Journée d'étude - Visualisations en SHS et textométrie
24/05/2013 – Créteil (CEDITEC)

Nuages arborés et analyse textuelle
Présentation de l'outil TreeCloud

Philippe Gambette

LIGM
Université Paris-Est
Marne-la-Vallée

Plan

- Nuages arborés, intérêts et limites
- Construction des nuages arborés
- Options de coloration
- Utilisation des nuages arborés
- Évaluation de qualité
- Perspectives

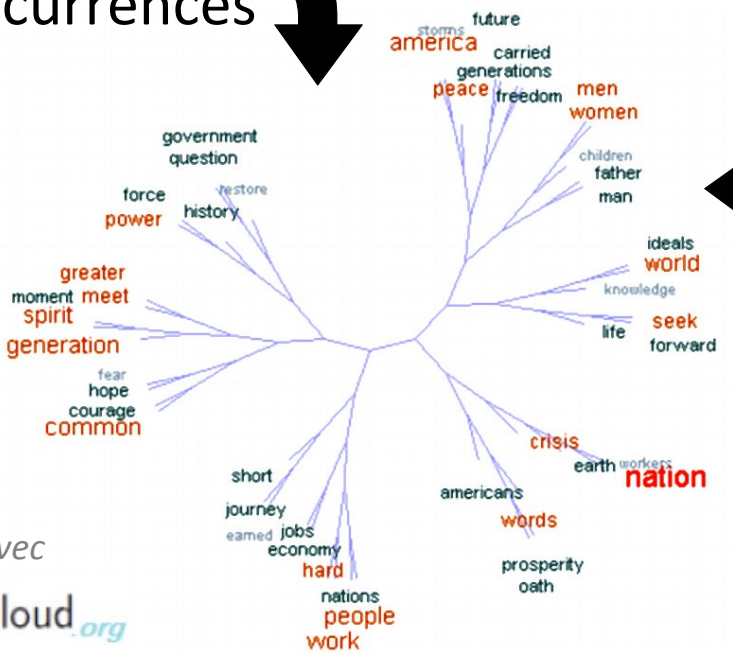
Plan

- Nuages arborés, intérêts et limites
- Construction des nuages arborés
- Options de coloration
- Utilisation des nuages arborés
- Évaluation de qualité
- Perspectives

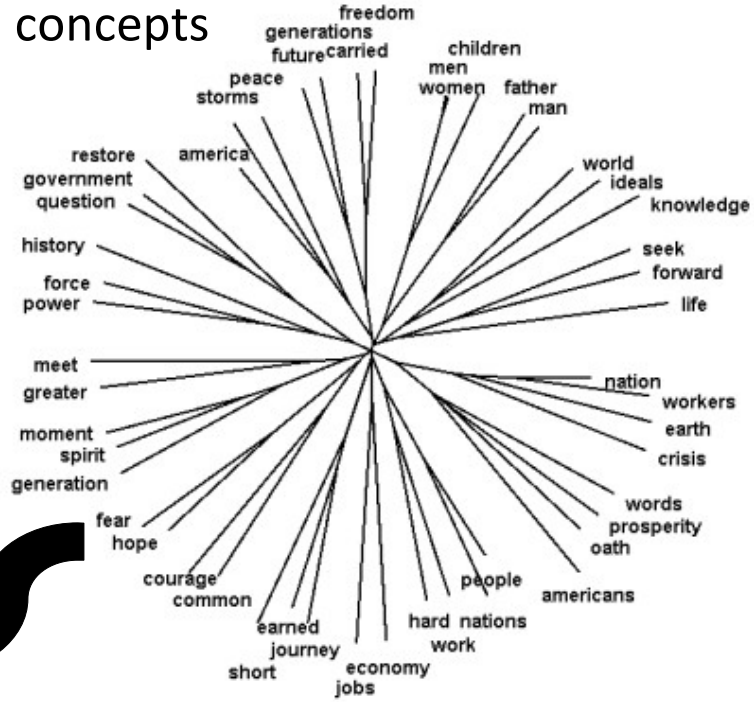
Nuage arboré, une information double

hiérarchie des mots

occurrences



hiérarchie des concepts



cooccurrences

Discours inaugural de Barack Obama

construit avec

TreeCloud.org

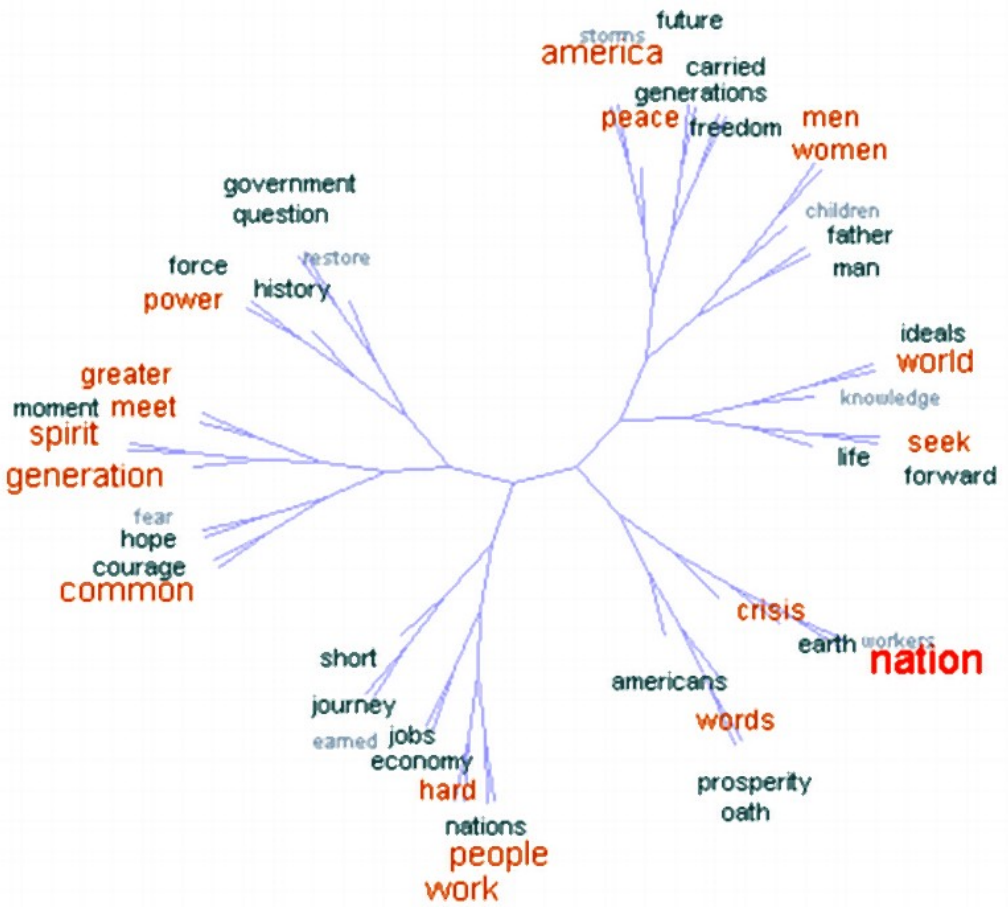
SplitsTree4

SplitsTree : Huson & Bryant, *Bioinformatics*, 2006

TreeCloud : Gambette & Véronis, *IFCS'09*

Intérêts de la visualisation arborée

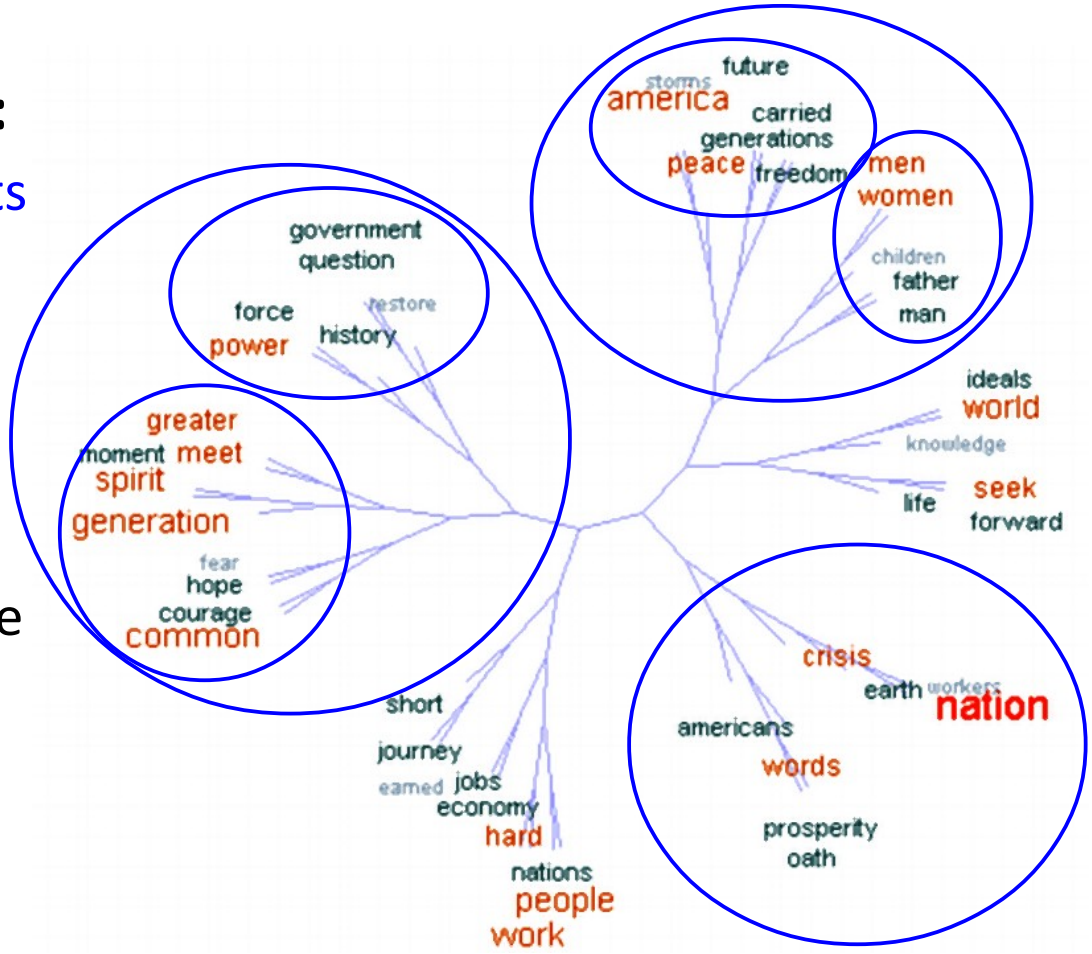
- **Quantité d'information :**
  - nombre de sacs de mots imbriqués linéaire (≠ réseaux, AFC)
- **Qualité d'information :**
  - prise en compte d'une information globale pour le rapprochement de mots dans l'arbre (≠ réseaux)



- **Lisibilité :**
  - dessin normalisé de l'arbre par méthode radiale (≠ réseaux)
  - placement des étiquettes sans chevauchement (≠ AFC)

Intérêts de la visualisation arborée

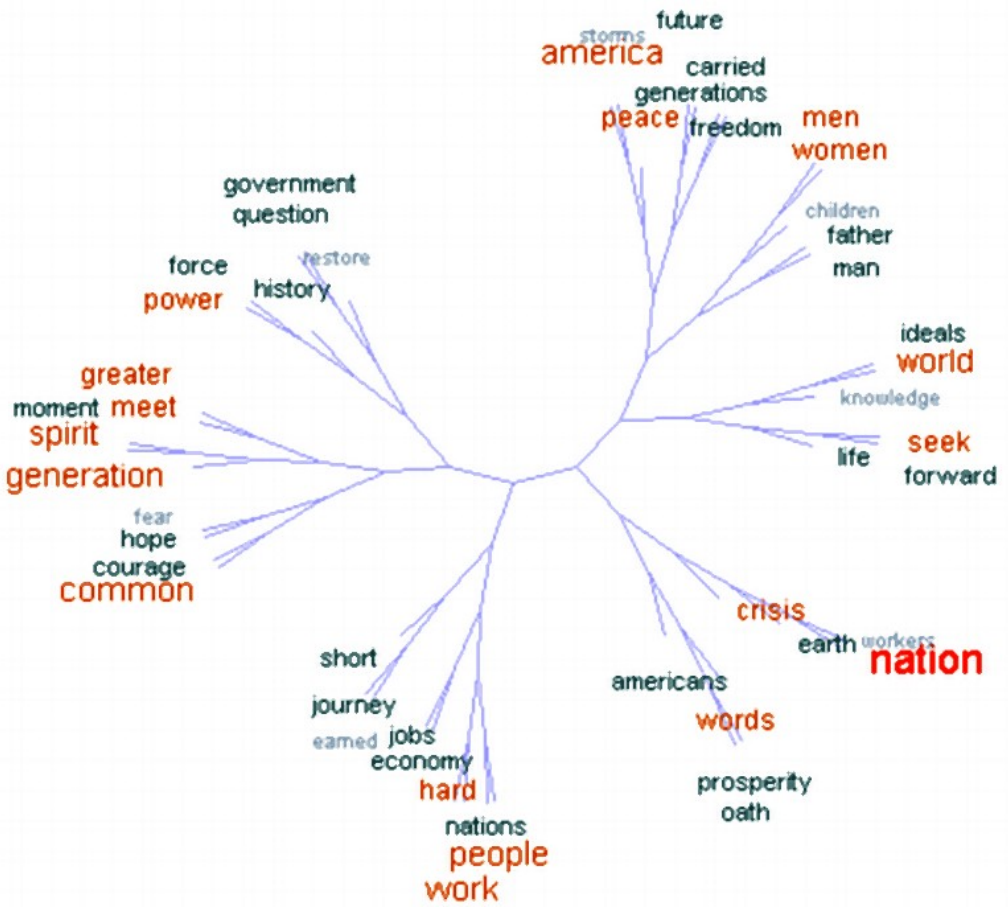
- **Quantité d'information :**
  - nombre de sacs de mots imbriqués linéaire (≠ réseaux, AFC)
- **Qualité d'information :**
  - prise en compte d'une information globale pour le rapprochement de mots dans l'arbre (≠ réseaux)



- **Lisibilité :**
  - dessin normalisé de l'arbre par méthode radiale (≠ réseaux)
  - placement des étiquettes sans chevauchement (≠ AFC)

Intérêts de la visualisation arborée

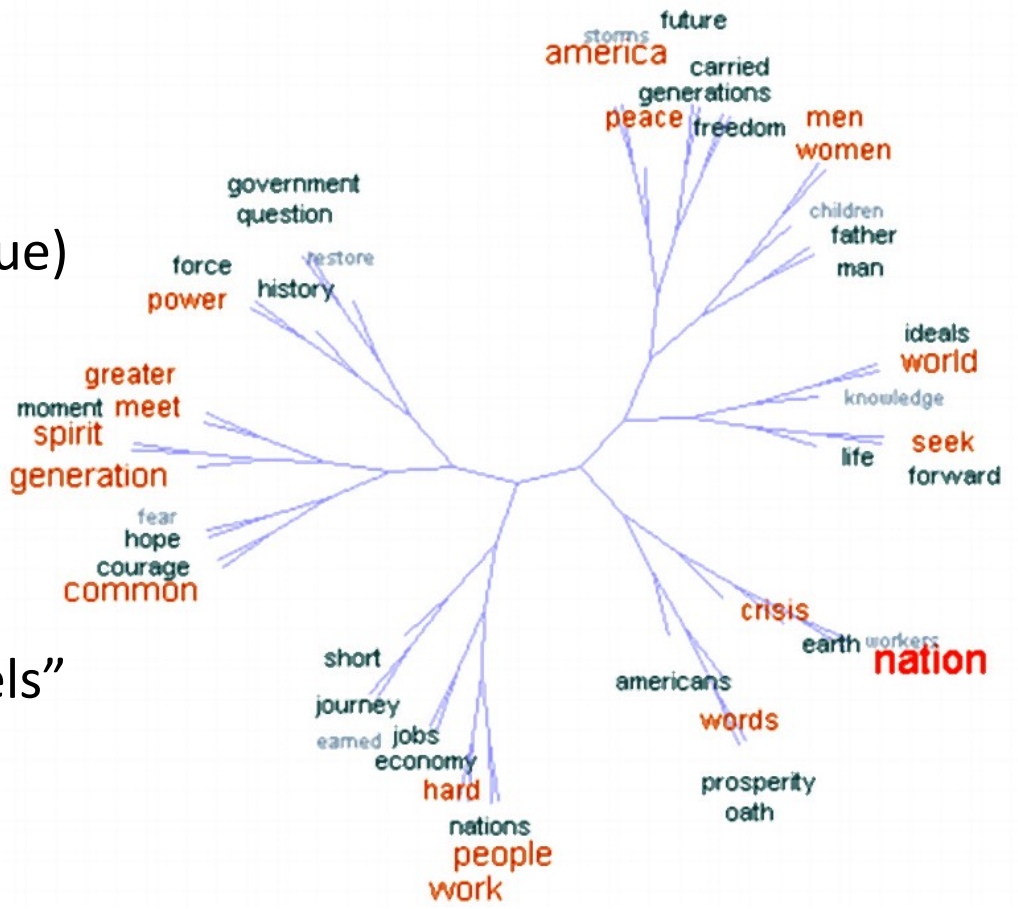
- **Quantité d'information :**
  - nombre de sacs de mots imbriqués linéaire (≠ réseaux, AFC)
- **Qualité d'information :**
  - prise en compte d'une information globale pour le rapprochement de mots dans l'arbre (≠ réseaux)



