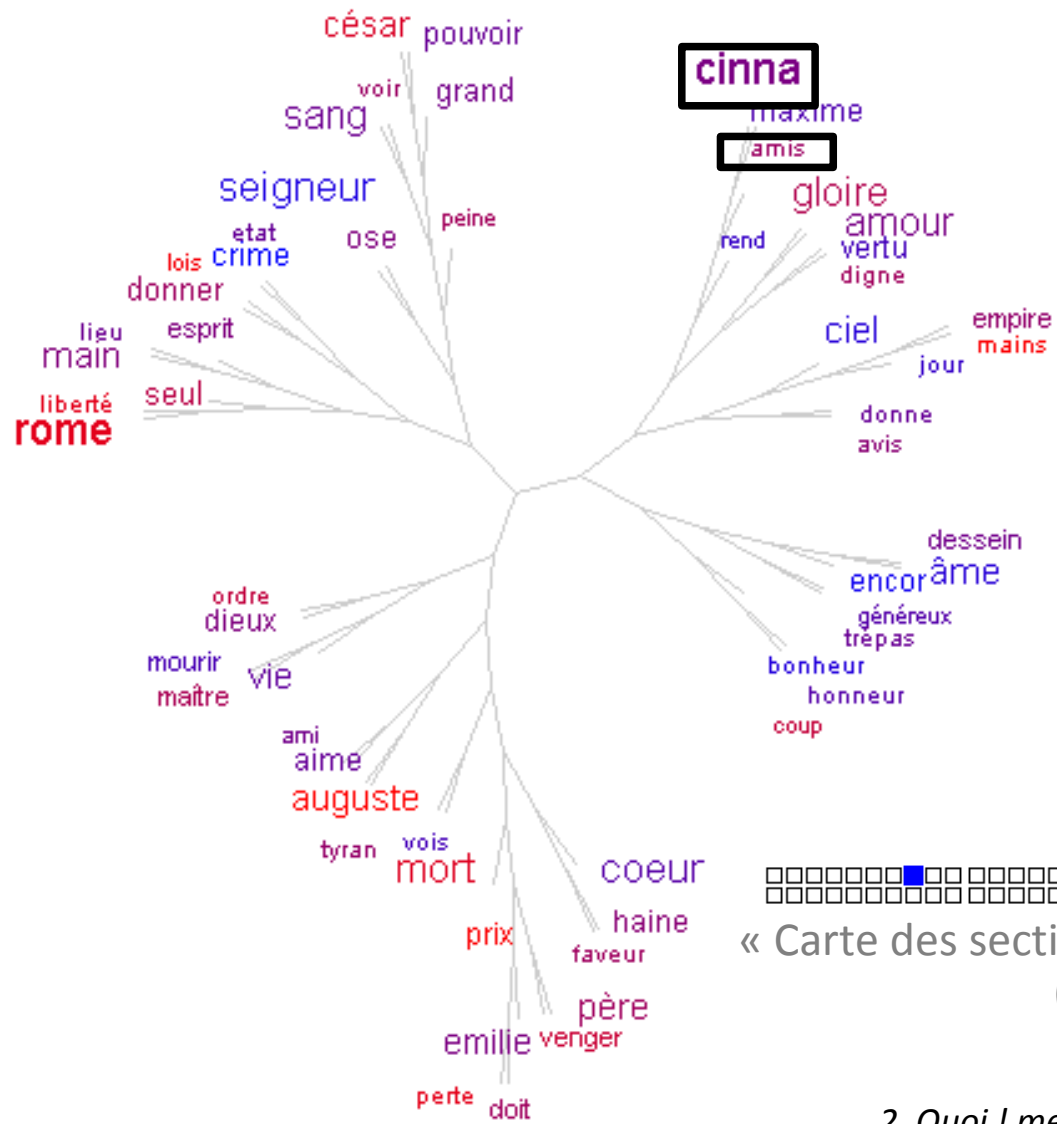
# Neighborhood of frequent words

Amstutz & Gambette, JADT 2010



# Neighborhood of frequent words

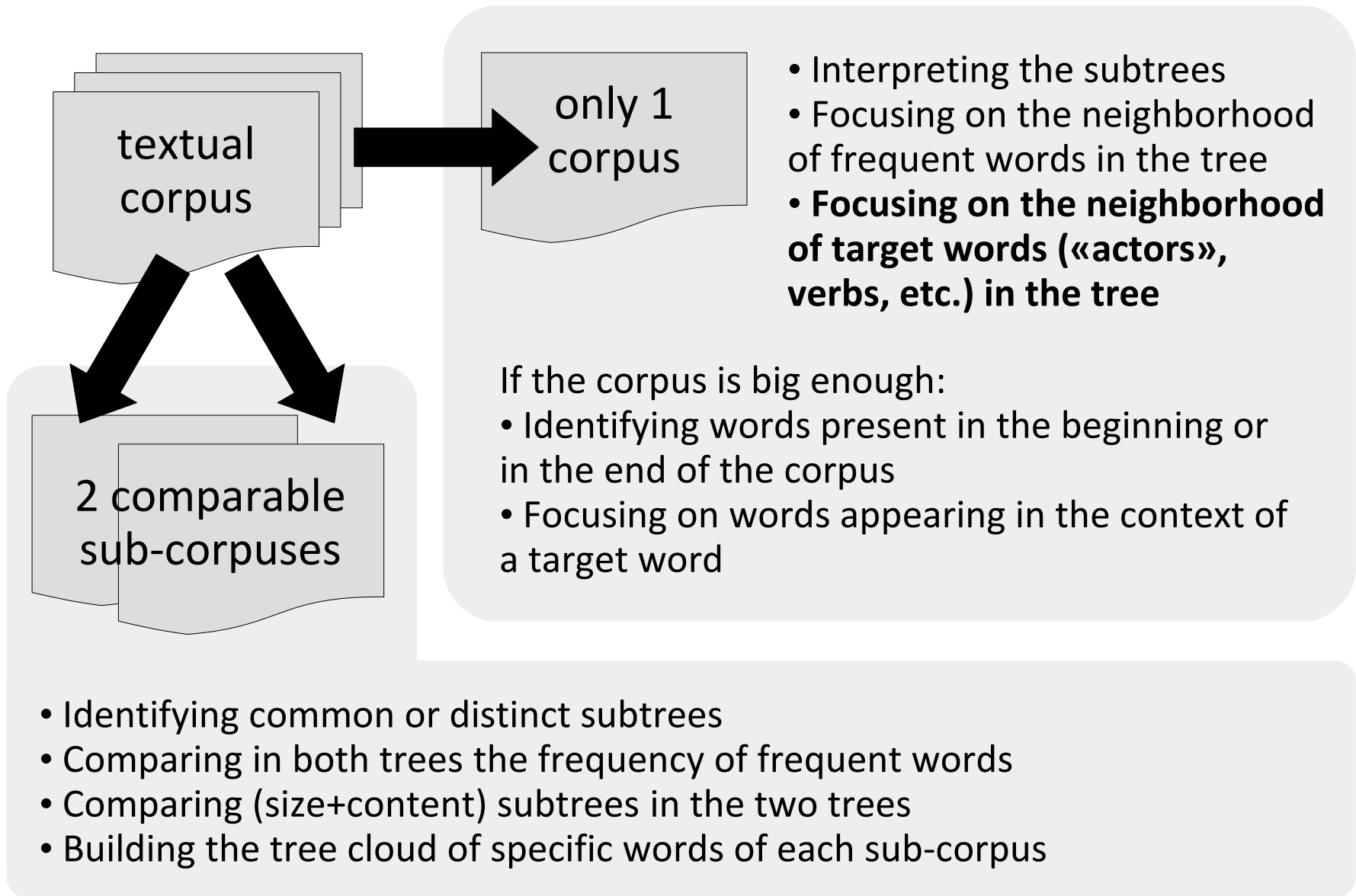
Amstutz & Gambette, JADT 2010



« Carte des sections » Lexico3 and contexts of « amis » (« friends ») in Auguste's lines in *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**.

# Corpus exploration with TreeCloud

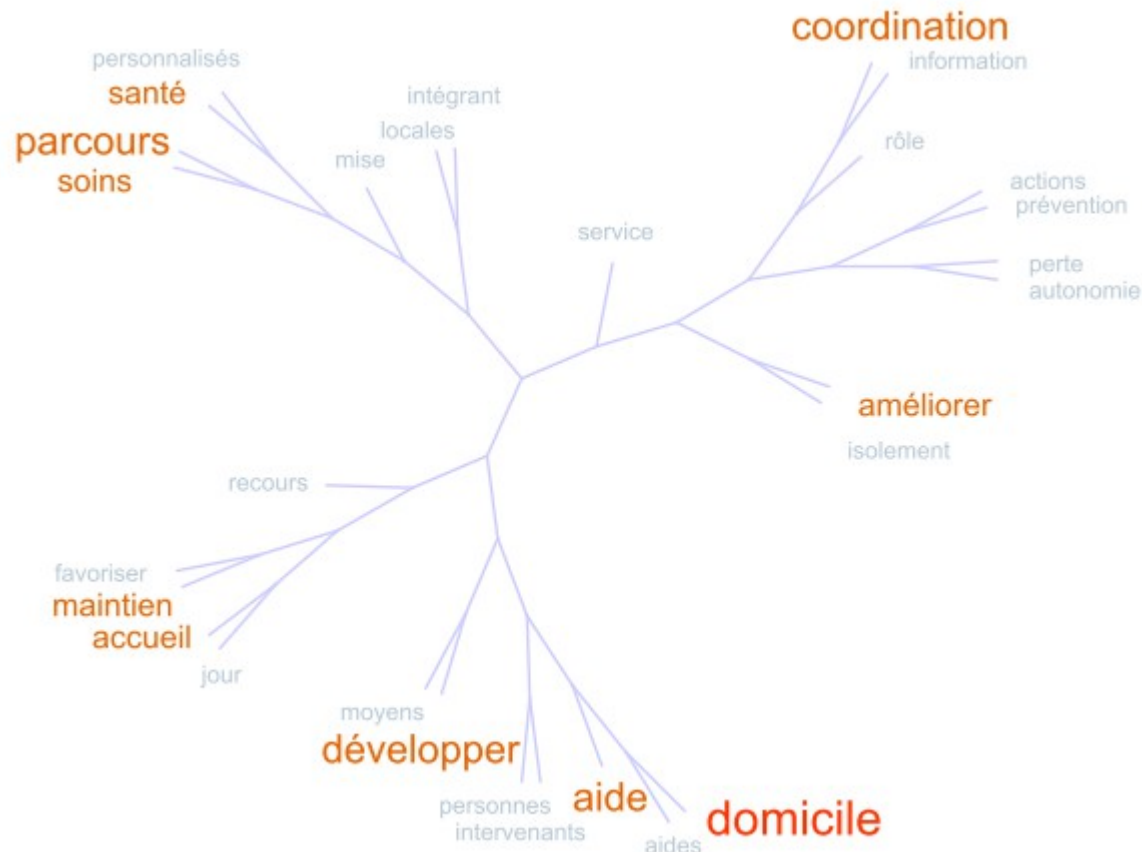




# Neighborhoods of words

Corpus: answers to open questions to health professionals, about the health path of old people in the south of France (Alpes de Haute-Provence)

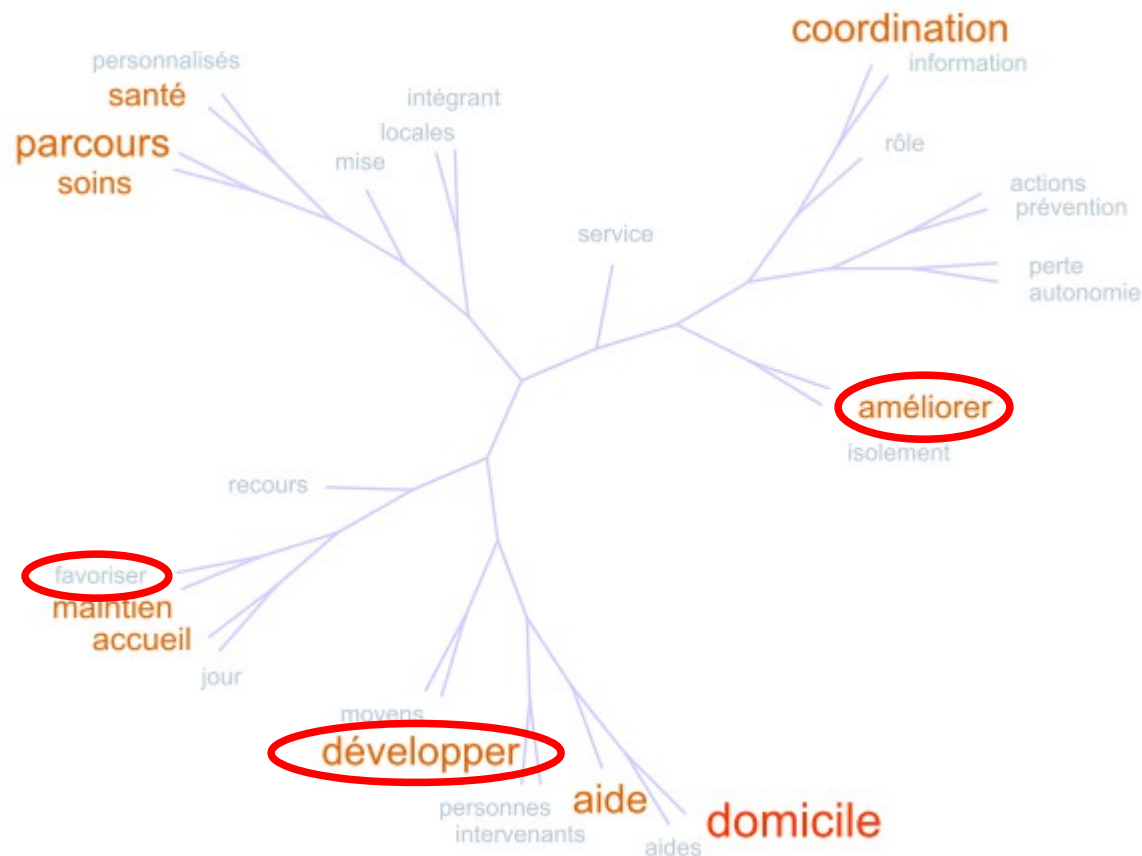
Suggestions for improvements:



