

24/06/2019 – Hôpital Necker (Paris)

Visualisation de réponses à des questions ouvertes avec TreeCloud

Philippe Gambette

LIGM
Université Paris-Est
Marne-la-Vallée



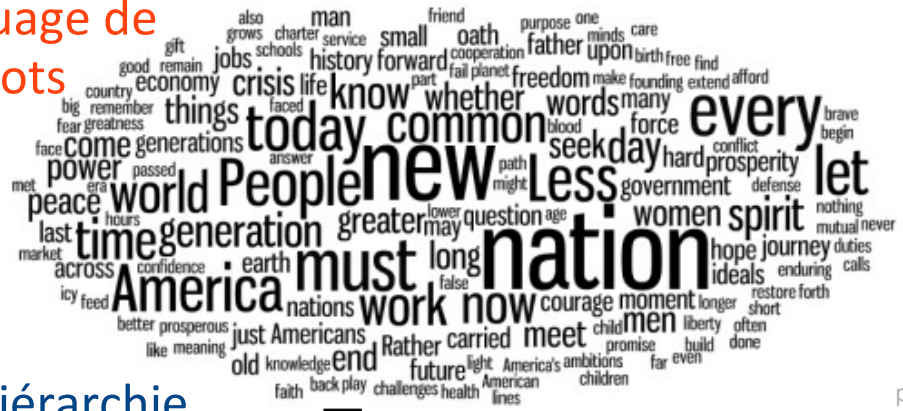
Plan

- Concept des nuages arborés
- Construction des nuages arborés
- Méthodologie et cas d'usage
- Outils disponibles
- Références

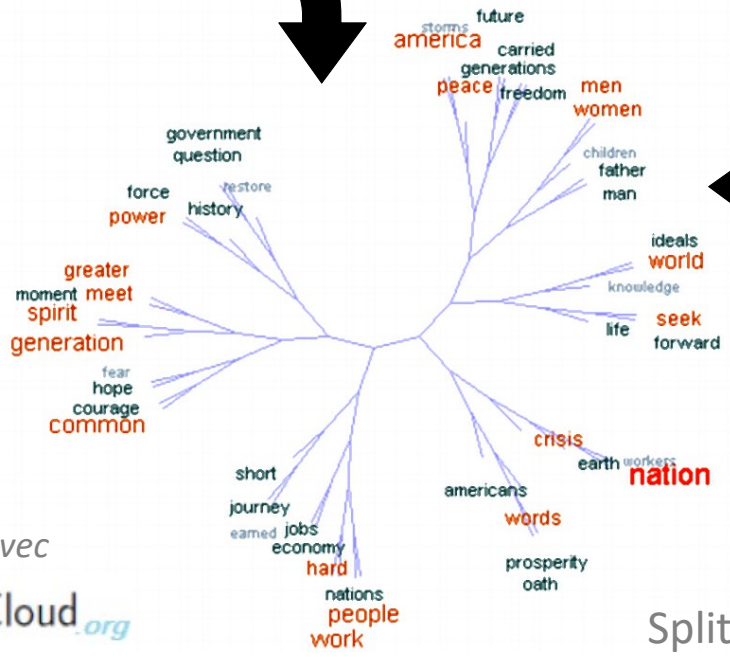
Concept des nuages arborés

Le « nuage arboré », une information double

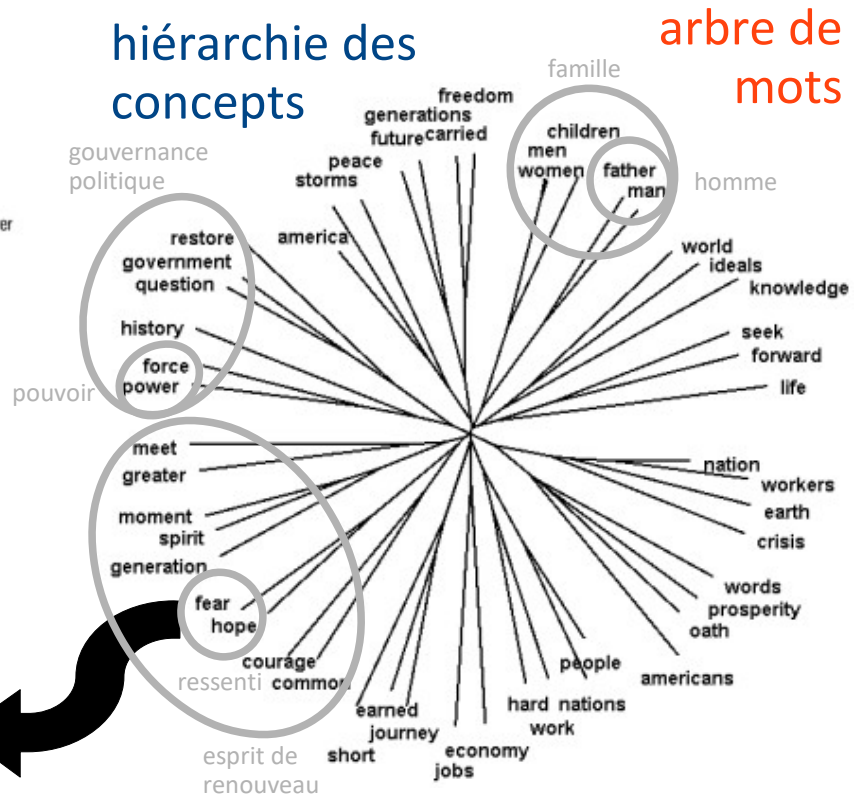
nuage de mots



hiérarchie des mots



hiérarchie des concepts



arbre de mots

Discours inaugural de Barack Obama

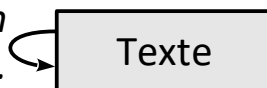
construit avec
TreeCloud.org
SplitsTree4

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

Construction des nuages arborés

Processus de construction

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



Suppression des mots vides

Sélection des mots

Recherche des cooccurrences

Calcul des distances entre mots

Construction de l'arbre

Tailles des mots

Couleurs des mots

Dessin du nuage arboré

“fenêtre glissante” :
on considère que deux mots
sont cooccurents s'ils font
partie d'un même fenêtre
glissante de 10 mots, par
exemple, c'est-à-dire s'ils sont
séparés par au plus 8 mots
(largeur de la fenêtre glissante
paramétrable)

**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 glissante 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*

Méthodologie et cas d'usage

Le « nuage arboré », pour quoi faire ?

