

Journées de rencontre ANR Biographes
07/11/2014 – Fondation Maison des Sciences de l'Homme, Paris

TreeCloud pour la visualisation et l'analyse de données textuelles

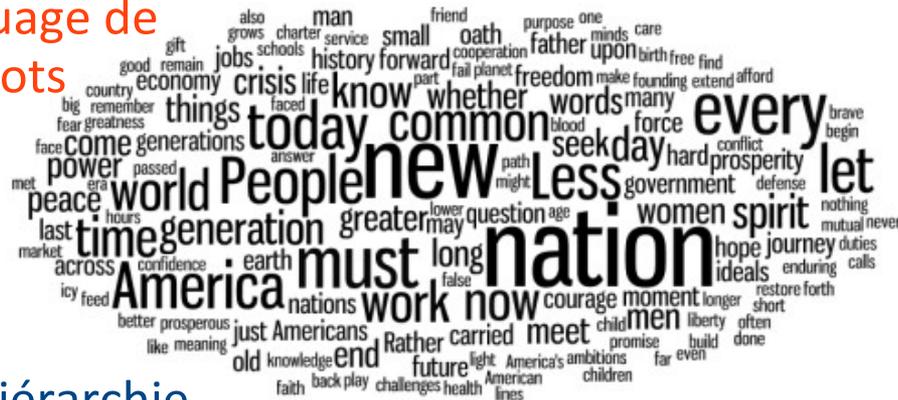
Philippe Gambette

LIGM
Université Paris-Est
Marne-la-Vallée



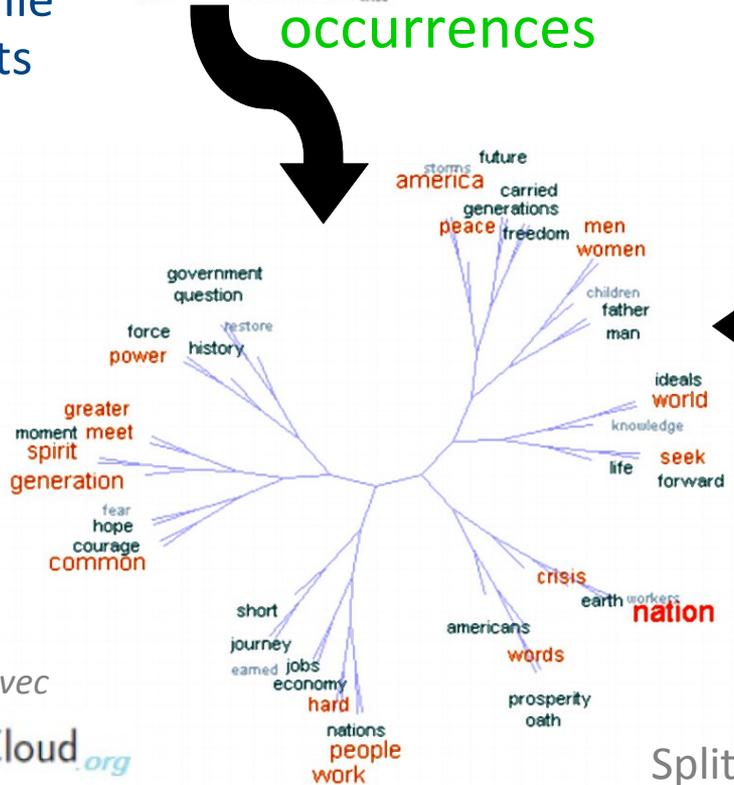
Le « nuage arboré », une information double

nuage de mots



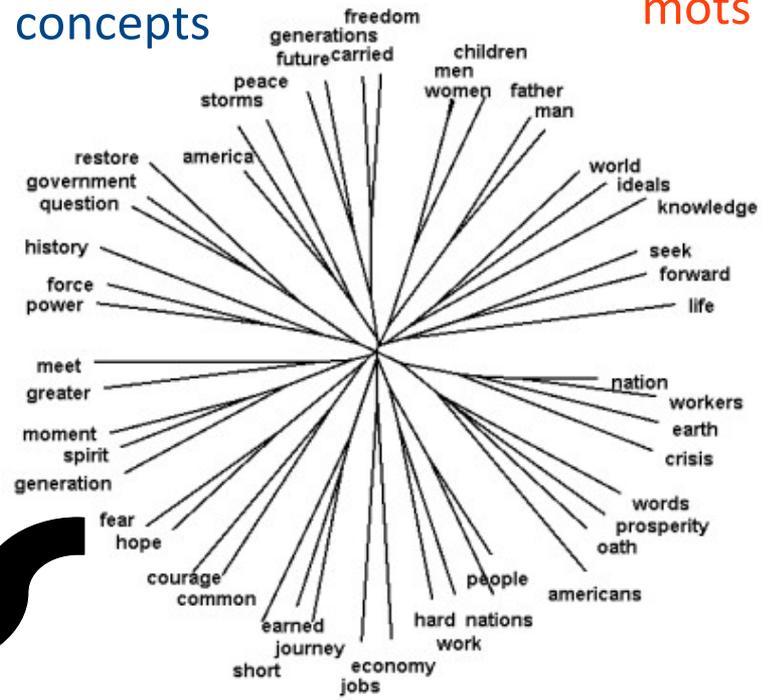
hiérarchie des mots

occurrences



hiérarchie des concepts

arbre de mots



cooccurrences

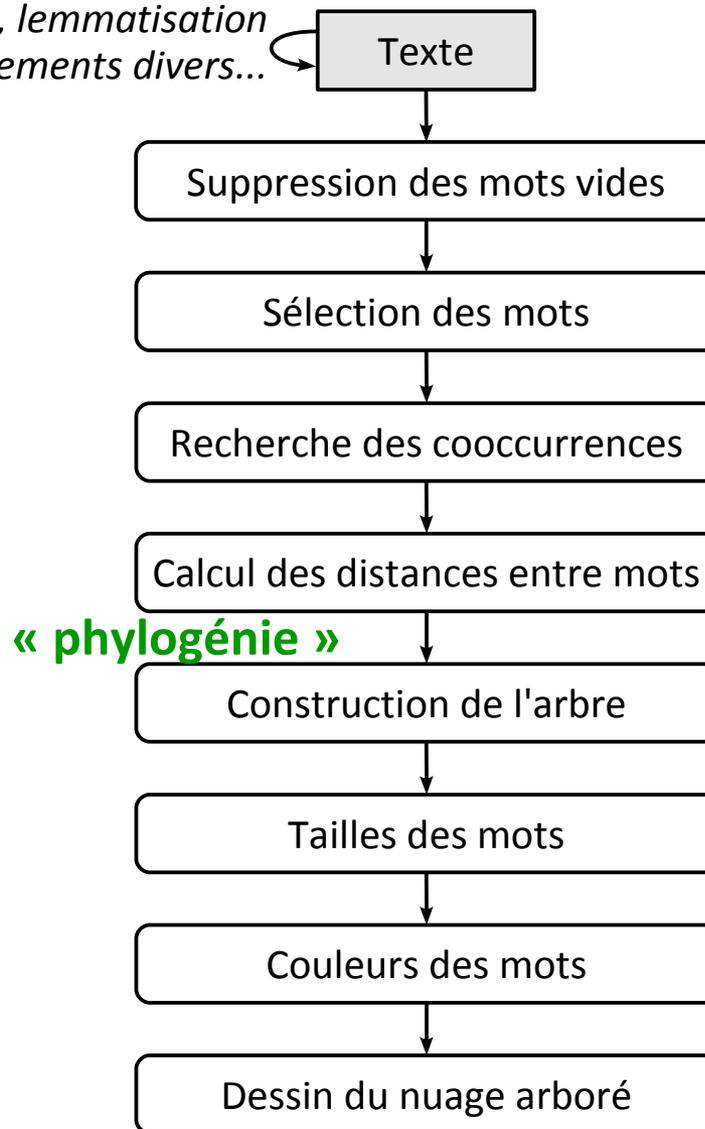
Discours inaugural de Barack Obama

construit avec
 TreeCloud.org
 SplitsTree4

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

Processus de construction

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



**Proposé dans la version
téléchargeable de TreeCloud**

antidico anglais, français

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

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

12 formules de distance de cooccurrence

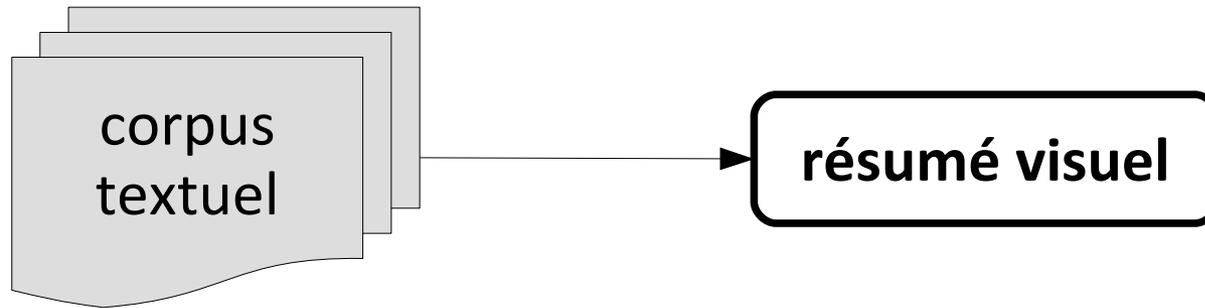
*Appel transparent au logiciel
SplitsTree*

Fréquences ou valeurs personnalisées

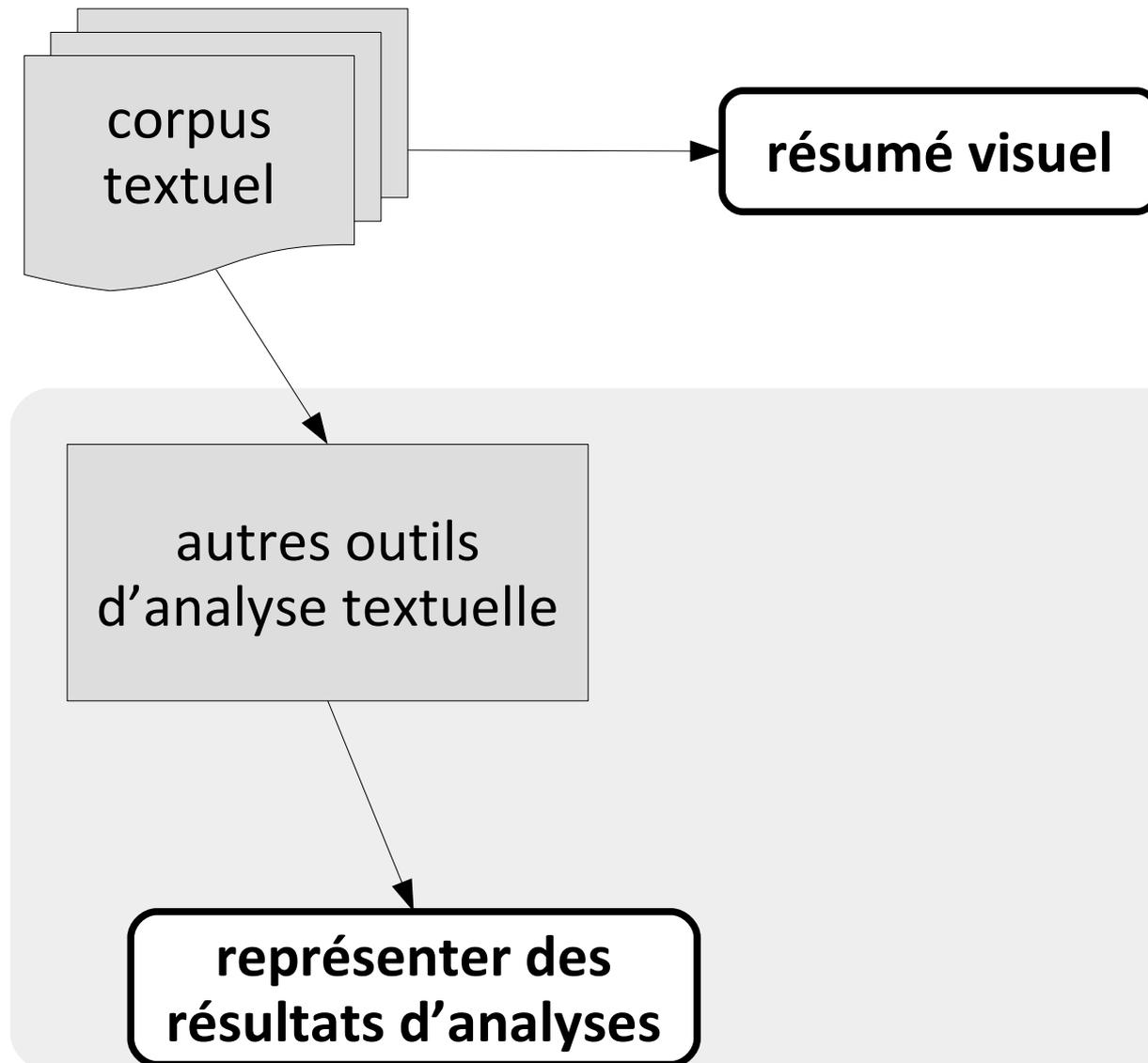
*Fréquences, chronologie, dispersion,
ciblées sur la cooccurrence d'un mot,
ou valeurs personnalisées*

*Appel transparent au logiciel SplitsTree
ou Dendroscope*

Le « nuage arboré », pour quoi faire ?