- **Lisibilité :**
  - dessin normalisé de l'arbre par méthode radiale (≠ réseaux)
  - placement des étiquettes sans chevauchement (≠ AFC)

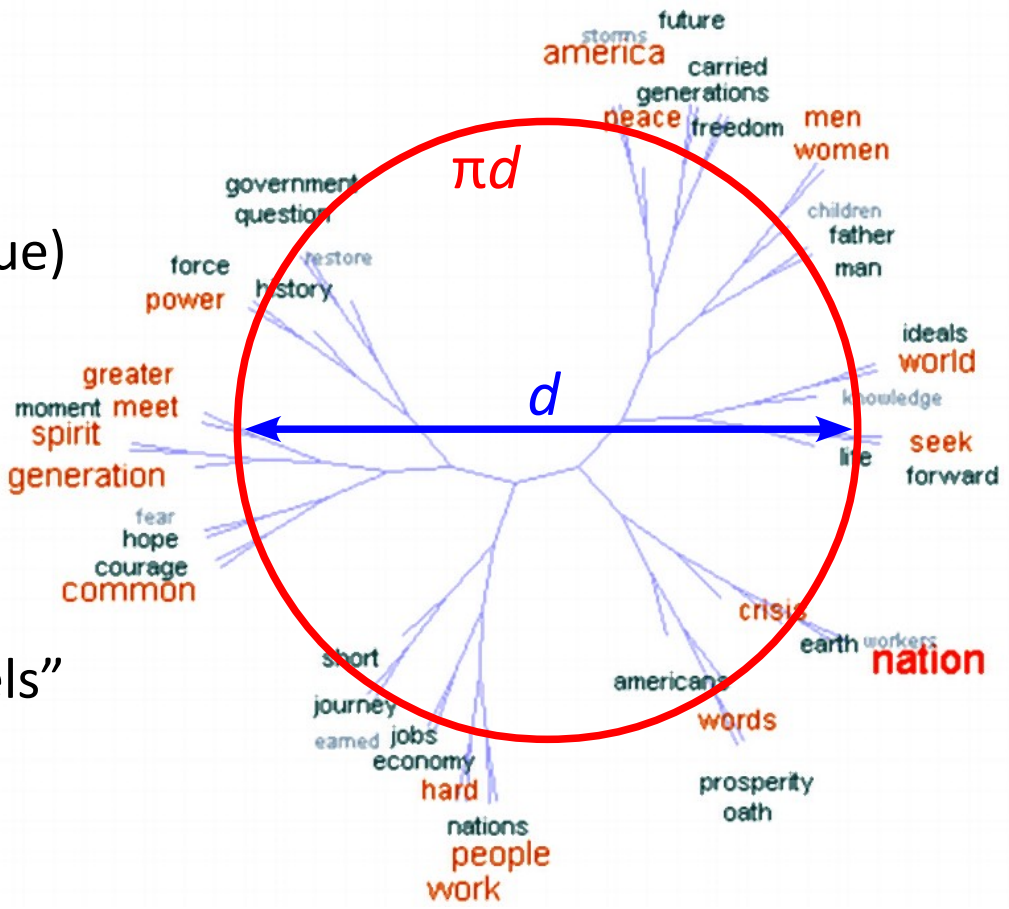
Limites de la visualisation arborée

- **Quantité d'information :**
  - nombre de mots linéaire en la largeur du dessin (≠ AFC, réseaux : quadratique)
- **Qualité d'information :**
  - artefacts de la méthode : pas d'arbre "parfait"
  - rapprochements "artificiels"
  - instabilité (≠ AFC)
- **Lisibilité :**
  - placement des étiquettes compliqué
  - problème des longueurs de branches
  - attention aux mauvaises interprétations



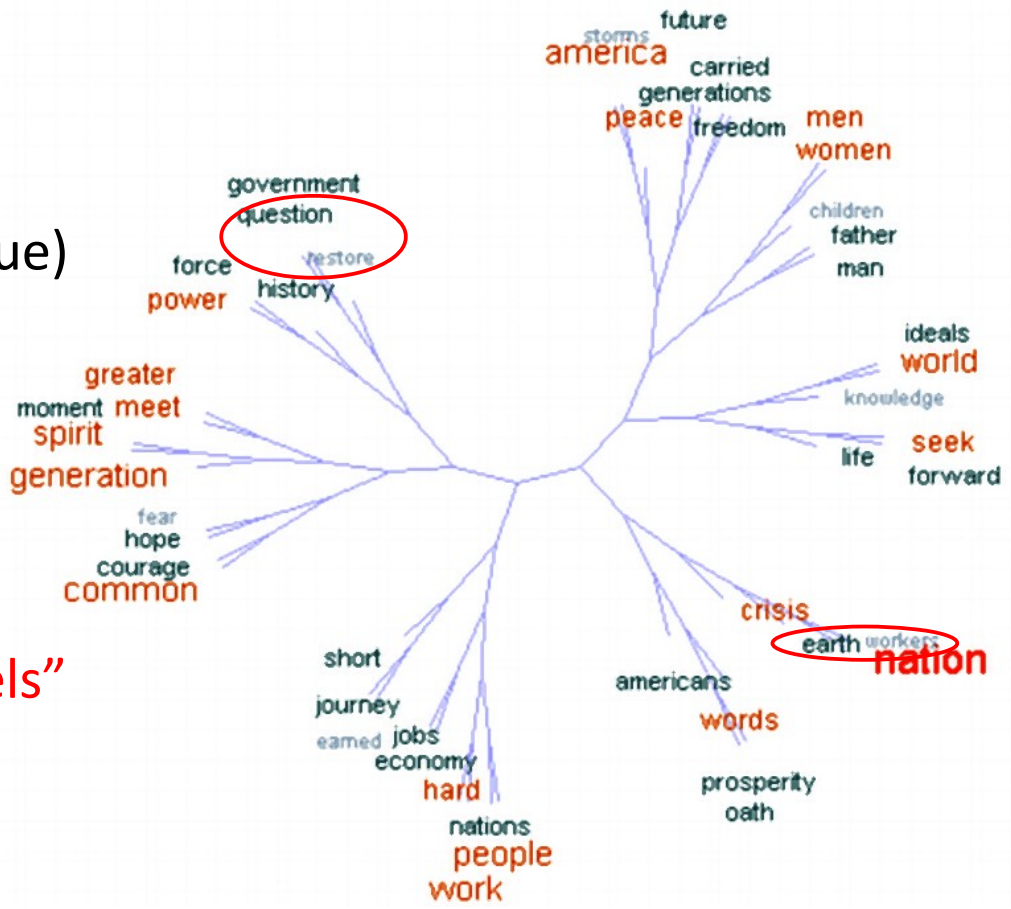
Limites de la visualisation arborée

- **Quantité d'information :**
  - nombre de mots linéaire en la largeur du dessin (≠ AFC, réseaux : quadratique)
- **Qualité d'information :**
  - artefacts de la méthode : pas d'arbre "parfait"
  - rapprochements "artificiels"
  - instabilité (≠ AFC)
- **Lisibilité :**
  - placement des étiquettes compliqué
  - problème des longueurs de branches
  - attention aux mauvaises interprétations



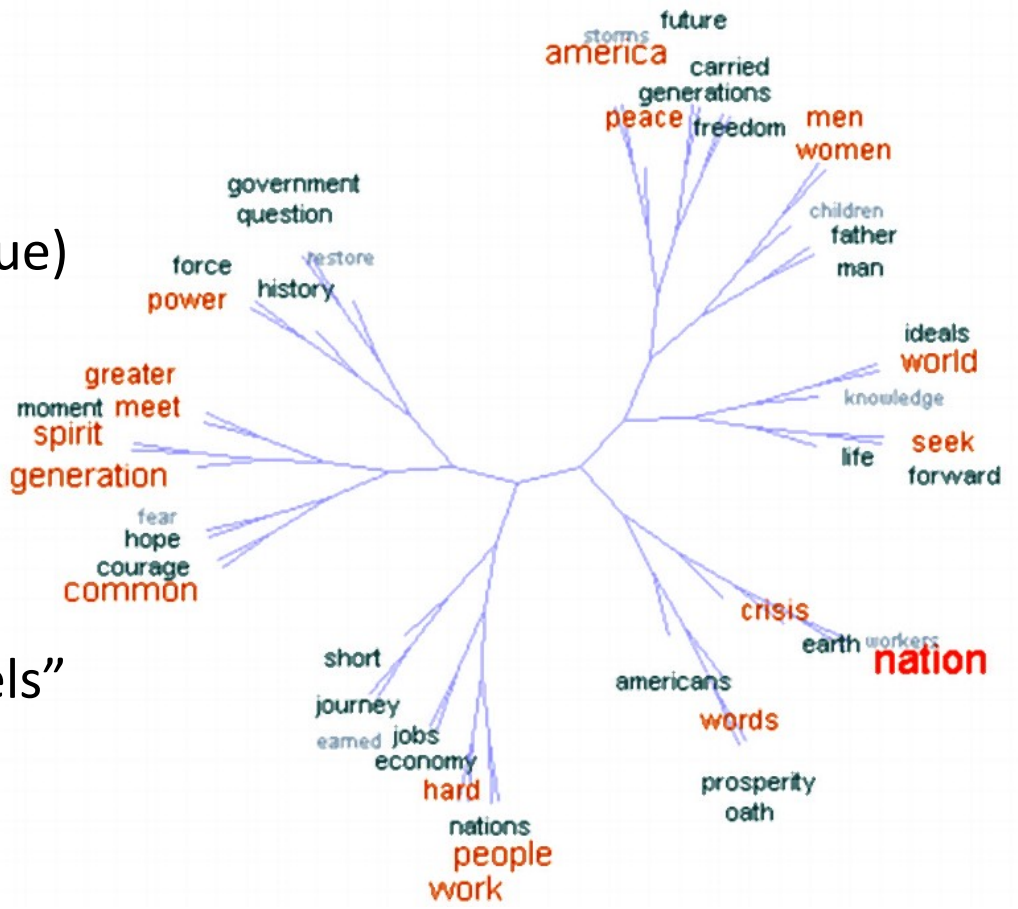
Limites de la visualisation arborée

- **Quantité d'information :**
  - nombre de mots linéaire en la largeur du dessin (≠ AFC, réseaux : quadratique)
- **Qualité d'information :**
  - artefacts de la méthode : pas d'arbre "parfait"
  - rapprochements "artificiels"
  - instabilité (≠ AFC)
- **Lisibilité :**
  - placement des étiquettes compliqué
  - problème des longueurs de branches
  - attention aux mauvaises interprétations



Limites de la visualisation arborée

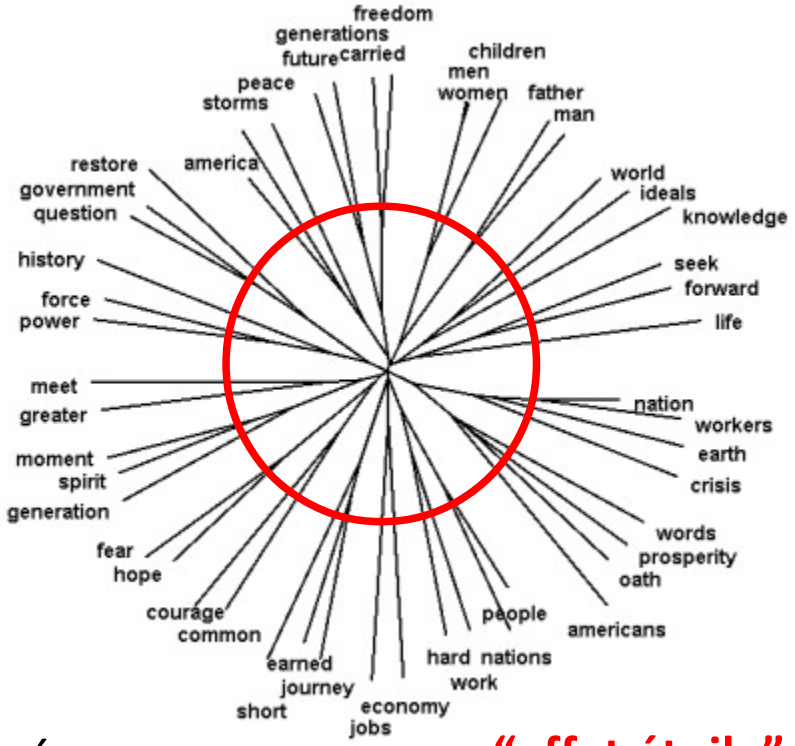
- **Quantité d'information :**
  - nombre de mots linéaire en la largeur du dessin (≠ AFC, réseaux : quadratique)
- **Qualité d'information :**
  - artefacts de la méthode : pas d'arbre "parfait"
  - rapprochements "artificiels"
  - instabilité (≠ AFC)
- **Lisibilité :**
  - placement des étiquettes compliqué
  - problème des longueurs de branches
  - attention aux mauvaises interprétations



Limites de la visualisation arborée

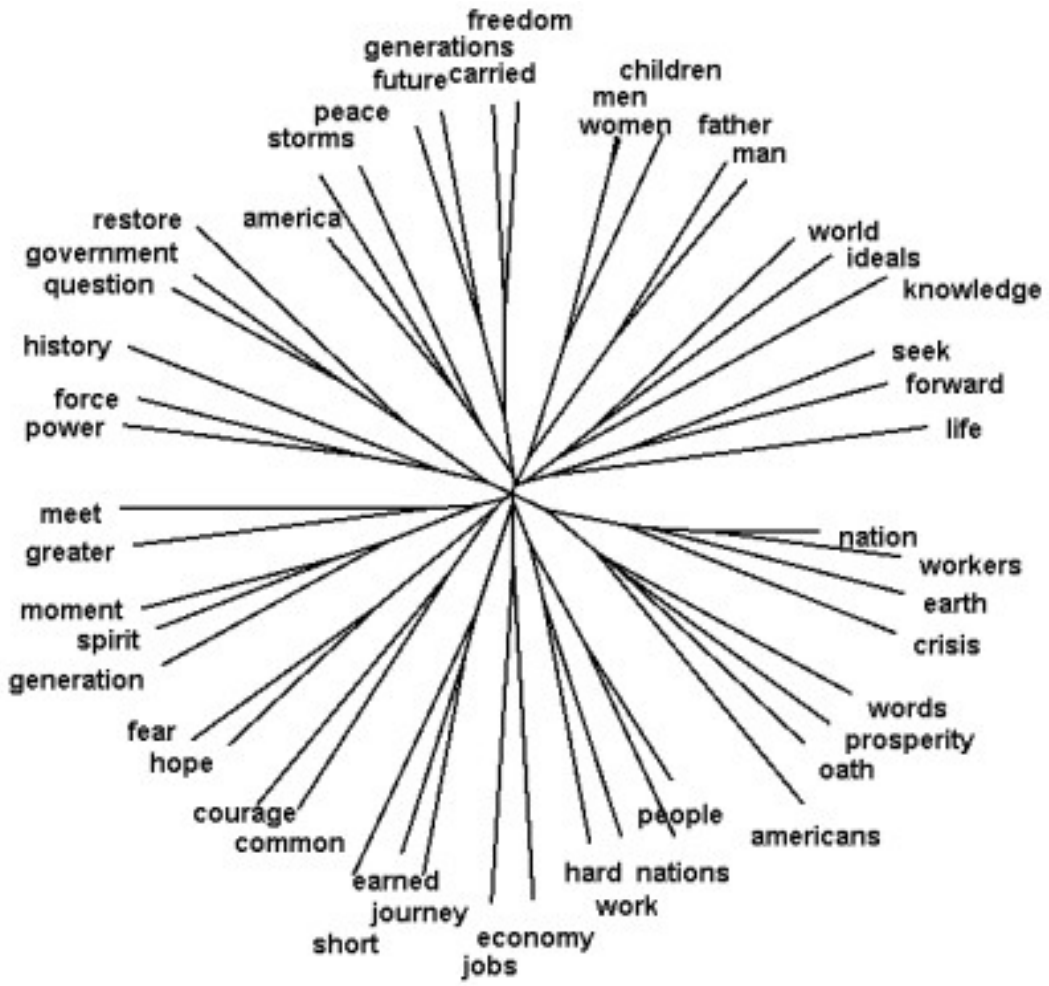
- **Quantité d'information :**
  - nombre de mots linéaire en la largeur du dessin (≠ AFC, réseaux : quadratique)
- **Qualité d'information :**
  - artefacts de la méthode : pas d'arbre "parfait"
  - rapprochements "artificiels"
  - instabilité (≠ AFC)
- **Lisibilité :**
  - placement des étiquettes compliqué
  - problème des longueurs de branches
  - attention aux mauvaises interprétations

Gambette, Gala, & Nasr,
Longueur de branches et arbres de mots,
Corpus 11:129-146, 2012.