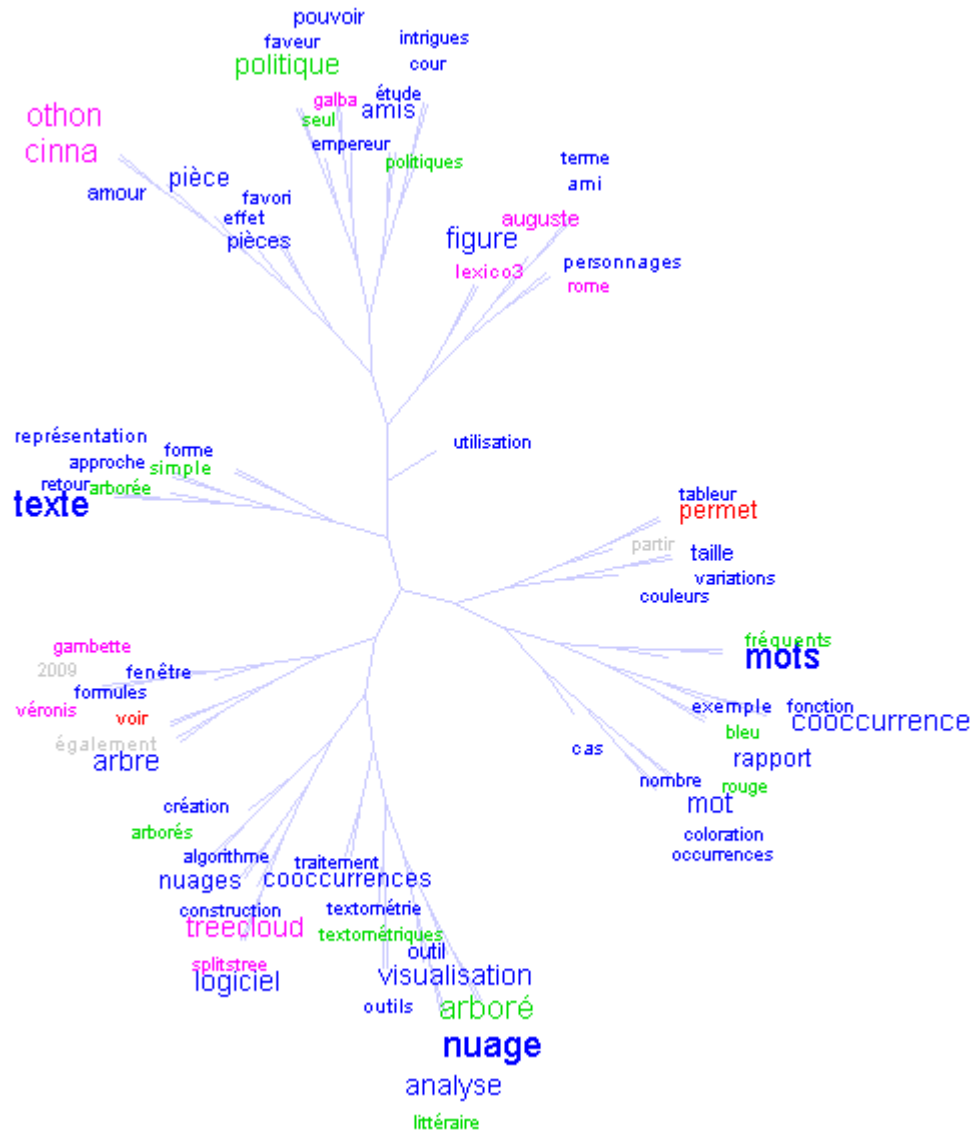
# Neighborhoods of words

Corpus: answers to open questions to health professionals, about the health path of old people in the south of France (Alpes de Haute-Provence)

Suggestions for improvements:



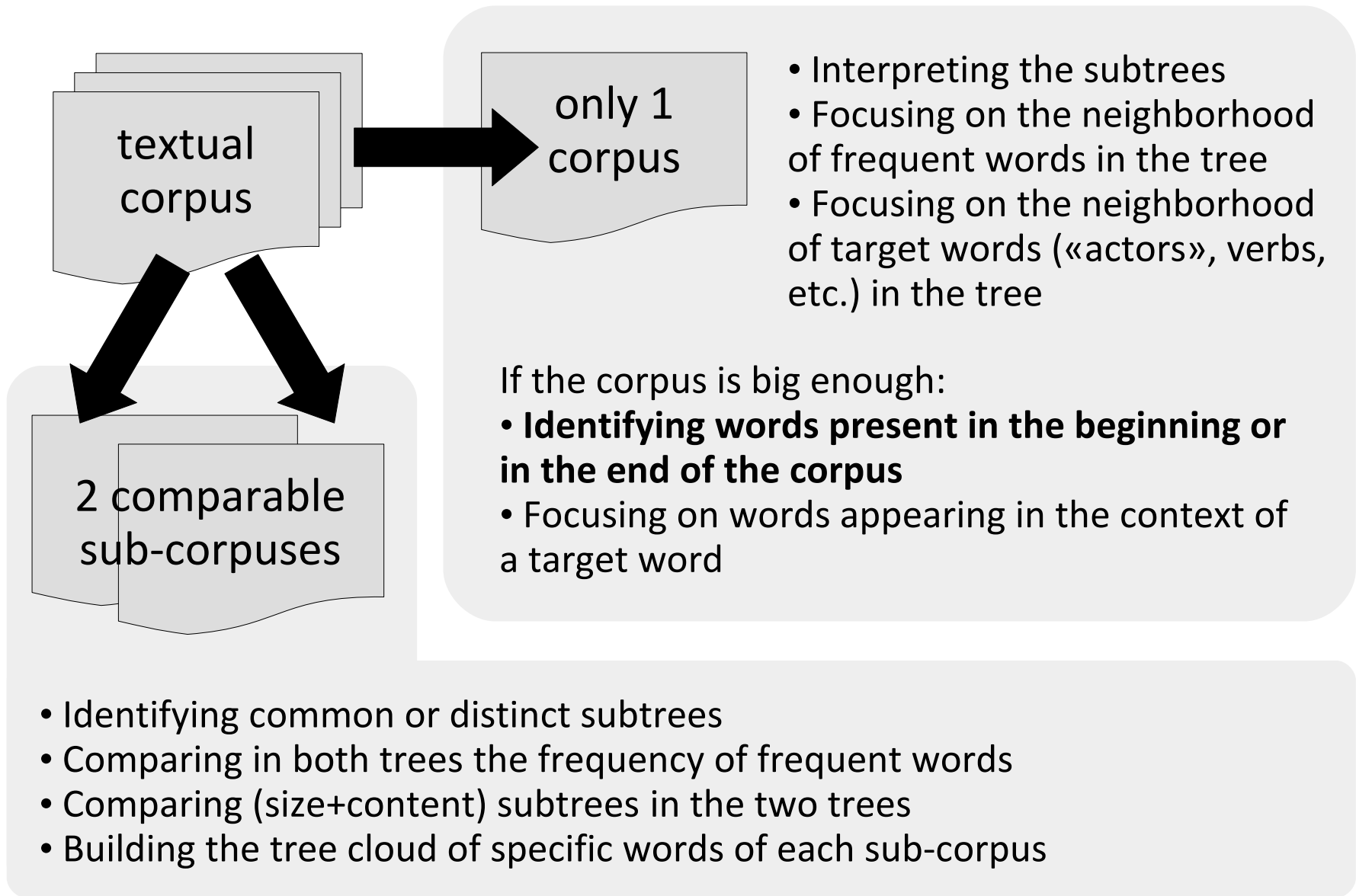
# Grammatical colouring



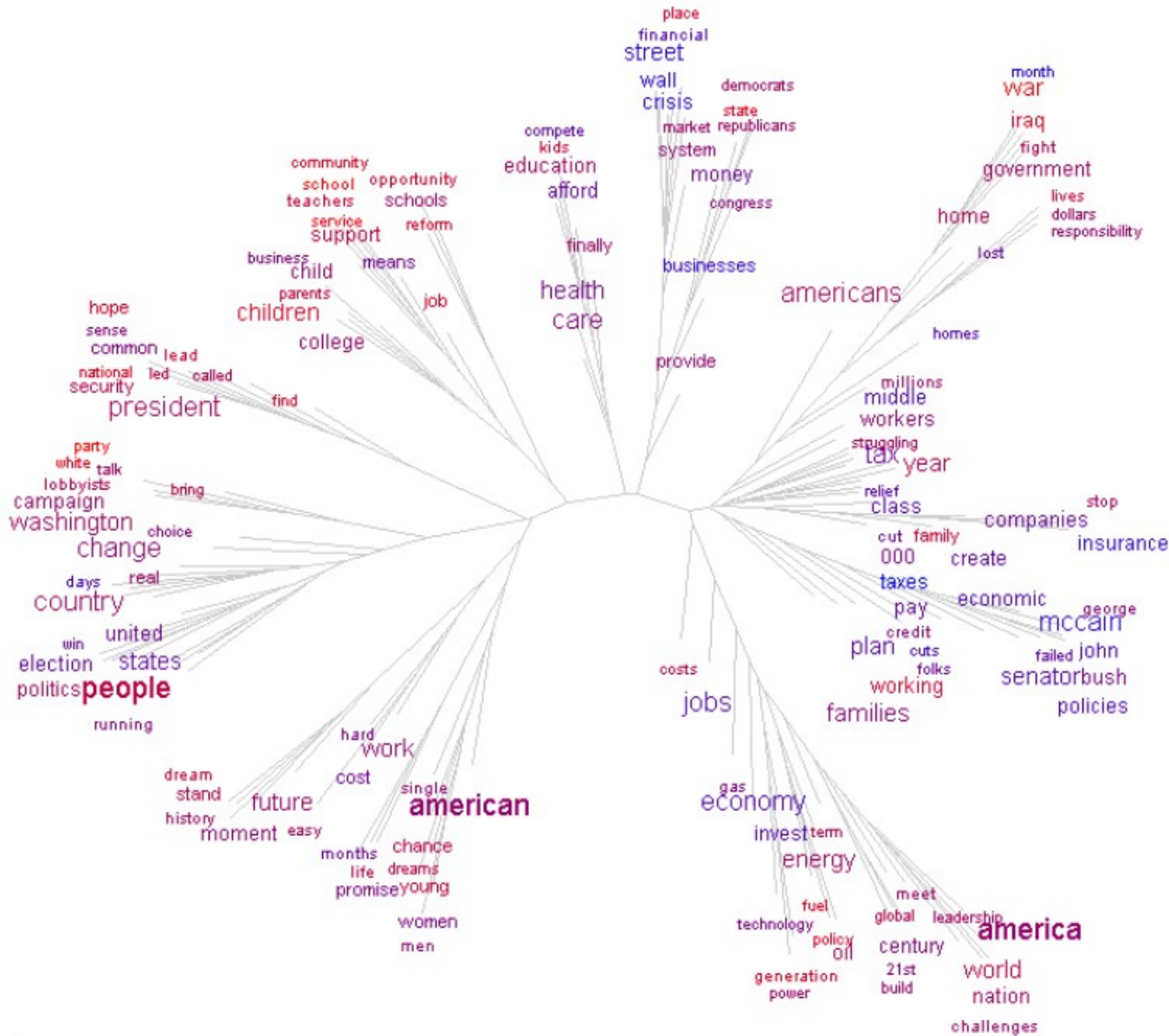
nouns  
adjectives  
verbs  
proper noun

Tree cloud of words present 5 times or more in the article of Amstutz & Gambette at JADT 2010, Liddell distance, 20 word window, imposed colouring deduced from a TreeTagger analysis of the text

# Corpus exploration with TreeCloud



# Words in the beginning or in the end



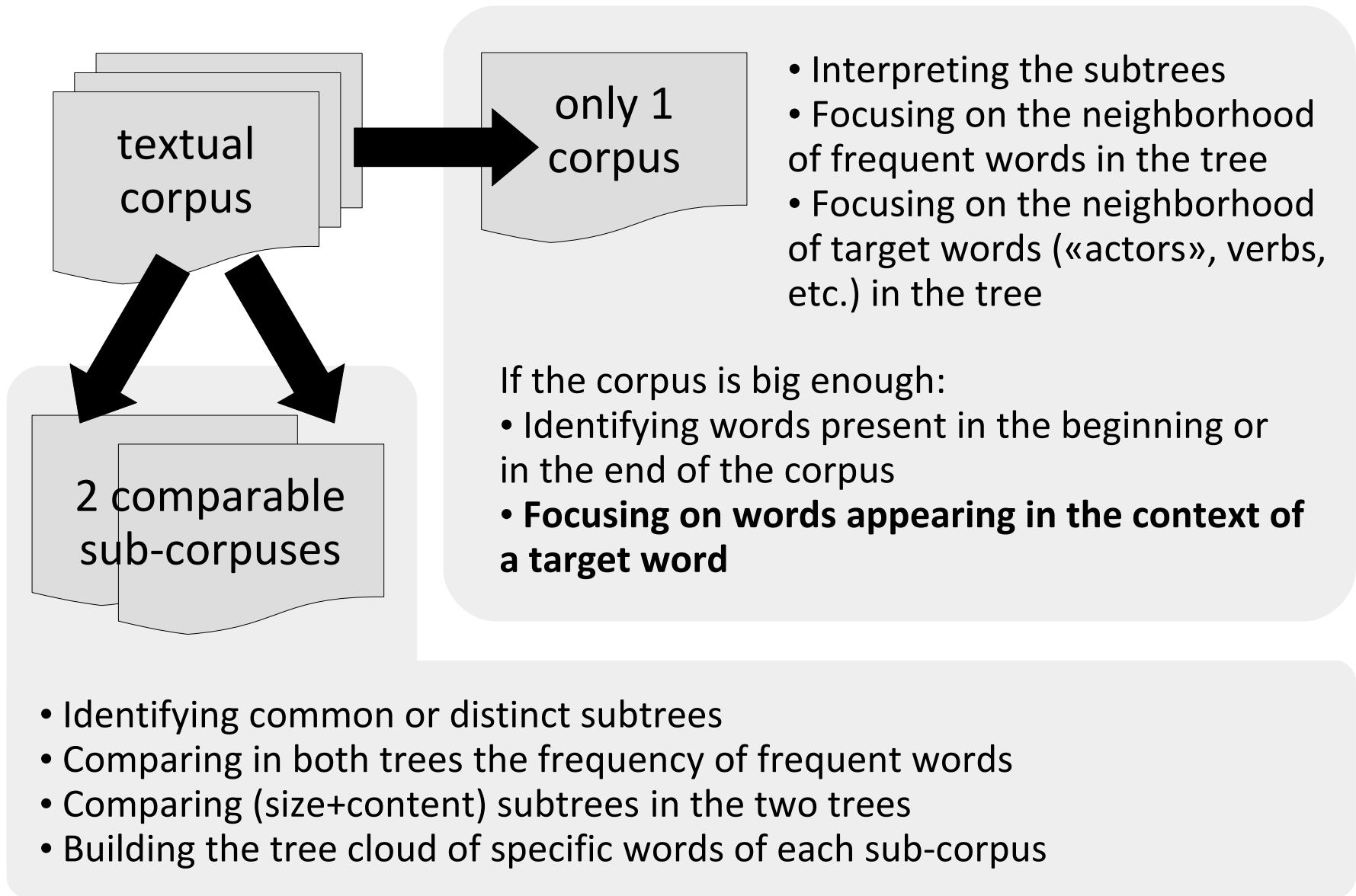
Tree cloud of all campaigning speeches of Barack Obama in 2008, chronological coloring