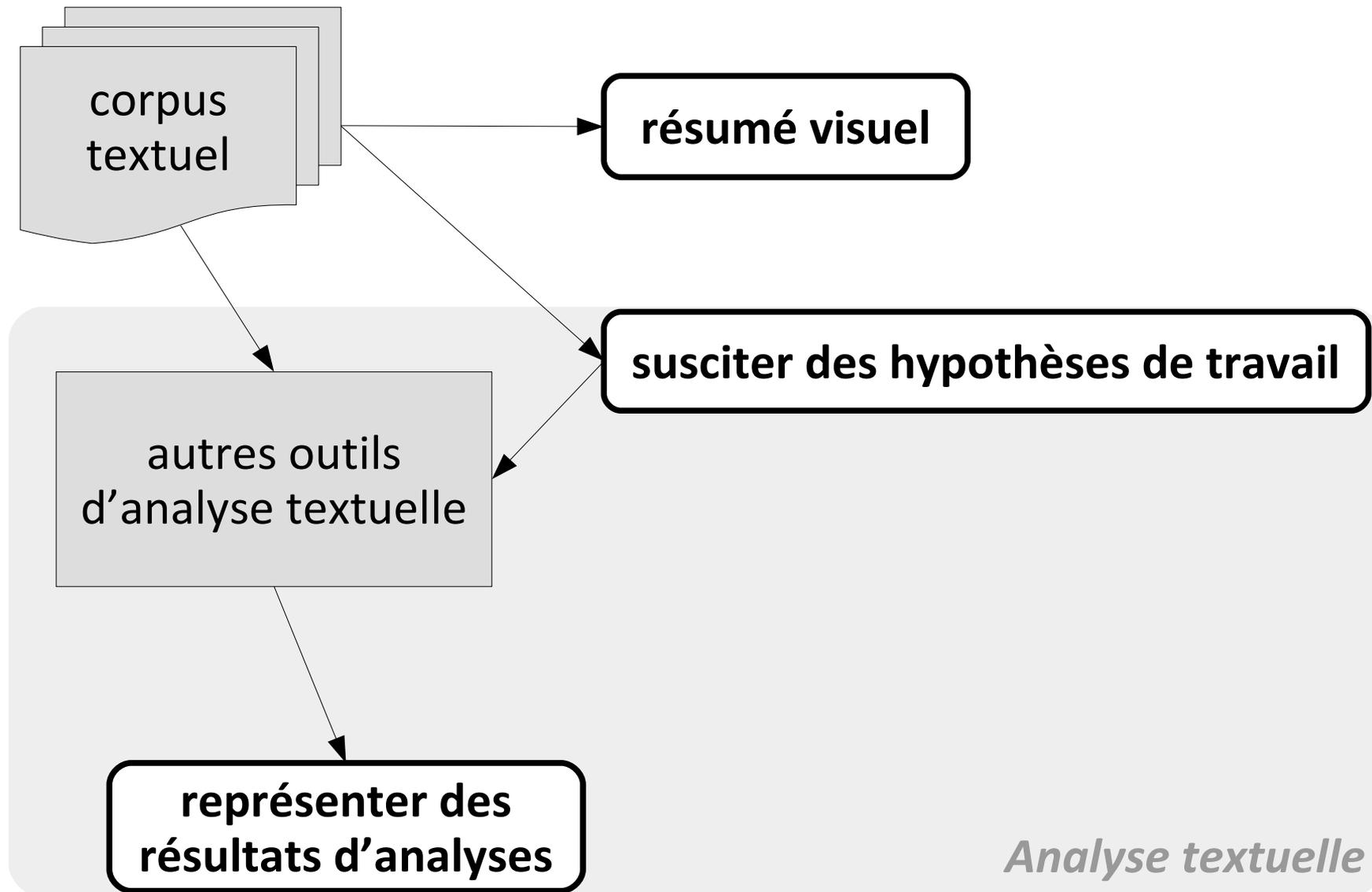


Le « nuage arboré », pour quoi faire ?



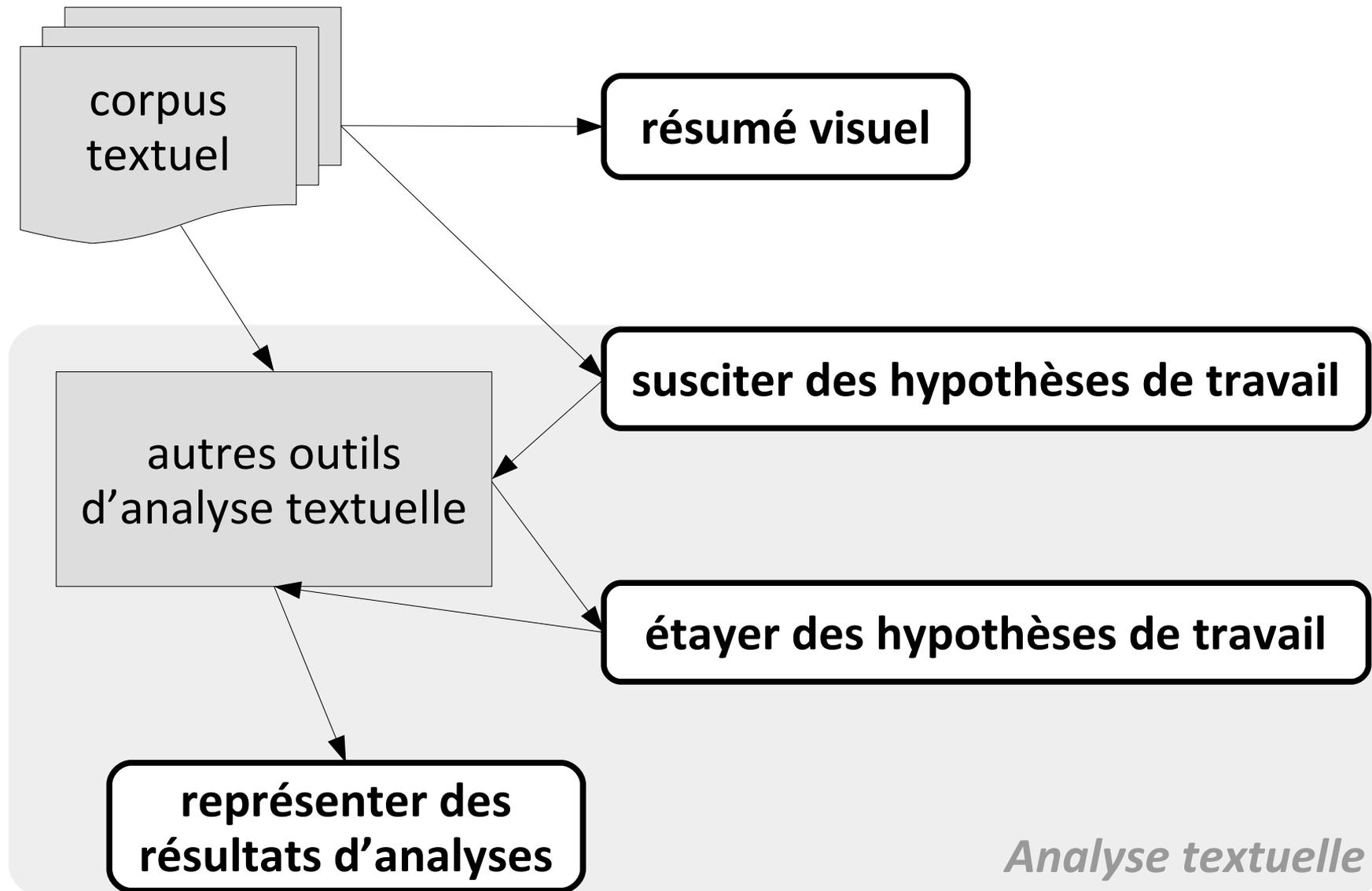
Analyse textuelle

Le « nuage arboré », pour quoi faire ?



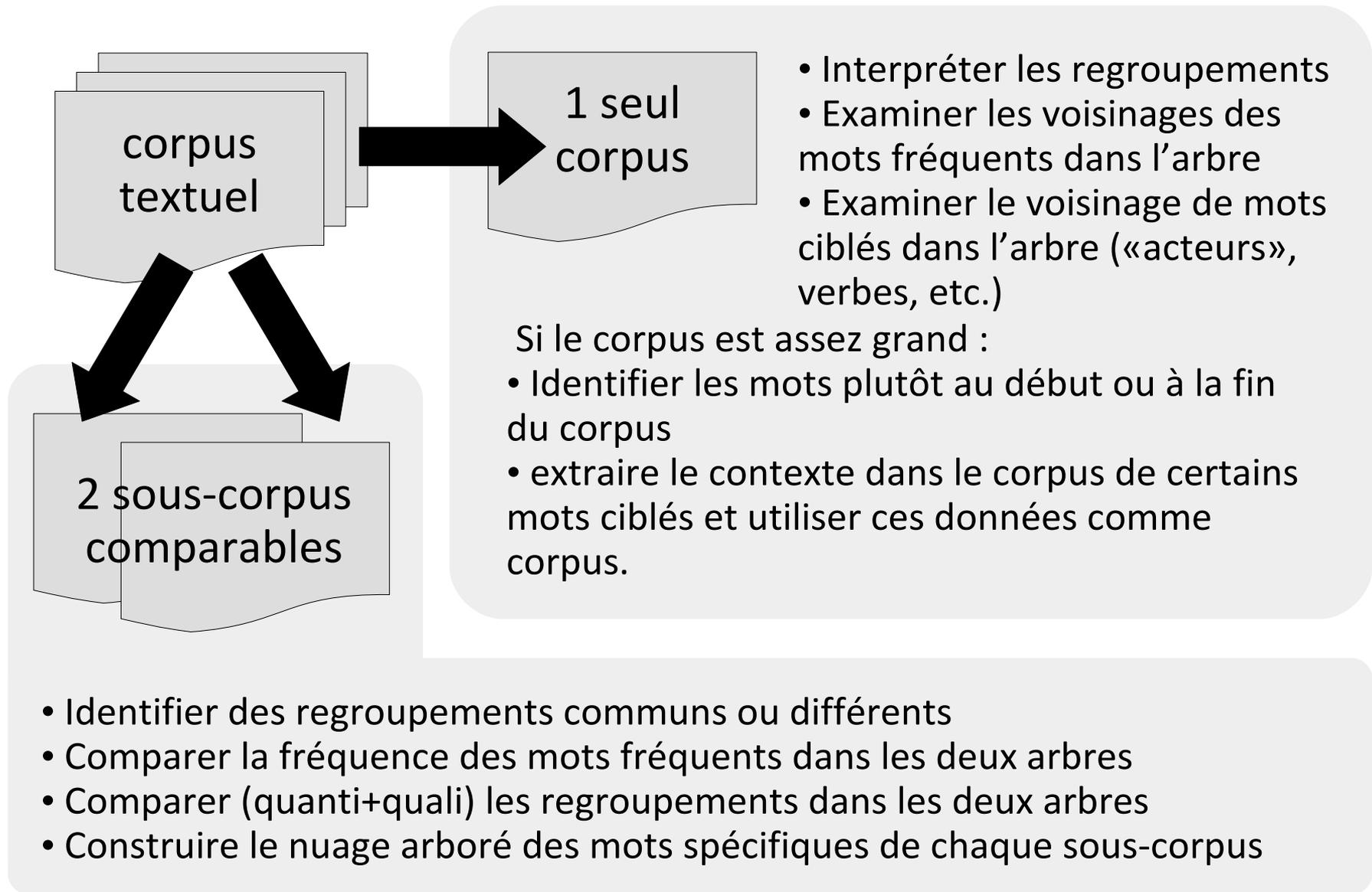
Analyse textuelle

Le « nuage arboré », pour quoi faire ?