- Outil exploratoire : susciter des hypothèses
 - analyse de toutes les réponses à une question
 - analyse des réponses regroupées par profil : un nuage arboré par profil
- Outil de communication :
 - résumer visuellement les résultats du questionnaire
 - soutenir une analyse

Le « nuage arboré », pour quoi faire ?

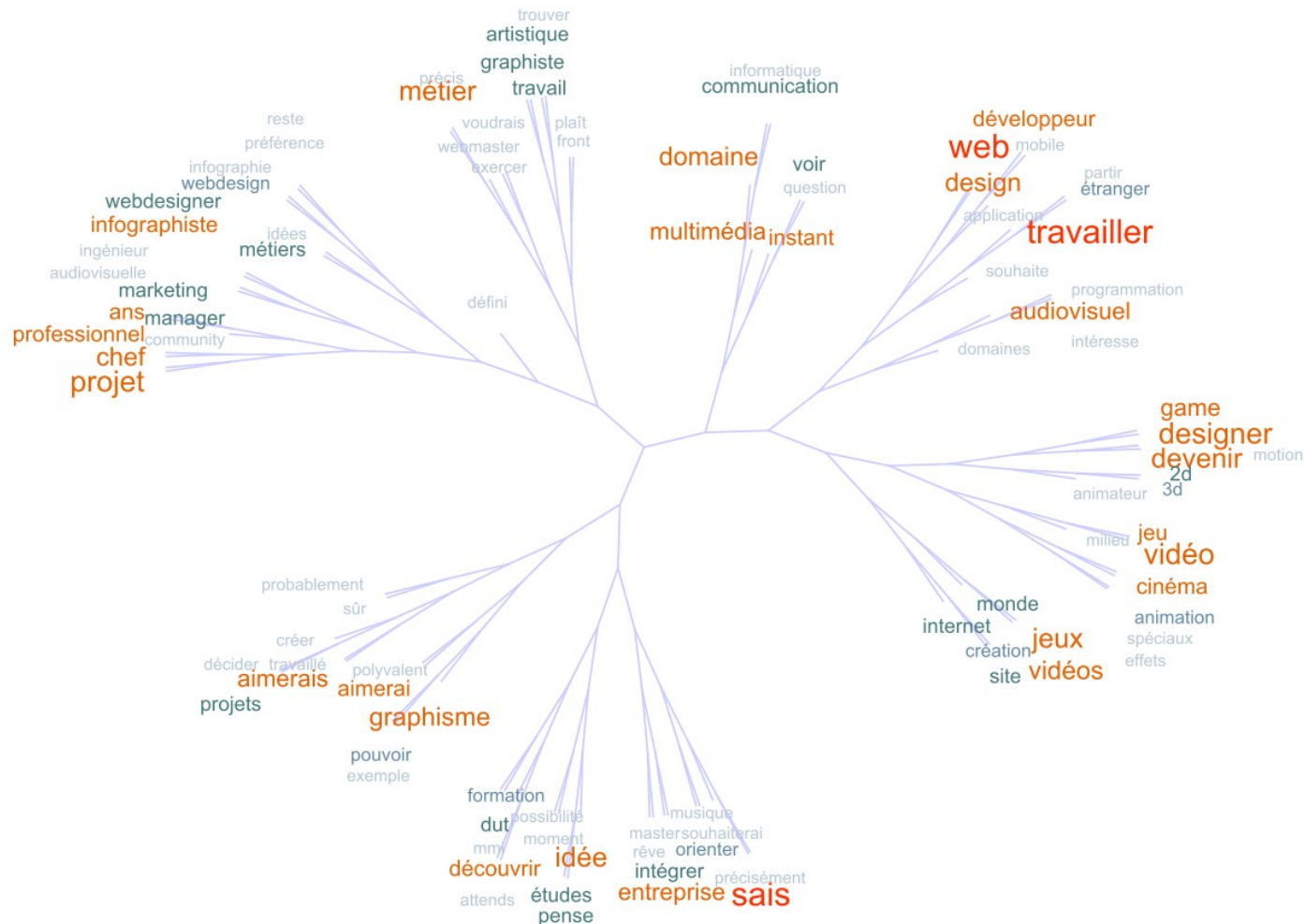
- Outil **exploratoire** : susciter des hypothèses
 - analyse de toutes les réponses à une question
 - analyse des réponses regroupées par profil : un nuage arboré par profil
- Outil de **communication** :
 - résumer visuellement les résultats du questionnaire
 - soutenir une analyse

Démarche fiable et honnête : nécessité de **retour au texte** !

Méthode : interpréter les regroupements

Dessiner des « patates »