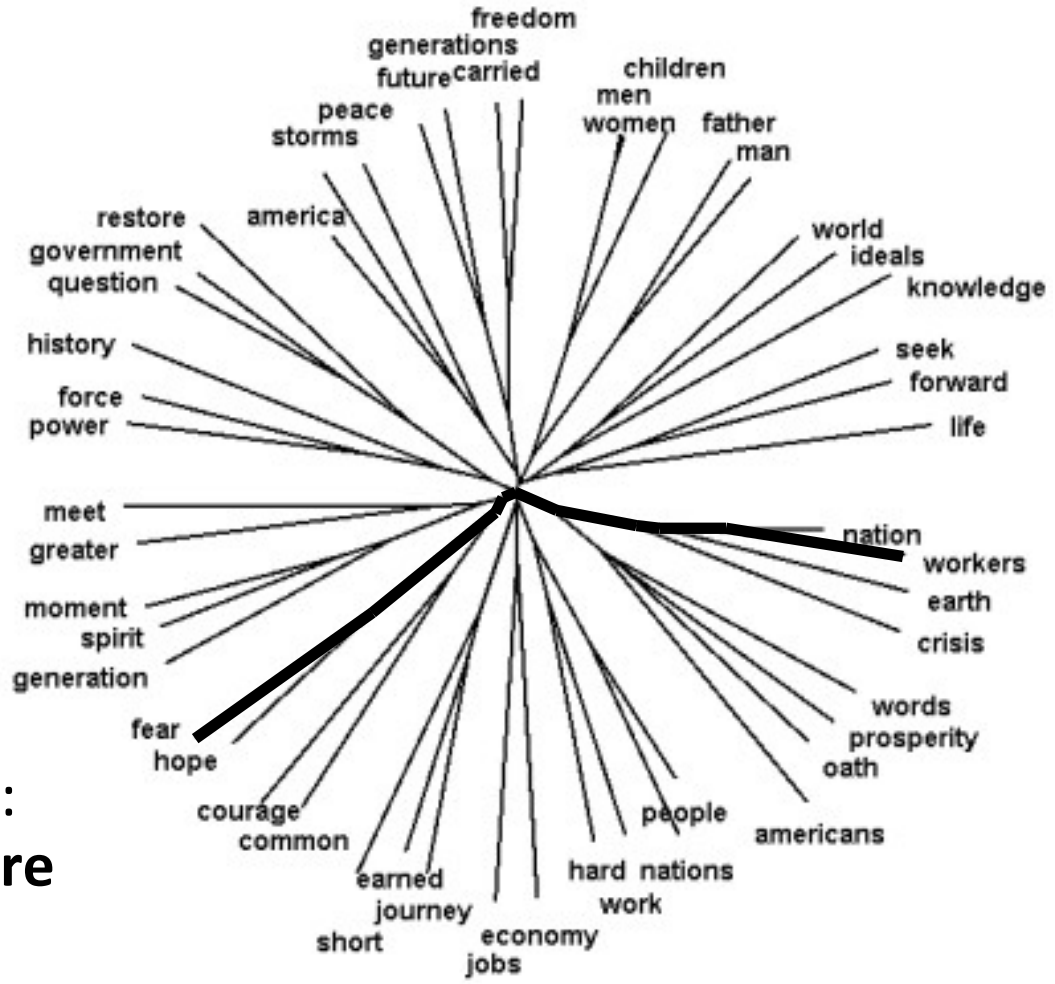
"effet étoile", centre illisible

Interprétation réelle



Principe de construction de l'arbre :
Les **distances dans l'arbre** entre deux mots reflètent *au mieux* le **degré de cooccurrence** entre ces deux mots

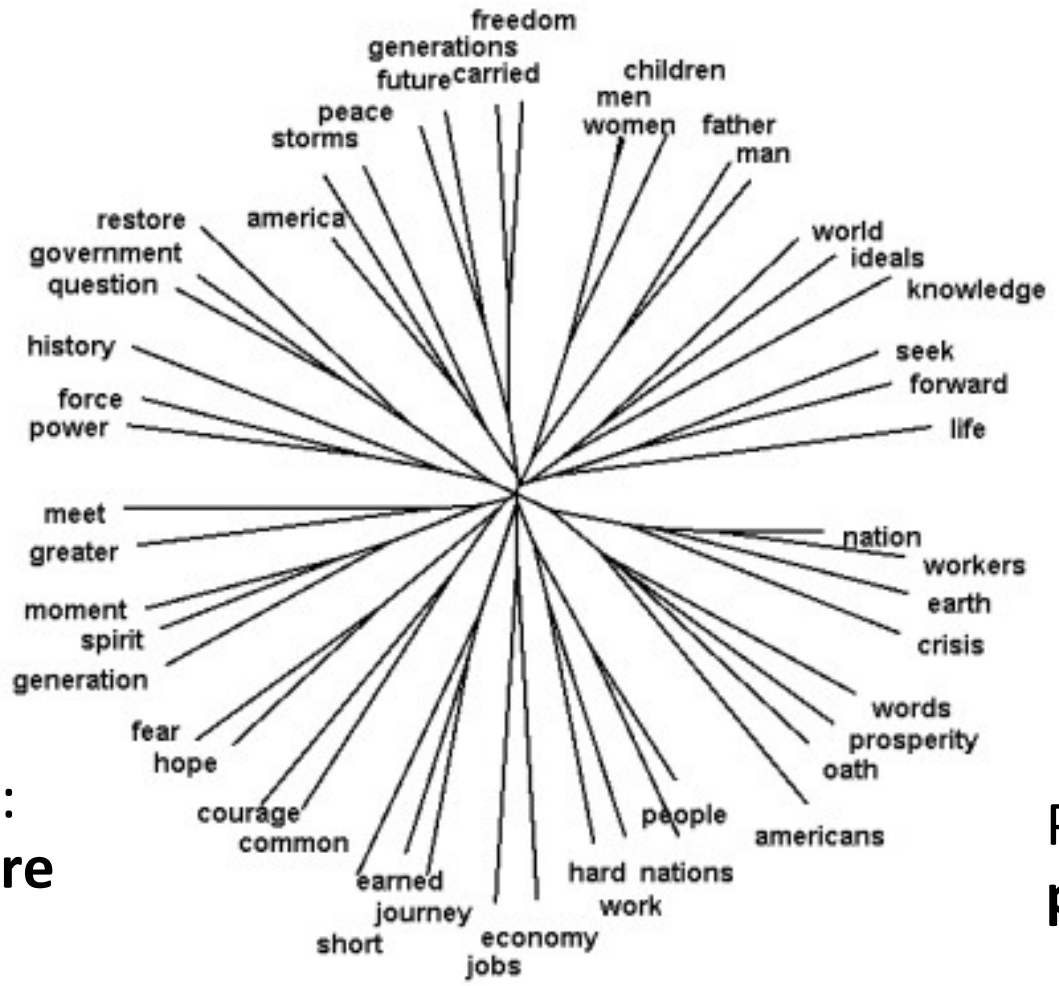
Interprétation réelle



Problème 1 :
difficiles à lire

Les **distances dans l'arbre** entre deux mots reflètent *au mieux* le **degré de cooccurrence** entre ces deux mots

Interprétation réelle

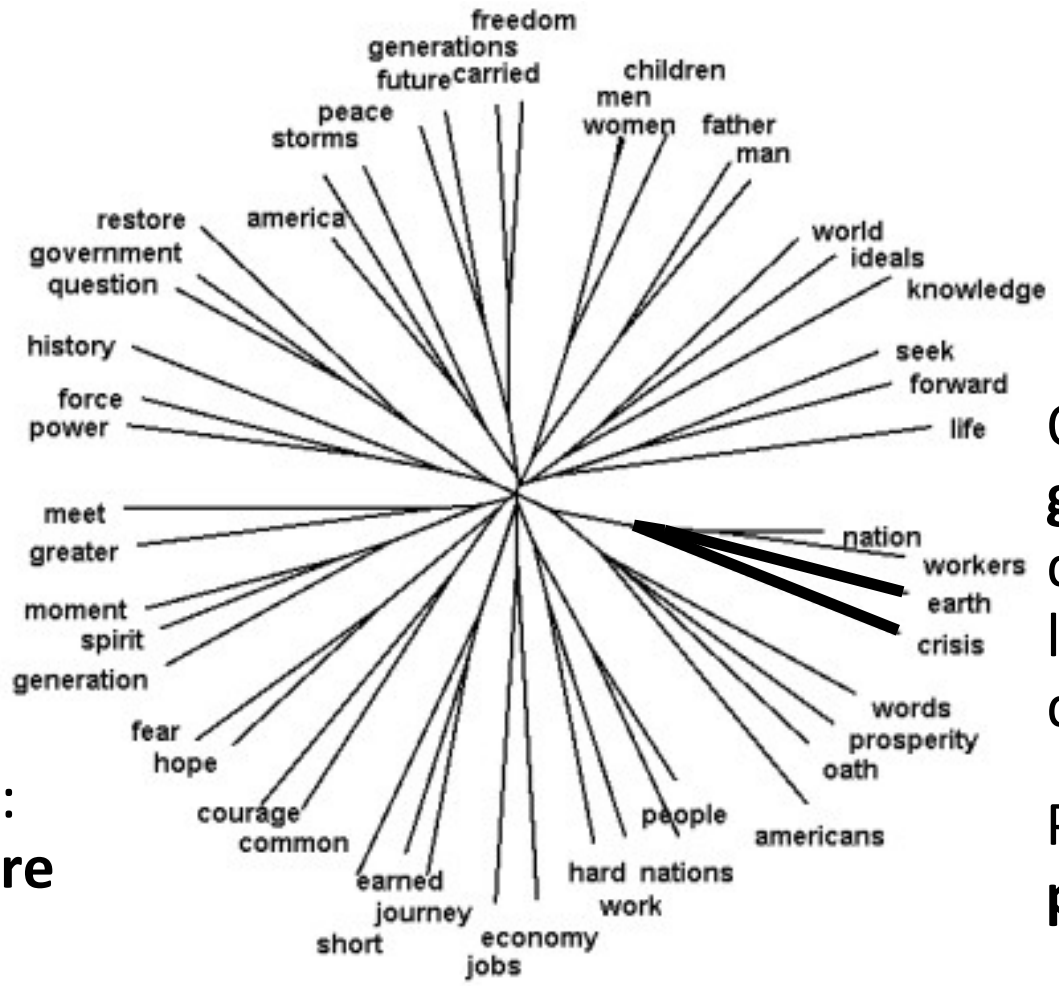


Problème 1 :
difficiles à lire

Problème 2 :
peu fiables

Les **distances dans l'arbre** entre deux mots reflètent *au mieux* le **degré de cooccurrence** entre ces deux mots

Interprétation réelle



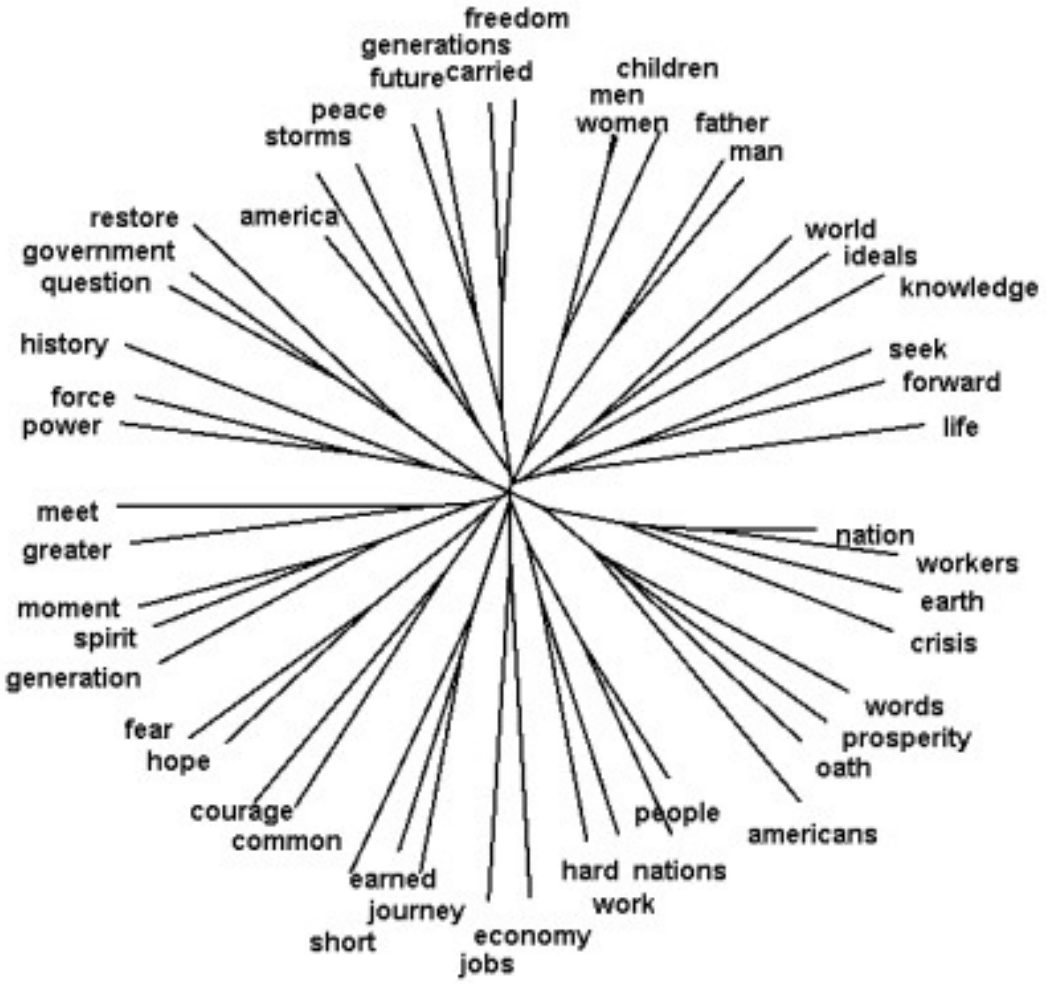
Optimisation **globale**, pas de garanties locales de qualité

Problème 1 : **difficiles à lire**

Problème 2 : **peu fiables**

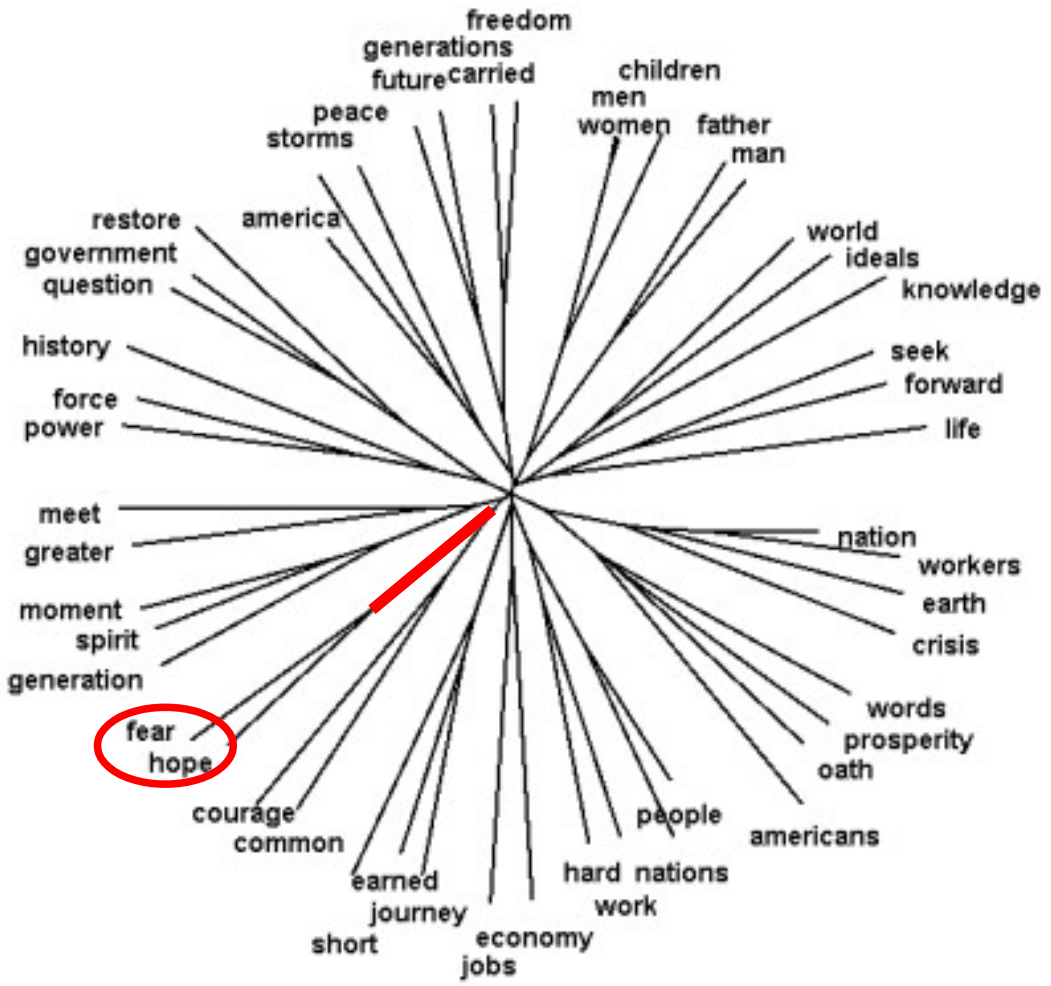
Les **distances dans l'arbre** entre deux mots reflètent *au mieux* le **degré de cooccurrence** entre ces deux mots