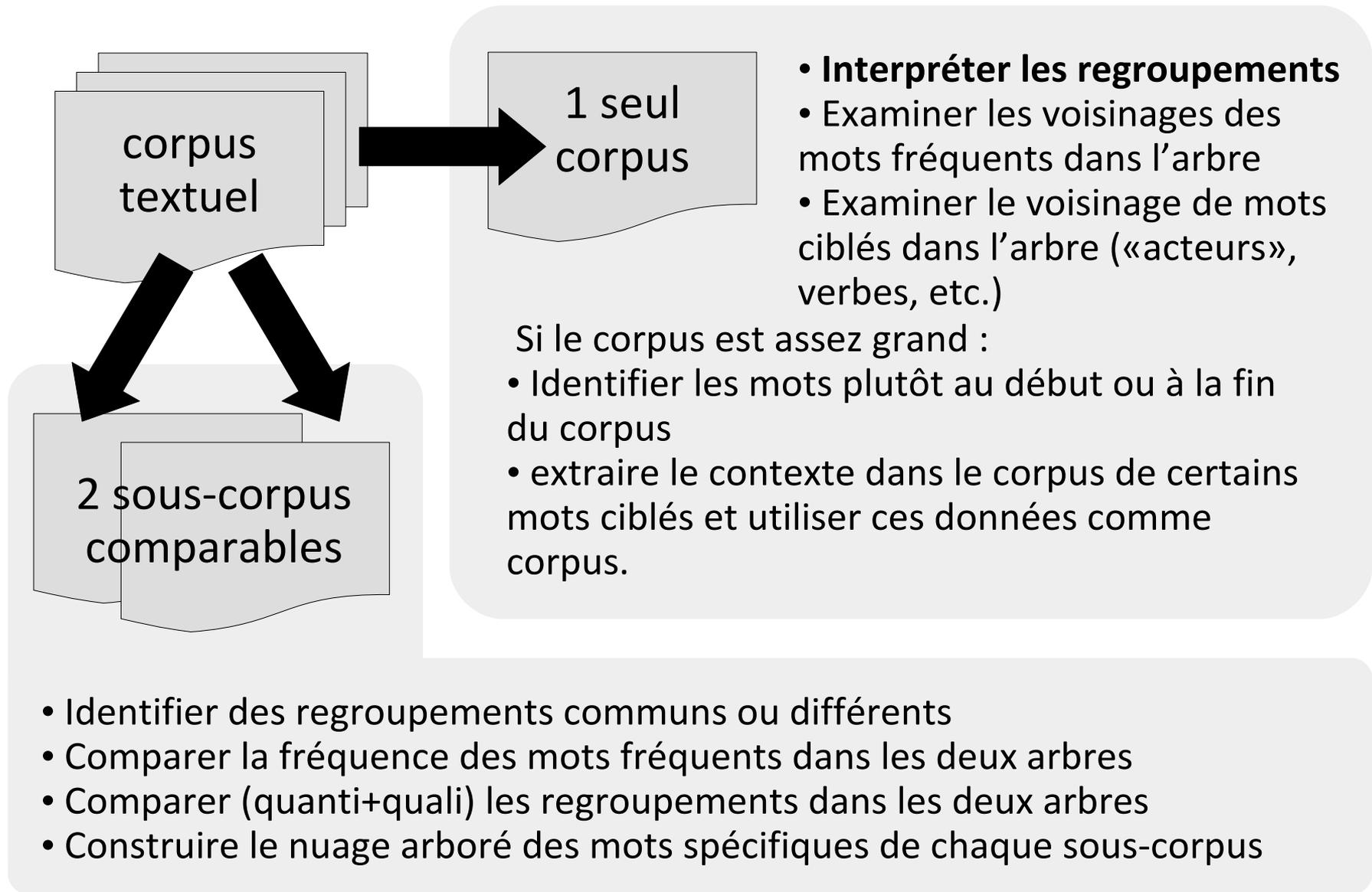


Analyse textuelle

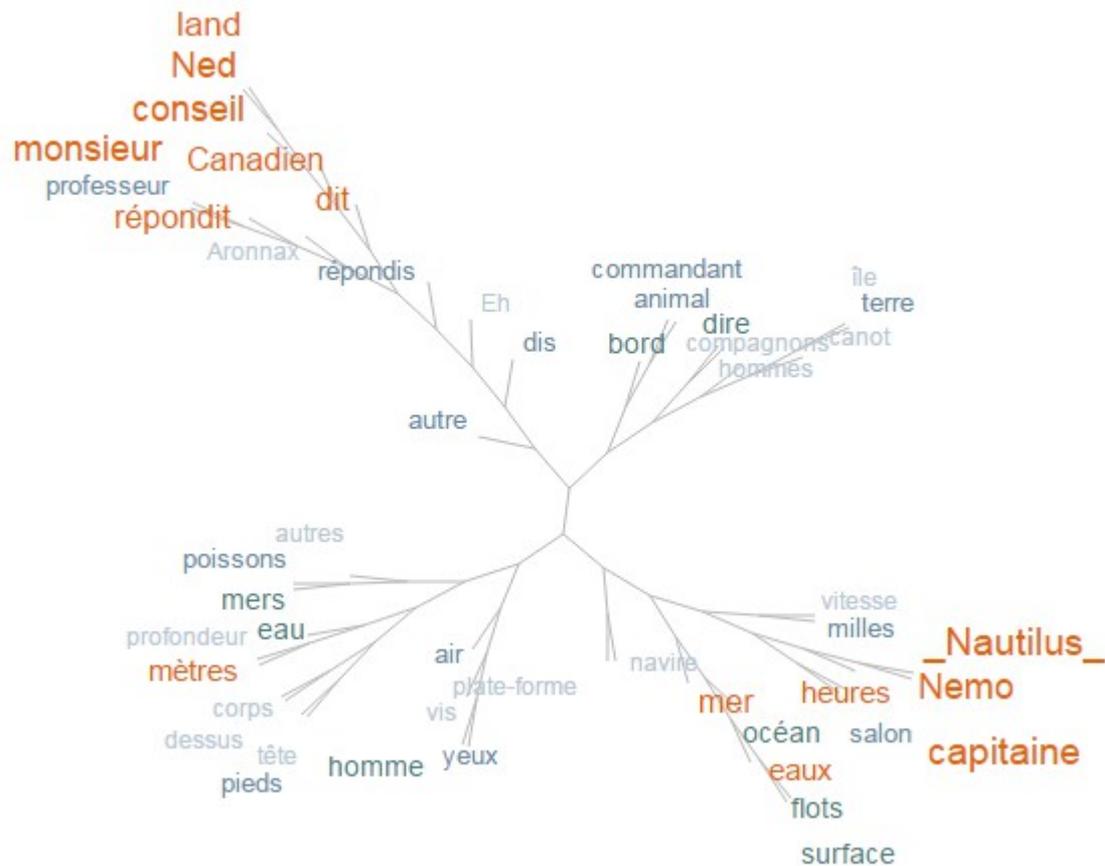
Exploration de corpus avec TreeCloud



Exploration de corpus avec TreeCloud



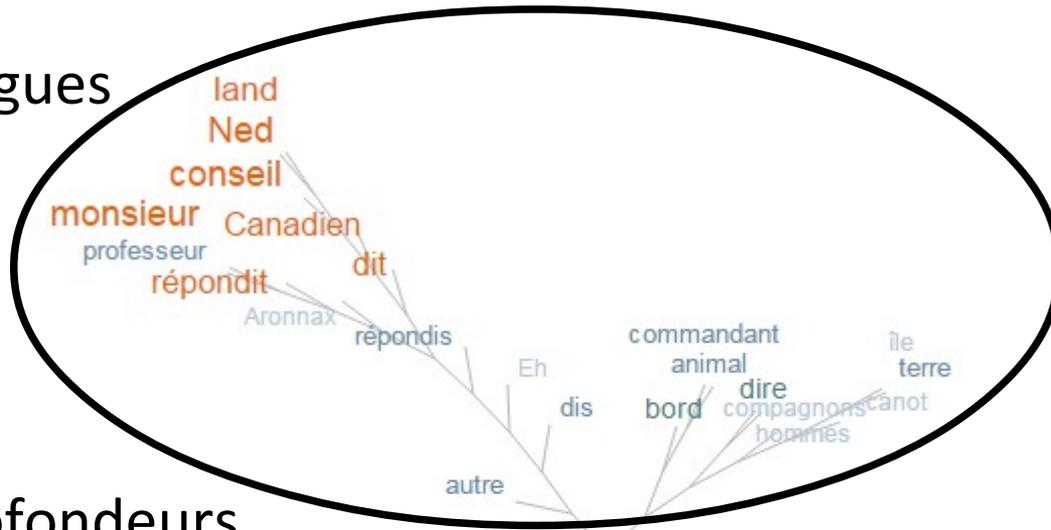
Méthode : interpréter les regroupements



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

Méthode : interpréter les regroupements

dialogues



profondeurs
marines

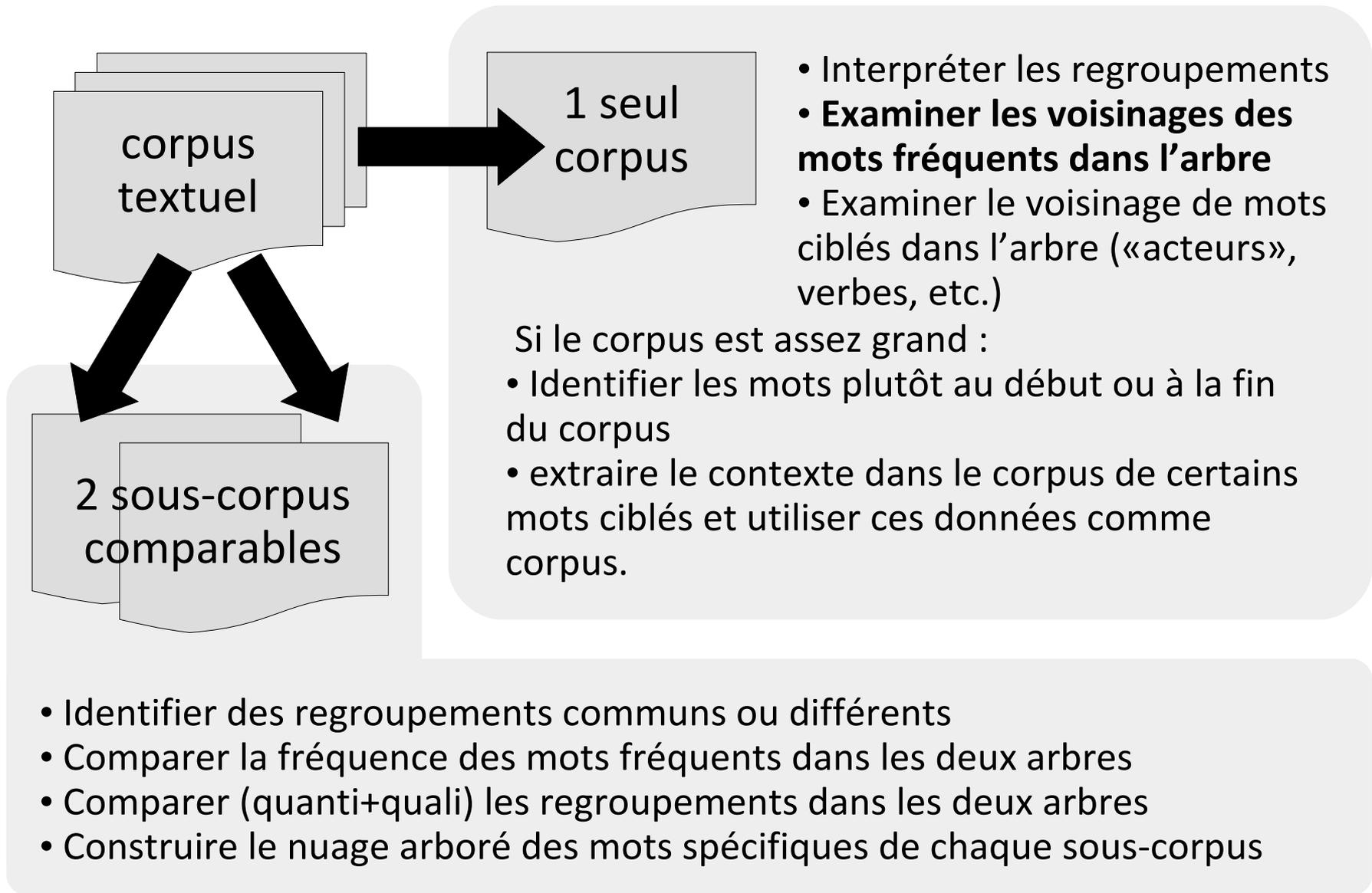


corps

océan

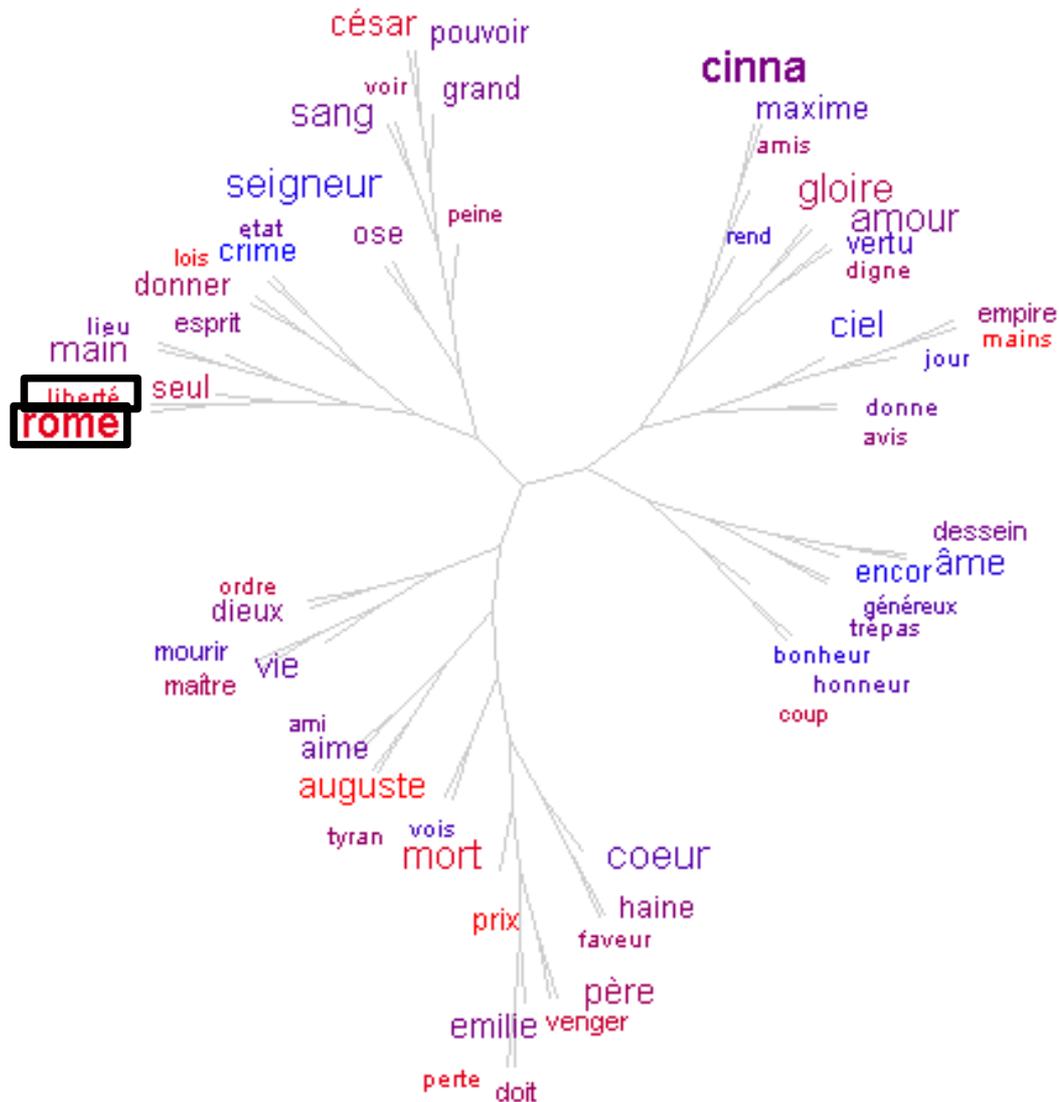
Nuage arboré
construit sur
Treecloud.org du
roman *Vingt mille
lieues sous les
mers* de Jules
Verne (distance
Jaccard, 50 mots
les plus fréquents)

Exploration de corpus avec TreeCloud



Méthode : voisinage des mots fréquents

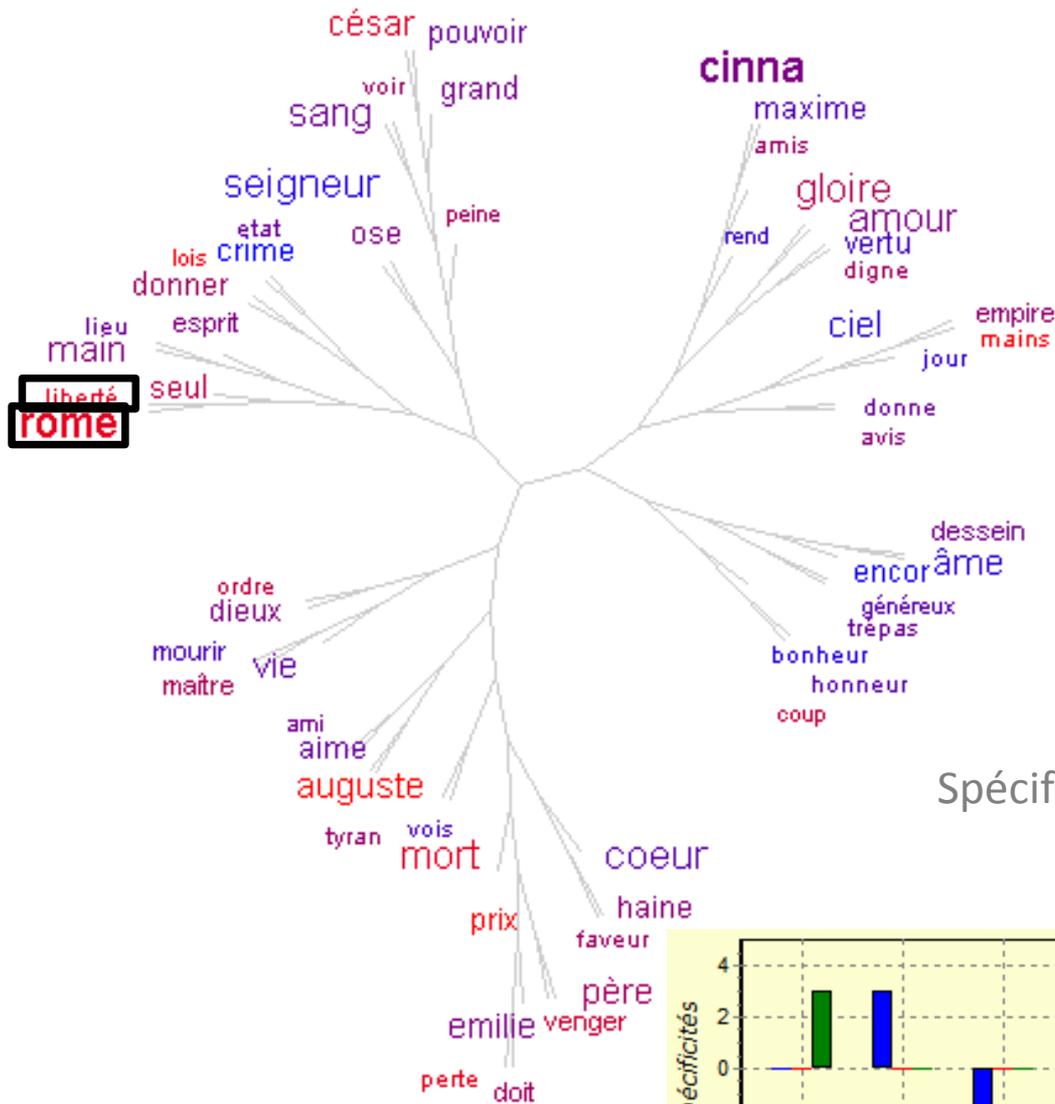
Amstutz & Gambette, JADT 2010



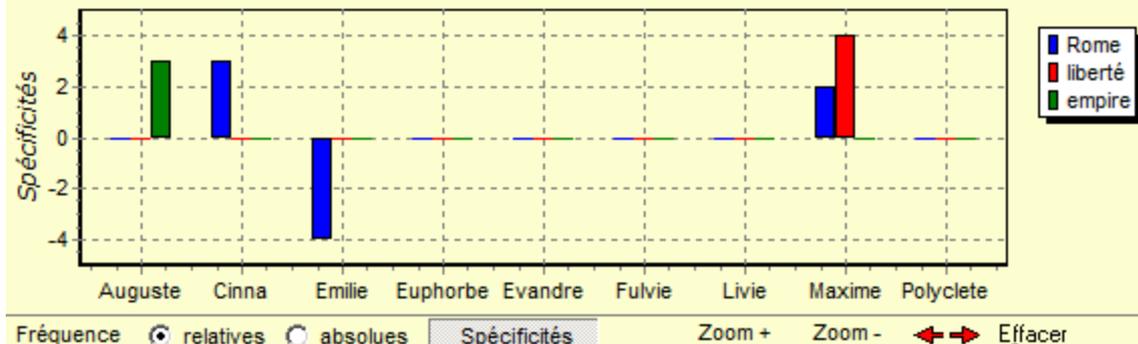
Nuage arboré global des 60 mots les plus fréquents dans *Cinna* de Corneille (distance Liddell, fenêtre de largeur 20), colorés chronologiquement (rouge au début, bleu à la fin)