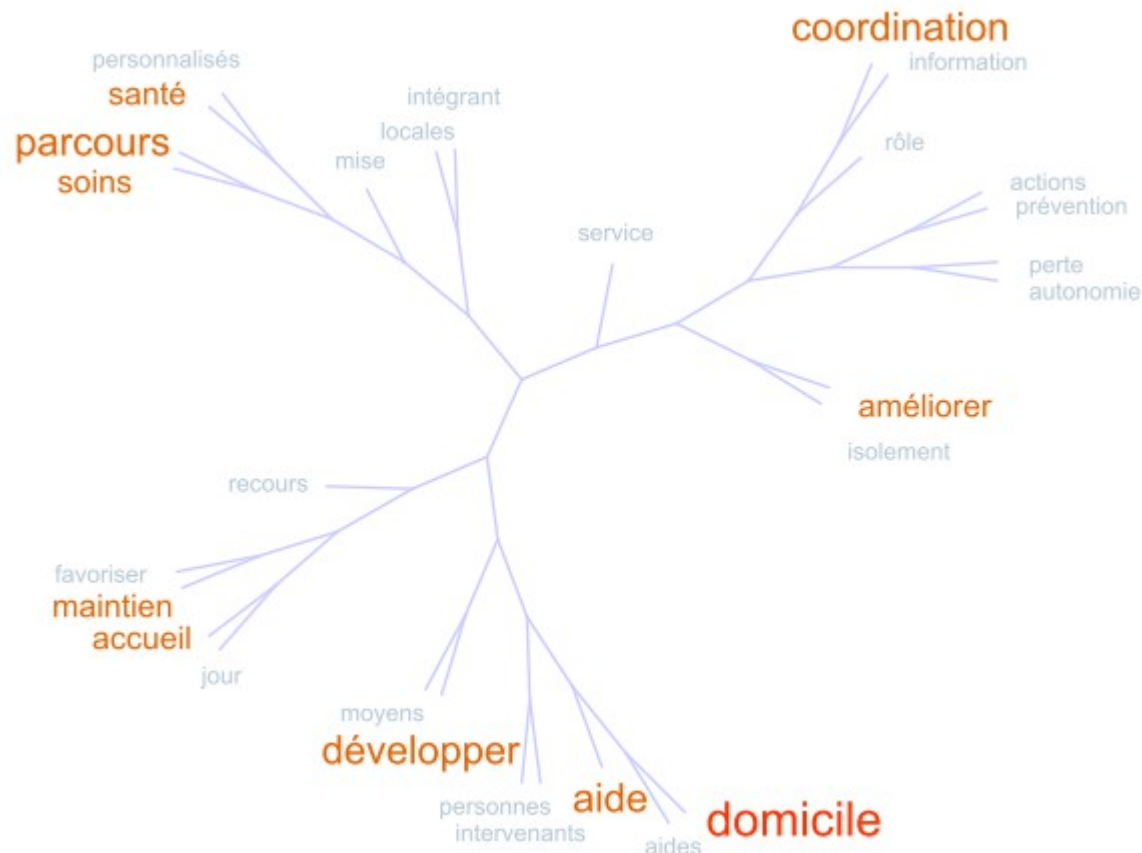
Corpus : 280 réponses d'étudiantes et étudiants en DUT SRC/MMI sur leur projet professionnel en début de première année de DUT.



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

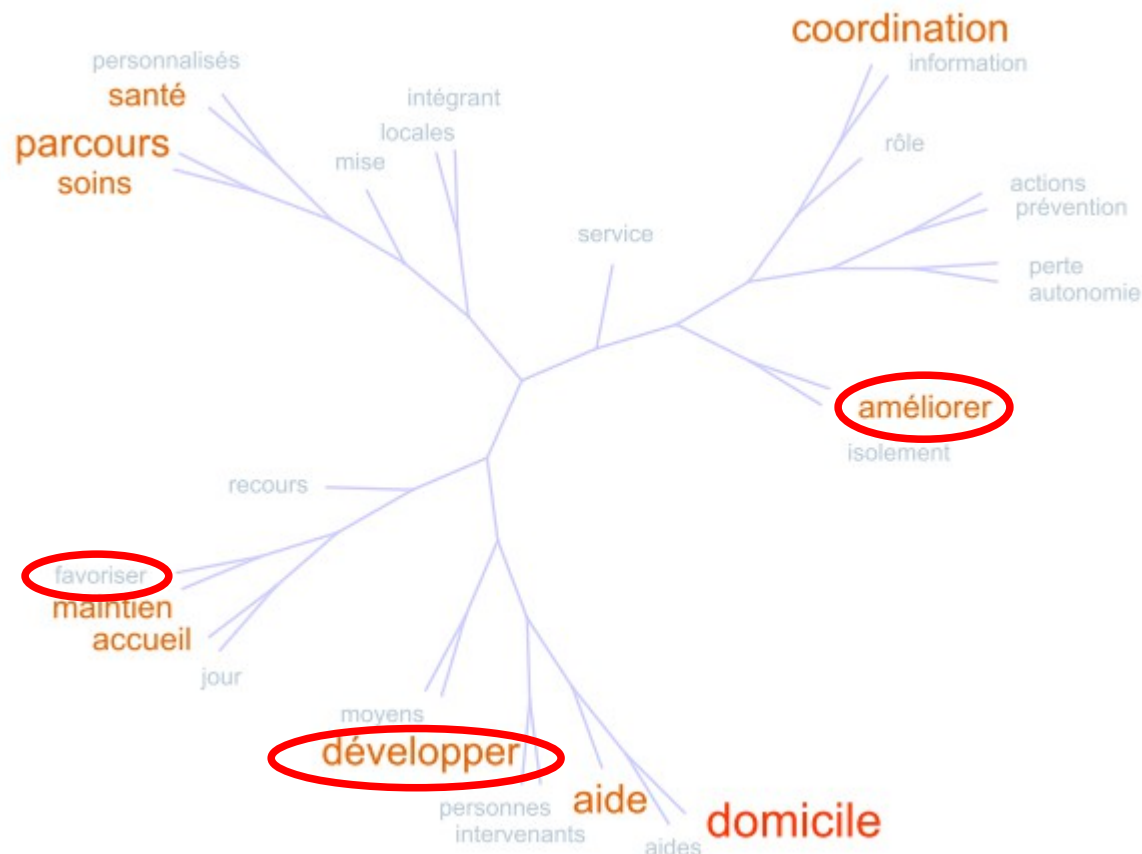
Suggestions d'améliorations :



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 :