Interprétation pratique



arbre de distances utilisé comme **classification**

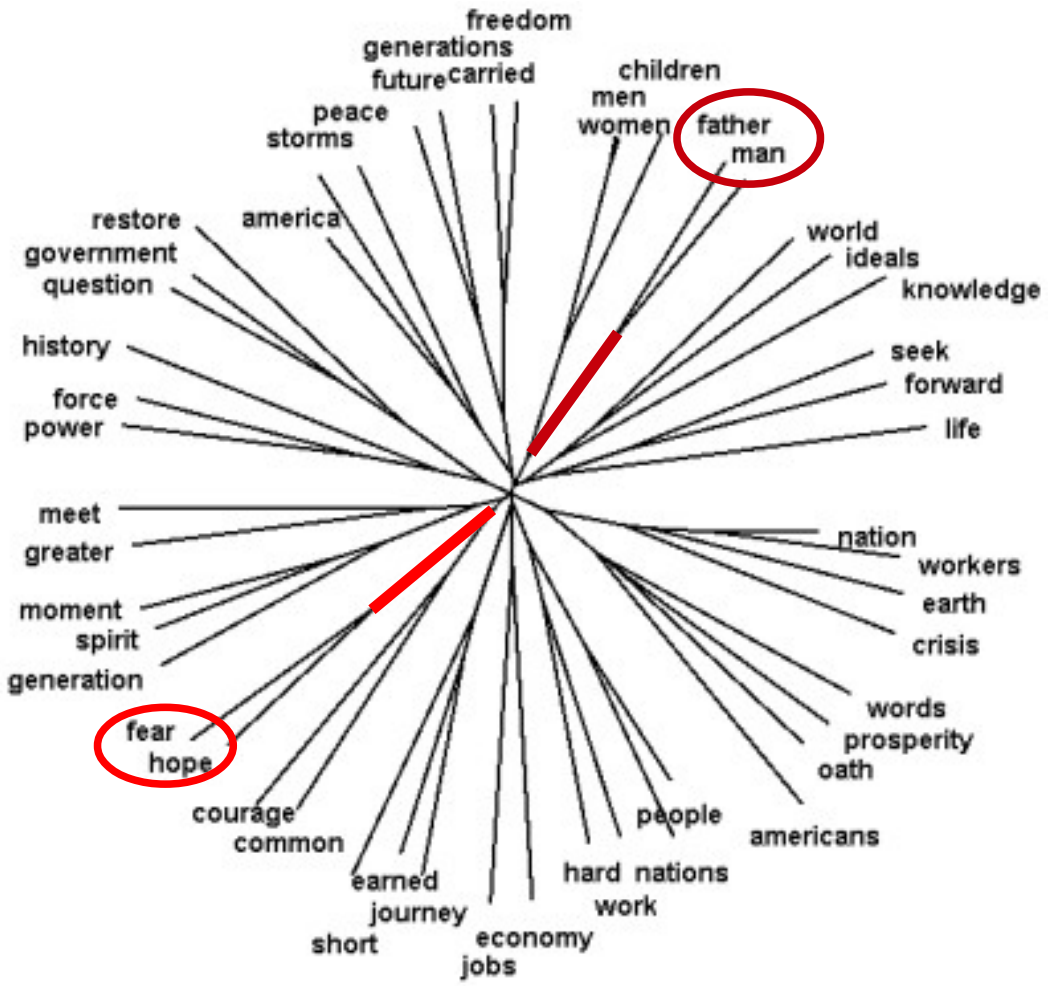
Interprétation pratique



arbre de distances utilisé comme **classification**

Les mots d'un **même sous-arbre bien séparé** du reste de l'arbre constituent une classe de mots

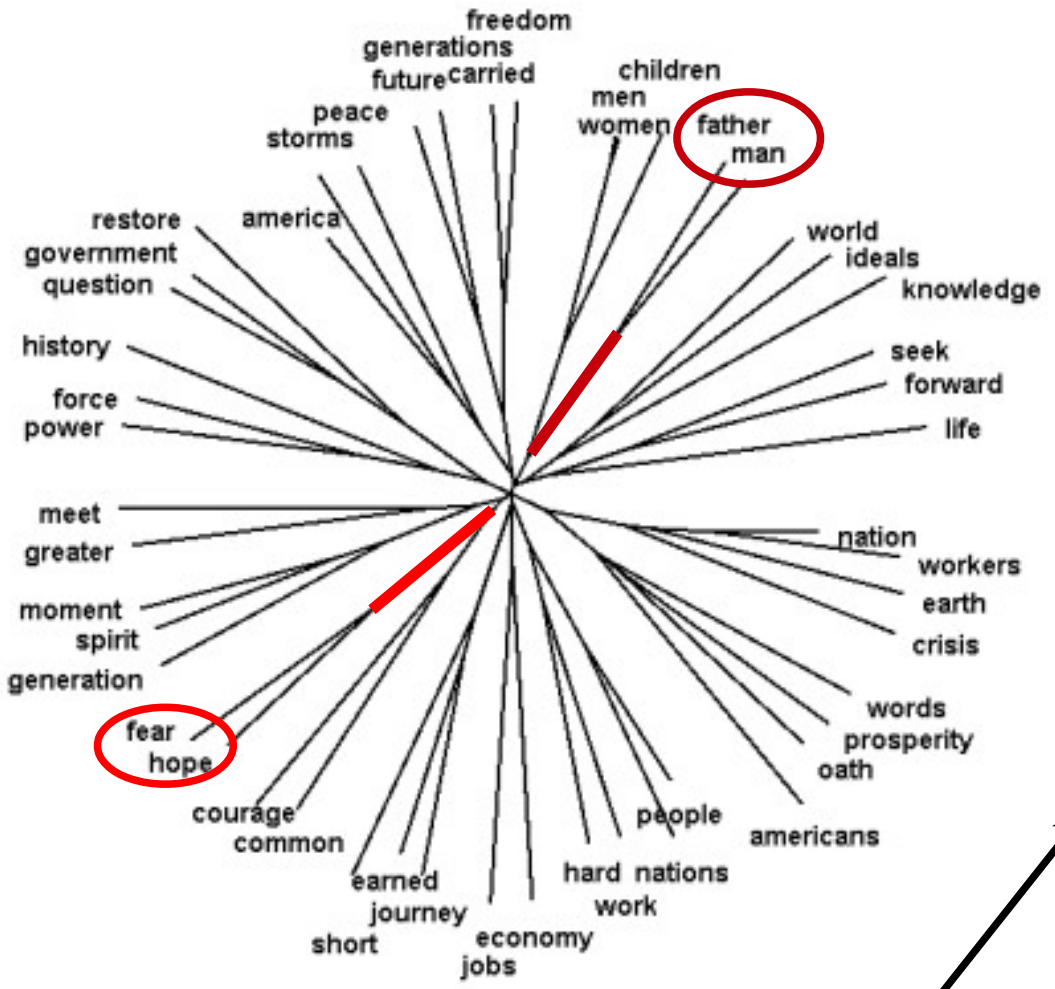
Interprétation pratique



arbre de distances utilisé comme **classification**

Les mots d'un **même sous-arbre bien séparé** du reste de l'arbre constituent une classe de mots

Interprétation pratique



arbre de distances utilisé comme **classification**

Problème : **toujours peu lisible** (longueur des arêtes externes) et **peu fiable**

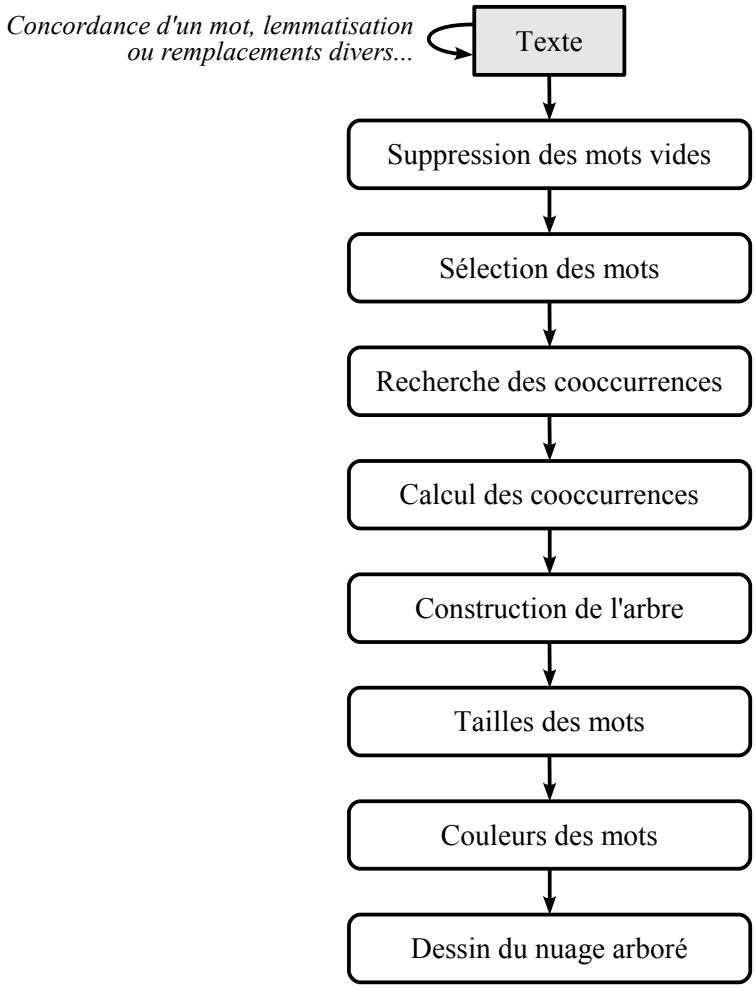
Les mots d'un **même sous-arbre bien séparé** du reste de l'arbre constituent une classe de mots

Plan

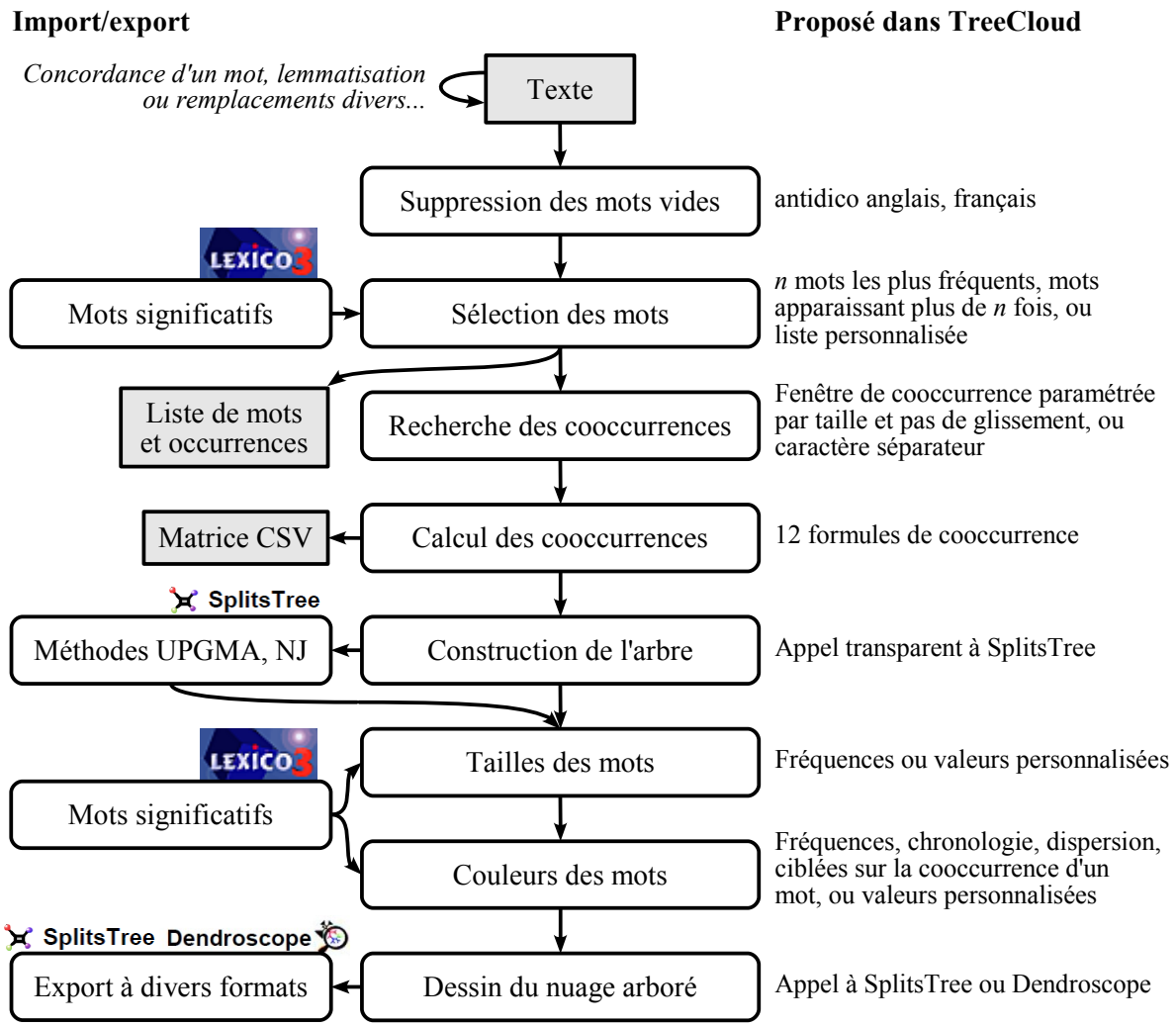
- Nuages arborés, intérêts et limites
- Construction des nuages arborés
- Options de coloration
- Utilisation des nuages arborés
- Évaluation de qualité
- Perspectives

Processus de construction

Import/export

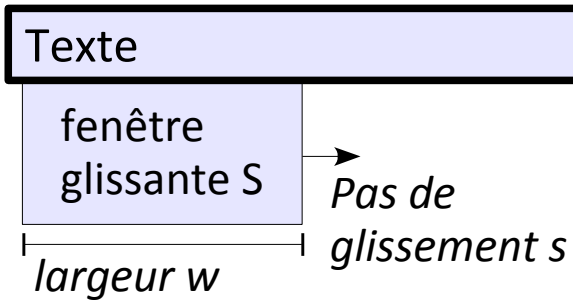


Processus de construction



Calcul des scores de cooccurrence

Calcul de la matrice de distance entre mots



La dissimilarité sémantique entre deux mots *u* et *v* dépend du nombre de fenêtres *S* où ils apparaissent ensemble.

matrices de cooccurrence
O_{11}, O_{12}, O_{21}, O_{22}

Pour 2 mots *u* et *v*	$v \in S$	$v \notin S$
$u \in S$	O_{11}	O_{12}
$u \notin S$	O_{21}	O_{22}

matrice de dissimilarité sémantique
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
Gambette, *User manual for TreeCloud,* 2009

Calcul des distances de cooccurrence

Les formules statistiques fournissent un score de similarité.

Comment obtenir des dissimilarités, dans l'intervalle [0,1] ?

dissimilarité = 1 – similarité normalisée sur [0,1]

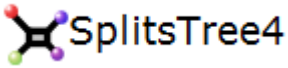
Normalisation des scores de similarité sur [0,1] :
- normalisation linéaire pour les matrices positives
- normalisation affines pour les matrices contenant des valeurs négatives, afin d'obtenir des distances dans l'intervalle [a,1] (a=0.1)

Construction de l'arbre

Plusieurs méthodes pour construire un arbre à partir d'une matrice de distances (classification hiérarchique) :

- Neighbor-Joining

Saitou & Nei, 1987



- Variantes d'Addtree

Barthelemy & Luong, 1987

- Heuristique des quadruplets

Cilibrasi & Vitanyi, 2007

Décoration de l'arbre

Tailles des mots :

- calculées directement à partir des **fréquences**
  (avec un log!)