Méthode : voisinage des mots fréquents

Amstutz & Gambette, JADT 2010

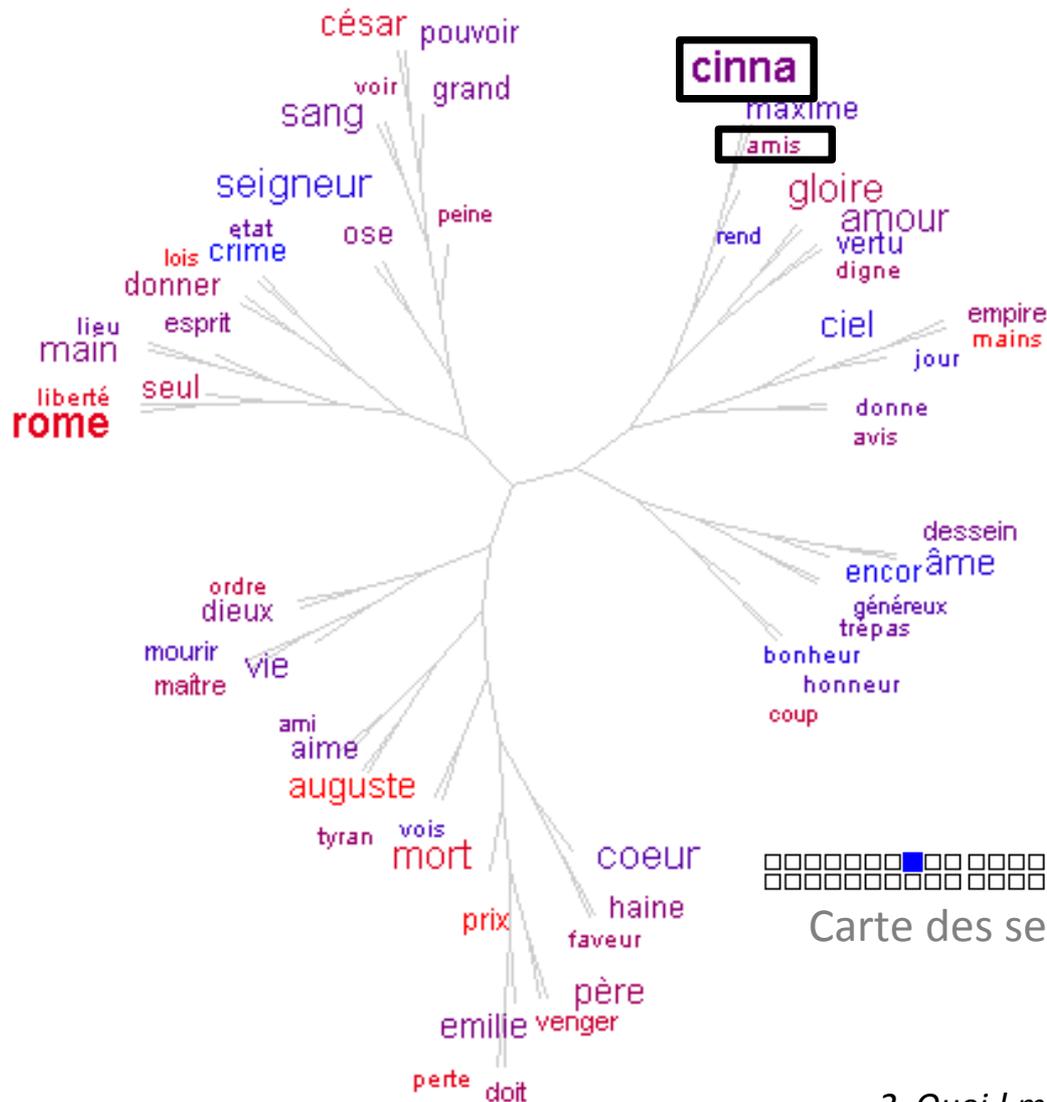


Spécificités d'emploi de « Rome », « liberté » et « empire », chez les différents personnages de Cinna, selon Lexico3



Méthode : voisinage des mots fréquents

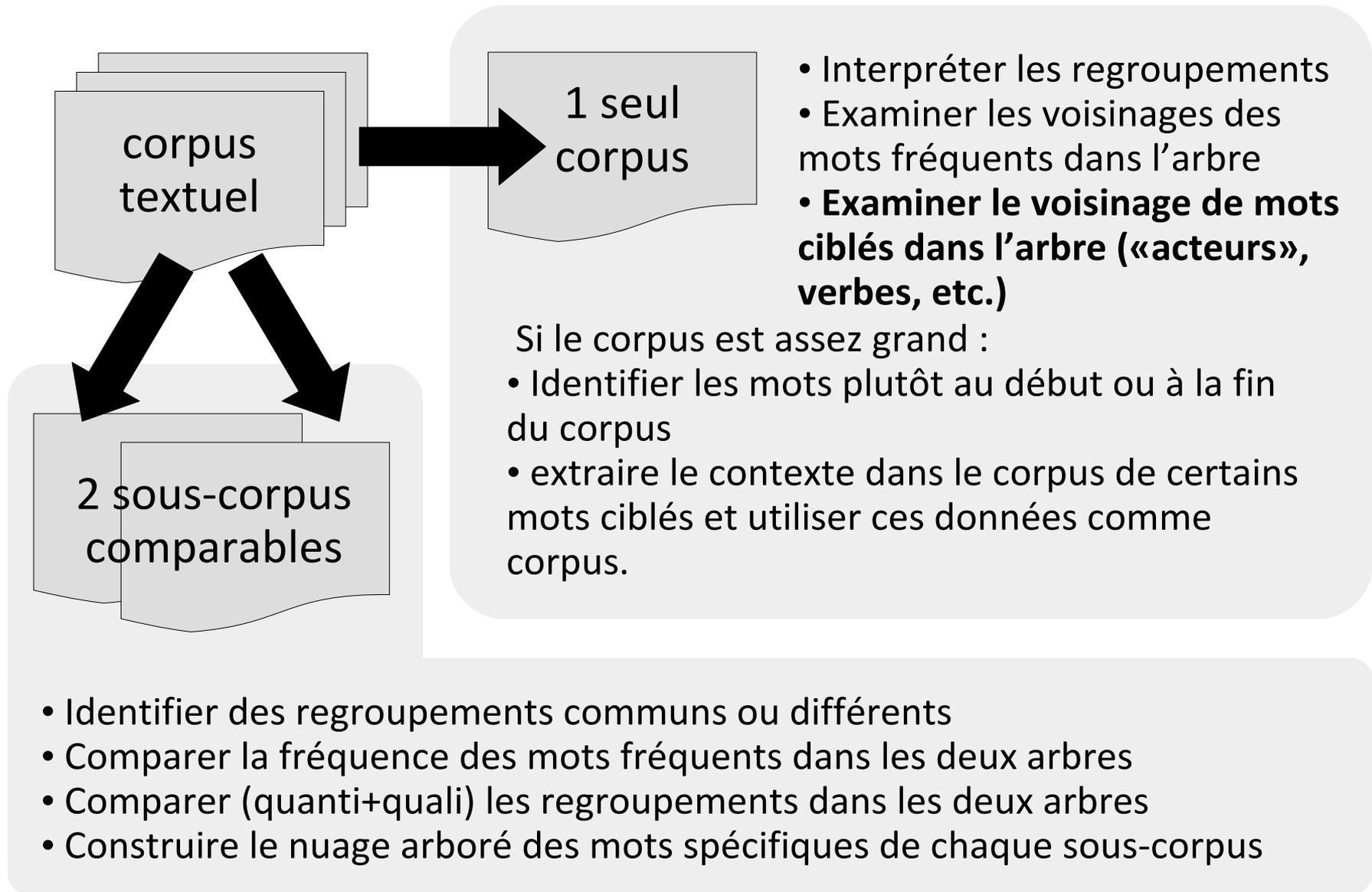
Amstutz & Gambette, JADT 2010



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

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

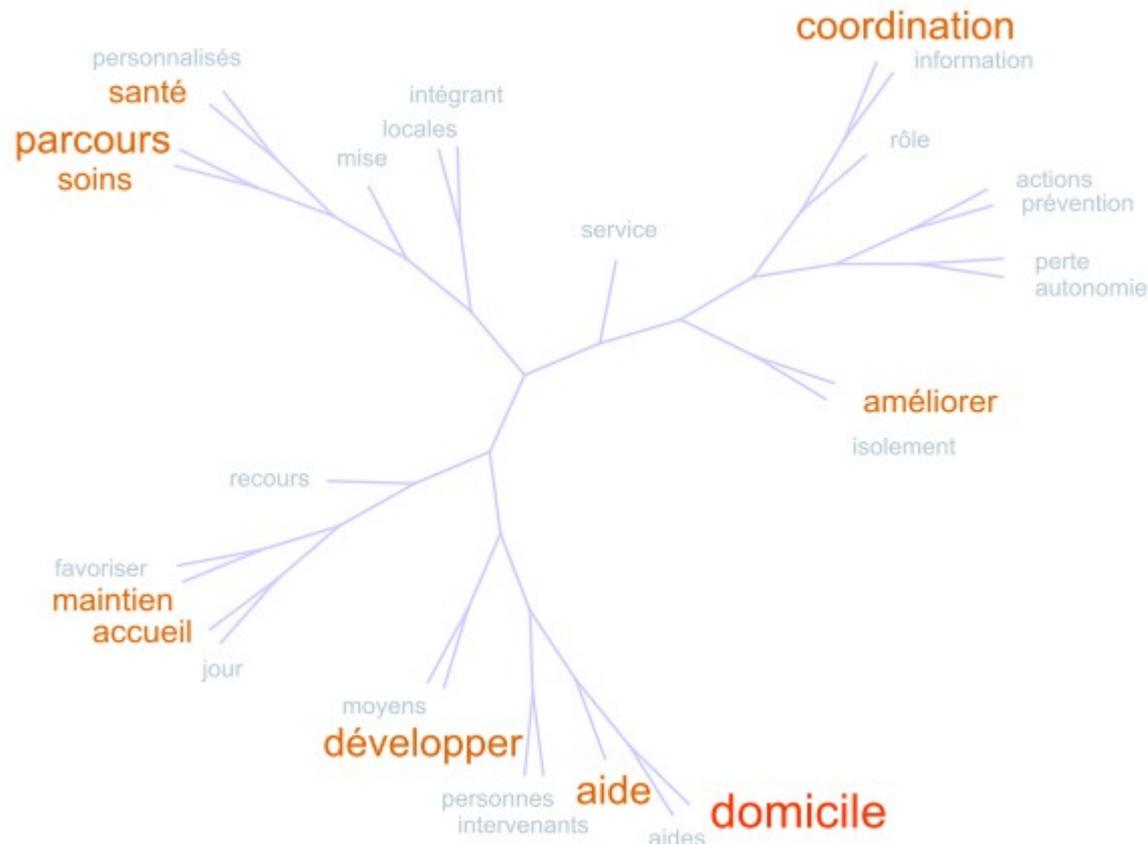
Exploration de corpus avec TreeCloud



Méthode : voisinage des verbes

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

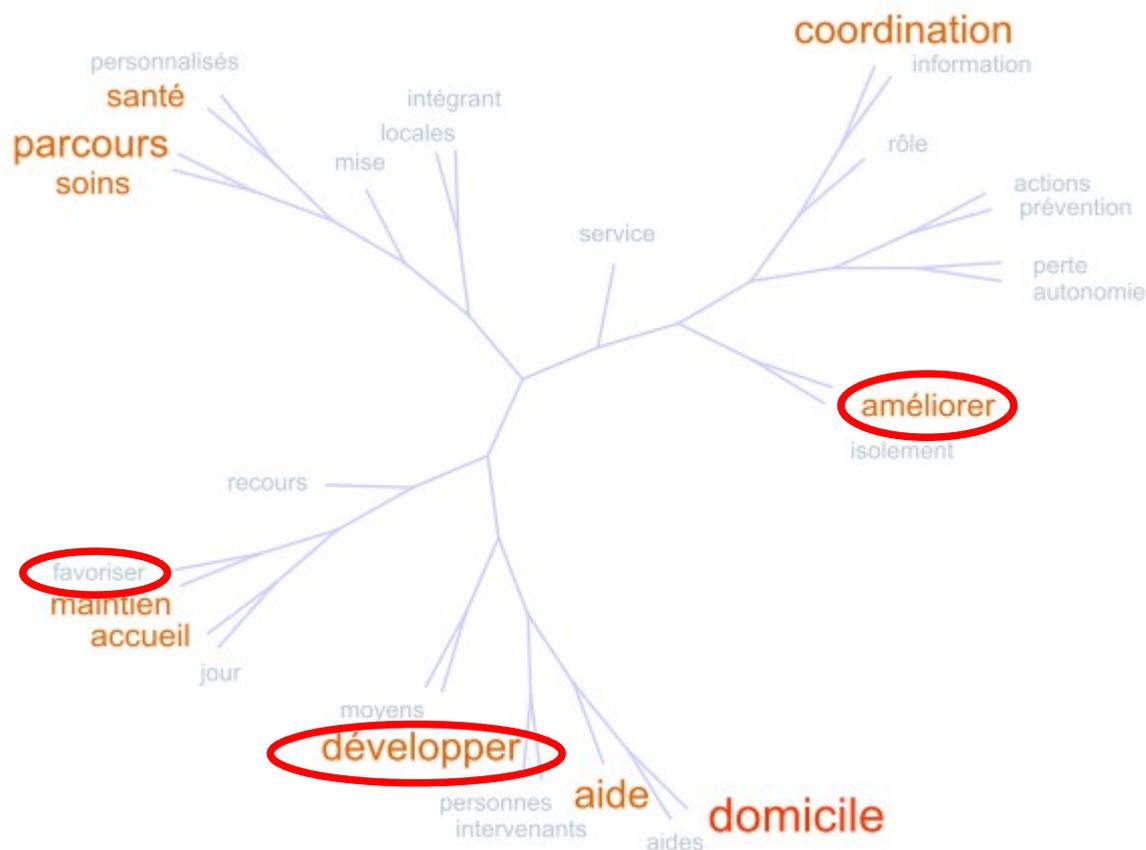
Suggestions d'améliorations :