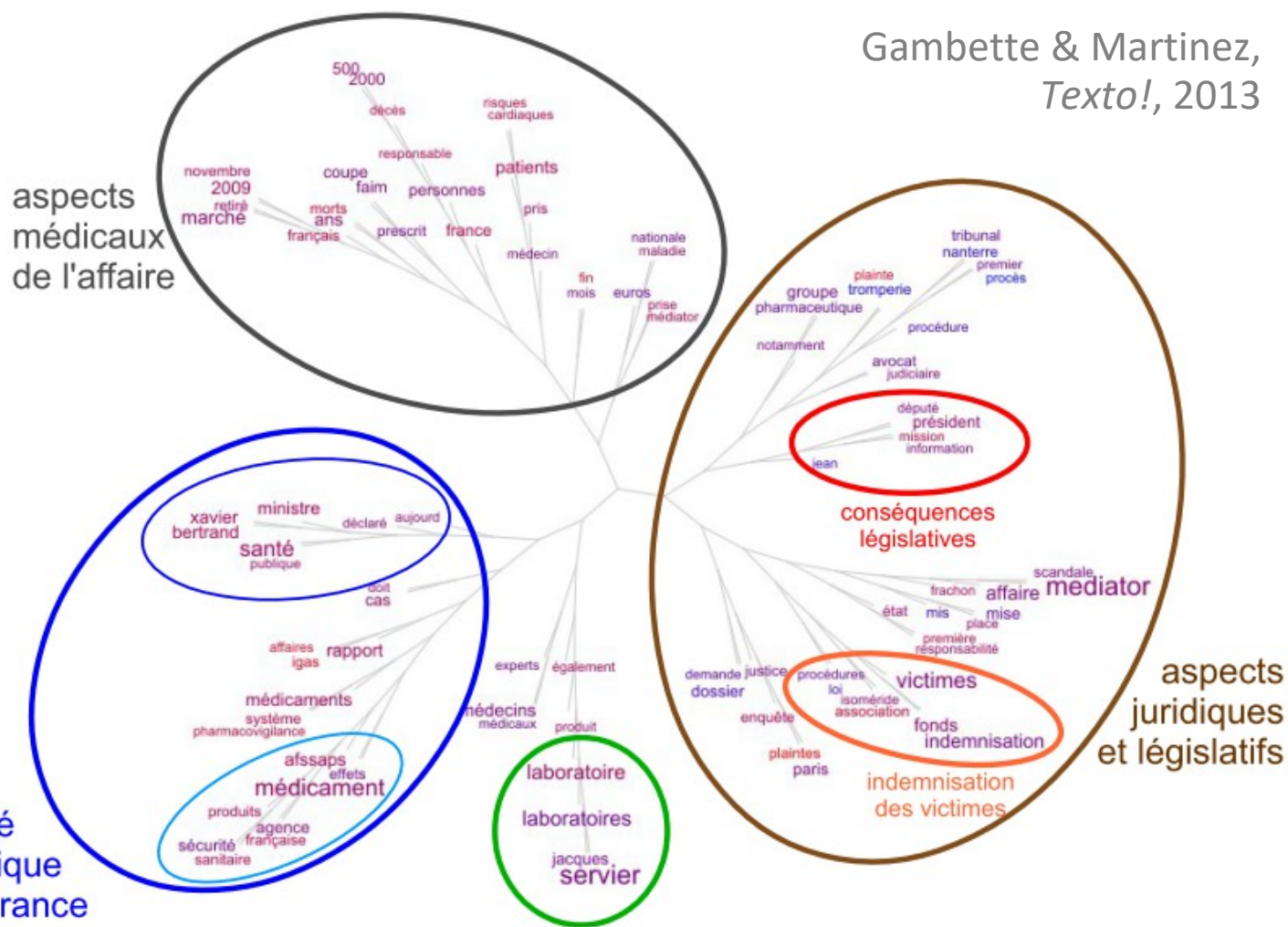


Comparer des sous-corpus

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

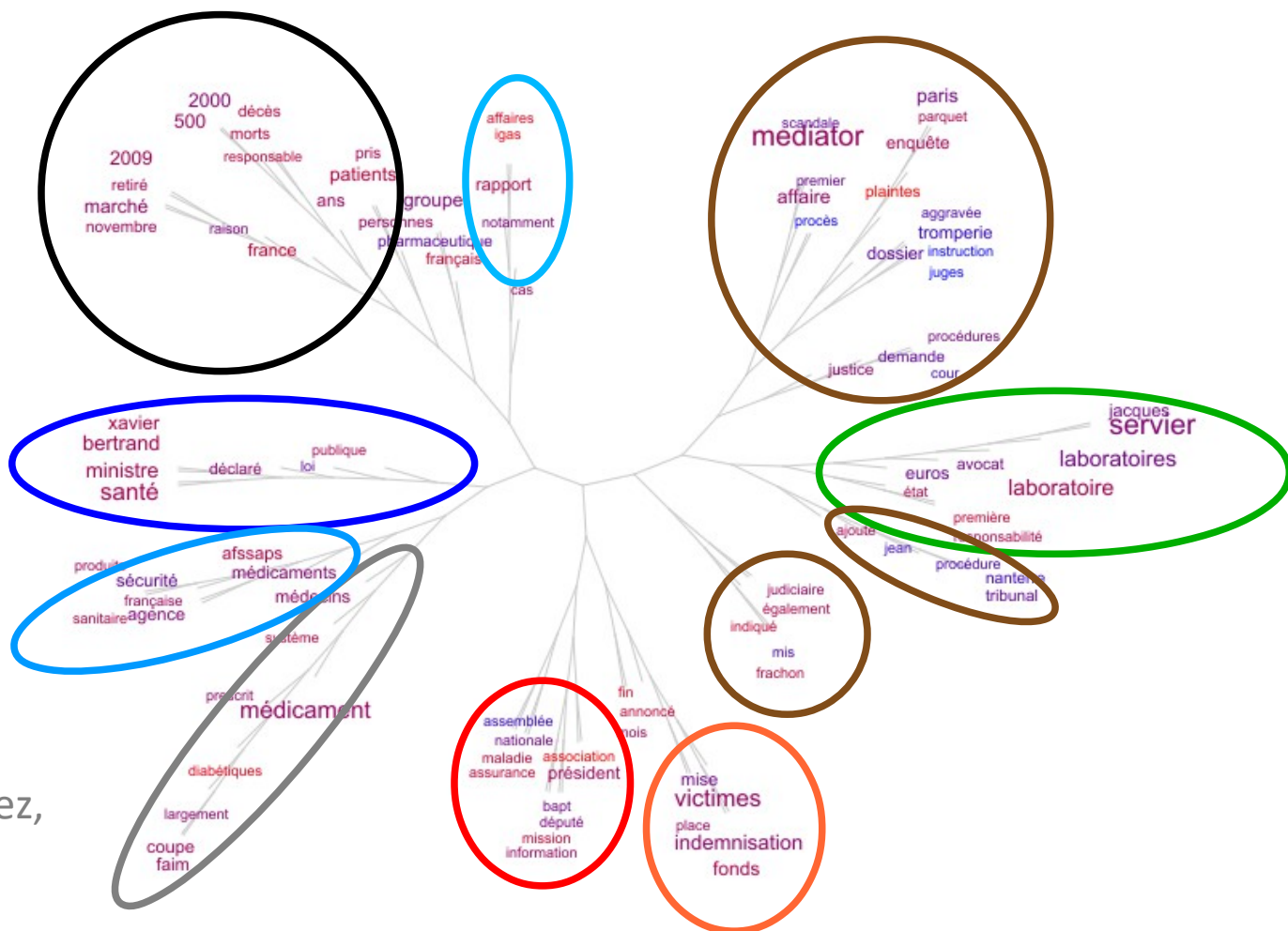


Comparer des sous-corpus

Comparer les articles d'agences et articles de journalistes