Beginning of the campaign

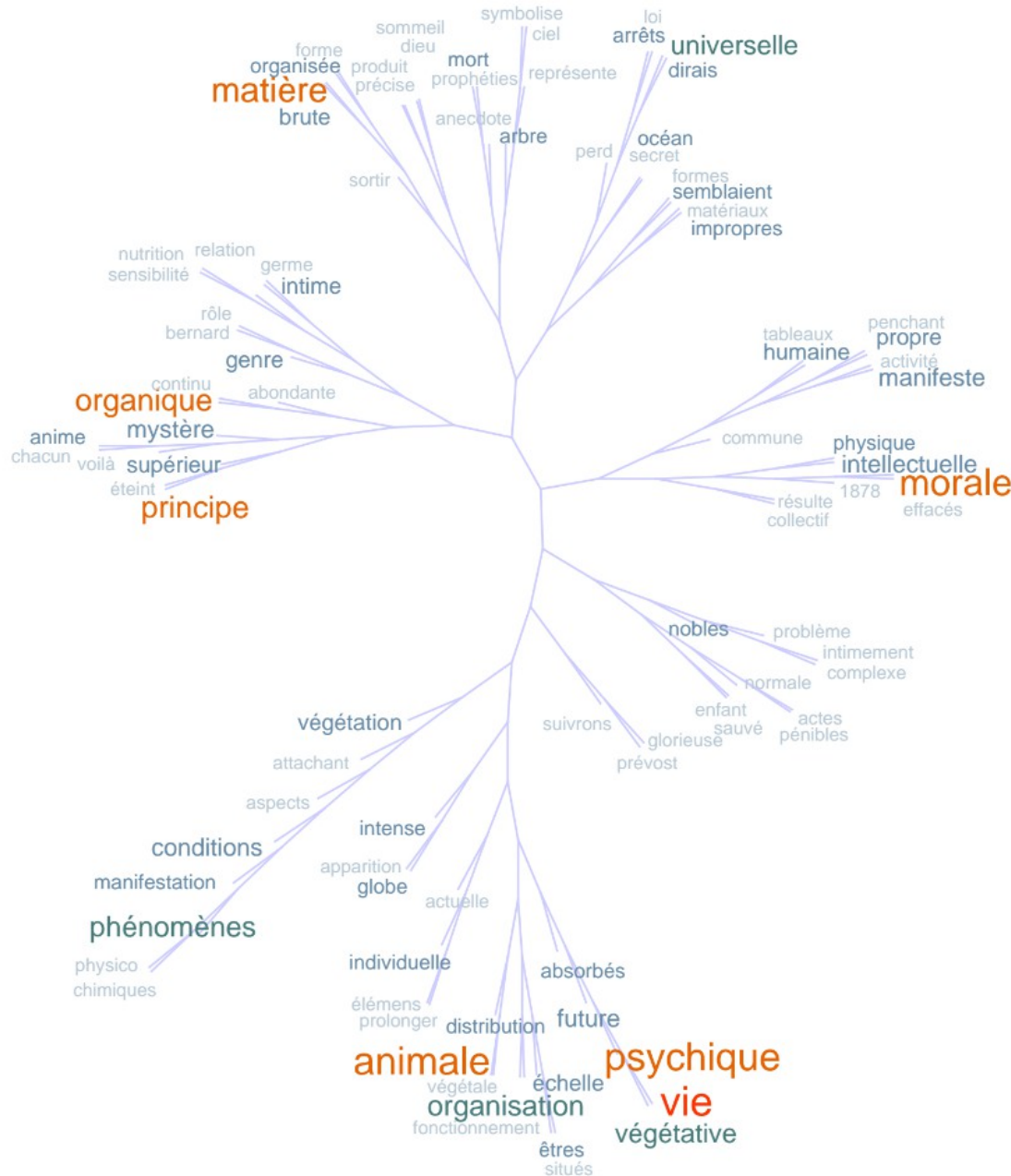
End of the campaign

Gambette & Véronis, IFCS 2009

# Corpus exploration with TreeCloud

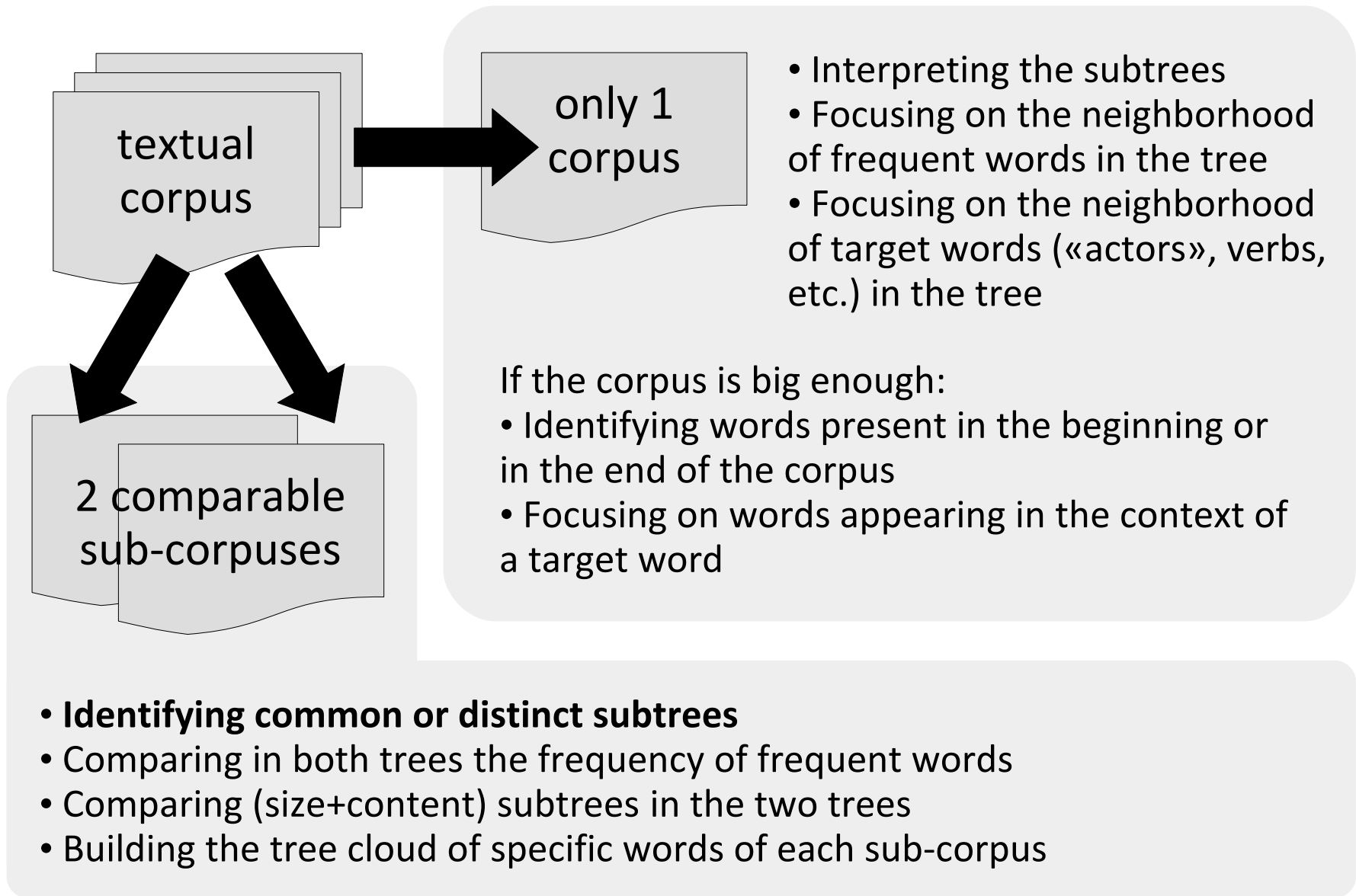


# Focusing on a target word



Tree cloud of the words cooccurring with «vie» («life»), in a corpus of articles by writers and philosophers in *La Revue des deux mondes* (19<sup>th</sup> century), colored according to cooccurrence score

# Corpus exploration with TreeCloud





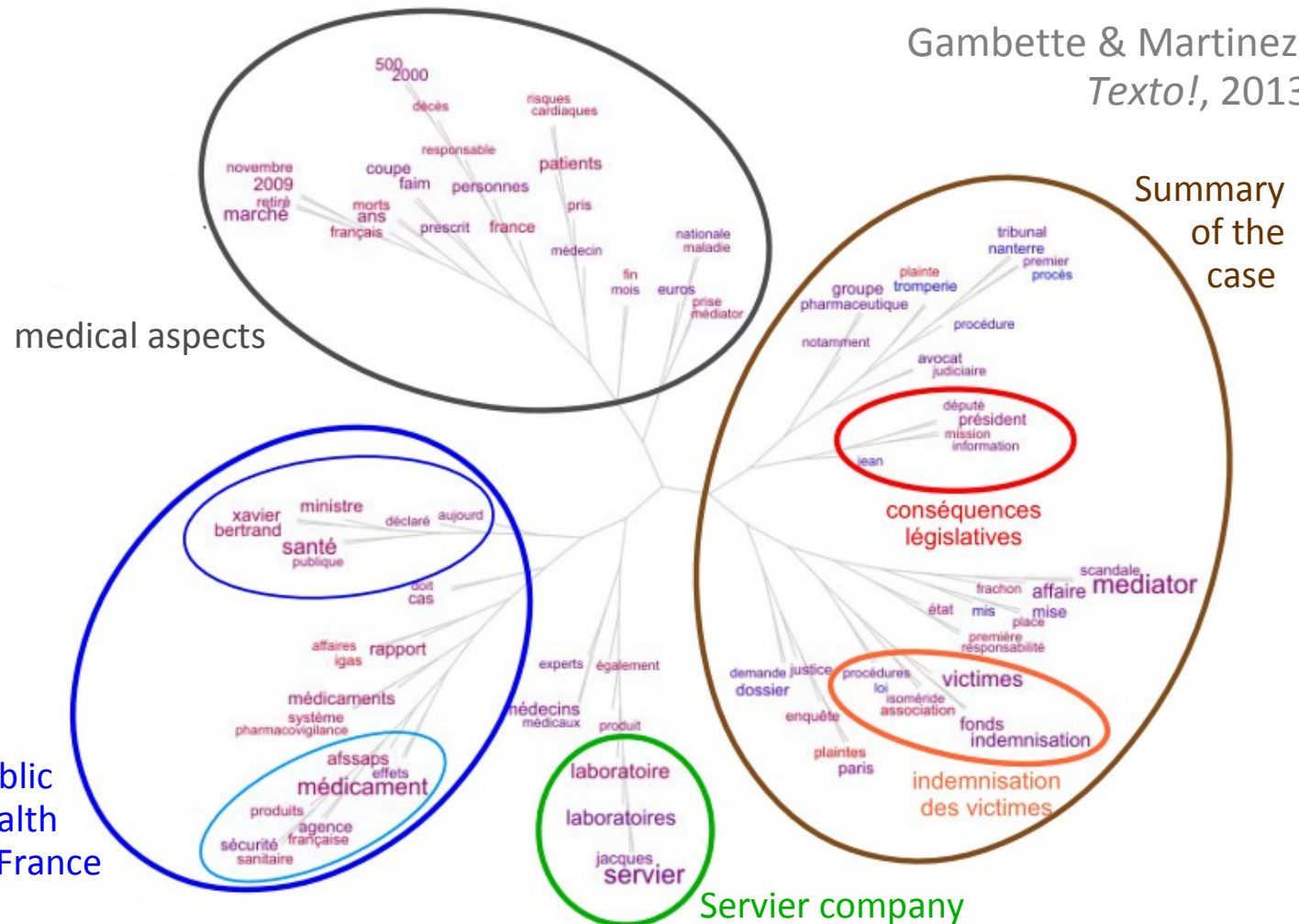
# Identifying distinct or common subtrees

## Comparing press agency and other journalist articles

Corpus: 595 press agency articles vs 1496 other journalist articles in 2011 about the Mediator case in the French press.