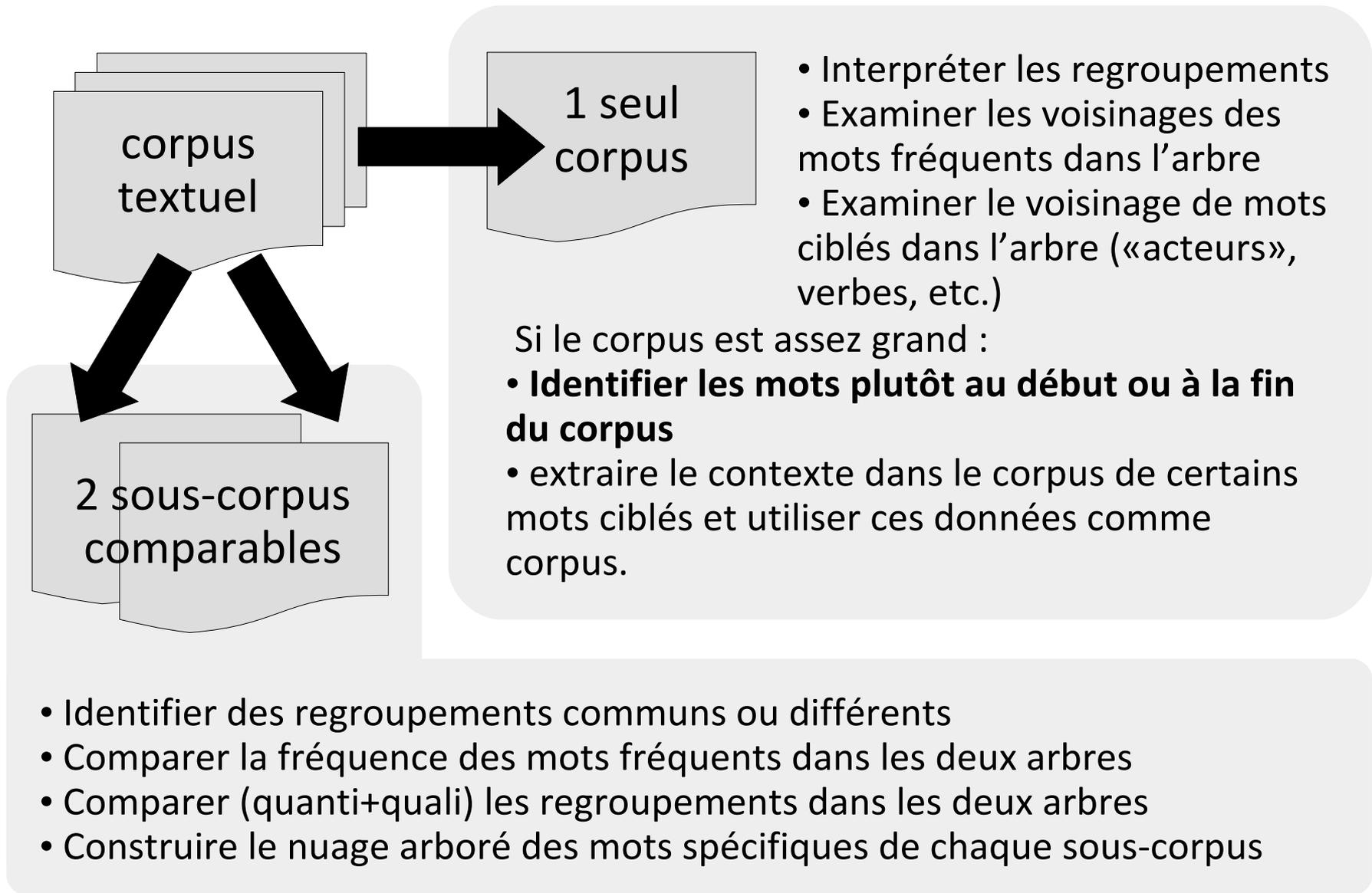
Méthode : voisinage des verbes

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

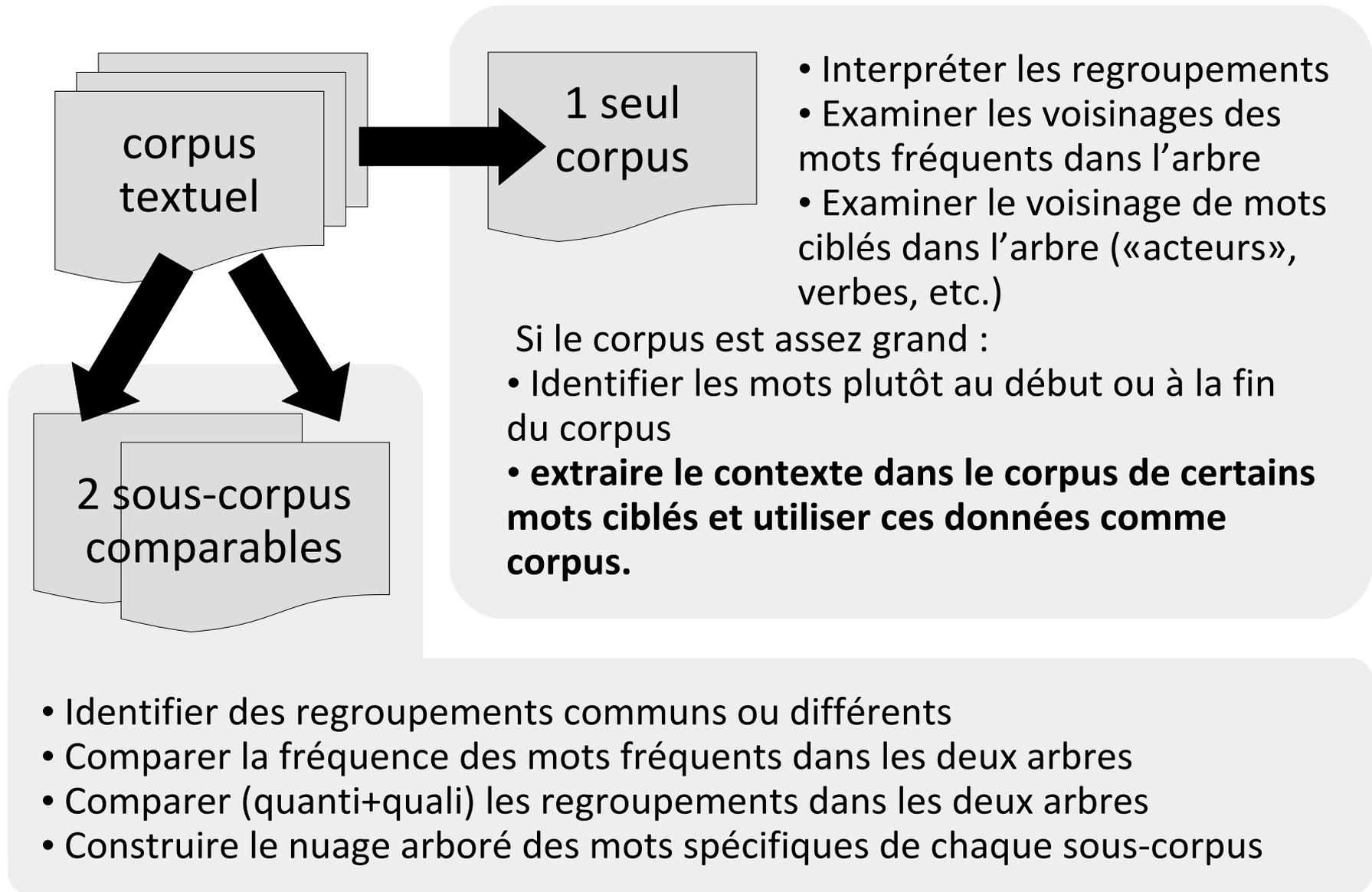
Suggestions d'améliorations :



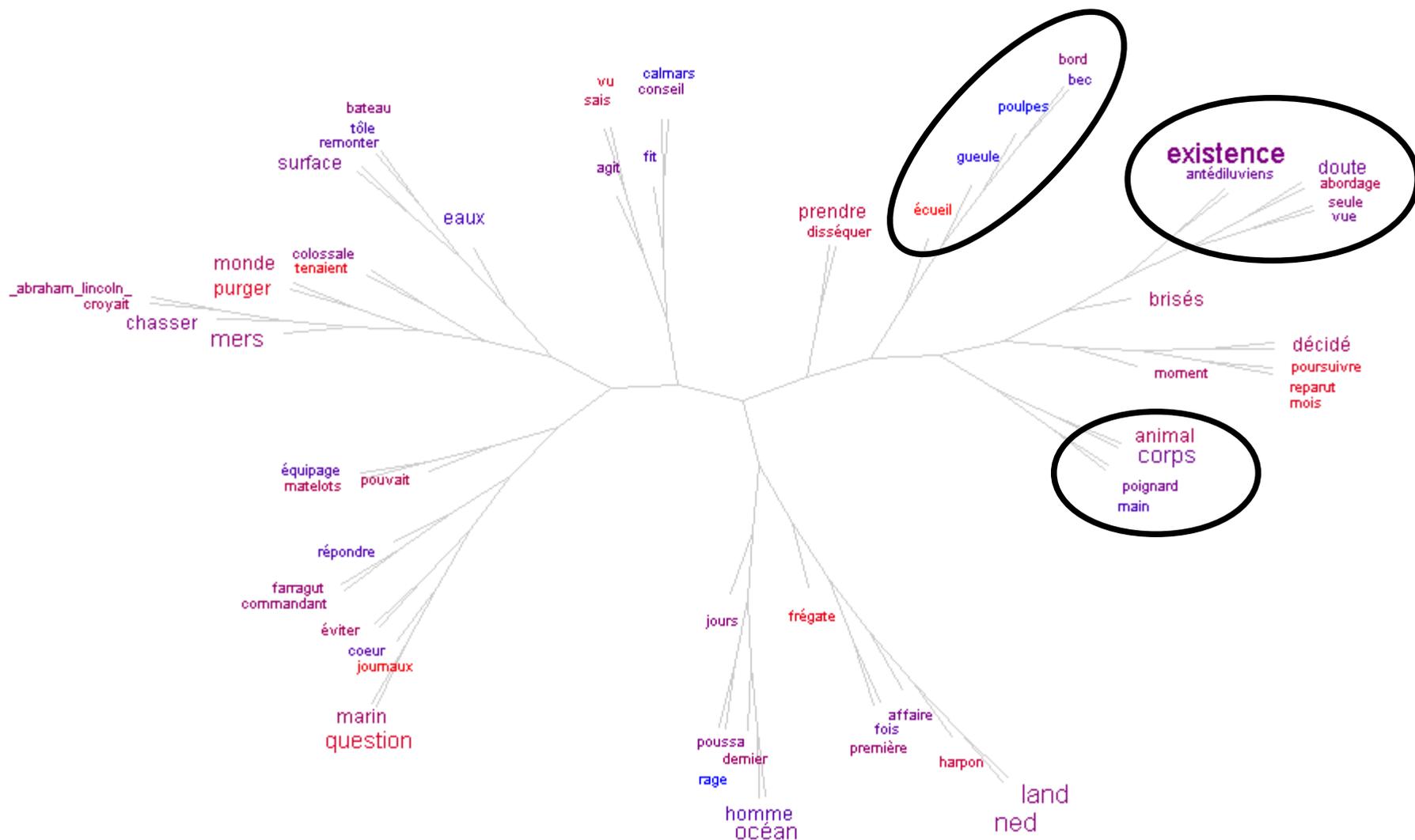
Exploration de corpus avec TreeCloud



Exploration de corpus avec TreeCloud



Nuages arborés des contextes de « monstre(s) »



Nuage arboré des contextes (10 mots à gauche et à droite) de « monstre » ou « monstres » dans *Vingt mille lieues sous les mers* (mots apparaissant au moins 2 fois)