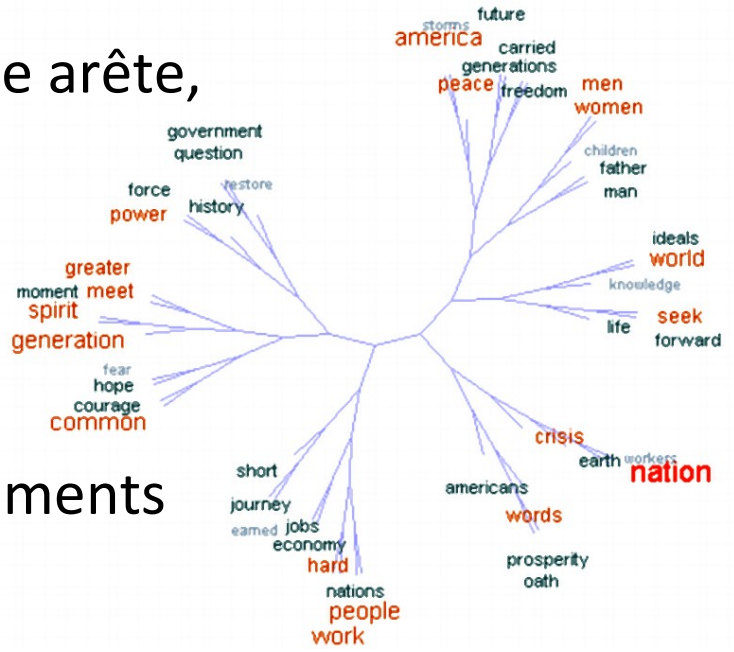
- calculées à partir des **rangs des fréquences**
  (distribution exponentielle)

- **score de spécificité** par rapport à un corpus de référence
  (TF-IDF, écart réduit...)

Dessin de l'arbre

Algorithme “equal angle” :

- montant pour calculer l'angle de chaque arête,
  en partant des feuilles
- descendant pour placer chaque arête
  en partant d'un sommet interne



Placement automatique des étiquettes :

→ heuristique pour éviter les chevauchements

Question des longueurs d'arêtes ?

Dessin de l'arbre

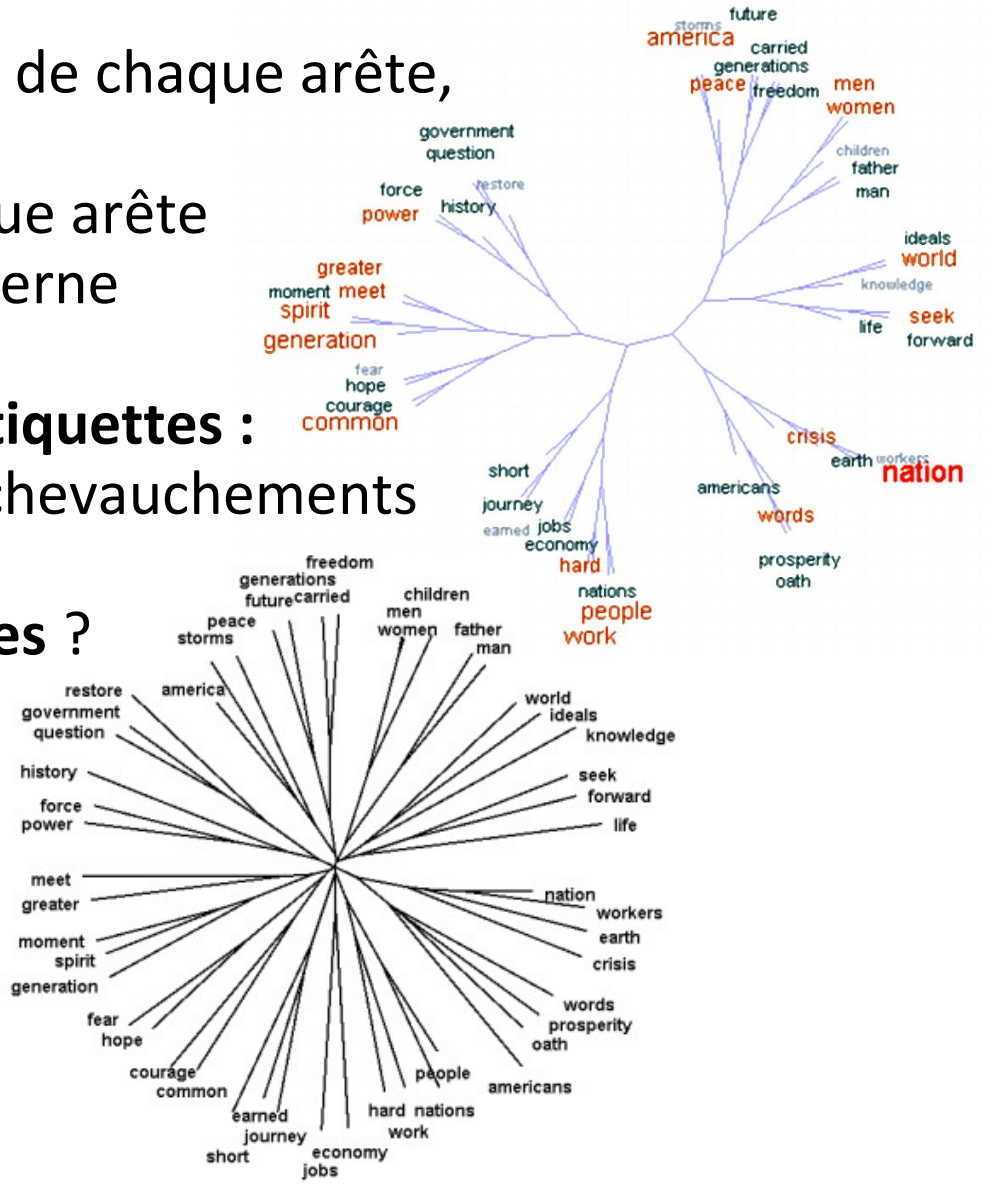
Algorithme “equal angle” :

- montant pour calculer l'angle de chaque arête,
  en partant des feuilles
- descendant pour placer chaque arête
  en partant d'un sommet interne

Placement automatique des étiquettes :

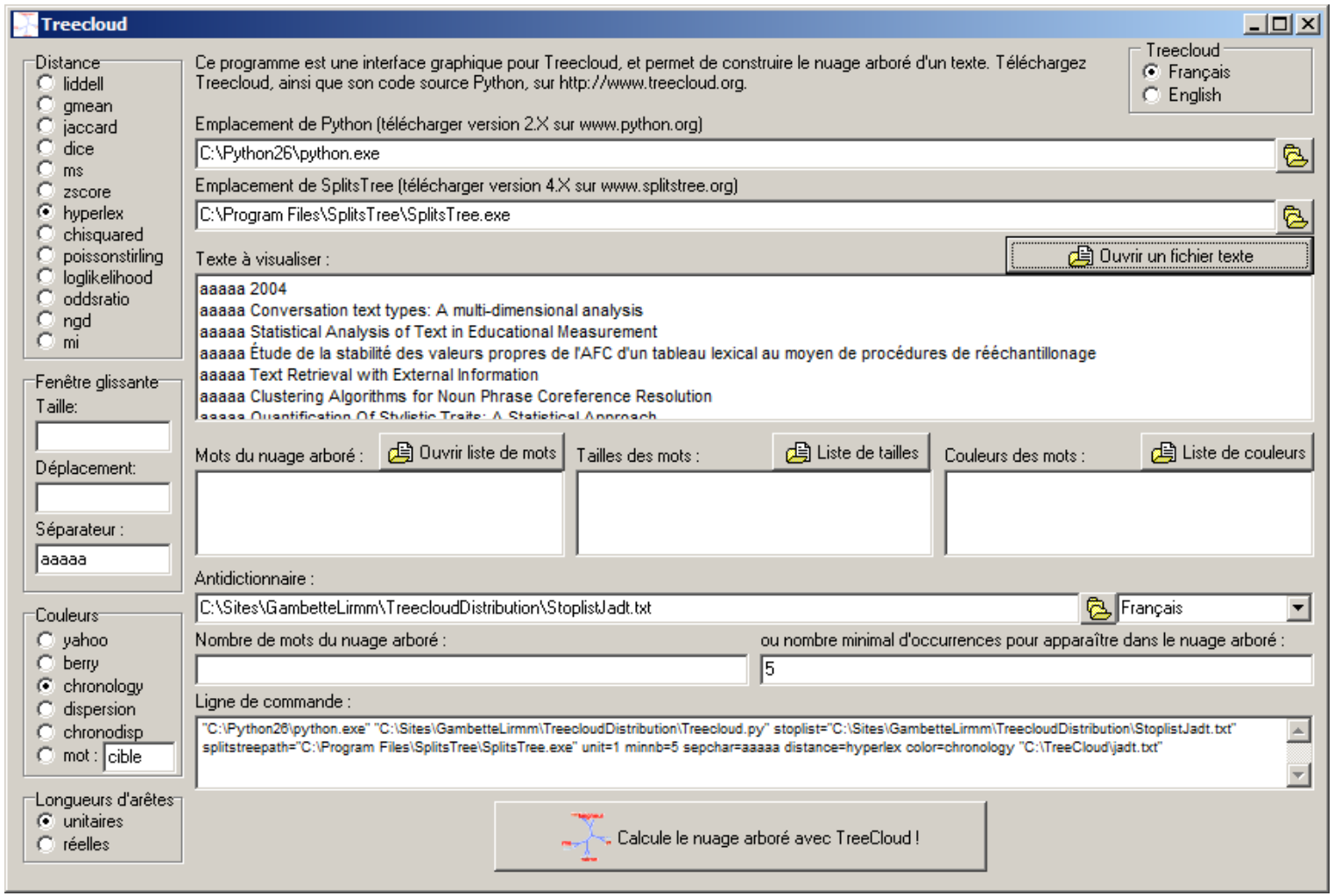
→ heuristique pour éviter les chevauchements

Question des longueurs d'arêtes ?



Implémentations

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

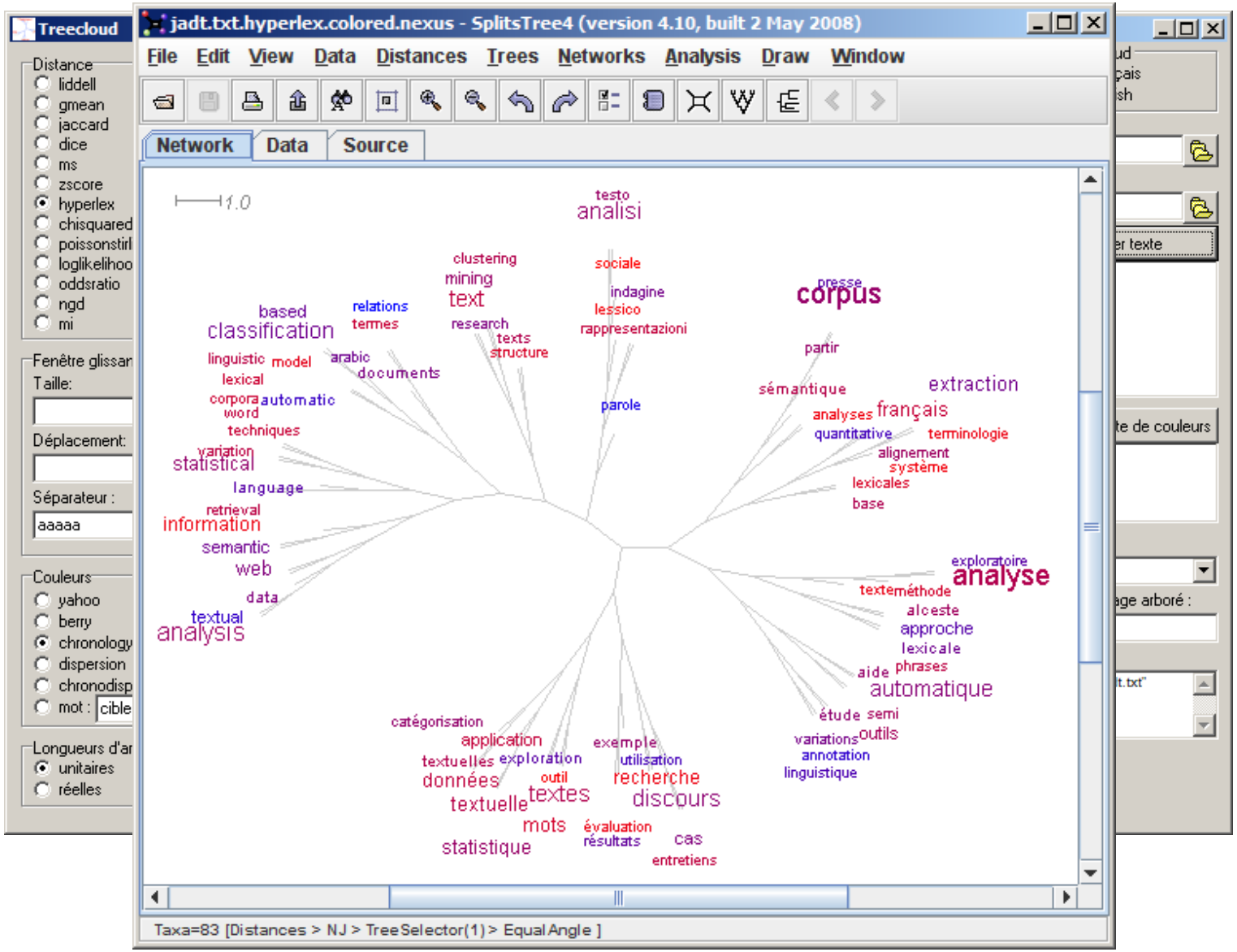


www.treecloud.org

www.splitstree.org

Implémentations

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



Interface web

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/

Interface web

www.treecloud.org

TreeCloud org
Nuage Arboré

Create! *Downloads* *Gallery* *Credits* *FAQ*
Créer! *Téléchargements* *Gallerie* *A propos* *FAQ*

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

Create your own tree cloud online

Ce site web vous permet de générer des n
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: *Visuali*
Classification as a Tool of Research, Proc. o
Societies), to appear, 2010 (supplementary m

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

TreeCloud org
Nuage Arboré

Créer! *Téléchargements* *Gallerie* *A propos* *FAQ*

Créez vos propres nuages arborés !

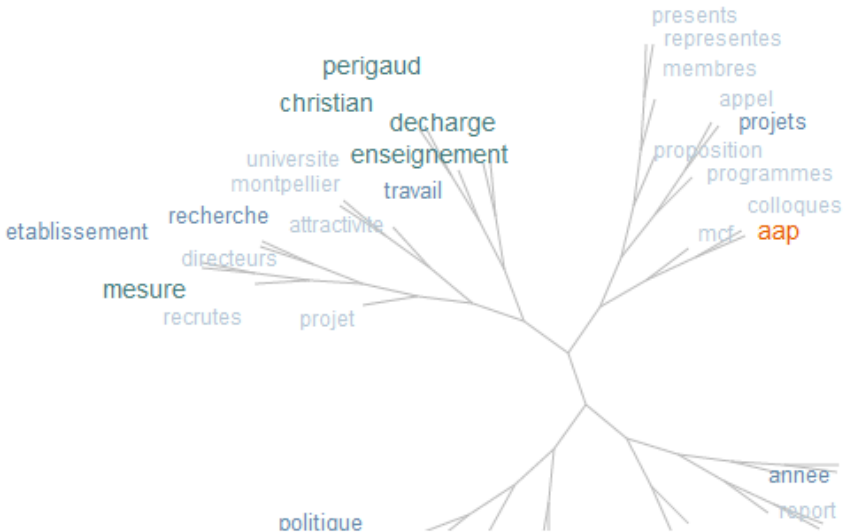
Collez votre texte dans le cadre ci-dessous, puis cliquez sur *Envoyer* ! Attention, l'utilisateur suivant verra votre texte quand il se connectera au site, si vous ne voulez pas faire apparaître vos textes, installez plutôt TreeCloud sur votre machine.