All articles

Gambette & Martinez,  
*Texto!*, 2013

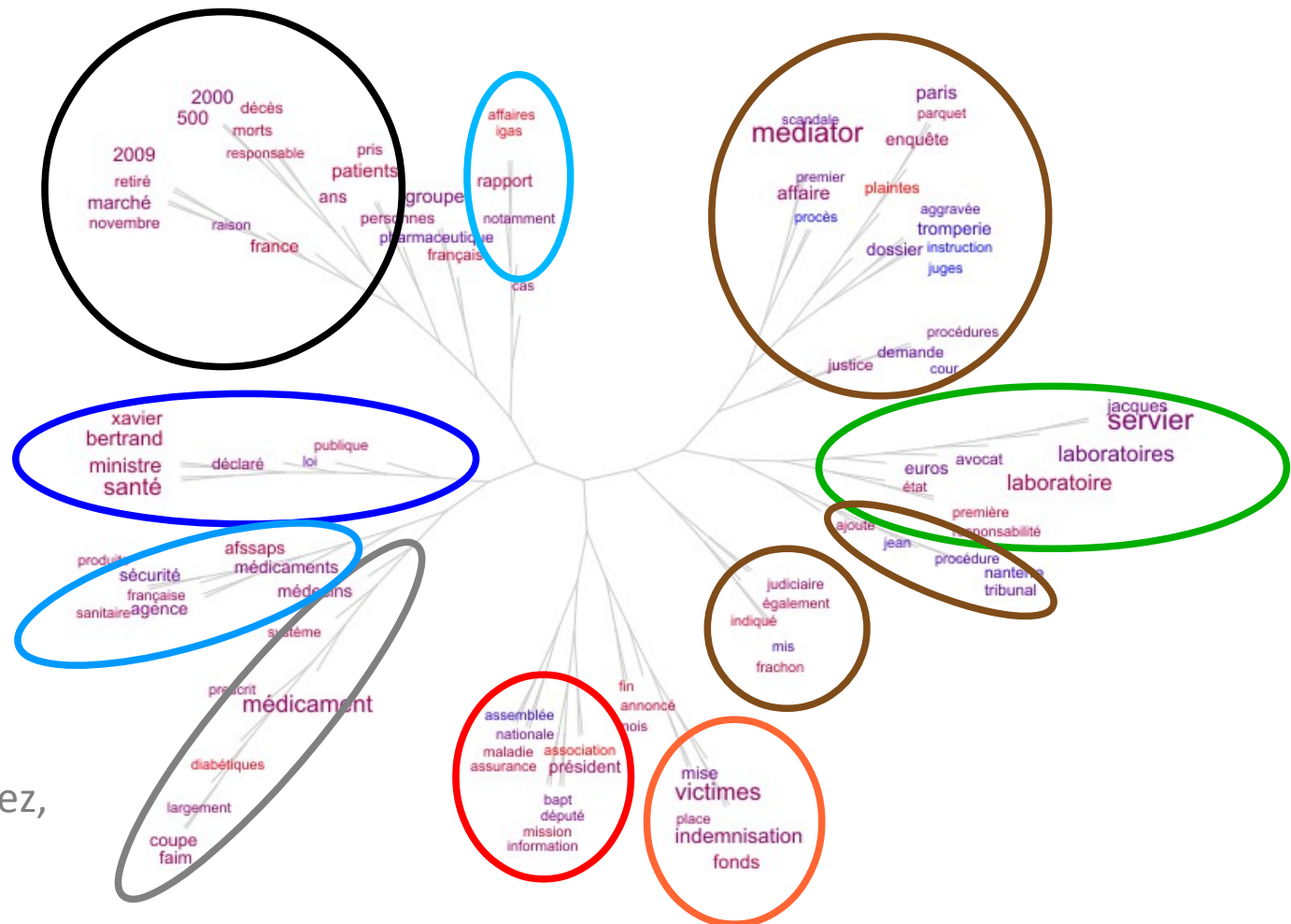


# Identifying distinct or common subtrees

## Comparing press agency and other journalist articles

Corpus: 595 press agency articles vs 1496 other journalist articles in 2011 about the Mediator case in the French press.

Press agency  
articles



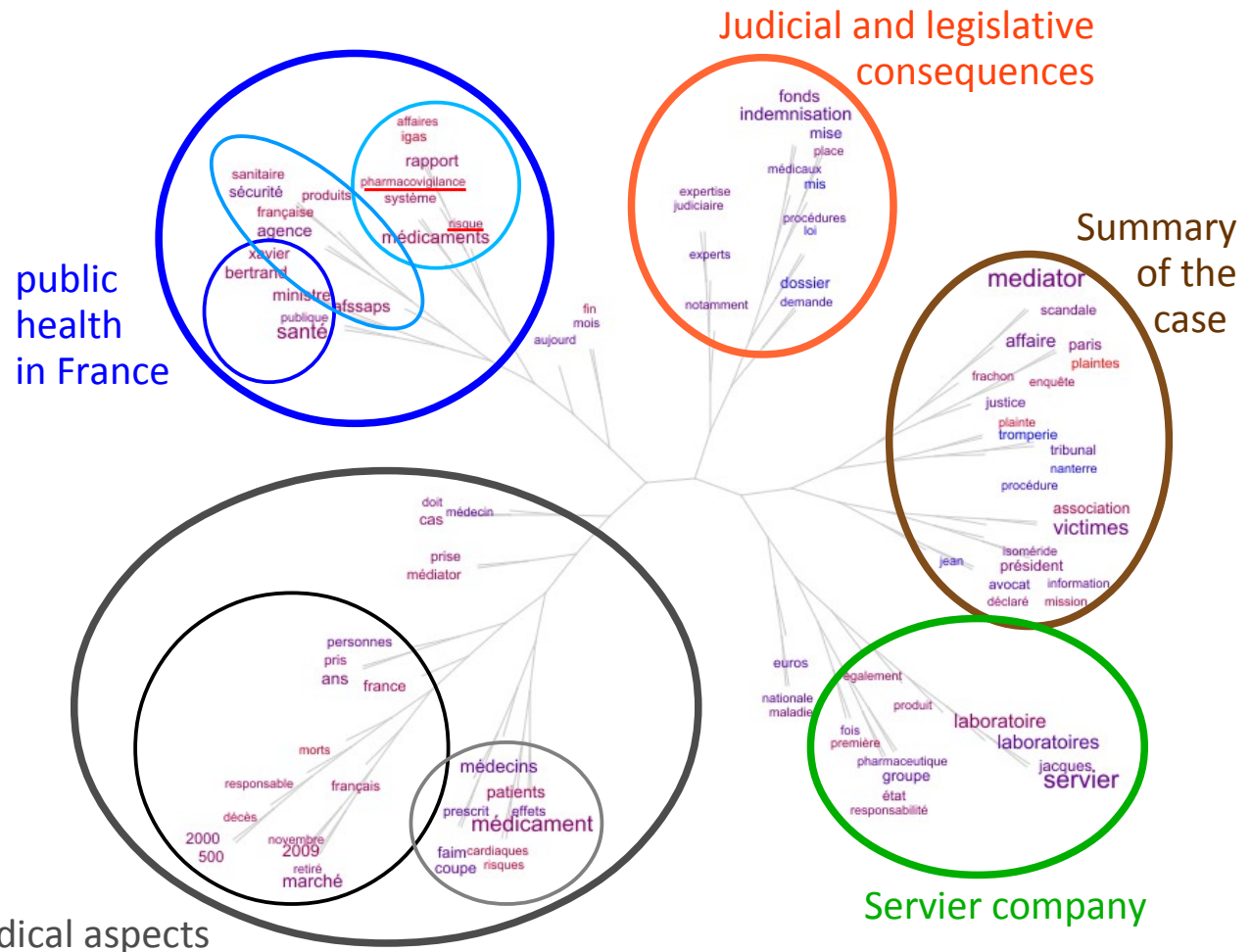
Gambette & Martinez,  
*Texto!*, 2013

# Identifying distinct or common subtrees

## Comparing press agency and other journalist articles

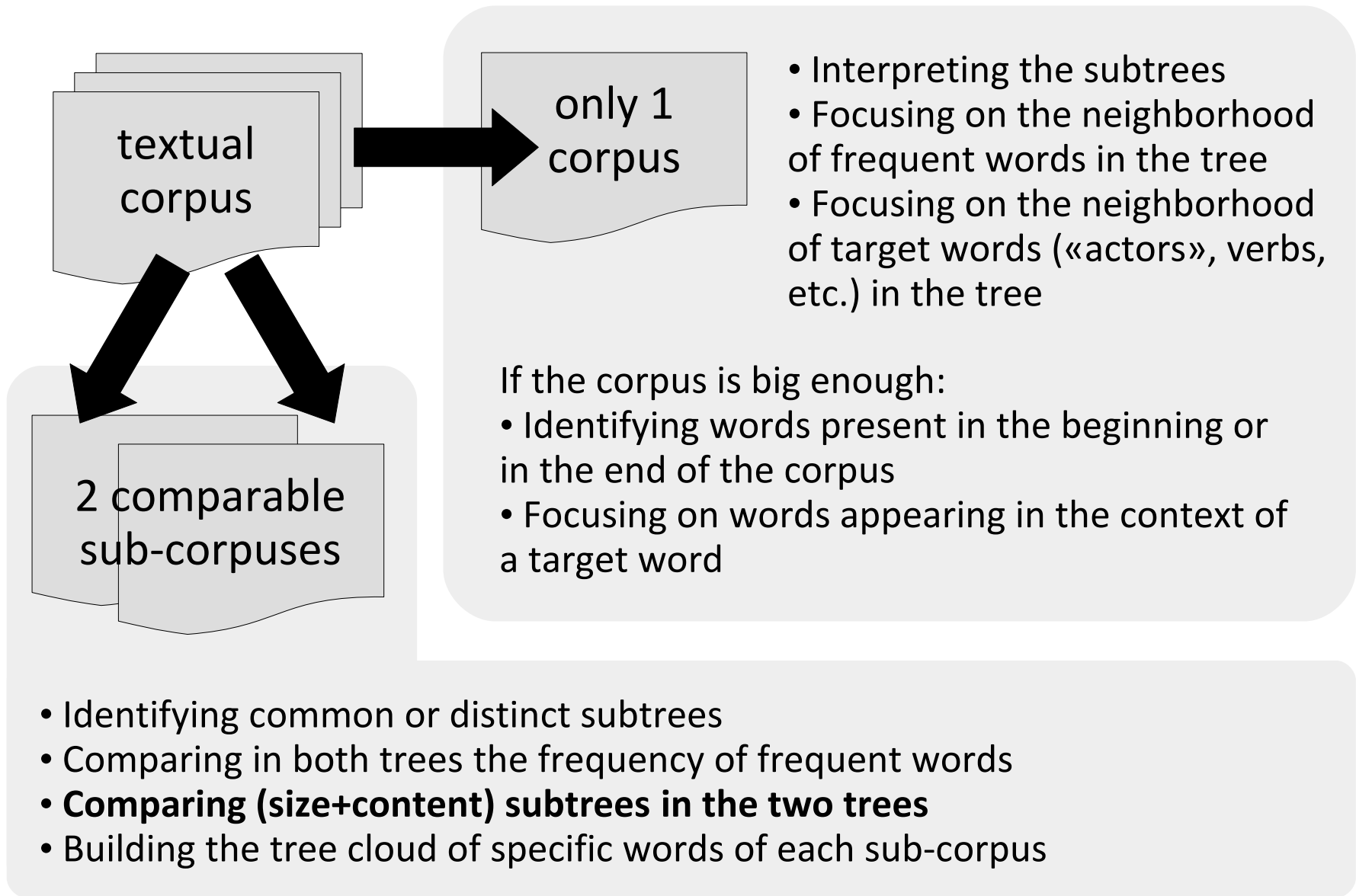
Corpus: 595 press agency articles vs 1496 other journalist articles in 2011 about the Mediator case in the French press.