Exploration de corpus avec TreeCloud

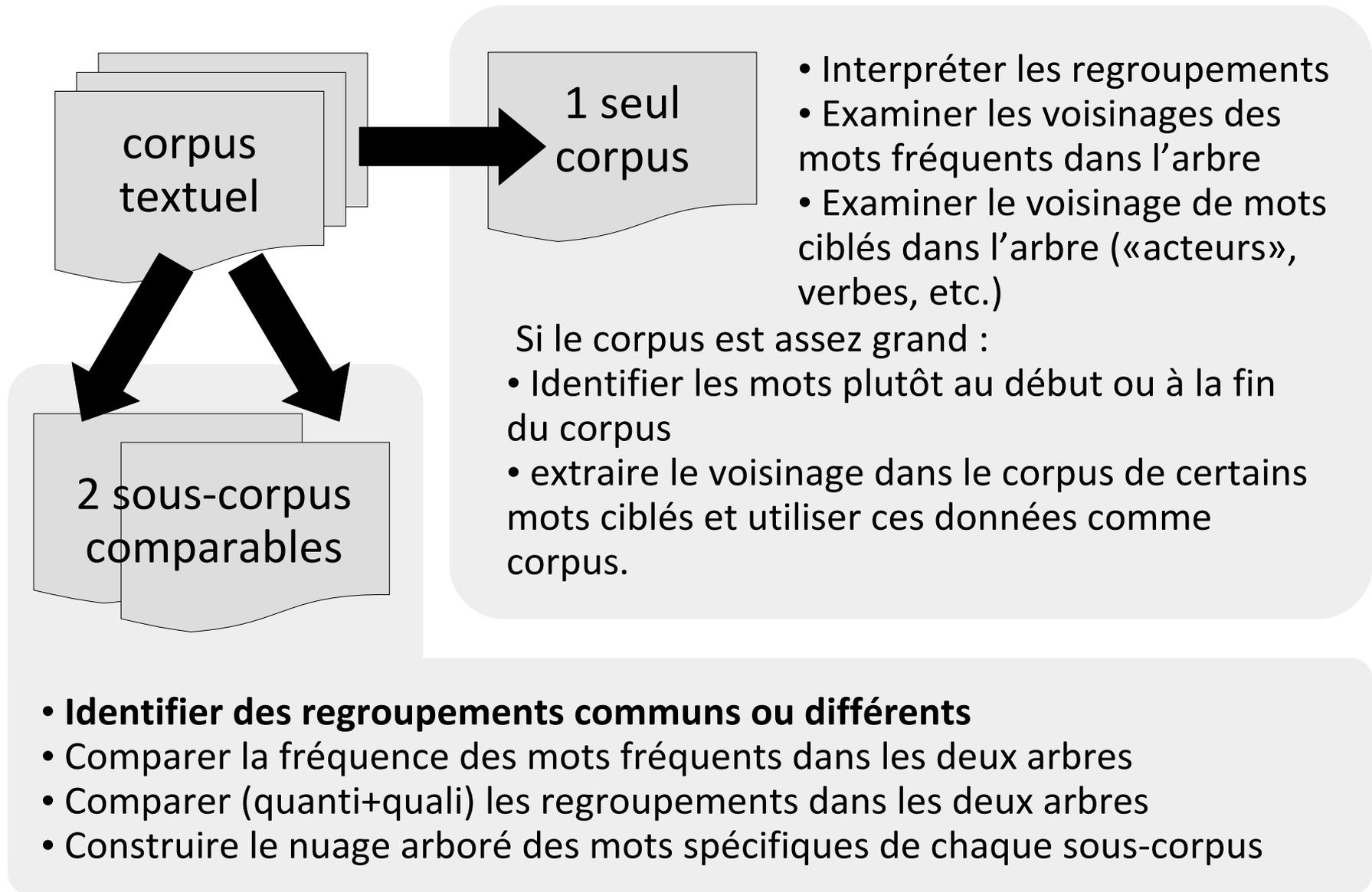


Illustration sur le corpus Mediator

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

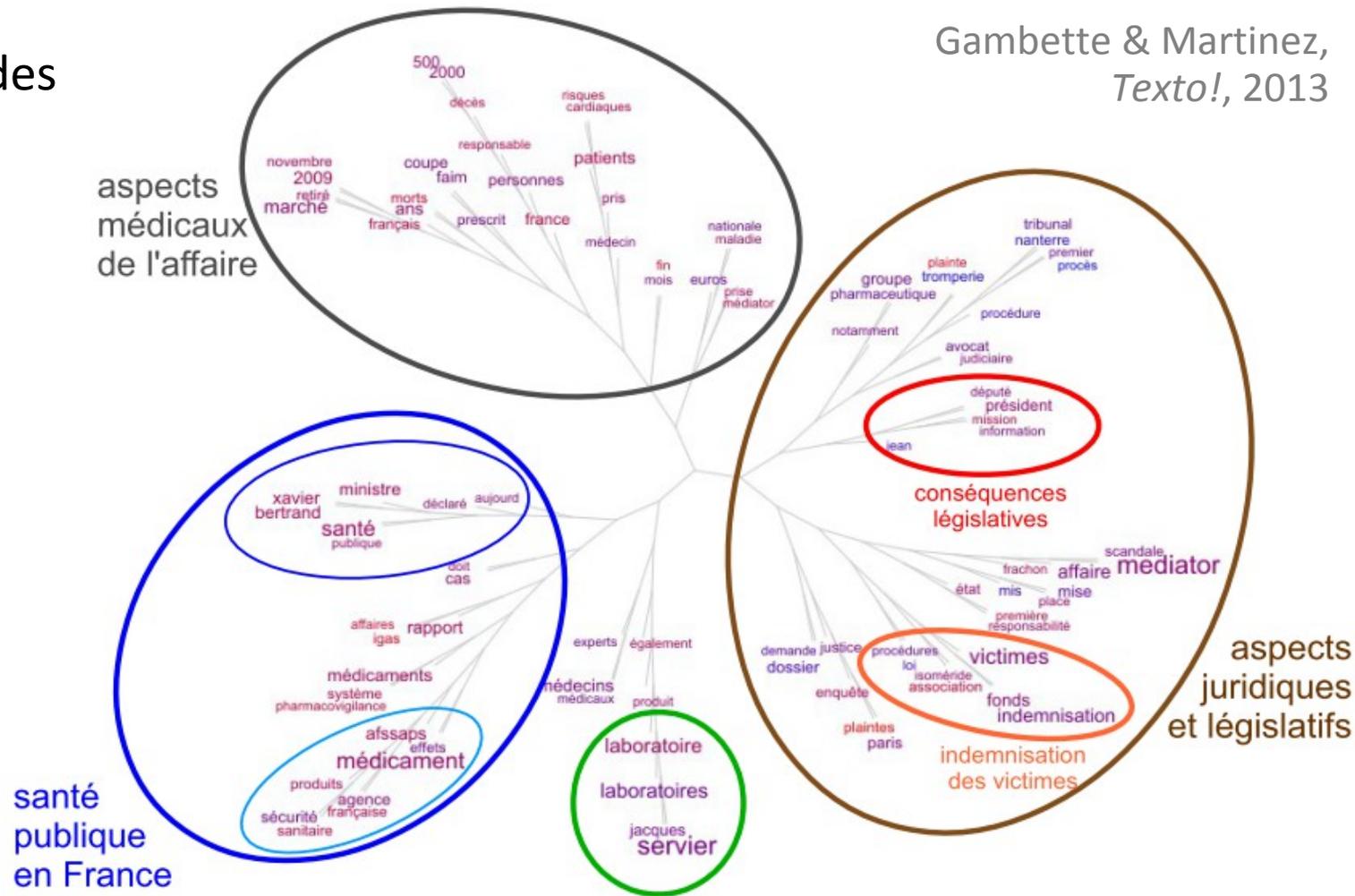
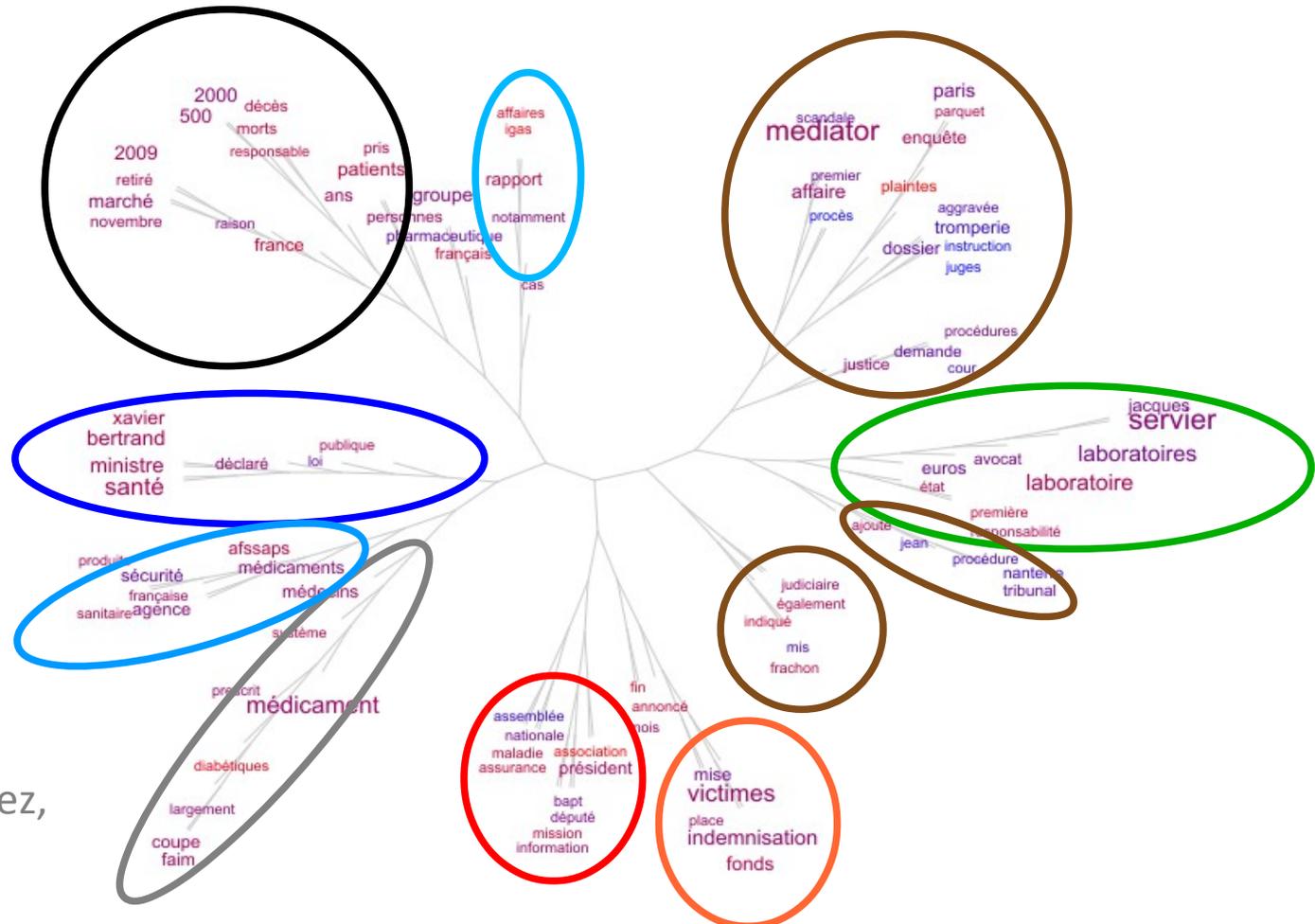


Illustration sur le corpus Mediator

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



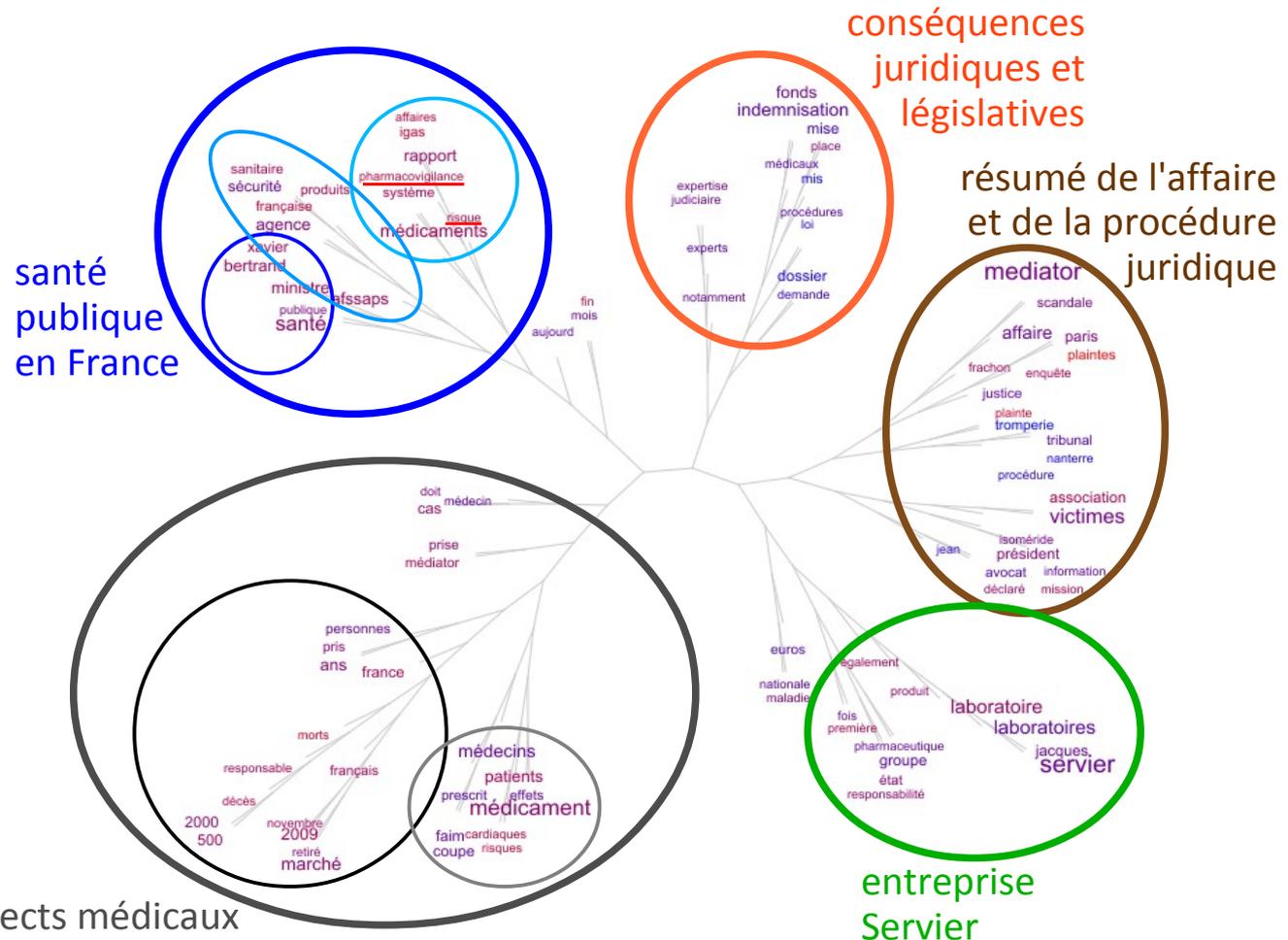
Gambette & Martinez,
Texto!, 2013

Illustration sur le corpus Mediator

Comparer les articles d'agences et articles de journalistes

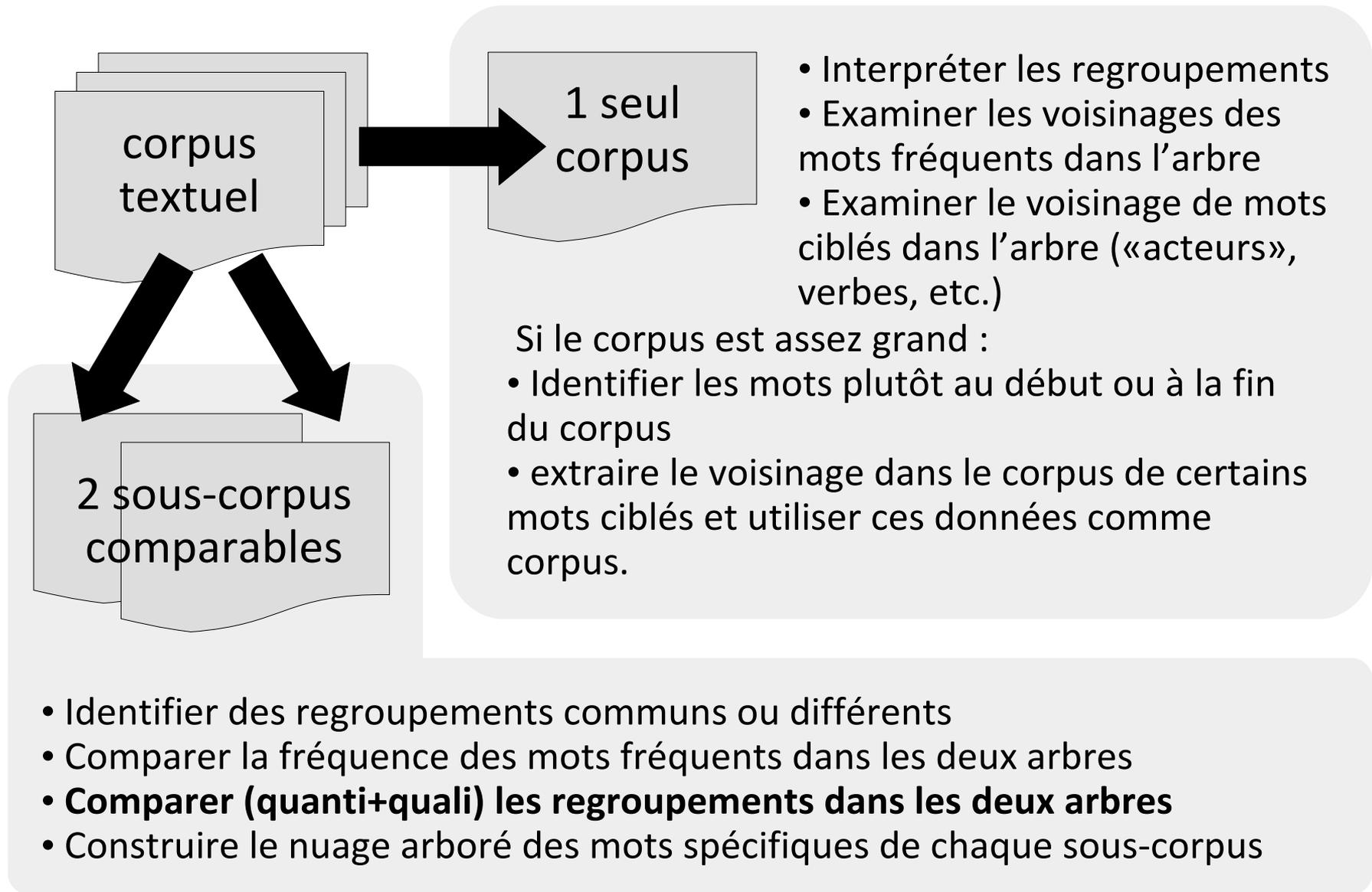
Corpus : 595 articles d'agences contre 1496 articles de journalistes de 2011 évoquant l'affaire du Mediator dans la presse française.