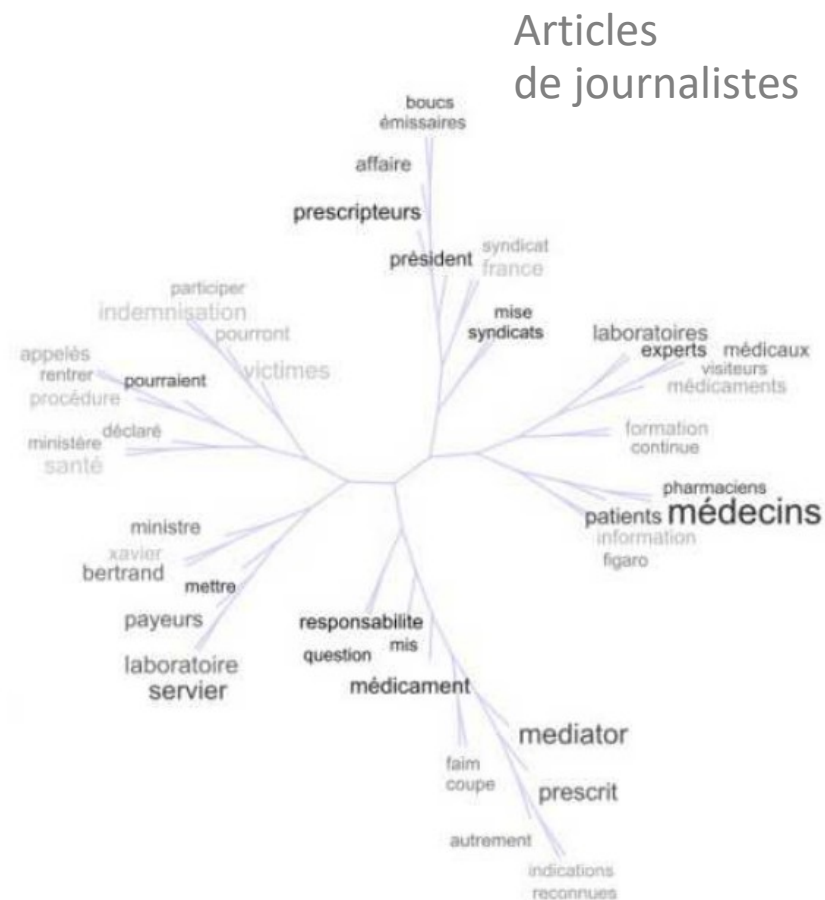
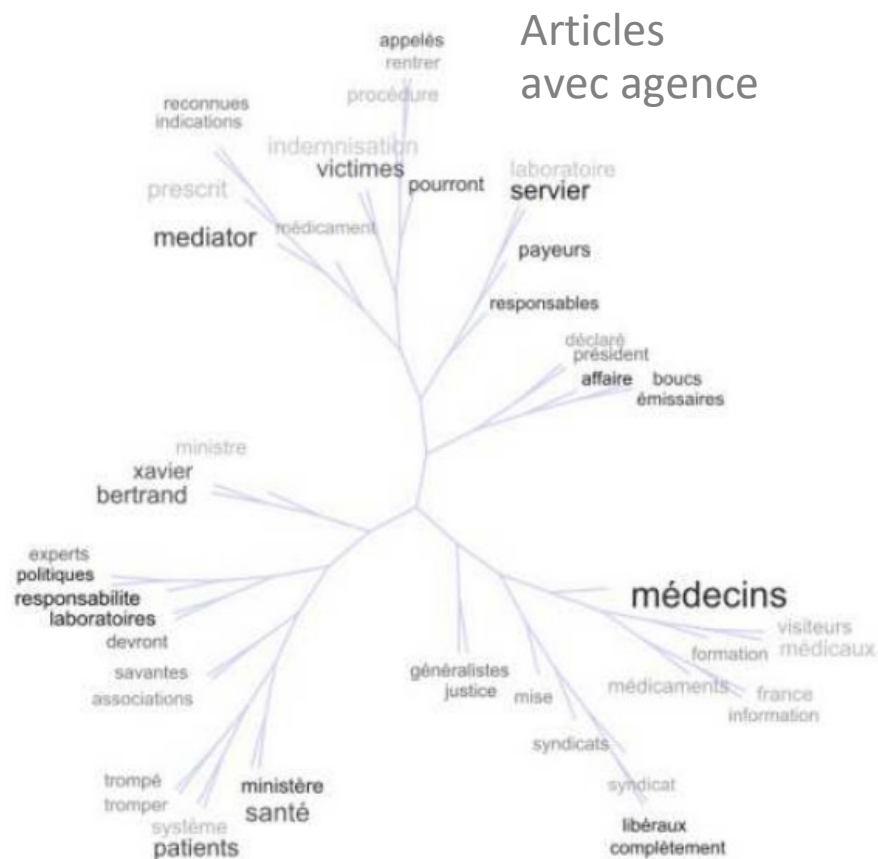
Corpus : 595 articles d'agences contre 1496 articles de journalistes de 2011 évoquant l'affaire du Mediator dans la presse française.