### Journalist articles

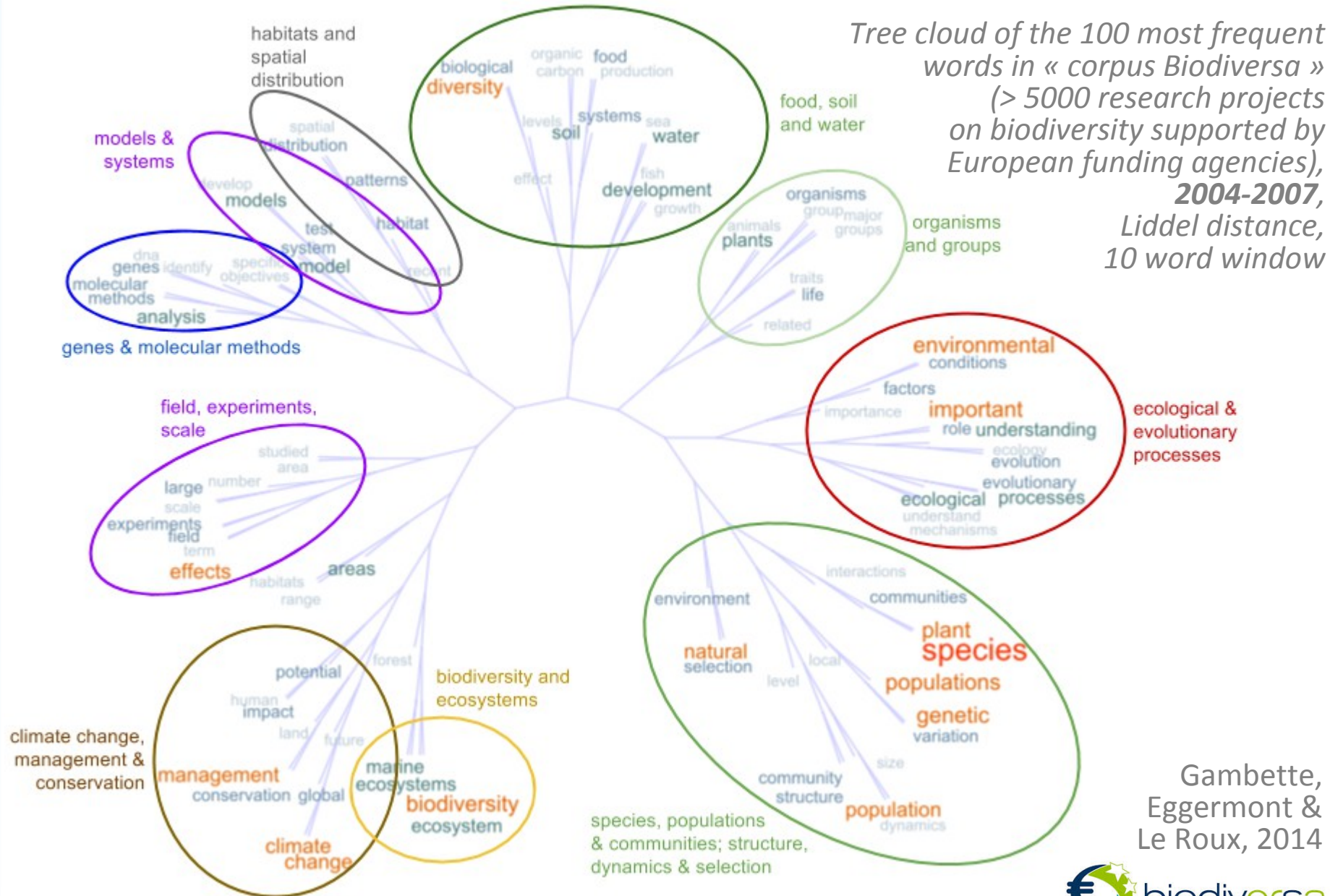


Gambette & Martinez,  
*Texto!*, 2013

# Corpus exploration with TreeCloud

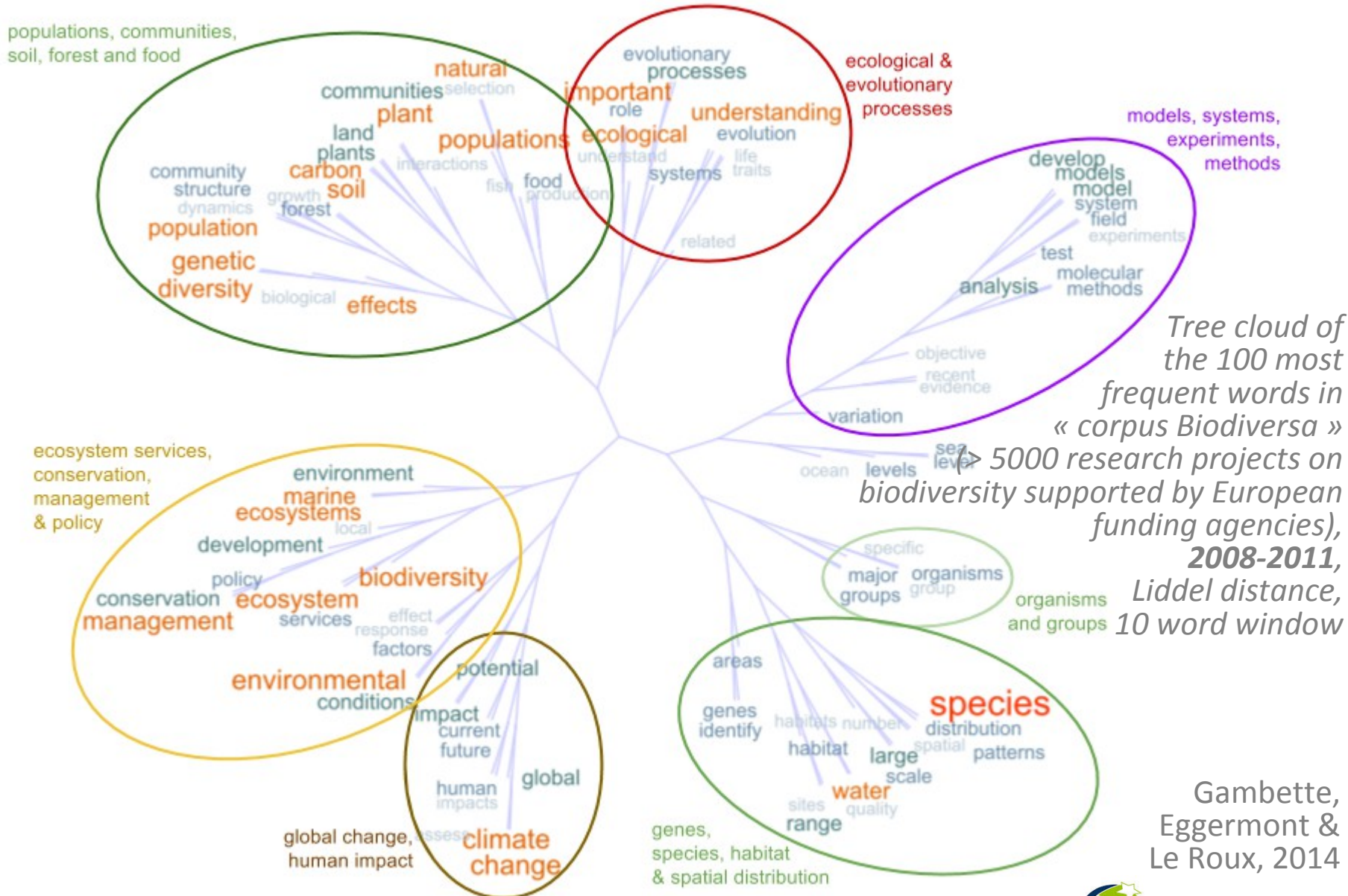


# Comparing subtrees

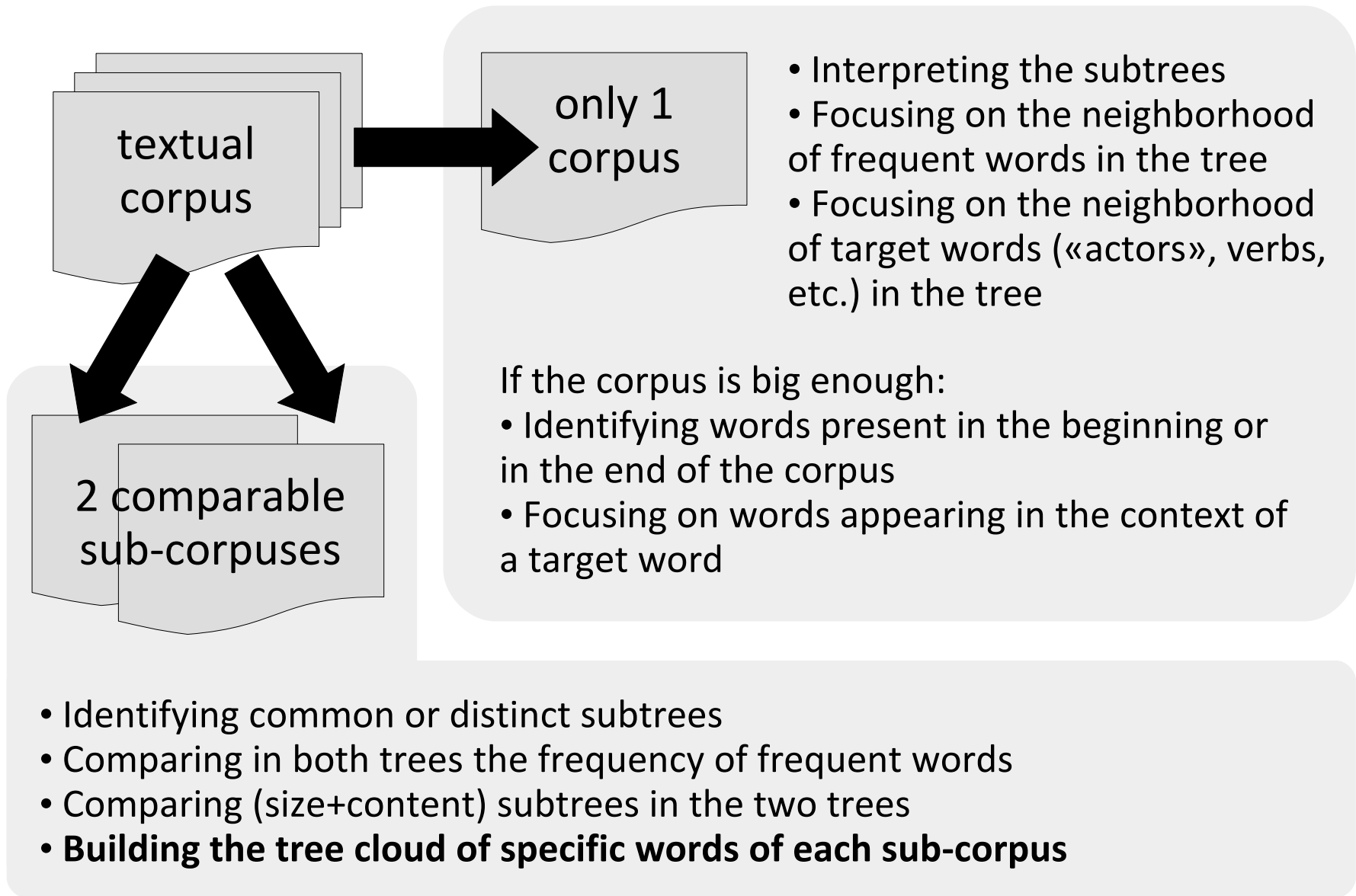


Gambette, Eggermont & Le Roux, 2014

# Comparing subtrees

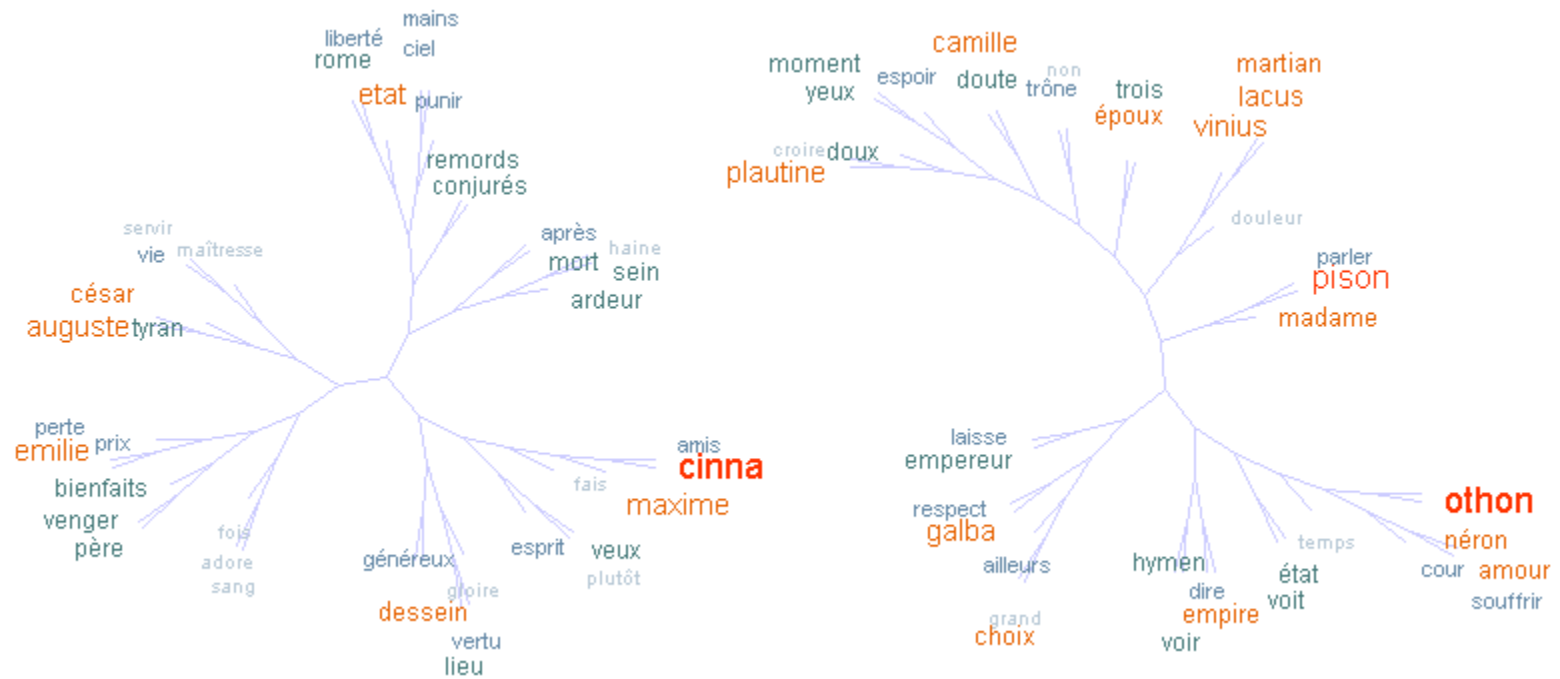


# Corpus exploration with TreeCloud



# Comparing specific words

Amstutz & Gambette,  
JADT 2010



*Tree clouds of the **specific words** in the theater plays *Cinna* et *Othon*, resized and colored according to their specificity score in Lexico 3.*