Texte :

conseil
scientifique
projet
proces
verbal

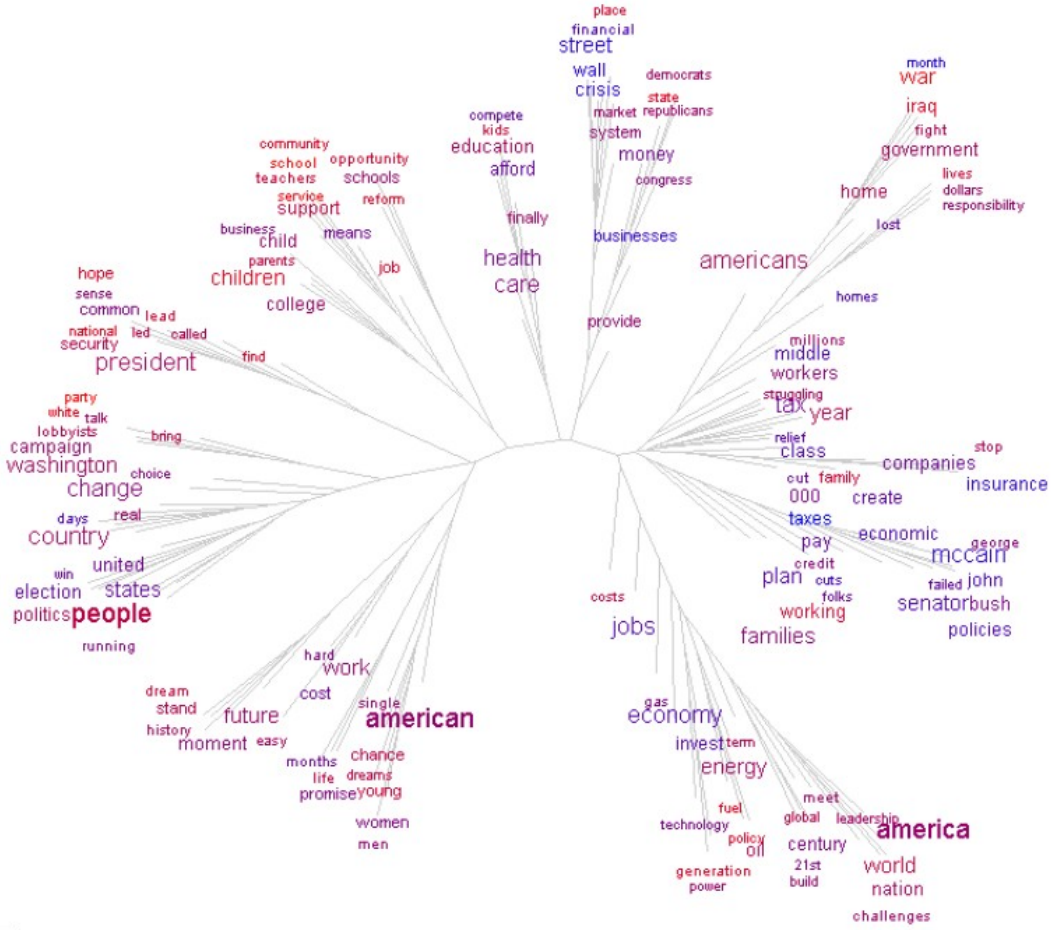
Envoyer

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.



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)

Interface web

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/

Plan

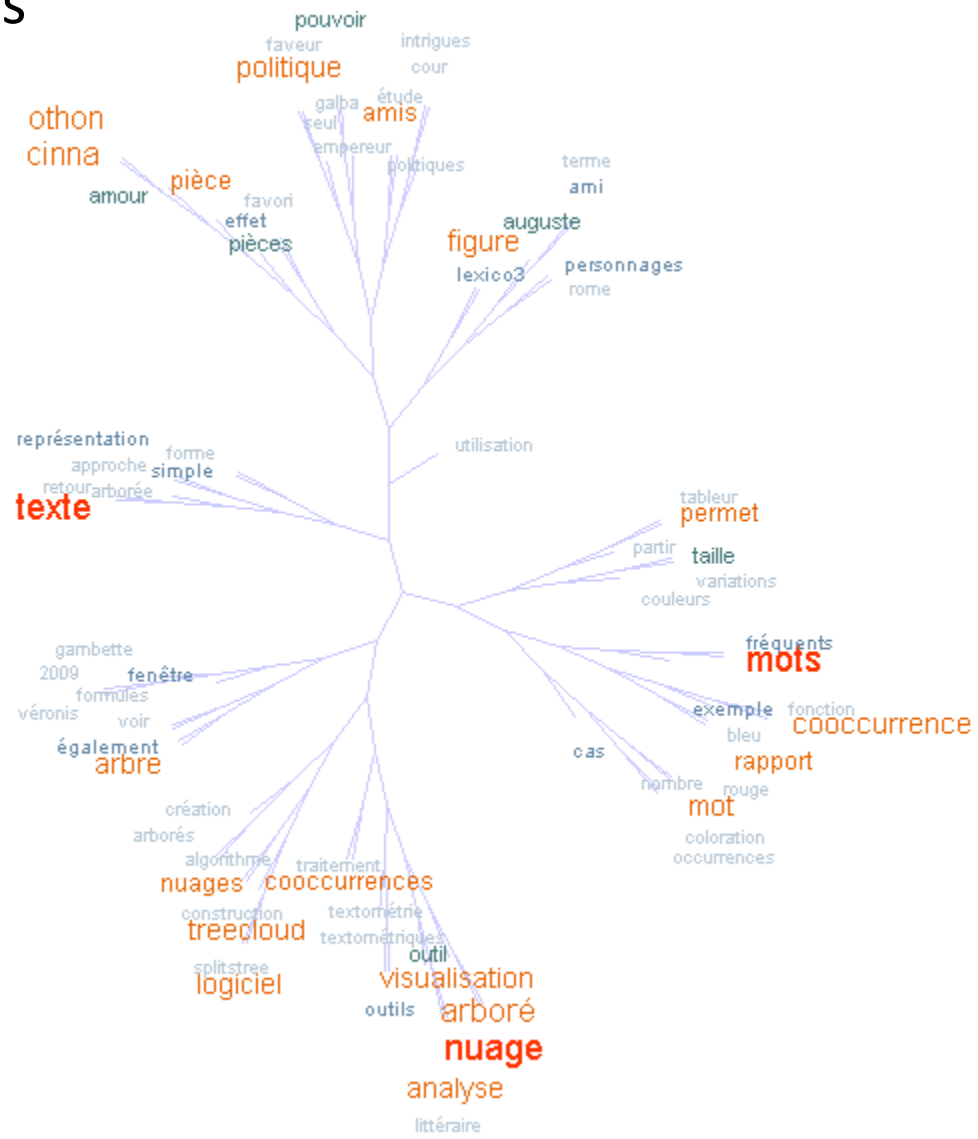
- Nuages arborés, intérêts et limites
- Construction des nuages arborés
- Options de coloration
- Utilisation des nuages arborés
- Évaluation de qualité
- Perspectives

Des couleurs pour guider la lecture

- coloration selon les fréquences
- coloration chronologique
- coloration de la dispersion
- coloration ciblée sur un mot
- coloration grammaticale

Des couleurs pour guider la lecture

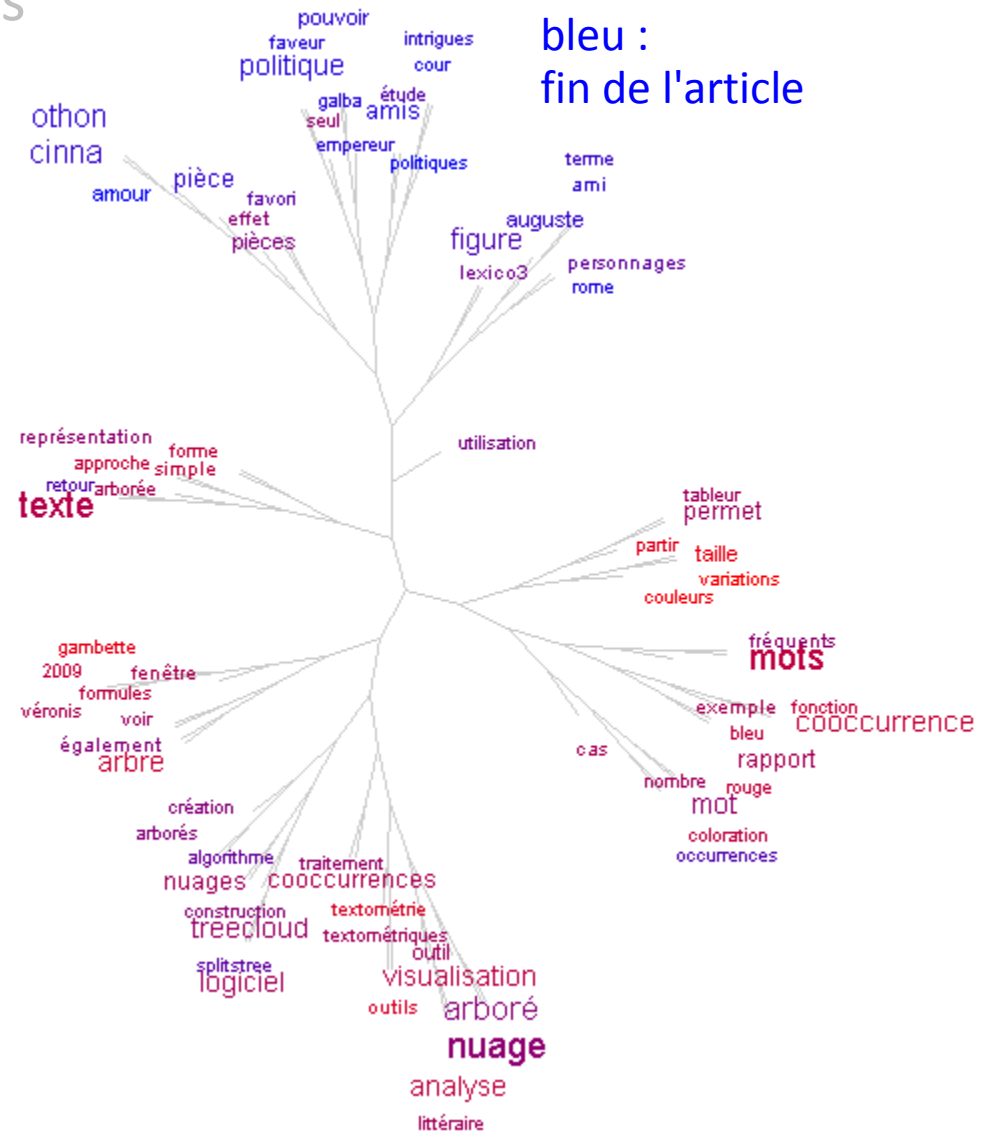
- coloration selon les fréquences
- coloration chronologique
- coloration de la dispersion
- coloration ciblée sur un mot
- coloration grammaticale



Nuage arboré des mots apparaissant 5 fois ou plus dans l'article d'Amstutz & Gambette, JADT 2010, distance Liddell, fenêtre de 20 mots, coloration Yahoo

Des couleurs pour guider la lecture

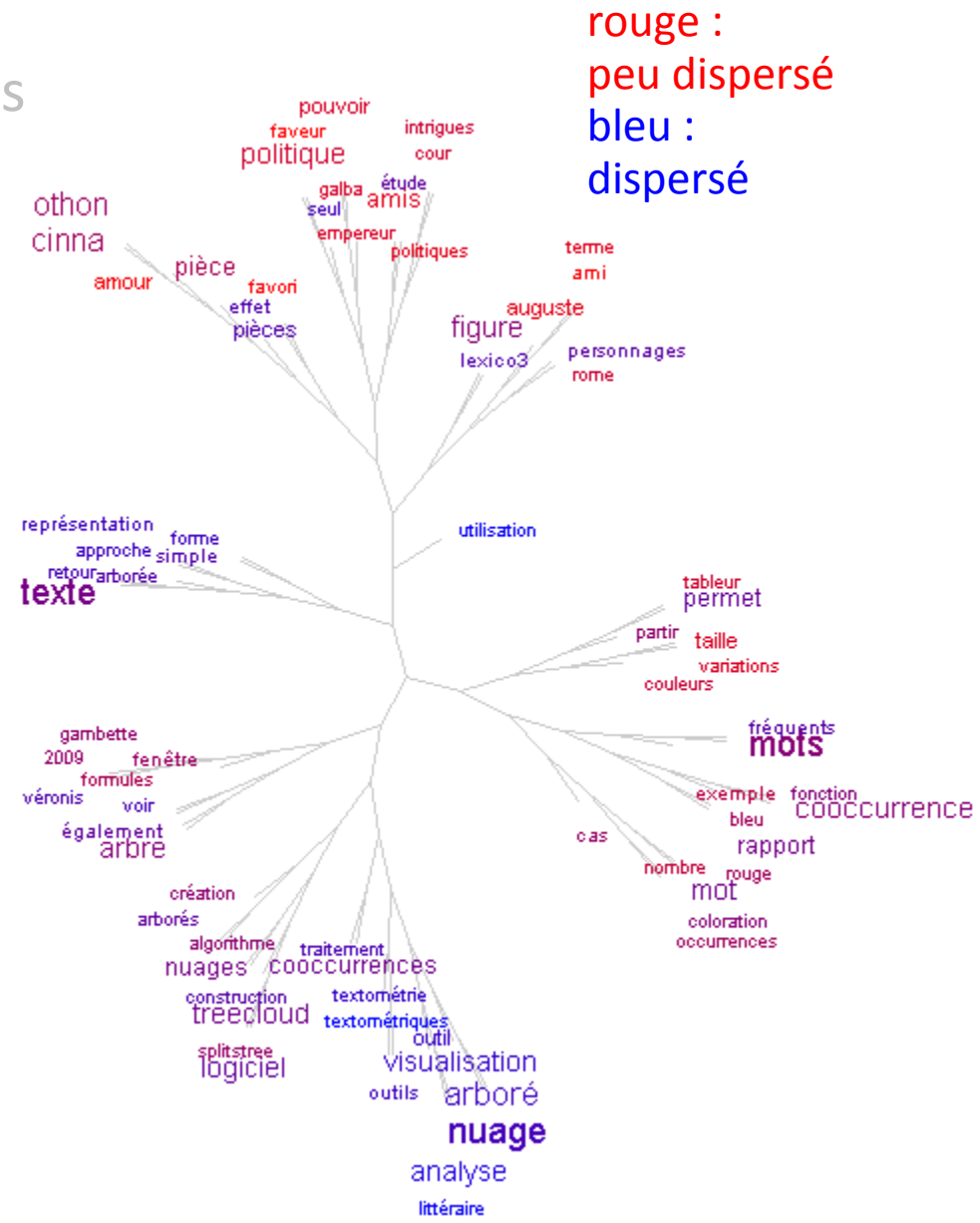
- coloration selon les fréquences
- coloration chronologique
- coloration de la dispersion
- coloration ciblée sur un mot
- coloration grammaticale



Nuage arboré des mots apparaissant 5 fois ou plus dans l'article d'Amstutz & Gambette, JADT 2010, distance Liddell, fenêtre de 20 mots, coloration chronologique

Des couleurs pour guider la lecture

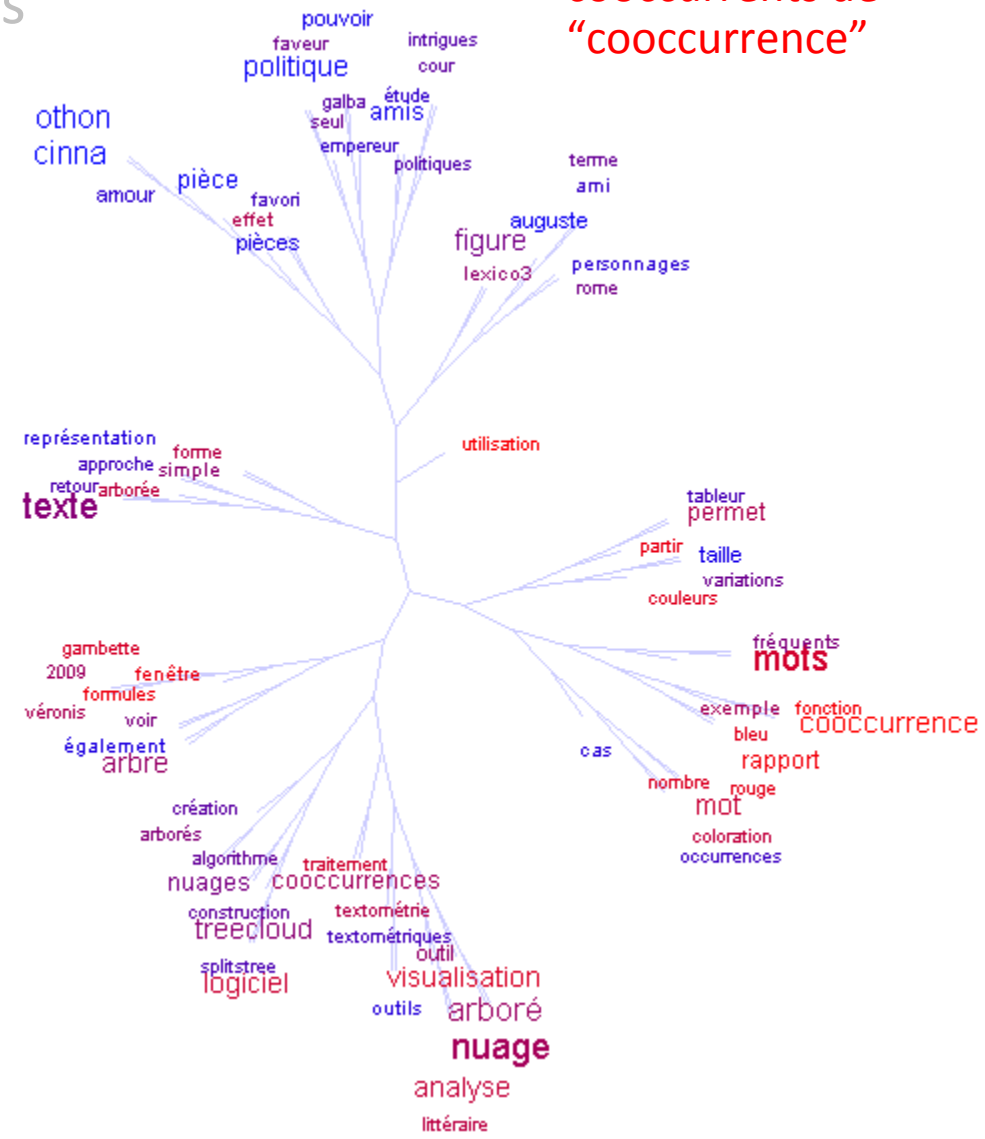
- coloration selon les fréquences
- coloration chronologique
- coloration de la dispersion
- coloration ciblée sur un mot
- coloration grammaticale