Articles
d'agences



Gambette & Martinez,
Texto!, 2013

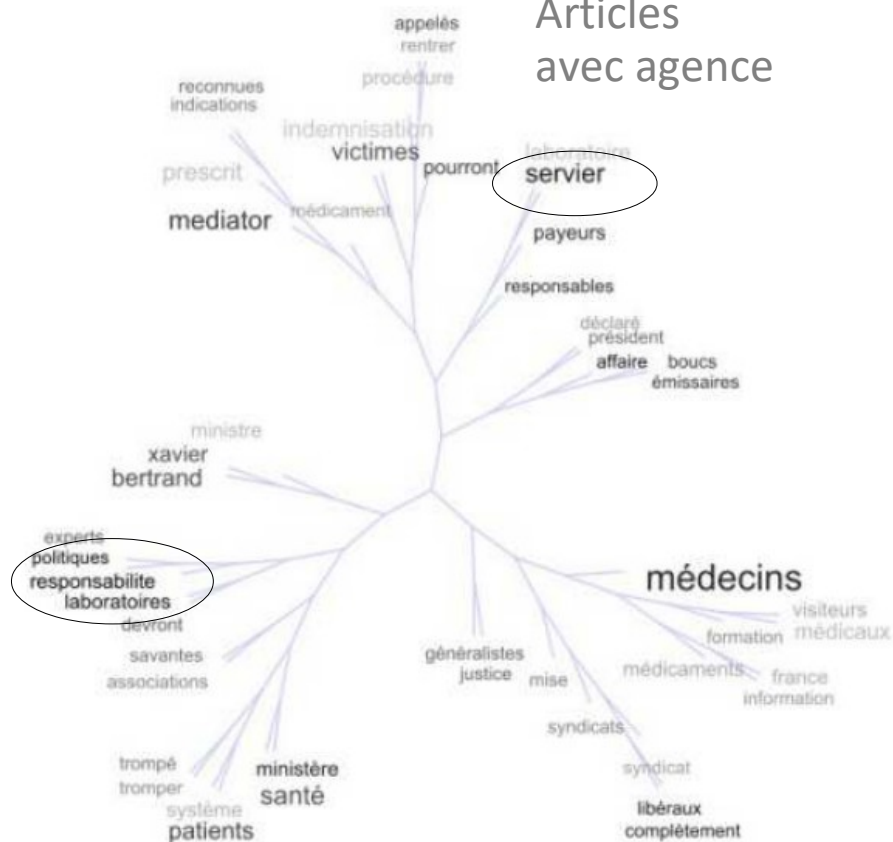
Nuages arborés des contextes de « médecins »



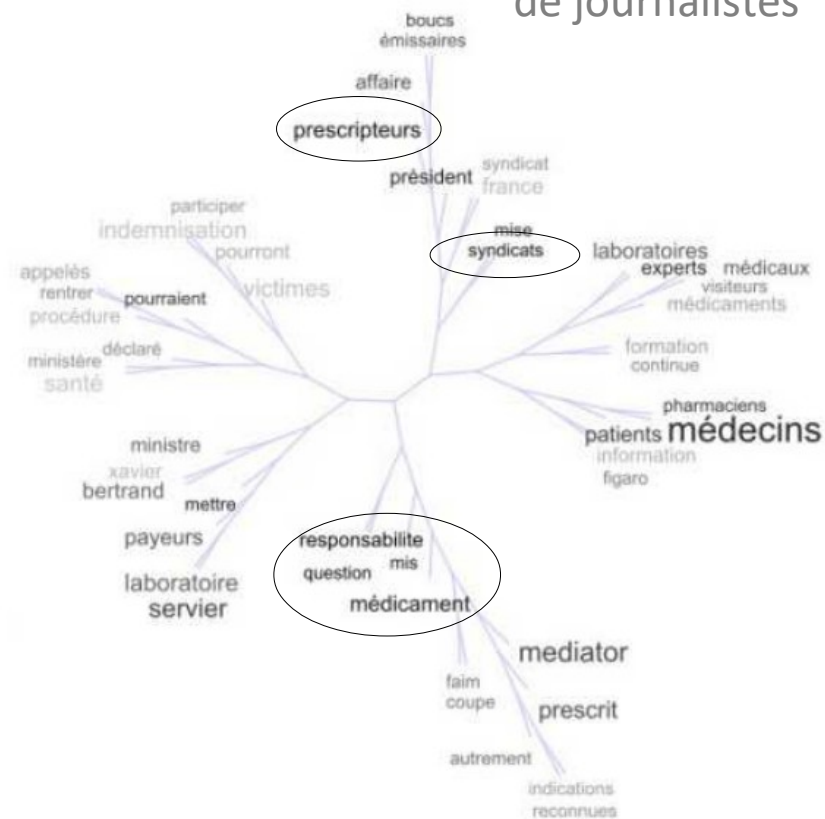
Nuage arboré des 50 mots les plus fréquents des contextes (10 mots avant et 10 mots après) du mot médecins dans le sous-corpus des articles sur le Mediator, colorés par le degré de cooccurrence avec le mot responsabilités (en noir pour les mots les plus cooccurents), construit par TreeCloud avec la formule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

Nuages arborés des contextes de « médecins »

Articles avec agence



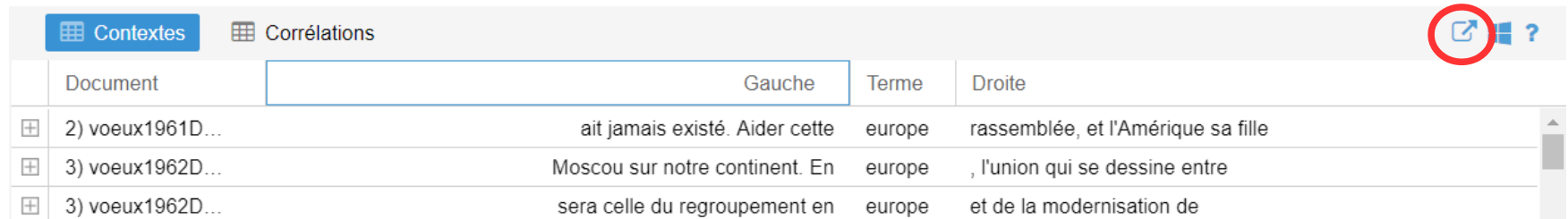
Articles de journalistes



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

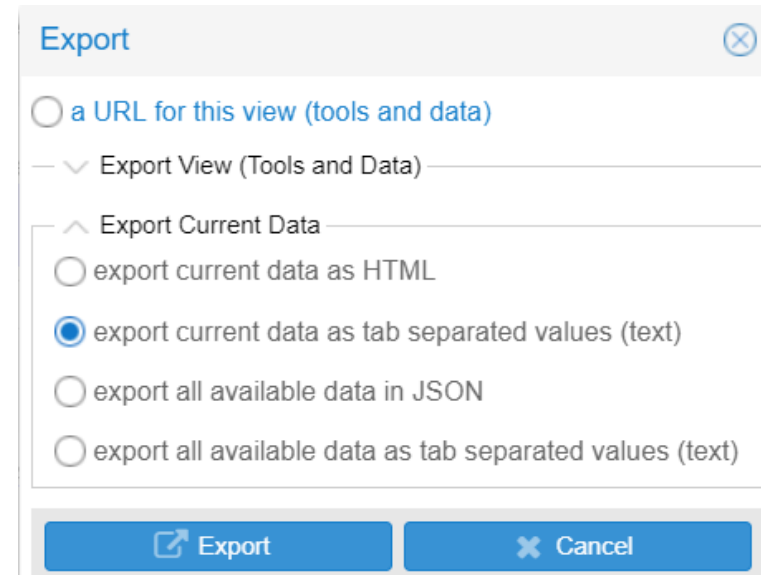
Extraction de contextes avec VoyantTools

- Charger le corpus dans <http://voyant-tools.org> (ex. : [corpus des vœux présidentiels](#), de Jean-Marc Leblanc)
- Charger les **contextes** de “europe”, par exemple, dans le cadre en bas à droite, puis exporter avec le bouton entouré en rouge :



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

- Dans la **fenêtre d'export**, choisir “Export Current Data”, “export current data as tab separated values (text)”
- Coller dans un **document tableur**
- Sélectionner uniquement les contextes gauches, droits, ou les deux, pour les charger dans TreeCloud.