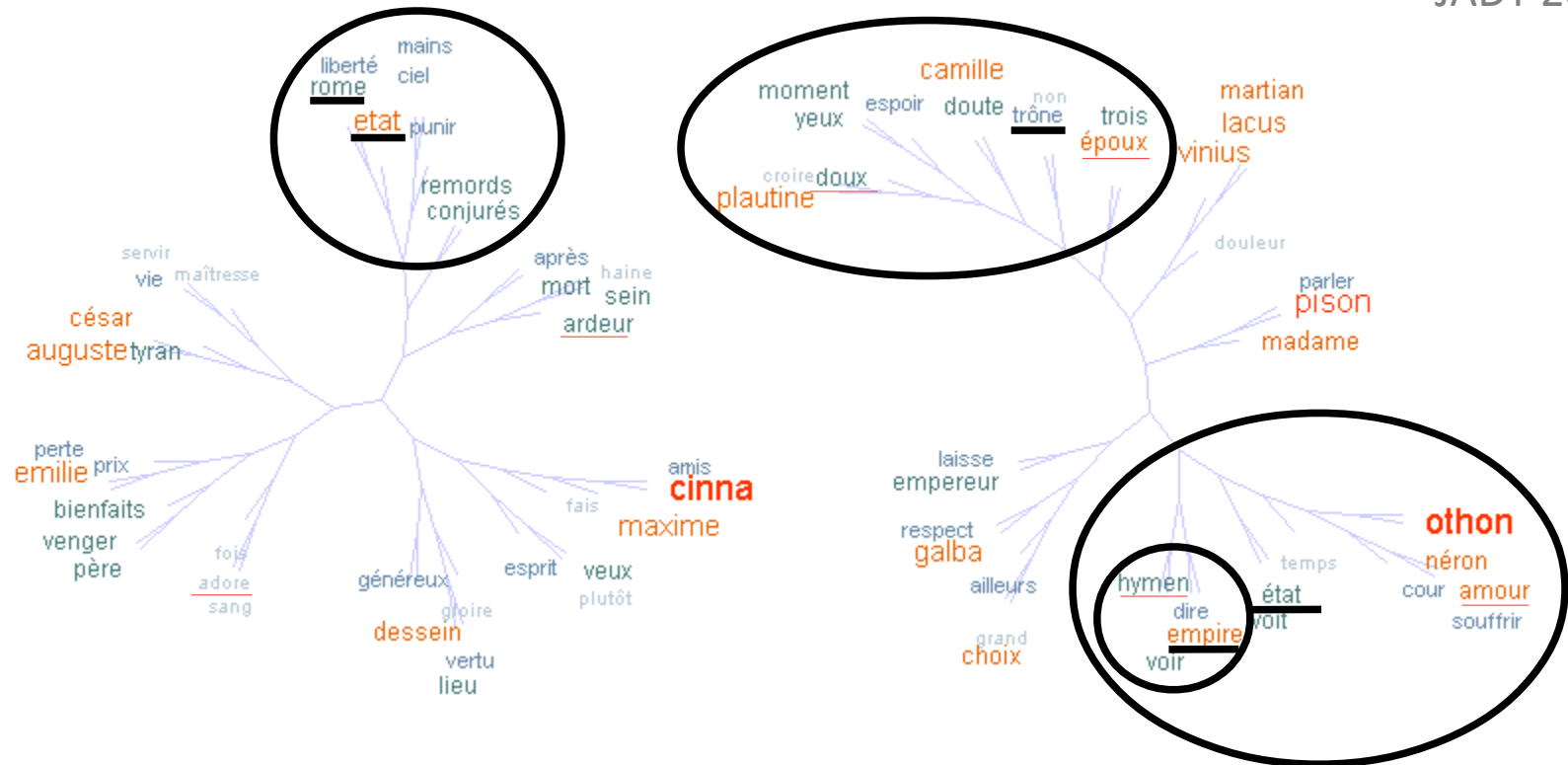


**What is political power based on?**



# Comparing specific words

Amstutz & Gambette,  
JADT 2010

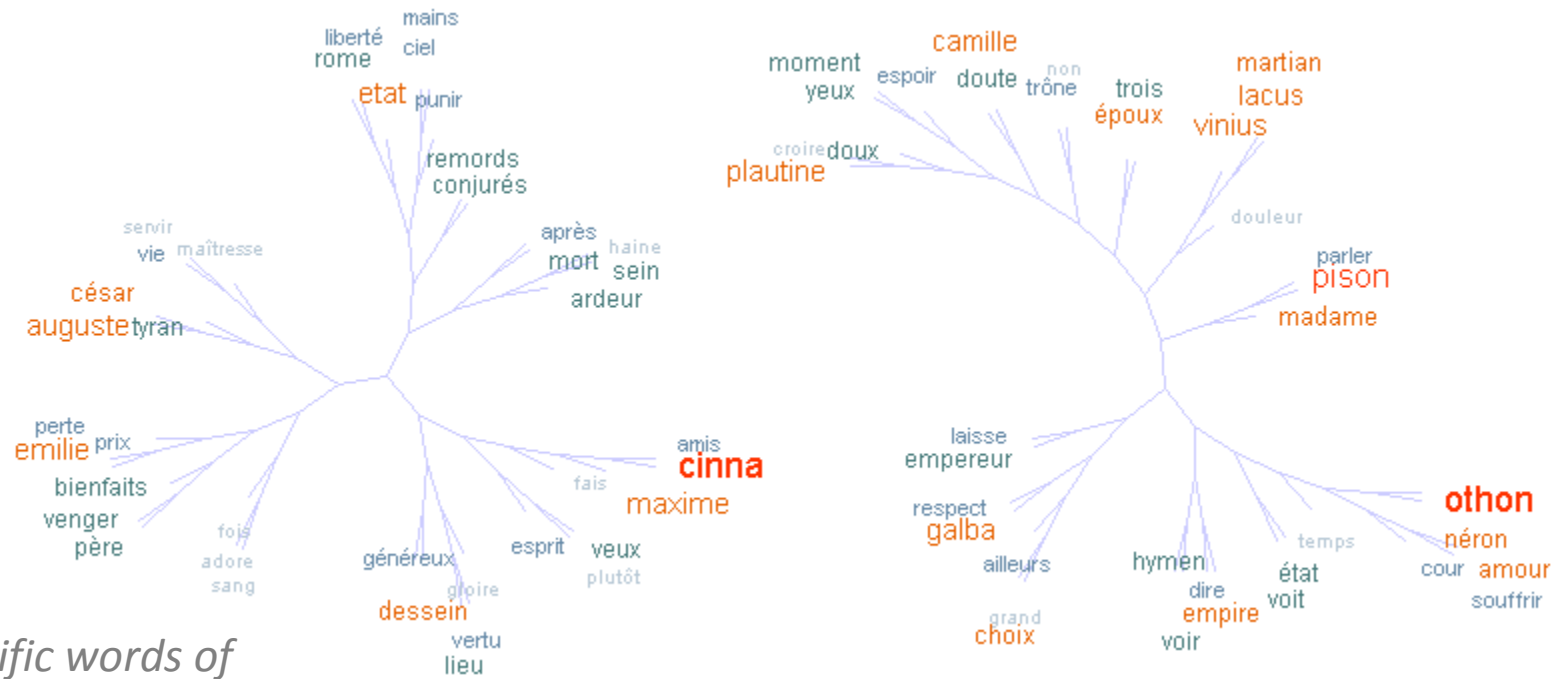


*Tree clouds of the **specific words** in the theater plays *Cinna* et *Othon*, resized and colored according to their specificity score in Lexico 3.*



**What is political power based on?**

# Comparing specific words

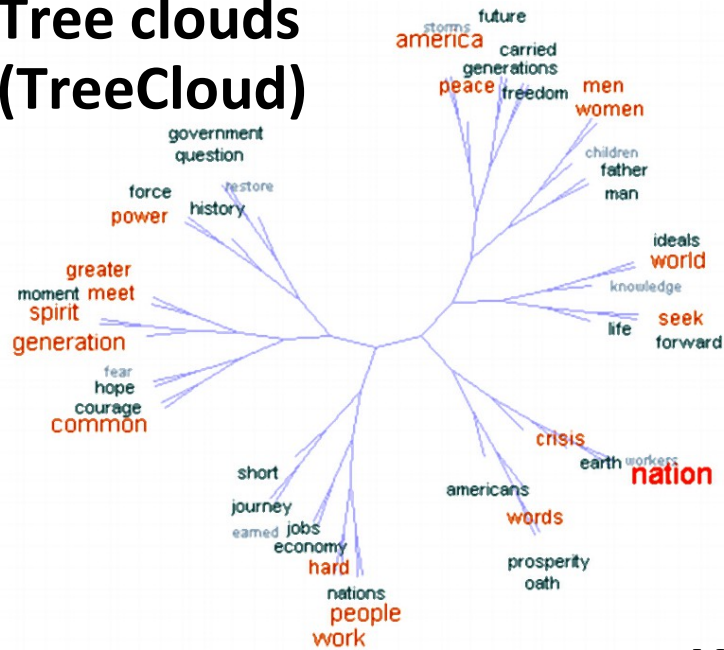


*Specific words of Cinna et Othon according to Lexico3*

	<i>Cinna</i>	<i>Othon</i>
Location of power (what the characters are fighting for)	Rome (« liberté »)	Empire (« trône »)
Reigning monarch	tyran (tyrant)	Empereur (emperor)
Characters with political influence	amis (friends)	maîtres / seigneurs (master / lord)
What political power is based on	gloire (glory)	amour matrimonial : « amour », « hymen », « choix » (love)
Characterization of the theater play	FOUNDATION	DYNASTIC SUCCESSION

# Comparing with other visualizations

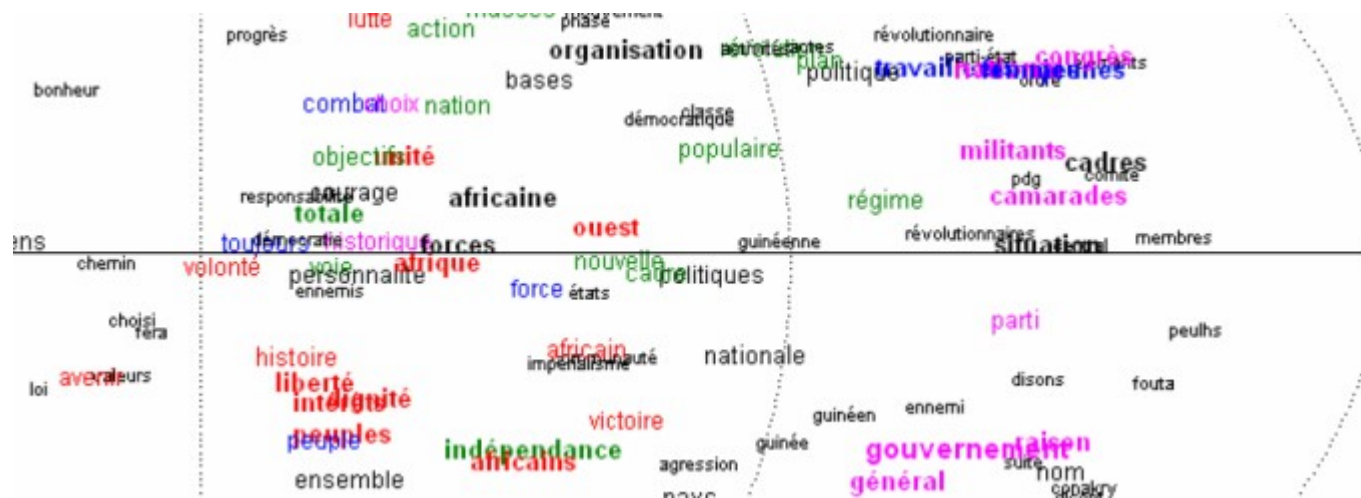
## Tree clouds (TreeCloud)



## Word networks (PhraseNet)



## Word projections (Astartex)



# Comparing with other visualizations

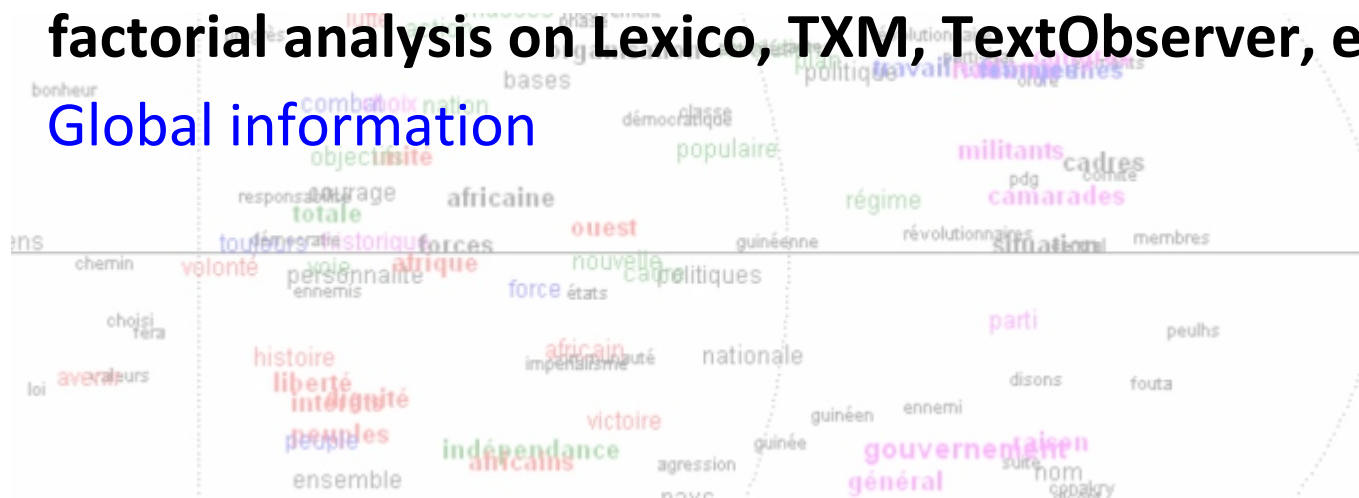
Tree clouds (TreeCloud, Hyperbase, etc.)



Word networks (PhraseNet by IBM ManyEyes, Tropes)

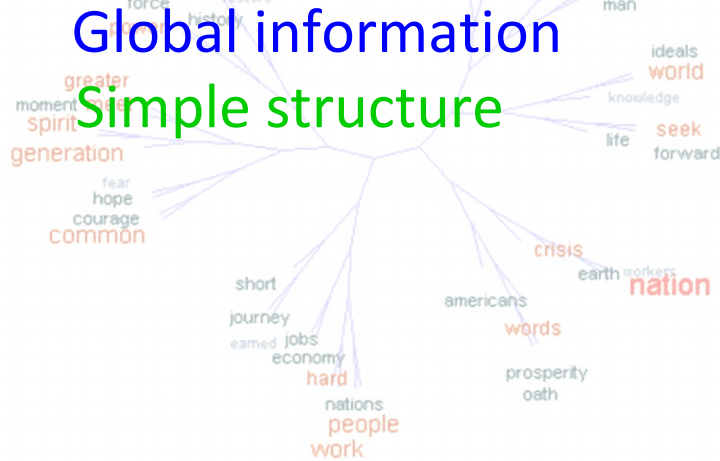


Word projections (Astartex, factorial analysis on Lexico, TXM, TextObserver, etc.)