Articles
de journalistes

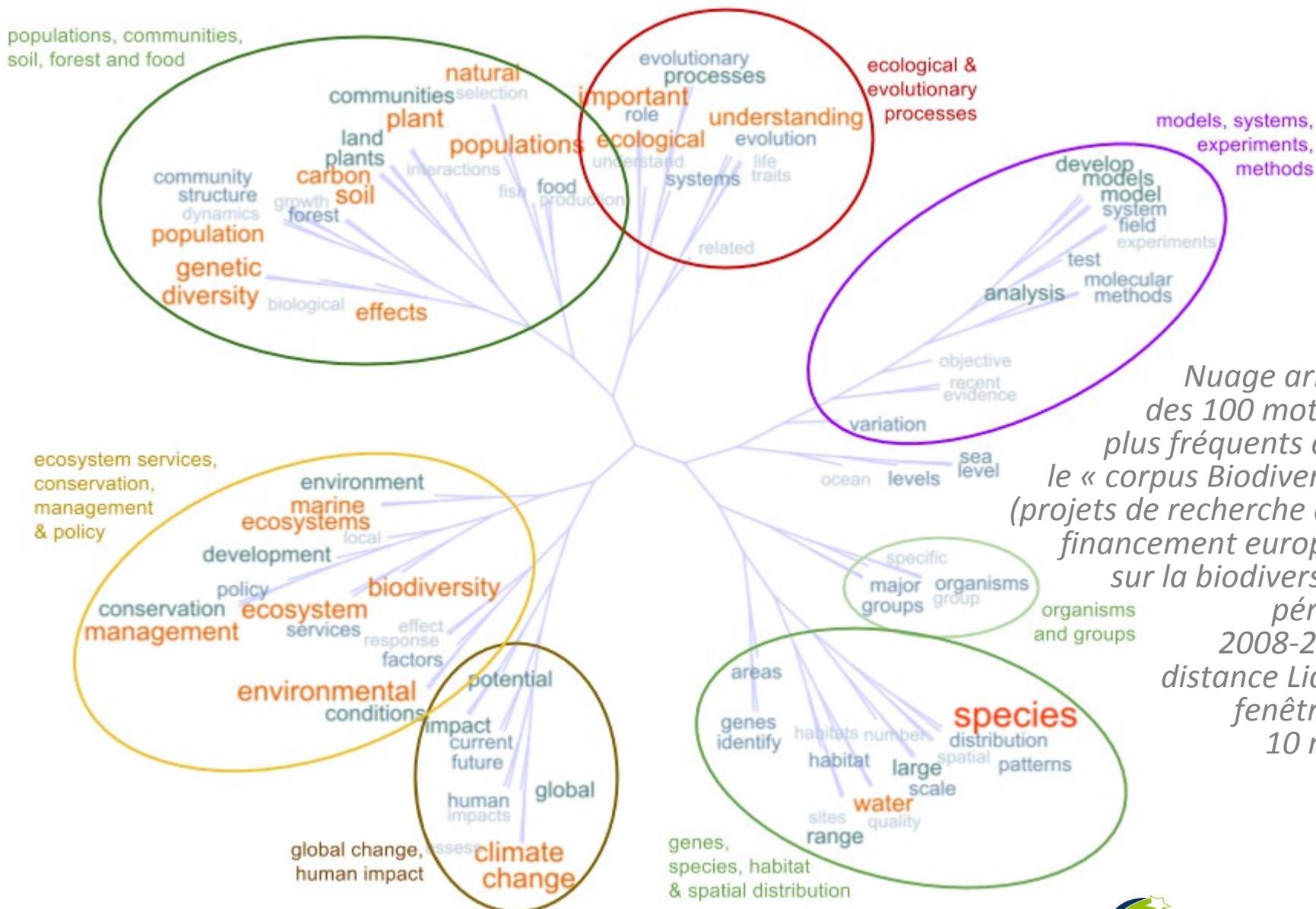


Gambette & Martinez,
Texto!, 2013

Exploration de corpus avec TreeCloud

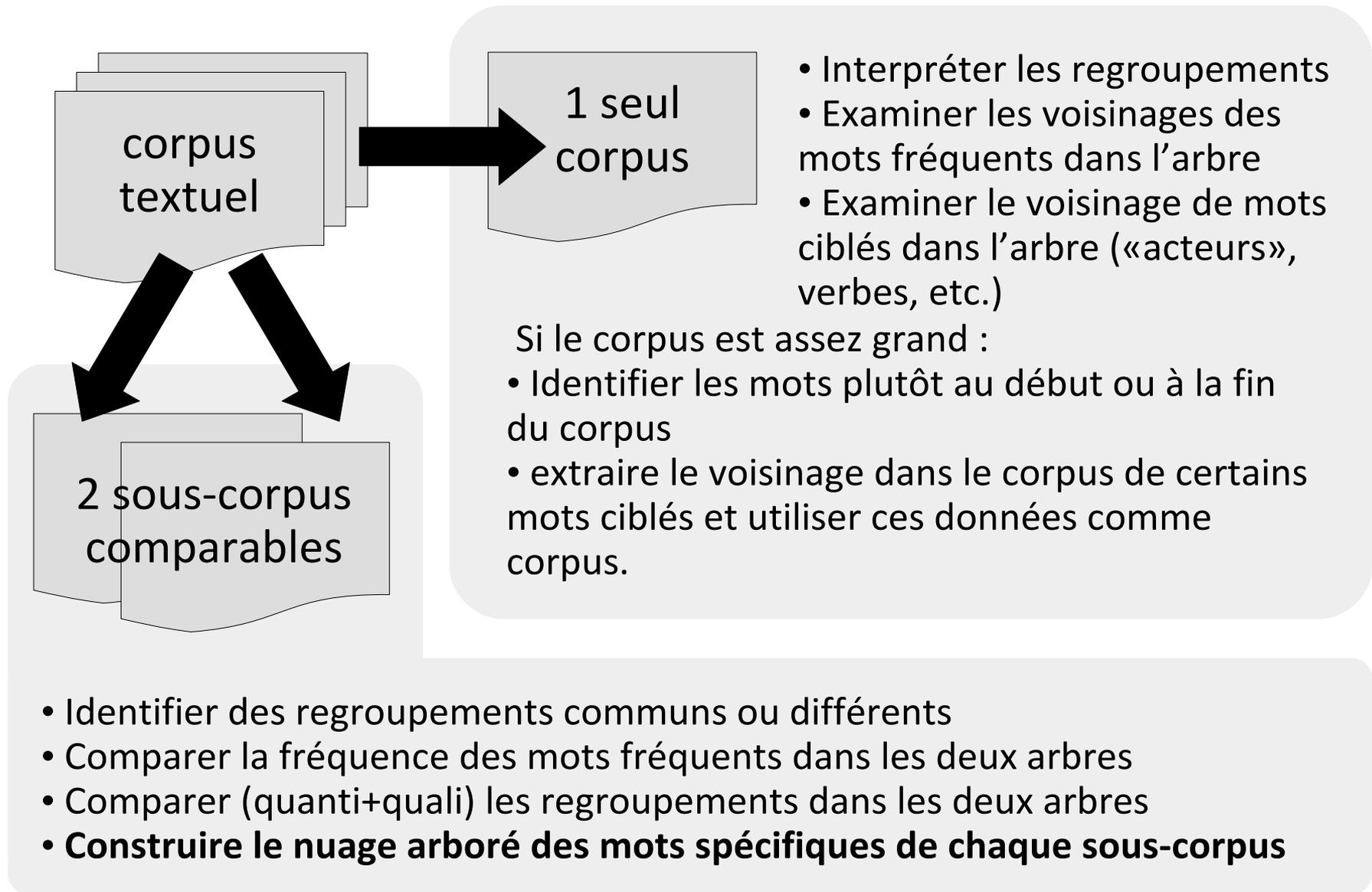


Méthode : comparaison de voisinages dans l'arbre



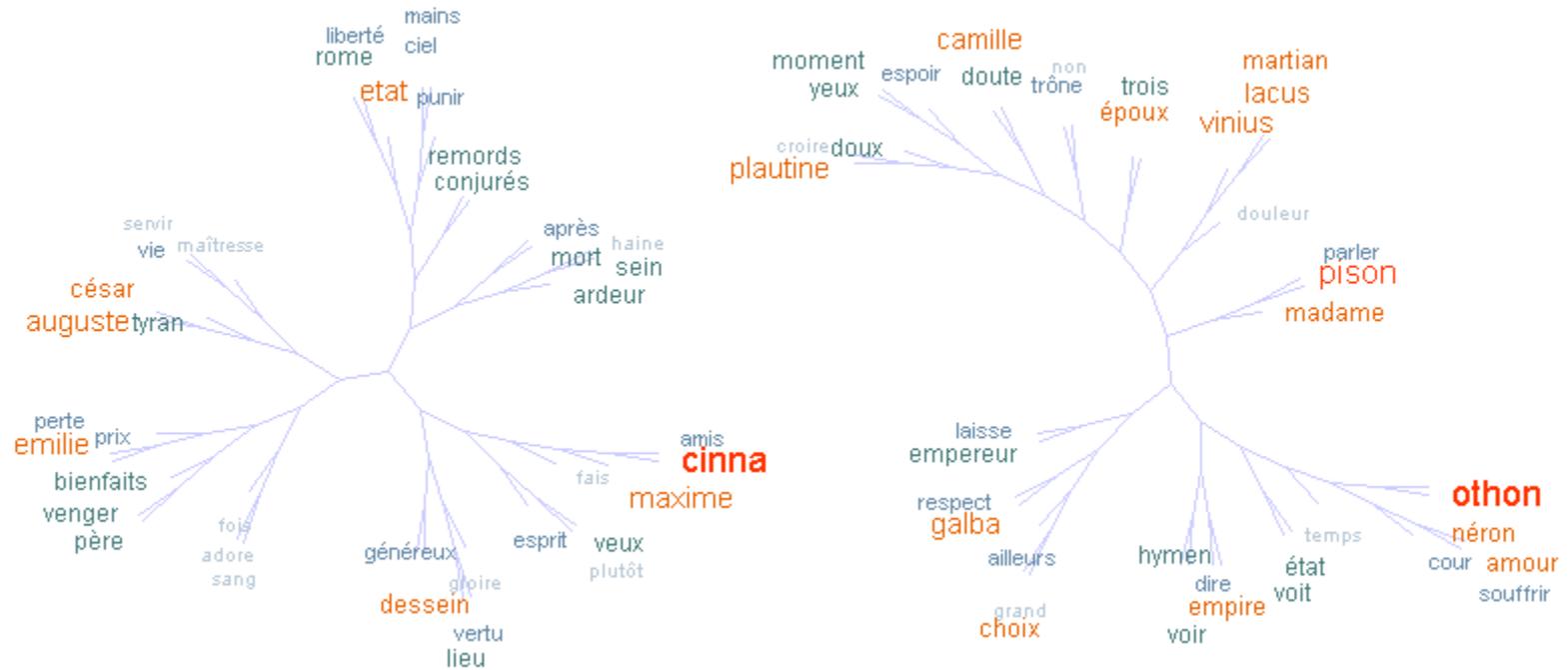
Nuage arboré des 100 mots les plus fréquents dans le « corpus Biodiversa » (projets de recherche avec financement européen sur la biodiversité), période 2008-2011, distance Liddel, fenêtre de 10 mots

Exploration de corpus avec TreeCloud



Méthode : comparaison des spécifiques

Amstutz & Gambette,
JADT 2010



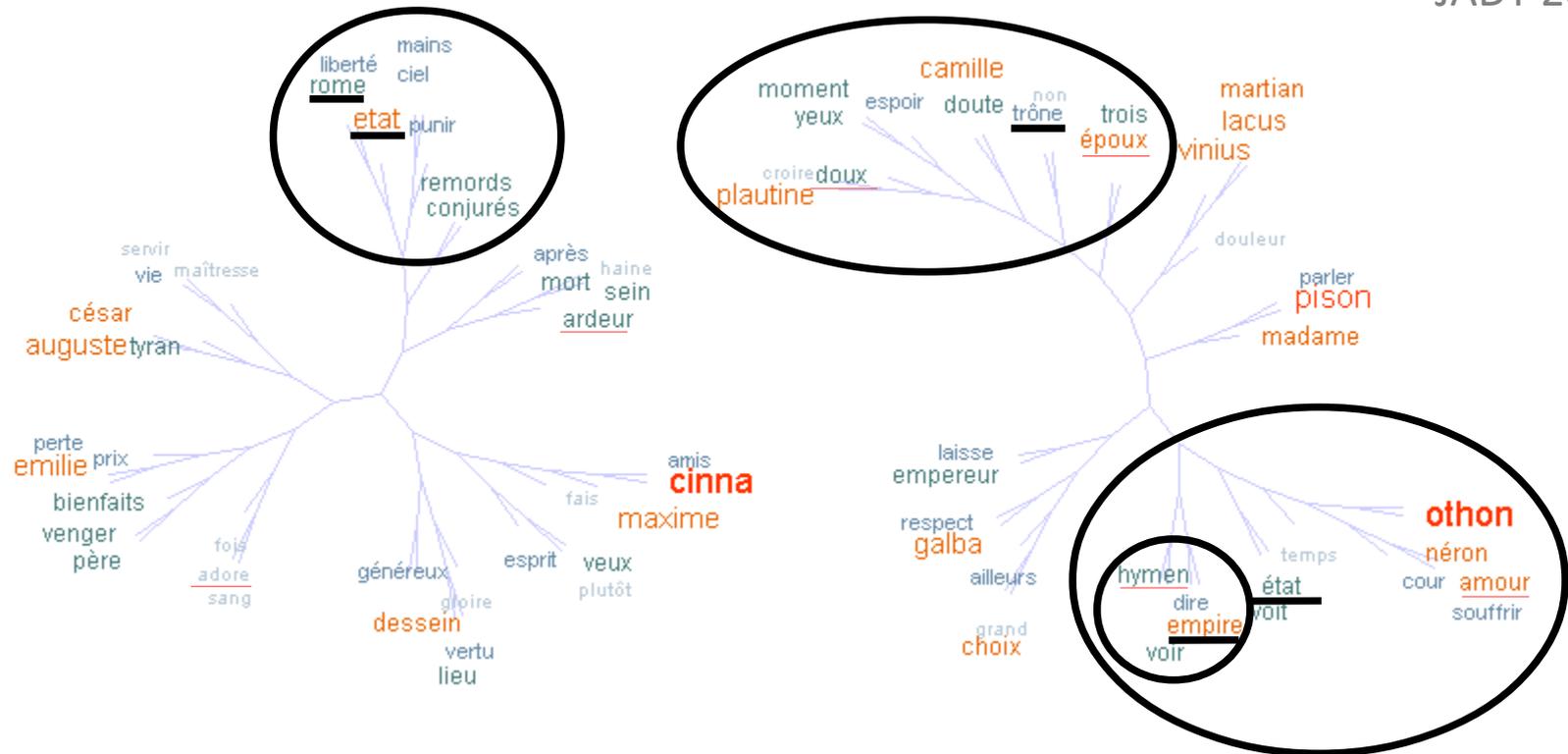
*Nuages arborés des **mots spécifiques** de Cinna et Othon, dimensionnés et colorés d'après leur spécificité calculée dans Lexico3.*



Quels moyens au service de la cause politique ?