Nuage arboré des mots apparaissant 5 fois ou plus dans l'article d'Amstutz & Gambette, JADT 2010, distance Liddell, fenêtre de 20 mots, coloration disperson

Des couleurs pour guider la lecture

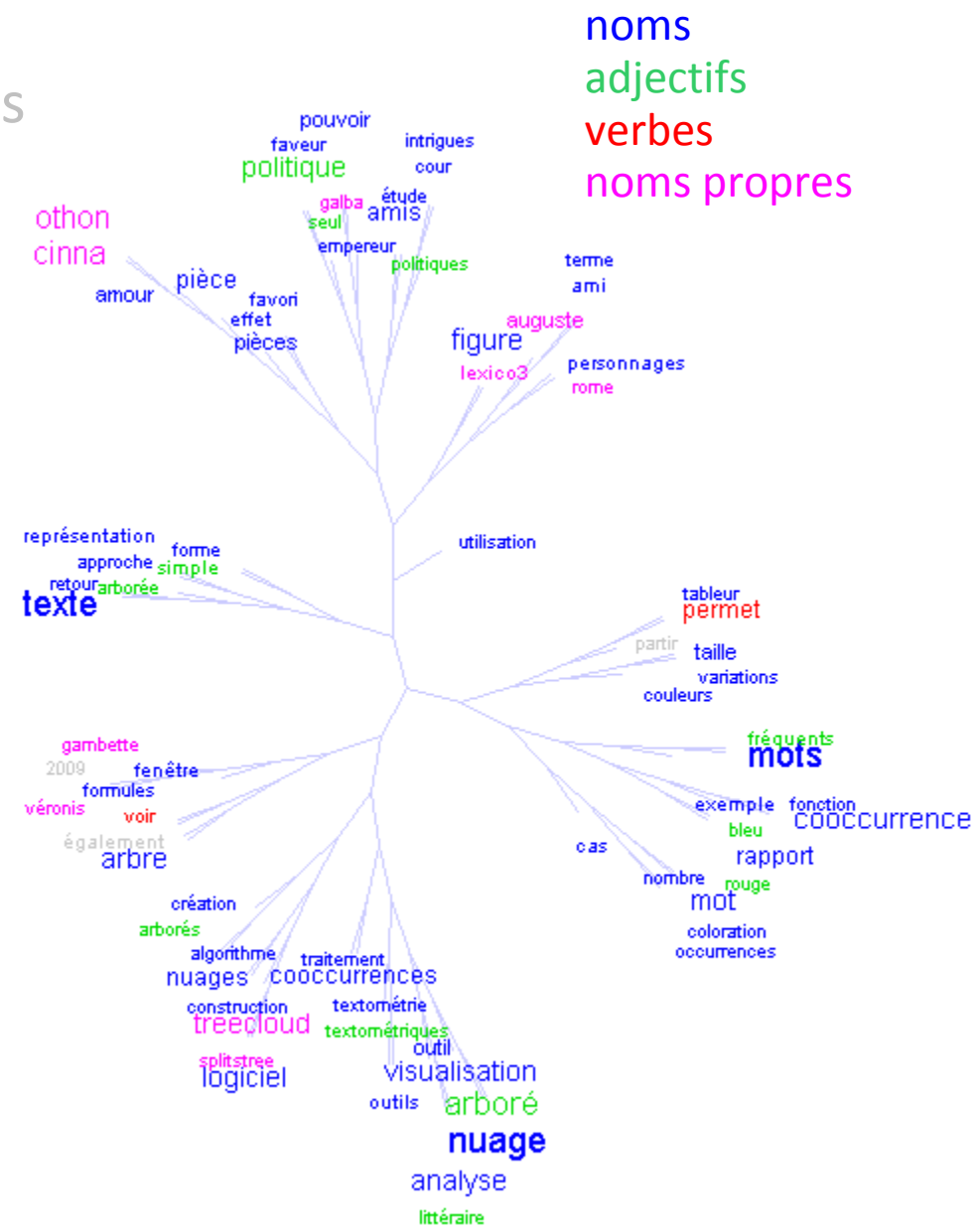
- coloration selon les fréquences
- coloration chronologique
- coloration de la dispersion
- coloration ciblée sur un mot
- coloration grammaticale



Nuage arboré des mots apparaissant 5 fois ou plus dans l'article d'Amstutz & Gambette, JADT 2010, distance Liddell, fenêtre de 20 mots, coloration ciblée sur le mot “cooccurrence”

Des couleurs pour guider la lecture

- coloration selon les fréquences
- coloration chronologique
- coloration de la dispersion
- coloration ciblée sur un mot
- coloration grammaticale



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

Plan

- Nuages arborés, intérêts et limites
- Construction des nuages arborés
- Options de coloration
- Utilisations des nuages arborés
- Évaluation de qualité
- Perspectives

Utilisations des nuages arborés

- **Résumé visuel** des thématiques d'un texte

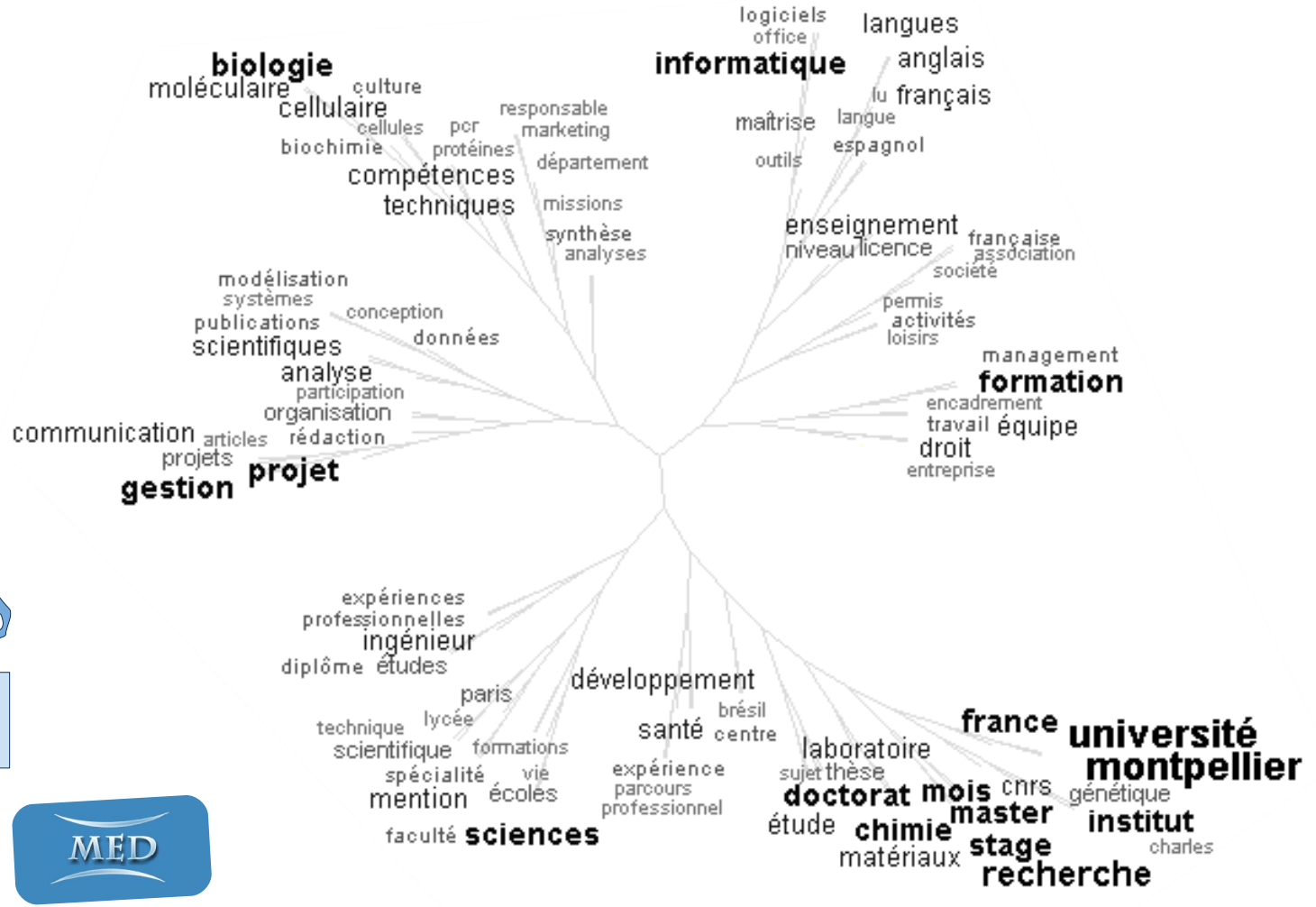
- Support de **médiation** et d'**argumentation** :
  - **appui visuel** d'une analyse subjective
  - **clarification** de rapports ou discours, lors de la rédaction

- En **analyse textuelle** (réponses aux questions ouvertes, romans, théâtre, corpus médiatique, etc.) :
  - susciter, formaliser et étayer des **hypothèses de travail**
  - **comparer des textes** selon leur représentation arborée
  - hiérarchiser l'utilisation d'**autres outils textométriques**
  - représenter les **résultats de l'analyse**

Support de médiation et d'argumentation

Présenter les compétences des docteurs

Corpus : CV soumis à une rencontre docteurs-entreprises



Rencontre Docteurs & Entreprises

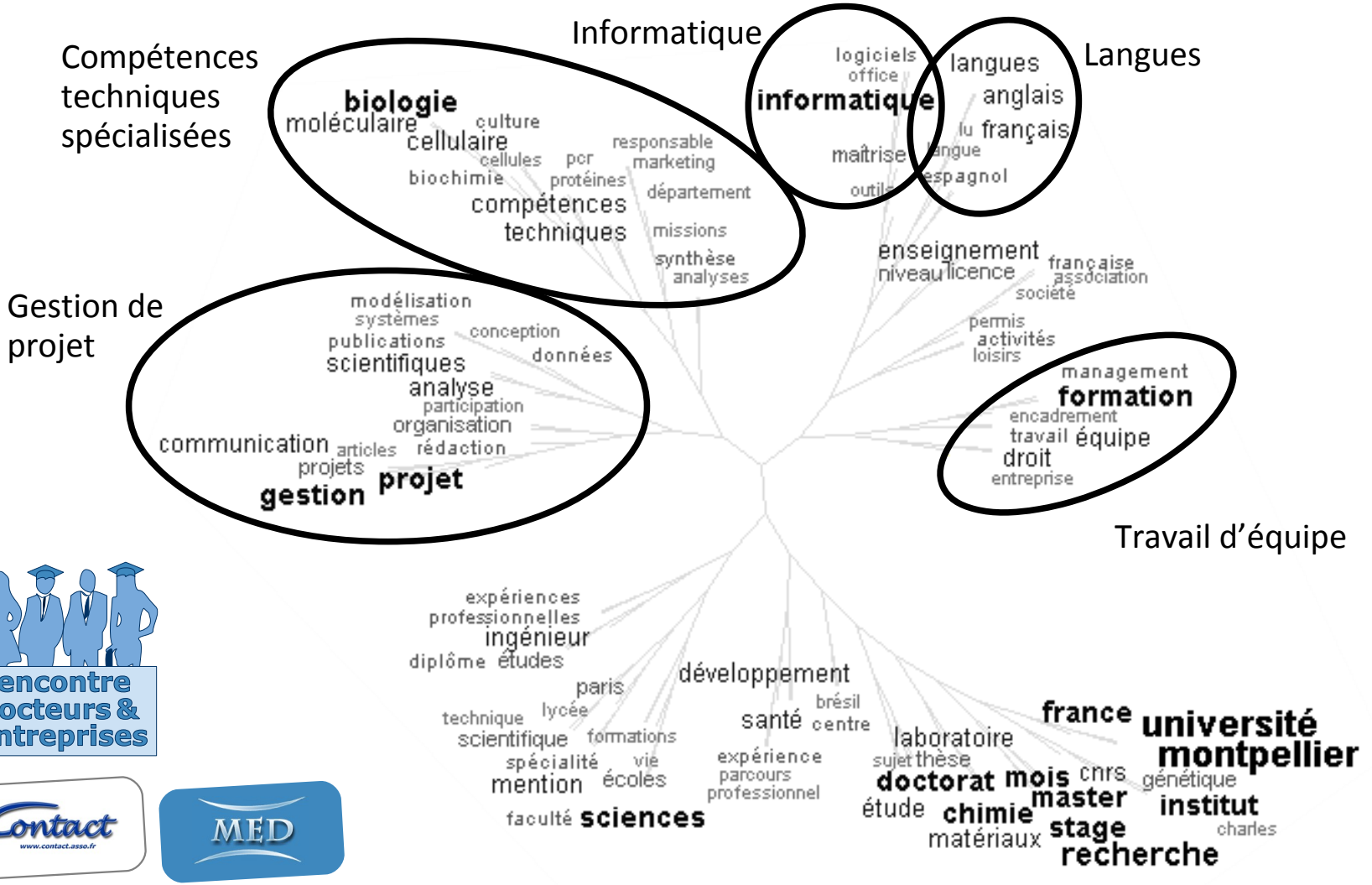
Contact
www.contact.asso.fr

MED

Support de médiation et d'argumentation

Présenter les compétences des docteurs

Corpus : CV soumis à une rencontre docteurs-entreprises



Rencontre Docteurs & Entreprises

Contact
www.contact.asso.fr

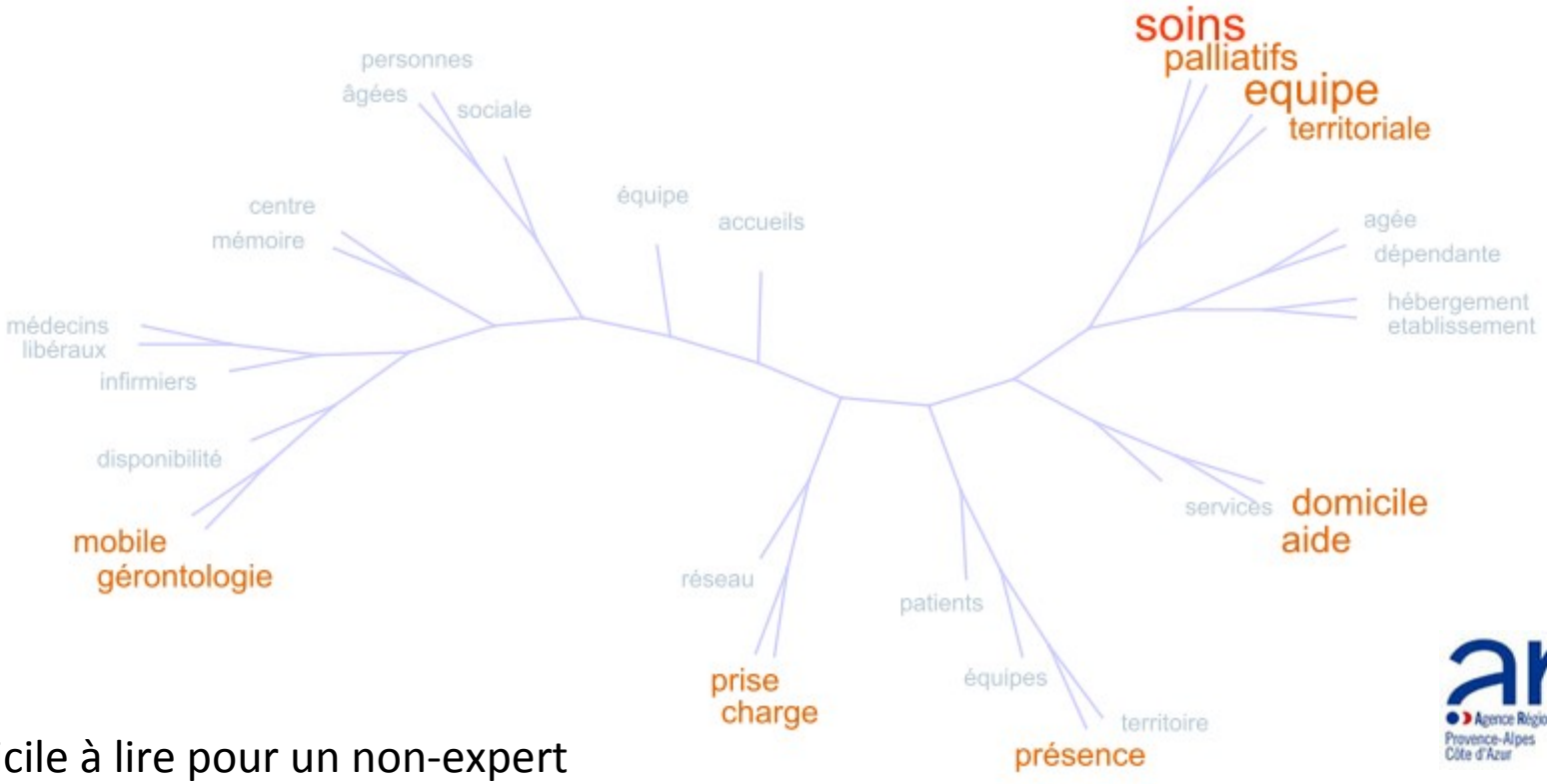
MED

Support de médiation et d'argumentation

Transmettre les résultats d'une consultation

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

Points forts de l'accueil dans le département



Difficile à lire pour un non-expert
Support cohérent d'accompagnement du discours