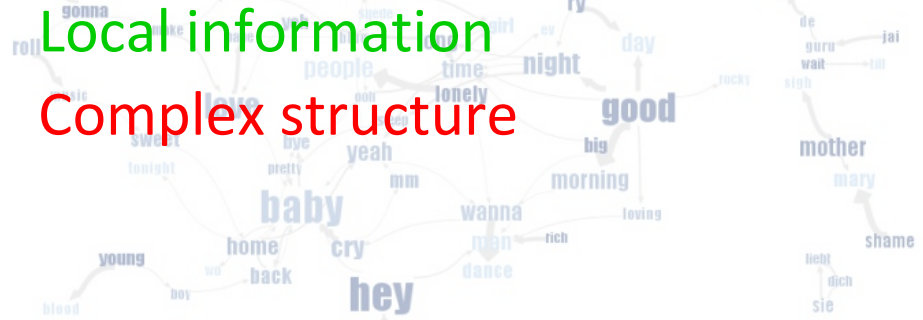


# Comparing with other visualizations

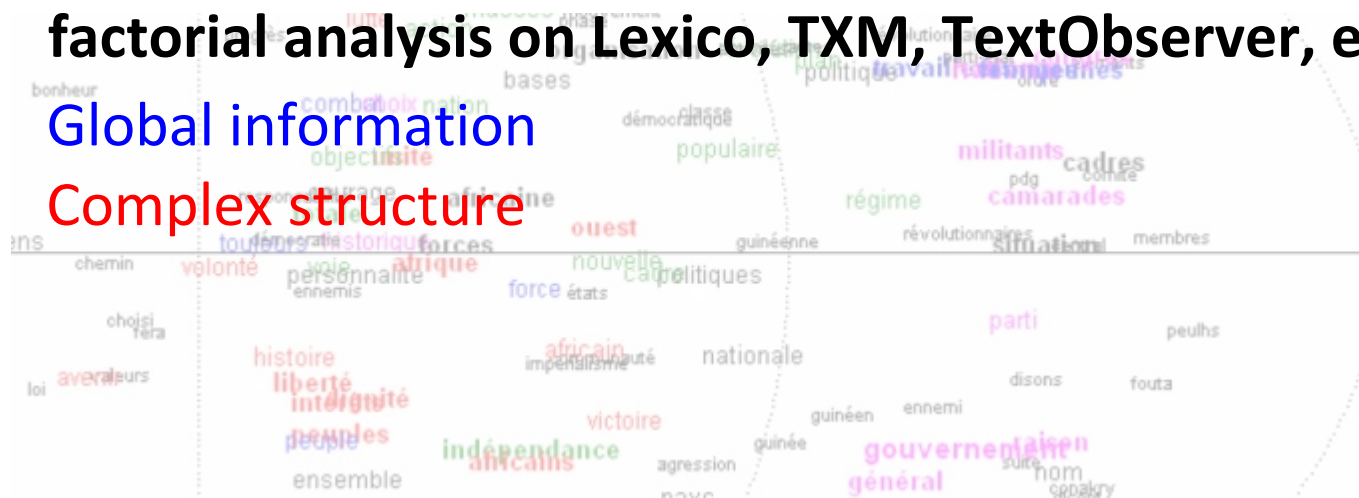
Tree clouds (TreeCloud, Hyperbase, etc.)



Word networks (PhraseNet by IBM ManyEyes, Tropes)



Word projections (Astartex, factorial analysis on Lexico, TXM, TextObserver, etc.)

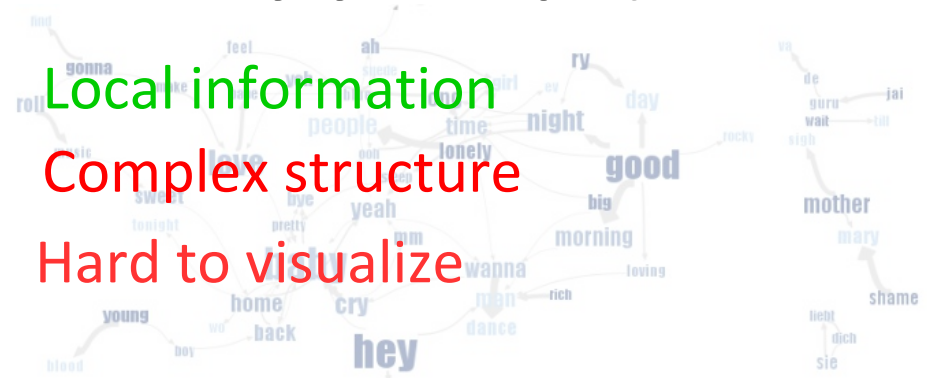


# Comparing with other visualizations

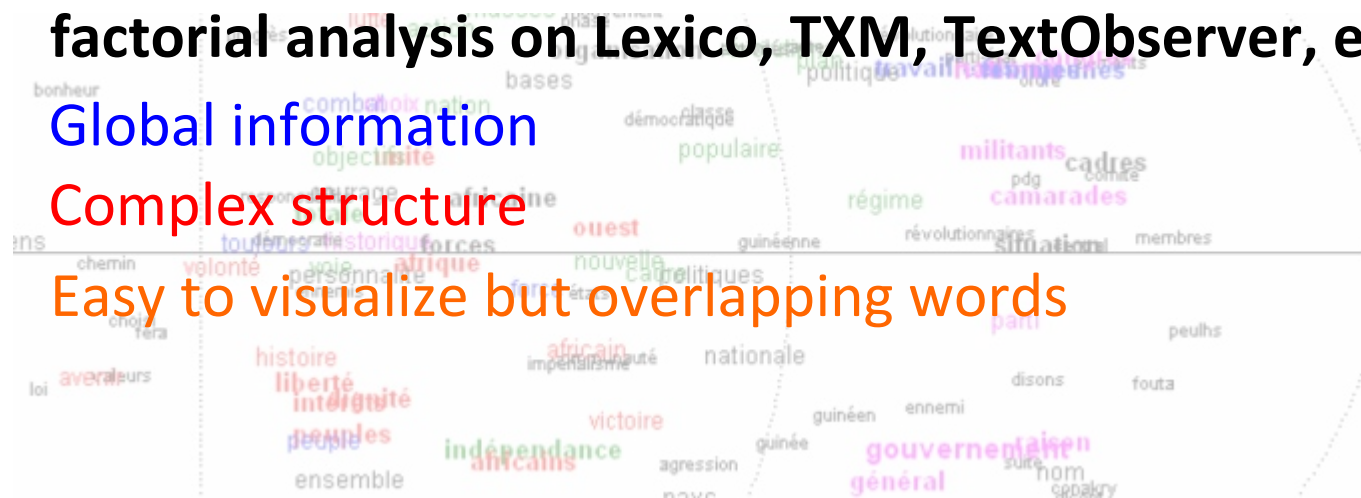
Tree clouds (TreeCloud, Hyperbase, etc.)



Word networks (PhraseNet by IBM ManyEyes, Tropes)

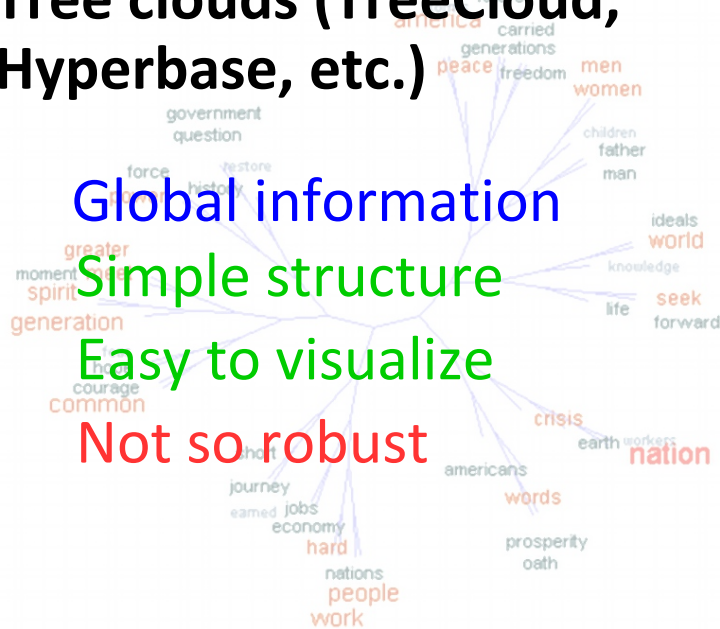


Word projections (Astartex, factorial analysis on Lexico, TXM, TextObserver, etc.)

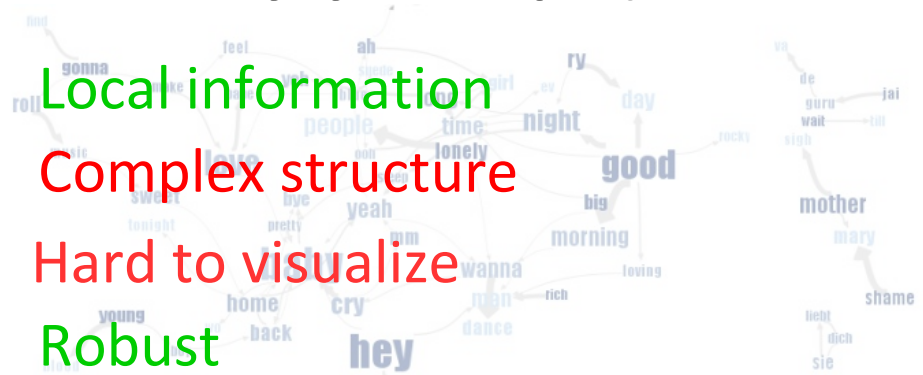


# Comparing with other visualizations

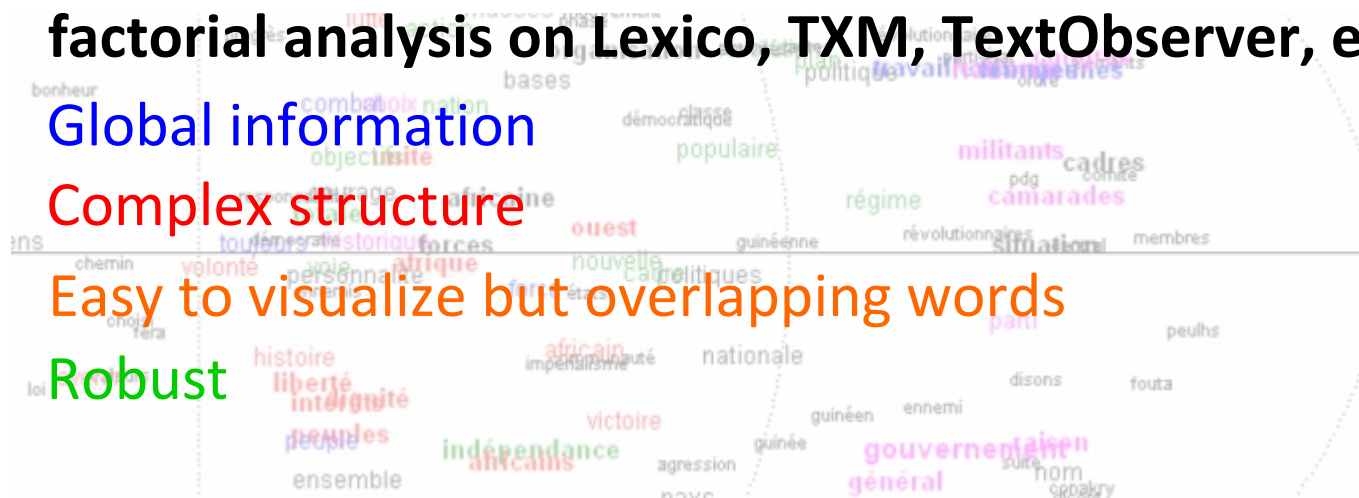
Tree clouds (TreeCloud, Hyperbase, etc.)



Word networks (PhraseNet by IBM ManyEyes, Tropes)

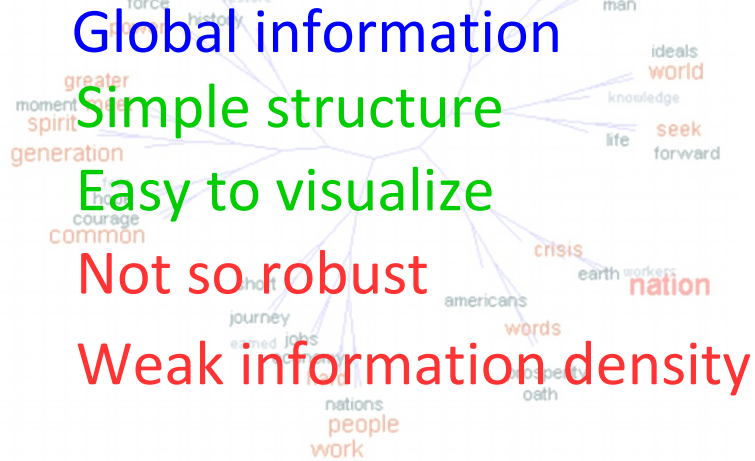


Word projections (Astartex, factorial analysis on Lexico, TXM, TextObserver, etc.)

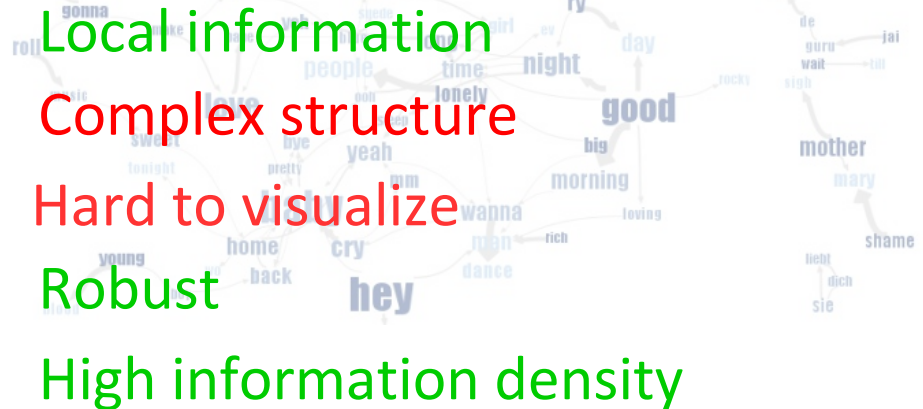


# Comparing with other visualizations

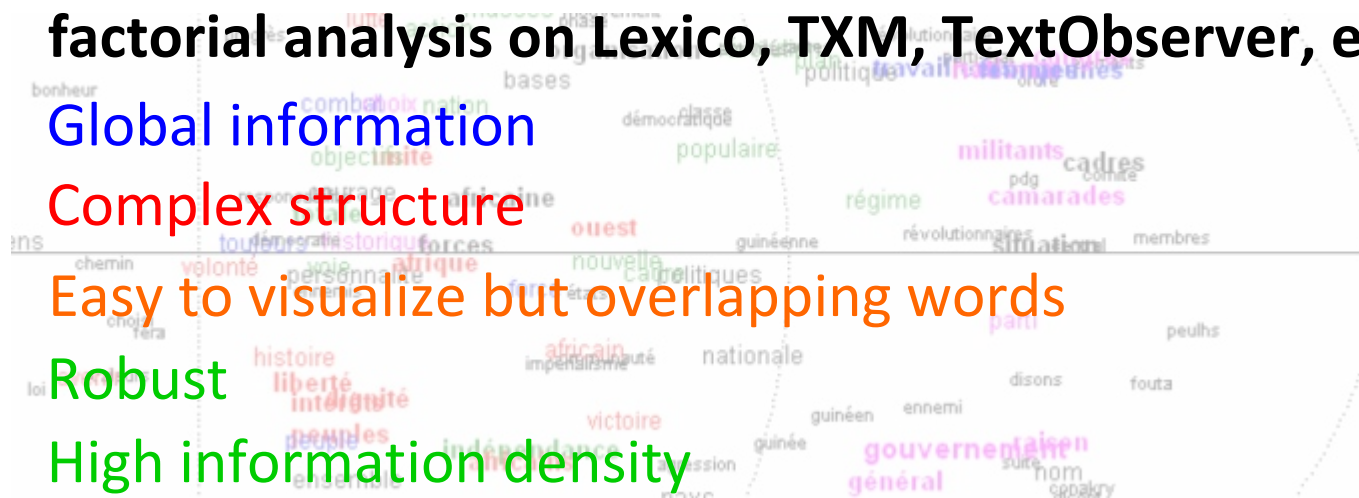
Tree clouds (TreeCloud, Hyperbase, etc.)