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)

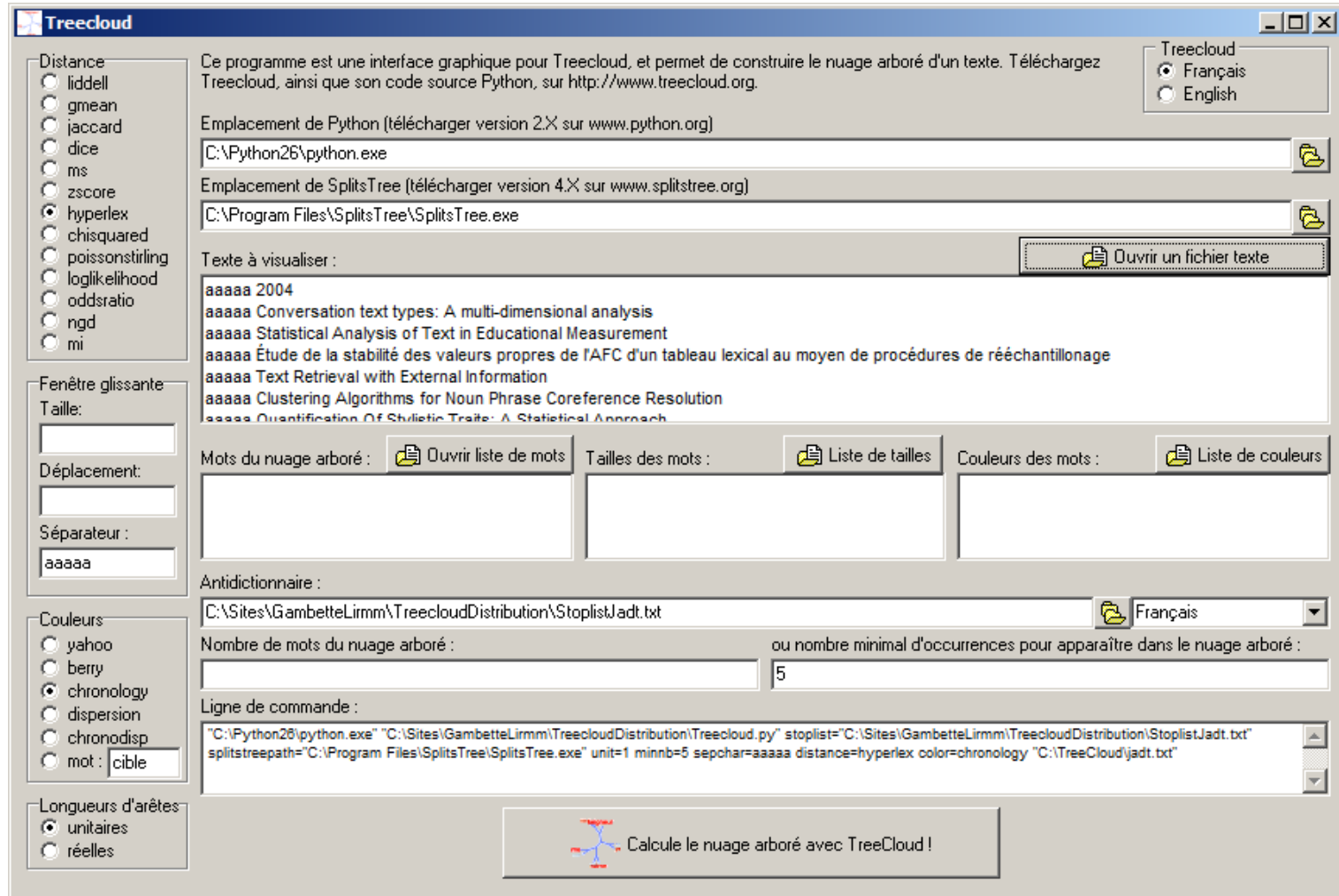
export all available data in JSON

export all available data as tab separated values (text)