Méthode : comparaison des spécifiques

Amstutz & Gambette,
JADT 2010



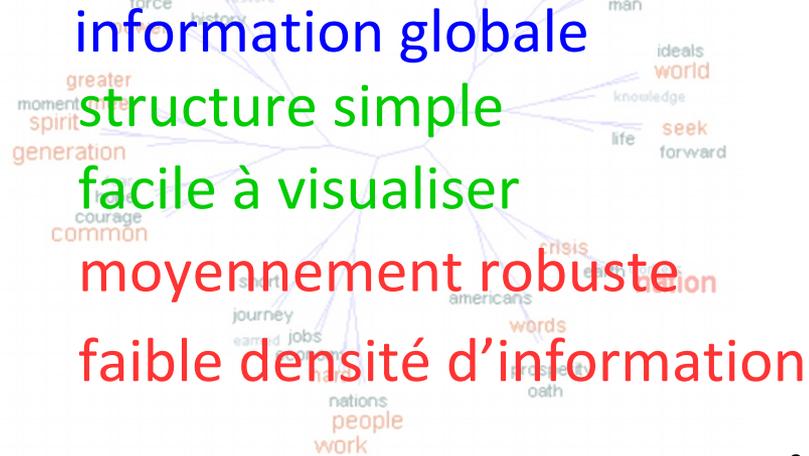
*Nuages arborés des **mots spécifiques** de Cinna et Othon, dimensionnés et colorés d'après leur spécificité calculée dans Lexico3.*



Quels moyens au service de la cause politique ?

Comparaison avec d'autres visualisations

nuage arboré (TreeCloud)



réseau de mots (PhraseNet d'IBM ManyEyes)

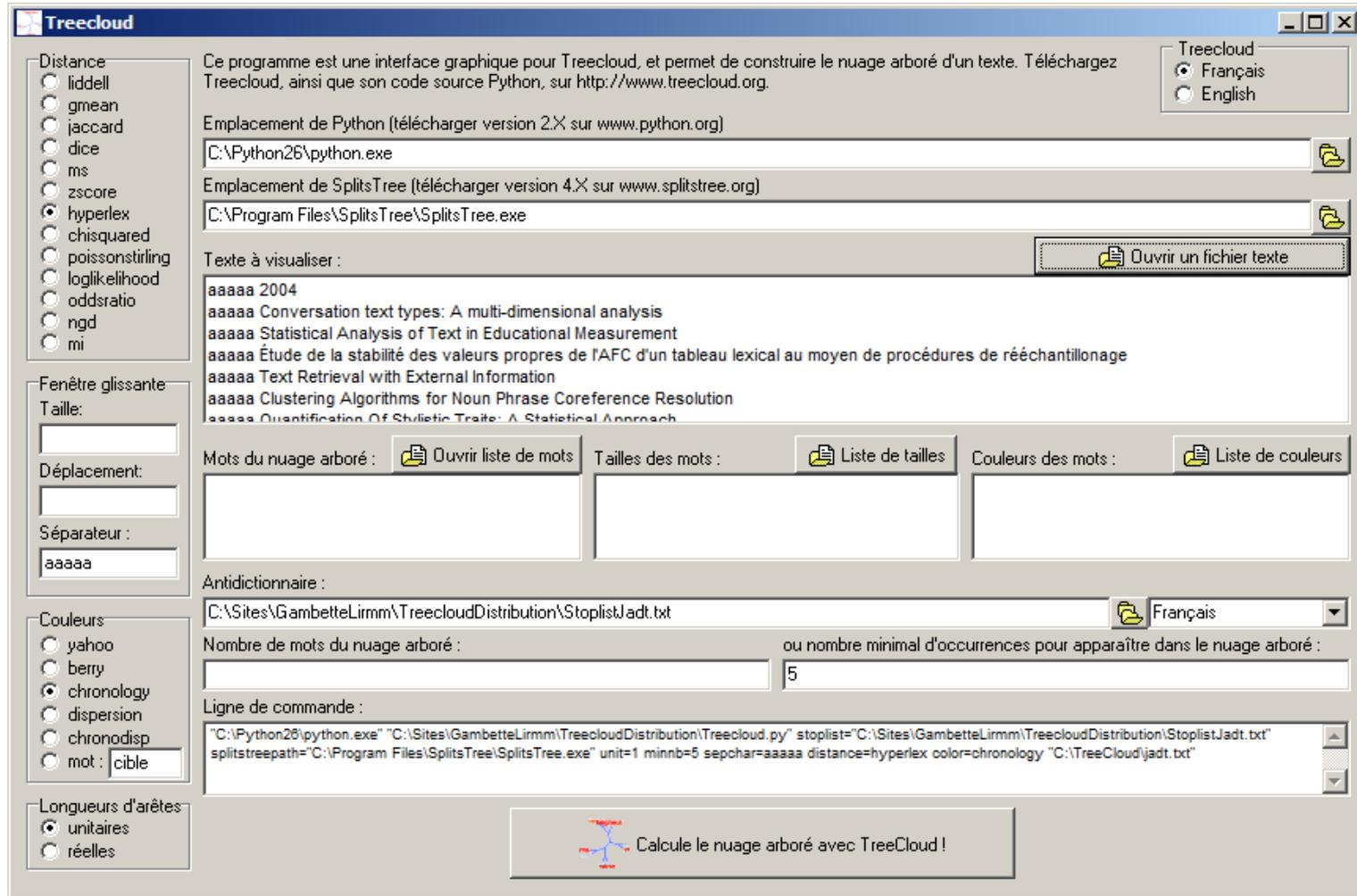


projection des mots (Astartex)



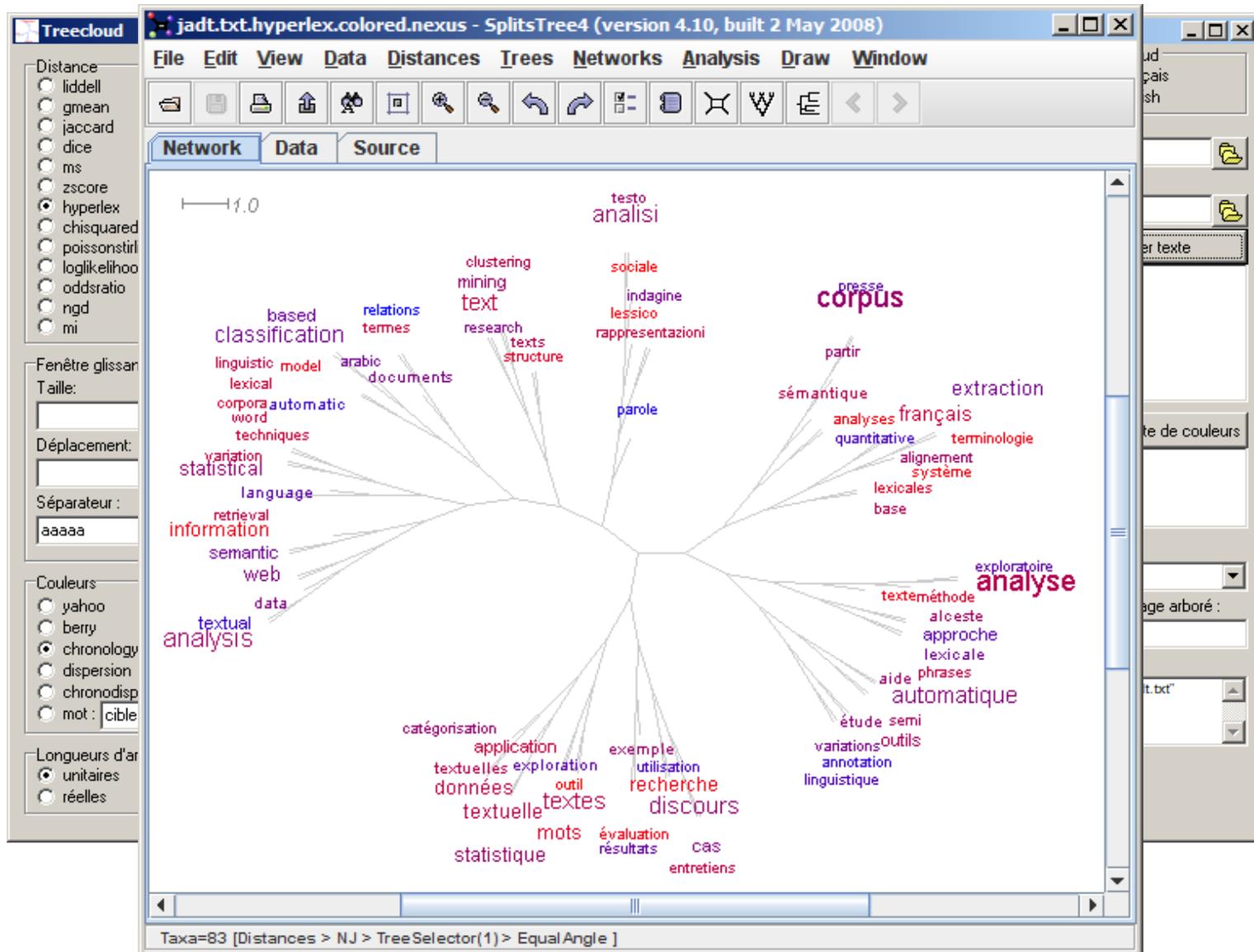
Implémentations

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



Implémentations

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



Interface web



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

www.treecloud.org

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

Create your own tree cloud online

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

Créez vos propres nuages arborés

Documents :



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

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



Créer! Téléchargements Galerie A propos FAQ

Créez vos propres nuages arborés !

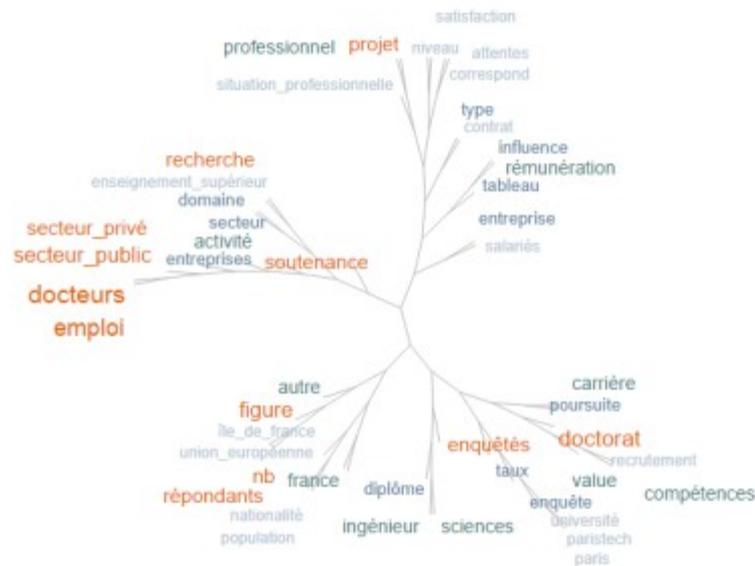
Collez votre texte dans le cadre ci-dessous, puis cliquez sur *Envoyer* ! Attention, l'utilisateur suivant verra votre texte quand il se connectera au site, si vous ne voulez pas faire apparaître vos textes, installez plutôt [TreeCloud](#) sur votre machine.

Texte :

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

Envoyer

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



Interface web

TreeCloud.org
Nuage Arboré

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

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

Create your own tree cloud online

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

Créez vos propres nuages arborés

Documents :

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

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

© 2007-2010 - Je

www.treecloud.org

TreeCloud.org
Nuage Arboré

Créer! Téléchargements Galerie A propos FAQ

Créez vos propres nuages arborés !

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

Texte :

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

Envoyer

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

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

Références (*treecloud.org*)

Philippe Gambette, Jean Véronis (2009)

Visualising a Text with a Tree Cloud, *IFCS'09, Studies in Classification, Data Analysis, and Knowledge Organization* 40, p. 561-570

<http://www.slideshare.net/PhilippeGambette/visualising-a-text-with-a-tree-cloud>

Delphine Amstutz & Philippe Gambette (2010)

Utilisation de la visualisation en nuage arboré pour l'analyse littéraire, JADT'10 (Proceedings of the 10th International Conference on statistical analysis of textual data),
Statistical Analysis of Textual Data, p. 227-238

<http://www.slideshare.net/PhilippeGambette/utilisation-de-la-visualisation-en-nuage-arbor-pour-lanalyse-littraire>

Philippe Gambette, Nuria Gala & Alexis Nasr (2012)

Longueur de branches et arbres de mots, *Corpus* 11:129-146

<http://www.slideshare.net/PhilippeGambette/longueur-de-branches-et-arbres-de-mots>

William Martinez & Philippe Gambette (2013)

L'affaire du Médiateur au prisme de la textométrie, *Texto!* XVIII(4)

<http://www.revue-texto.net/index.php?id=3318>

Philippe Gambette, Hilde Eggermont & Xavier Le Roux (2014)

Temporal and geographical trends in the type of biodiversity research funded on a competitive basis in European countries, *rapport BiodivERsa*

<http://www.biodiversa.org/700/download>

Co-auteurs des travaux en cours :

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