Word networks (PhraseNet by IBM ManyEyes, Tropes, etc.)



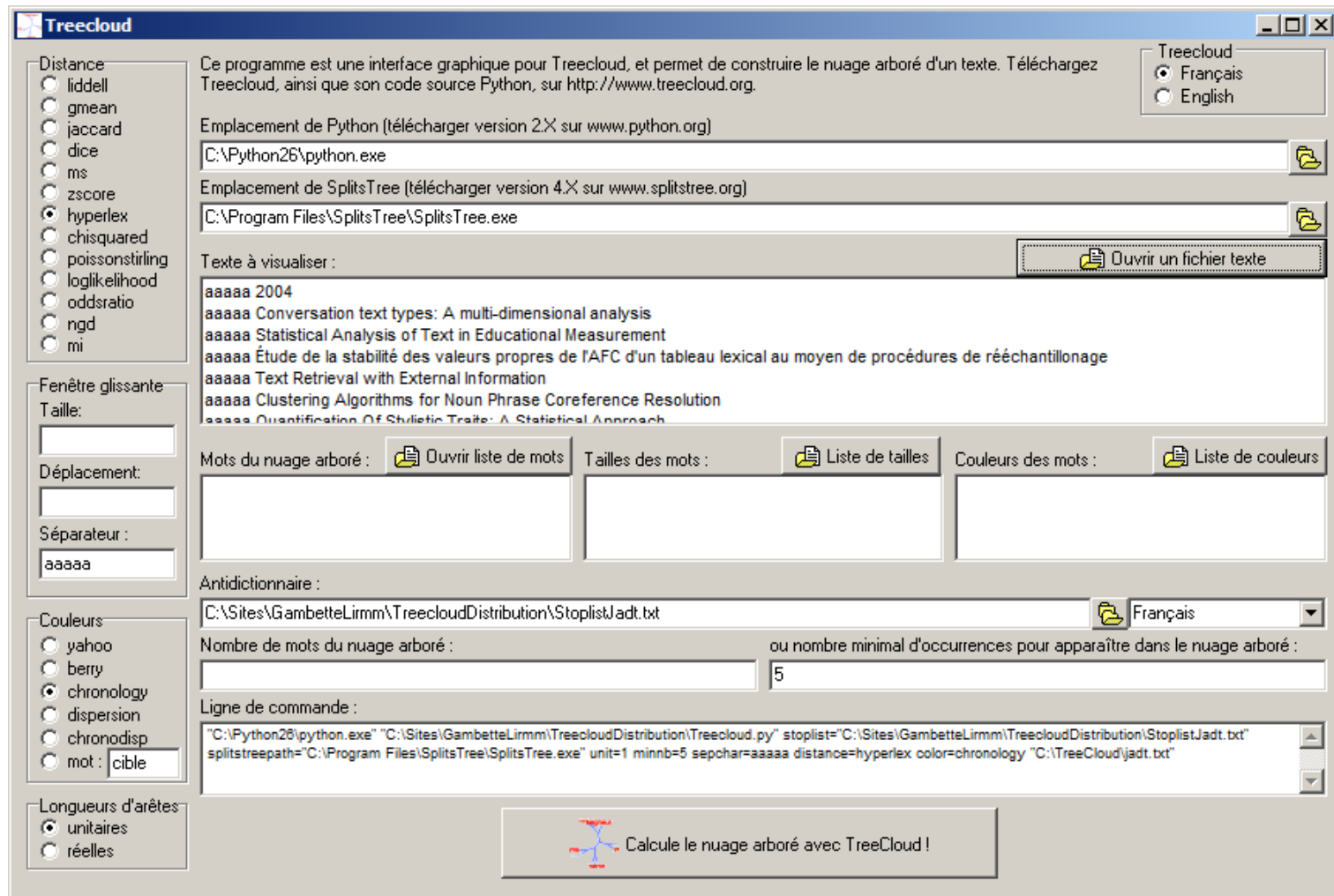
Word projections (Astartex, factorial analysis on Lexico, TXM, TextObserver, etc.)





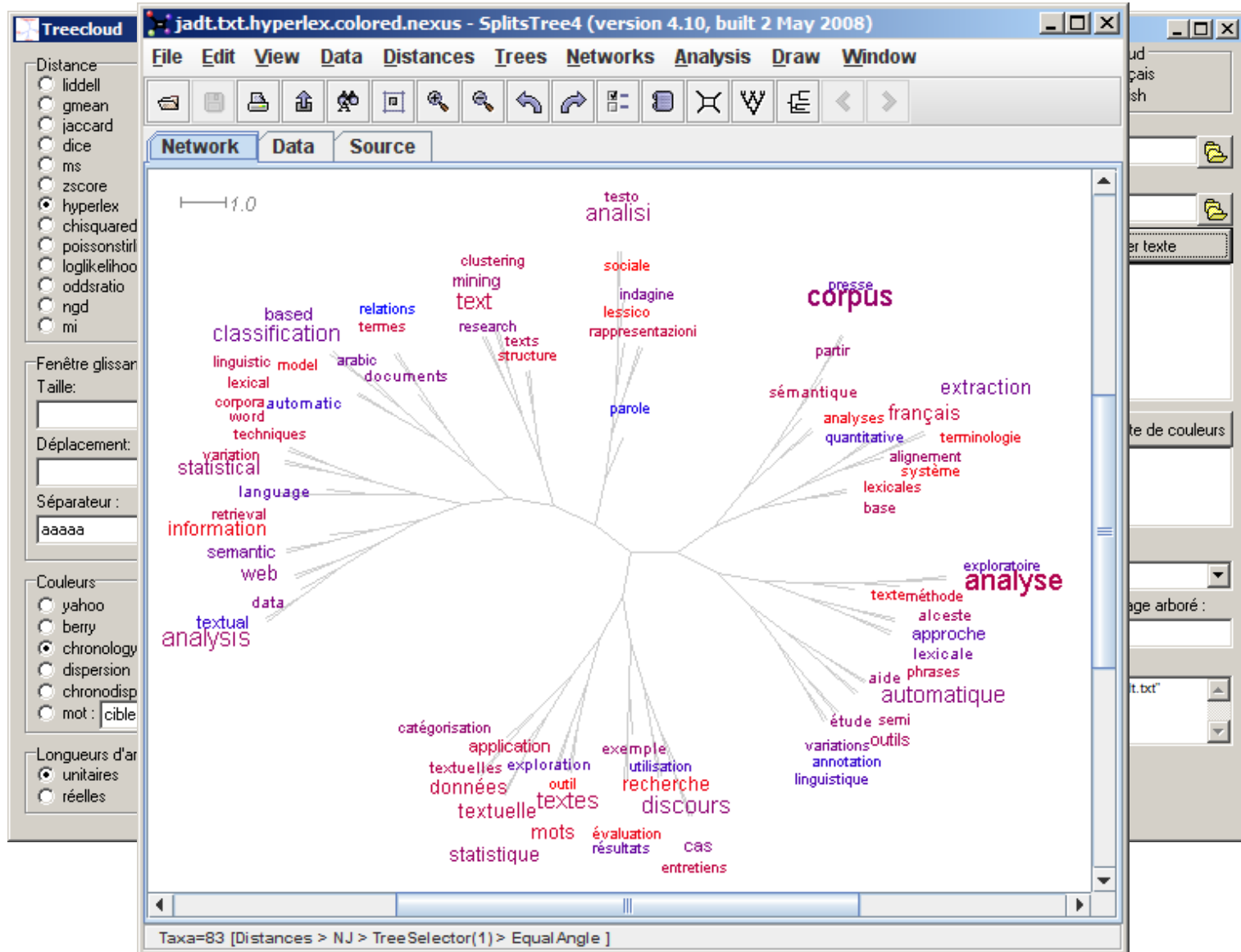
# Implementations

## Free software TreeCloud (Python/Delphi) + SplitsTree (Java)




# Implementations

Free software TreeCloud (Python/Delphi) + SplitsTree (Java)



# Web interface



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](#).

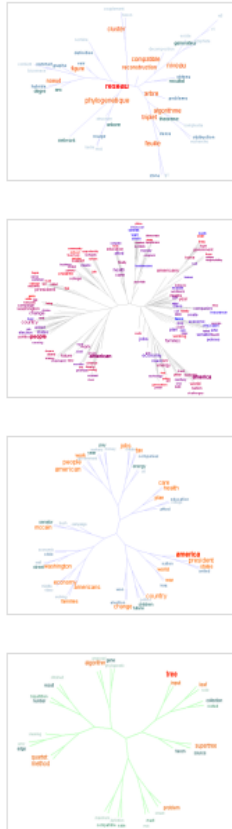
**Créez vos propres nuages arborés en ligne !**

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

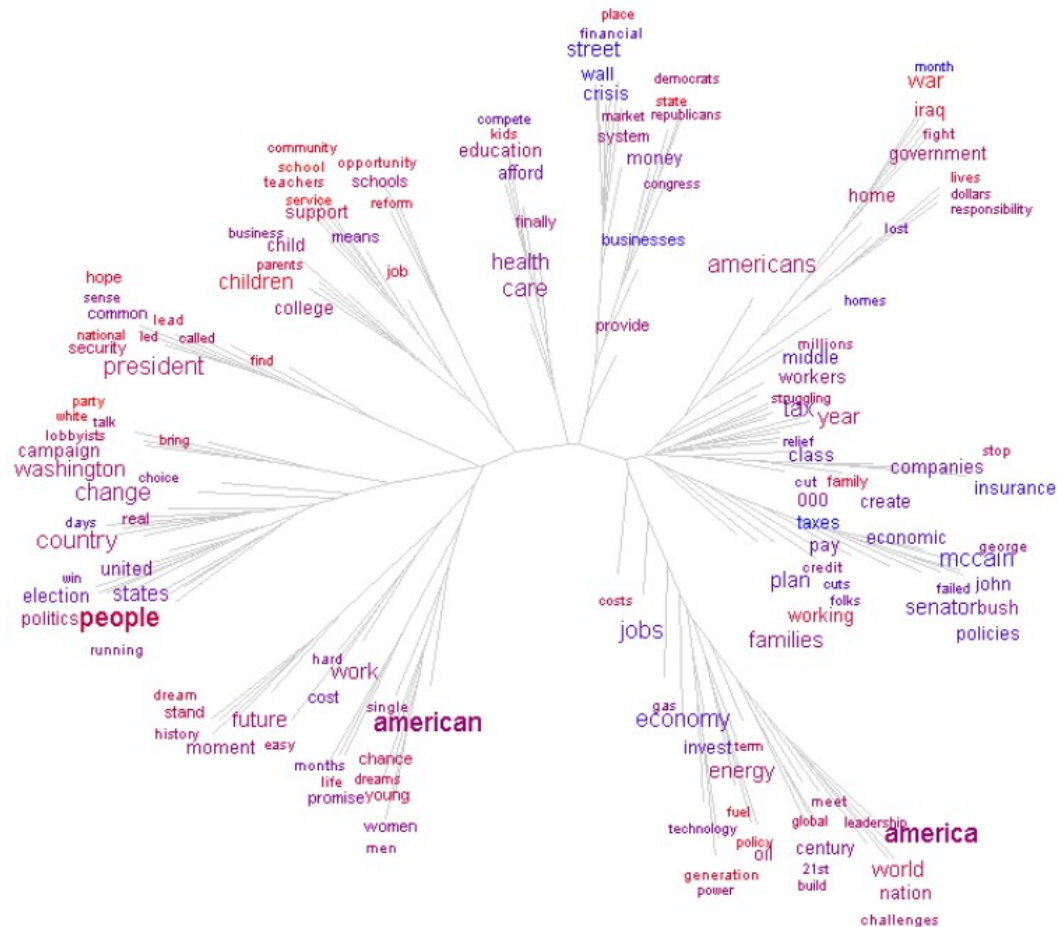
Interface based on the free software NuageArboré by Jean-Charles Bontemps (C, CGI/Python, JavaScript).

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

Integration of Unitex for the detection of compound words by Claude Martineau

# Execution time



Limits on the corpus size to use TreeCloud ?



30 seconds to build the tree cloud of Barack Obama 2008 campaign speeches (>300 000 words)

# Perspectives

## Integrating the visualization in other software

- Integration into Unitex  thanks to Google Summer of Code 
- Javascript treecloud visualization available in PhyloPlot  
<http://adamzy.github.io/PhyloPlot/>

## Helping users with the methodology

- Adding tools to compare the trees
- Adding dynamic processes:
  - \* adding words, removing words, etc.
  - \* going back to the full text

# References (available on *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>

Nadège Lechevrel & Philippe Gambette (2016)

**Une approche textométrique pour étudier la transmission des savoirs biologiques au XIXe siècle**, *Nouvelles perspectives en sciences sociales*, à paraître