Outils

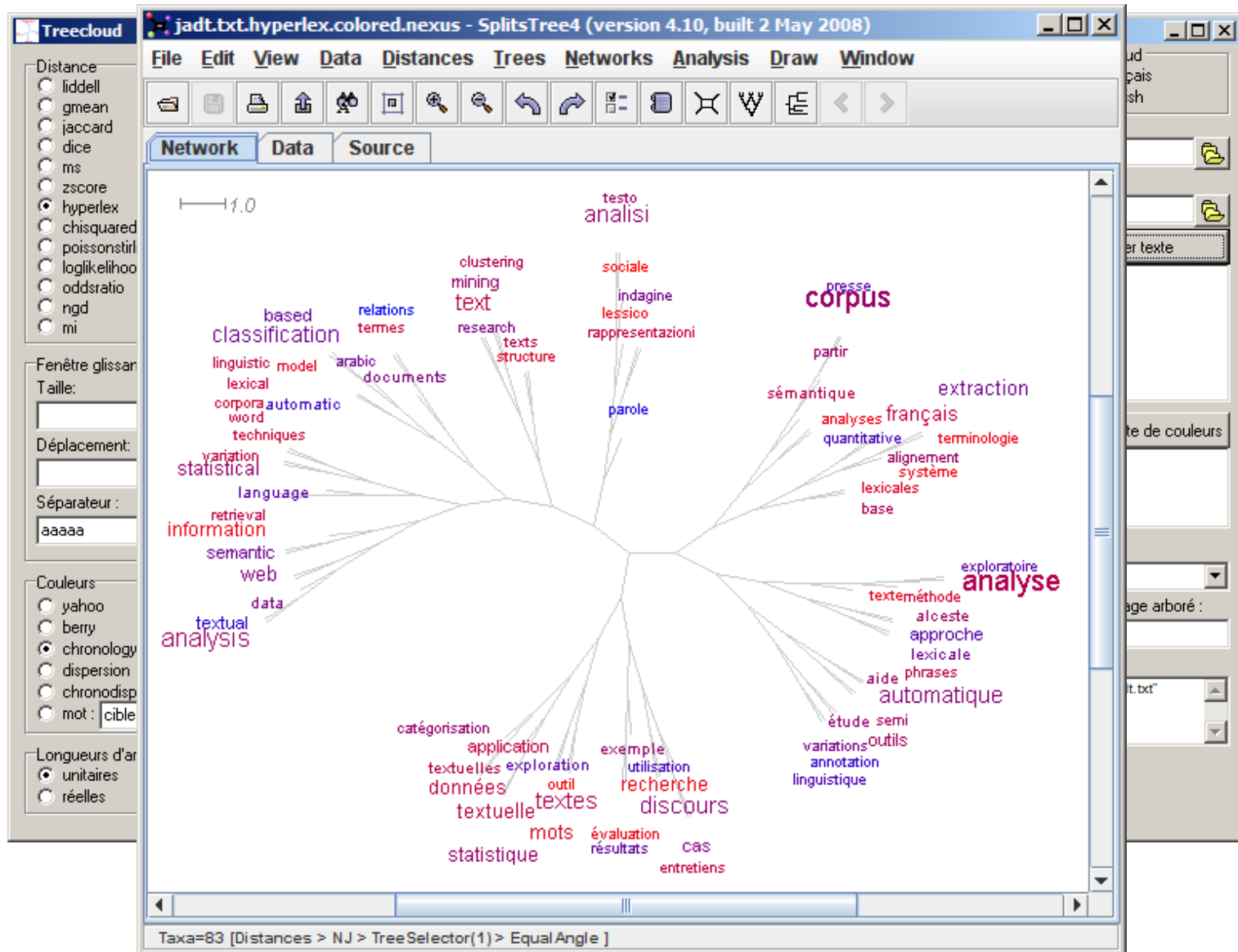
Implémentations

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



Implémentations

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



Implémentations

Version téléchargeable

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

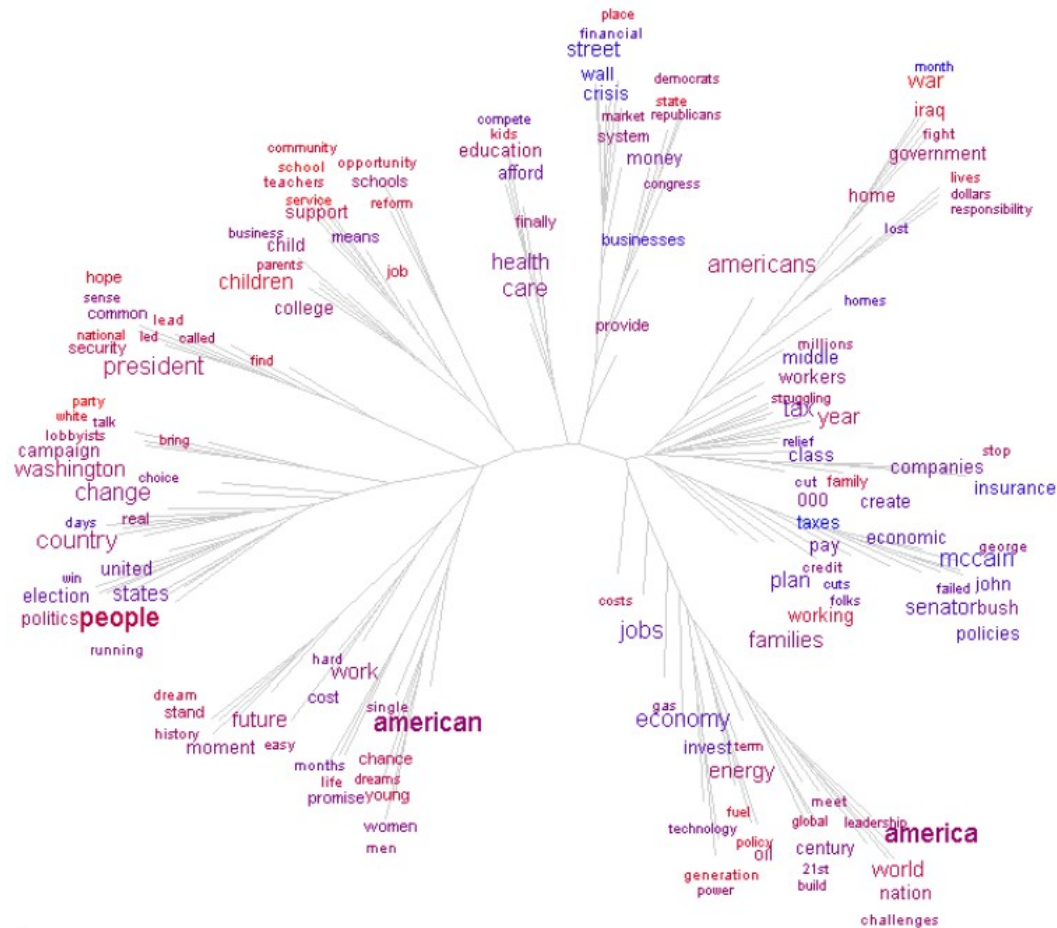
- [Tutoriel, manuel d'utilisation](#)
- Coloration de mots personnalisée
- Tailles de mots personnalisée
- Calcul des cooccurrences par blocs délimités par un séparateur

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

Temps d'exécution

Limites sur la taille du corpus pour utiliser TreeCloud ?



30 secondes pour la construction du nuage arboré de l'ensemble des discours de campagne de Barack Obama (>300 000 mots)

Références (*treecloud.org*)

Philippe Gambette, Jean Véronis (2009)

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

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

Delphine Amstutz & Philippe Gambette (2010)

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

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

Philippe Gambette, Nuria Gala & Alexis Nasr (2012)

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

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

William Martinez & Philippe Gambette (2013)

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

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

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

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

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

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