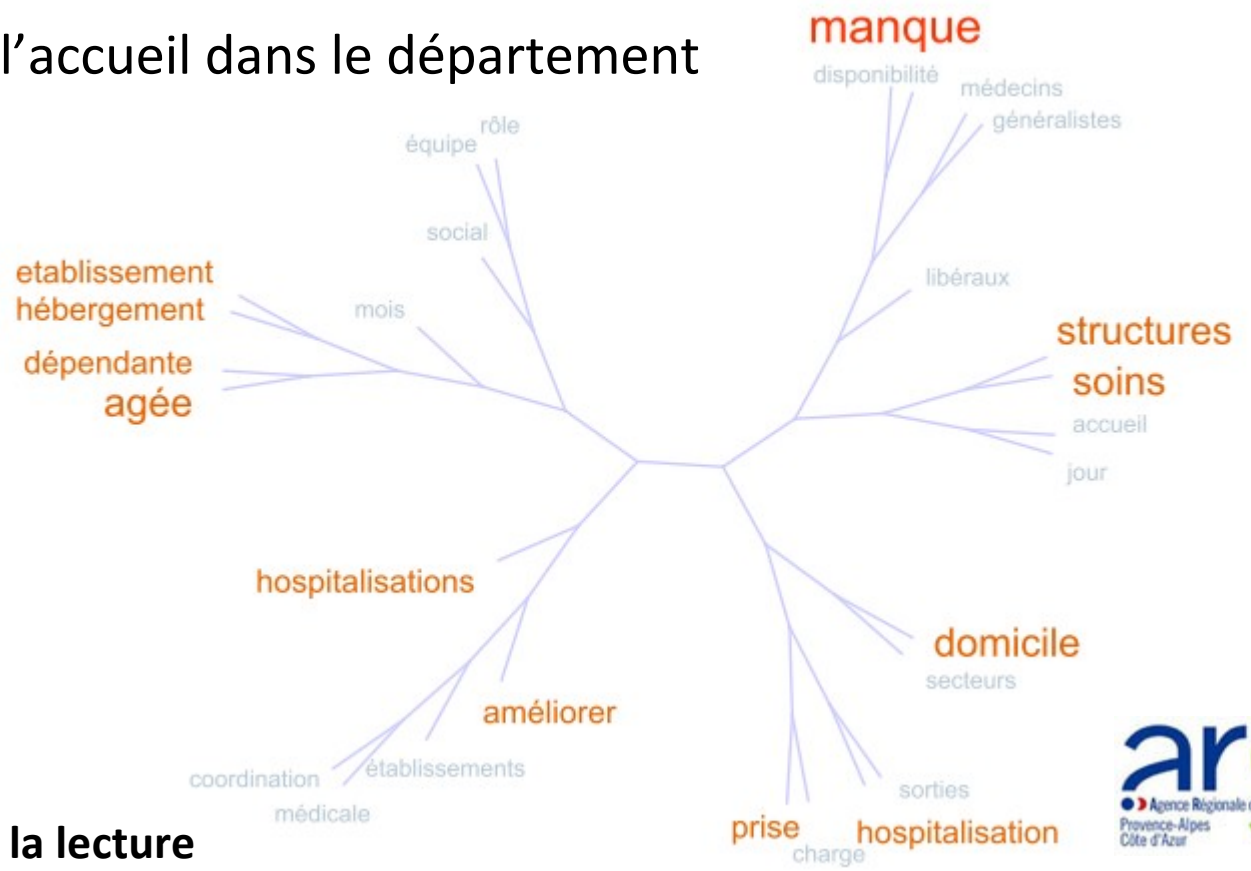
P. Grenier-Tisserand et
l'équipe projet de la DT 04

Support de médiation et d'argumentation

Transmettre les résultats d'une consultation

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

Points faibles de l'accueil dans le département



Modaux qui **guident la lecture**
Support cohérent d'accompagnement du discours

P. Grenier-Tisserand et
l'équipe projet de la DT 04

Support de médiation et d'argumentation

Transmettre les résultats d'une consultation

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

Points faibles de l'accueil dans le département

Modaux qui **guident la lecture**
Support cohérent d'accompagnement du discours

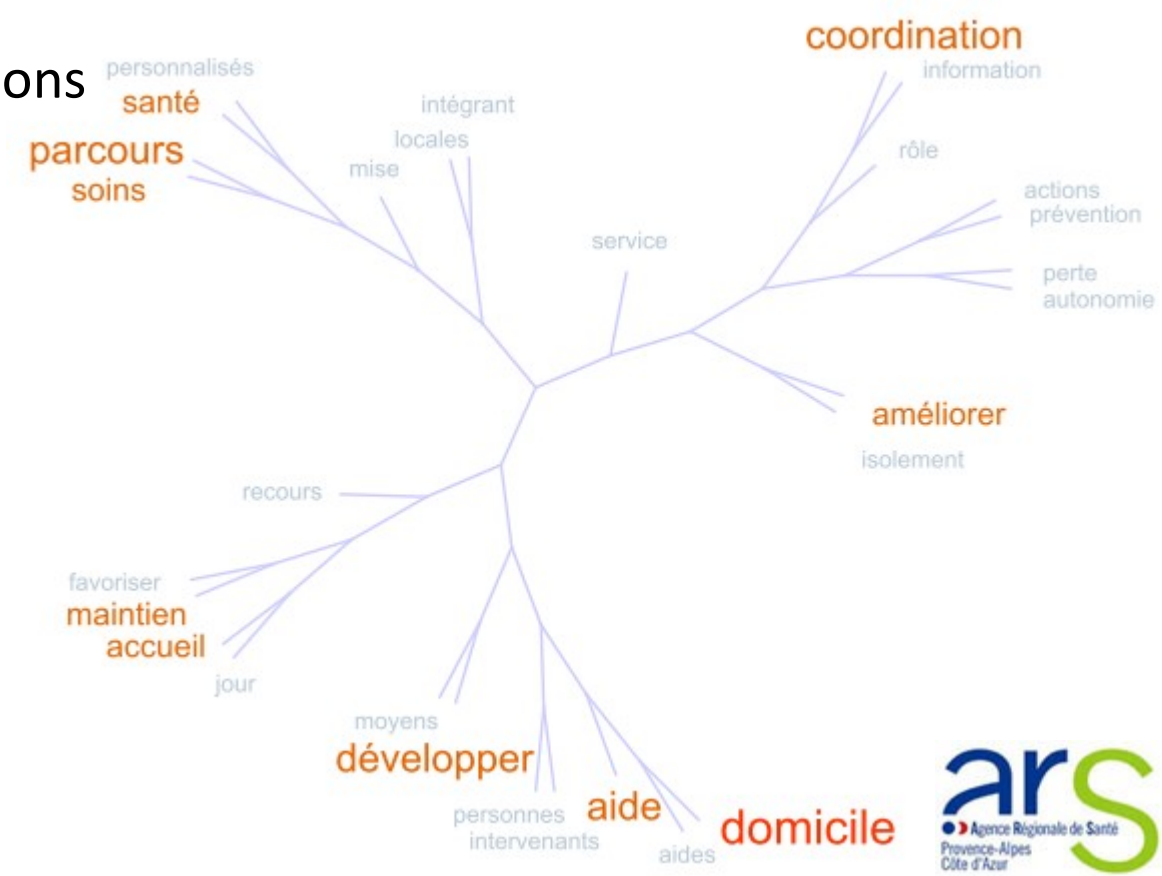
P. Grenier-Tisserand et
l'équipe projet de la DT 04

Support de médiation et d'argumentation

Transmettre les résultats d'une consultation

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



Verbes qui **guident la lecture**
Support cohérent d'accompagnement du discours

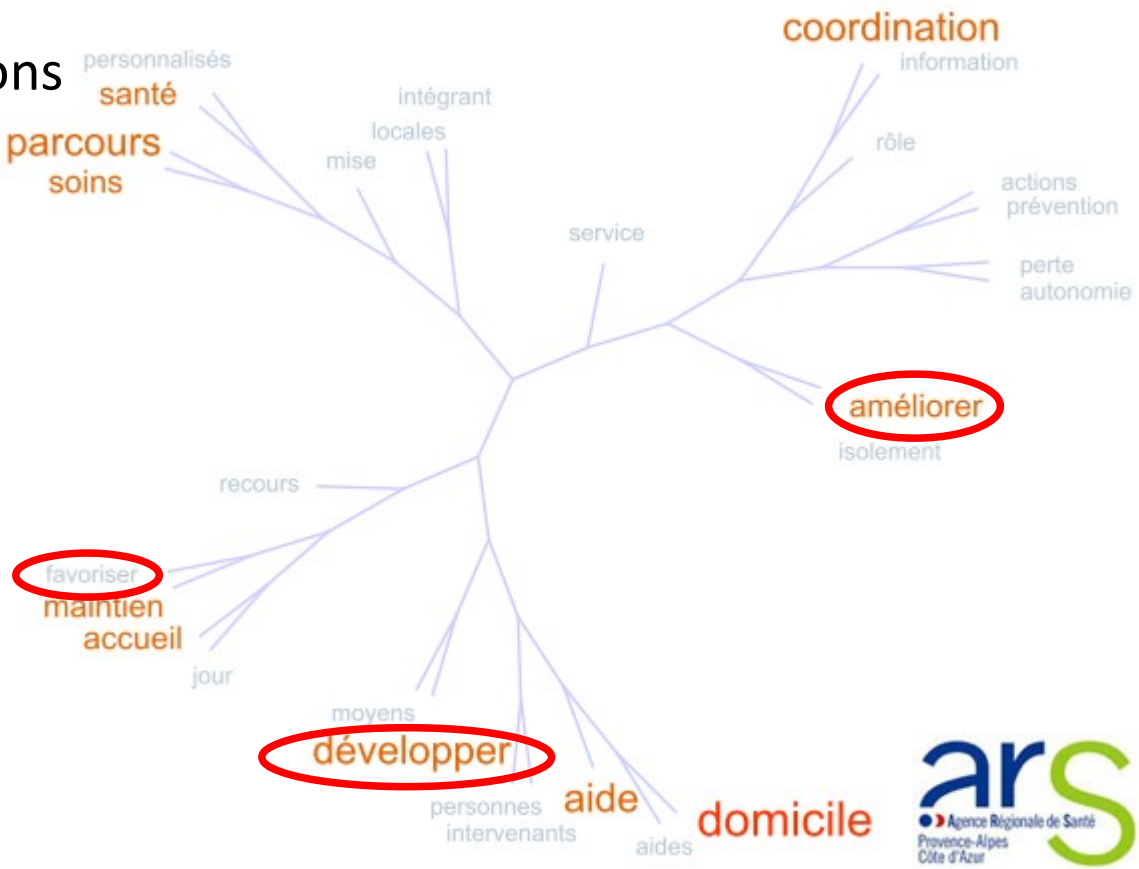
P. Grenier-Tisserand et
l'équipe projet de la DT 04

Support de médiation et d'argumentation

Transmettre les résultats d'une consultation

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

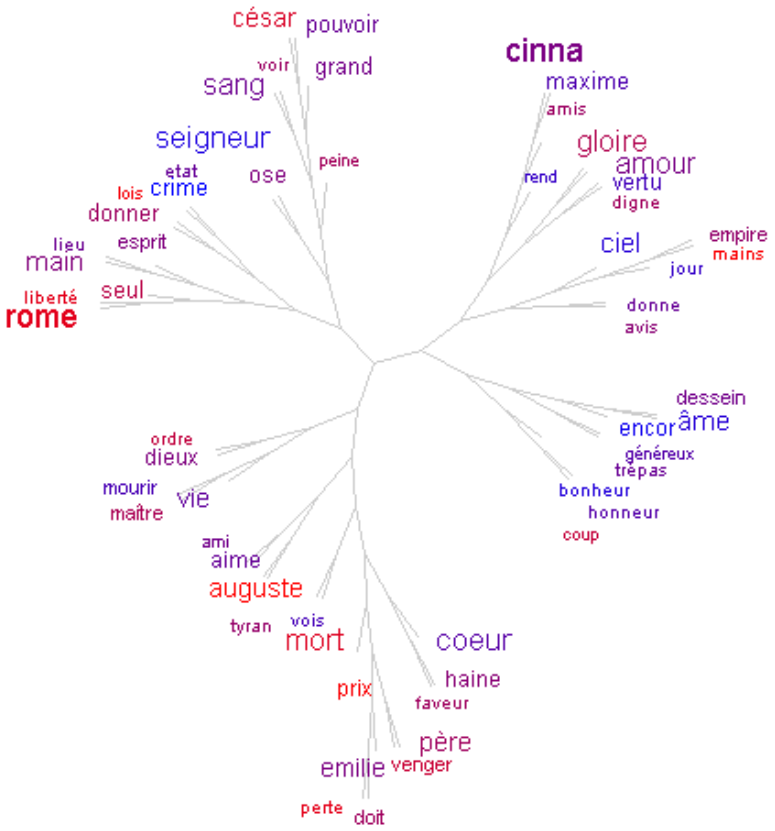
Suggestions d'améliorations



Verbes qui **guident la lecture**
Support cohérent d'accompagnement du discours

P. Grenier-Tisserand et
l'équipe projet de la DT 04

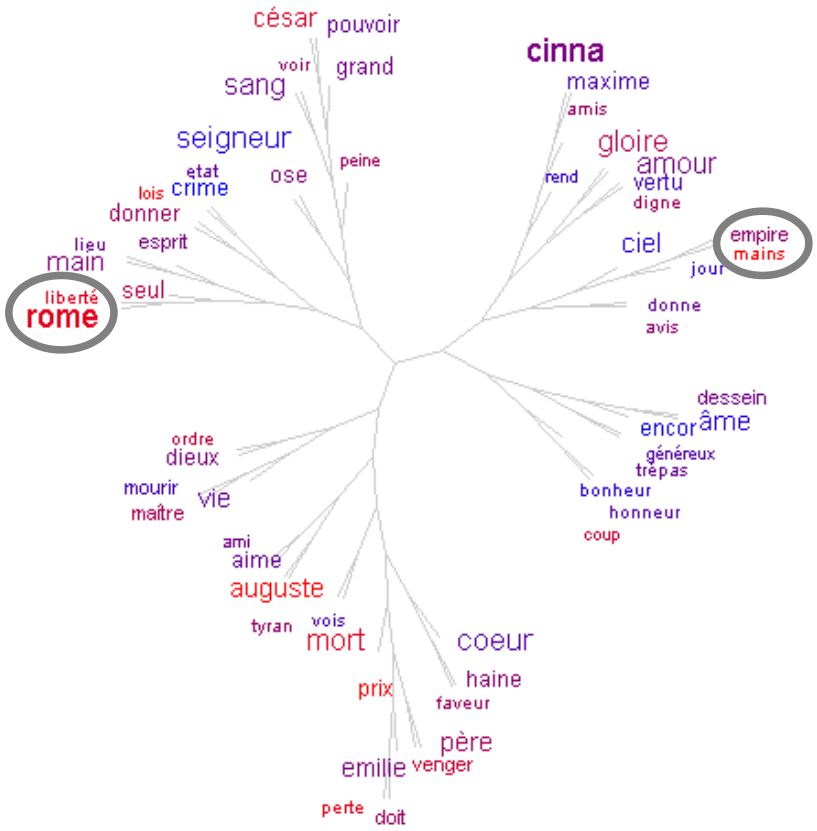
Analyse littéraire : illustration sur *Cinna*



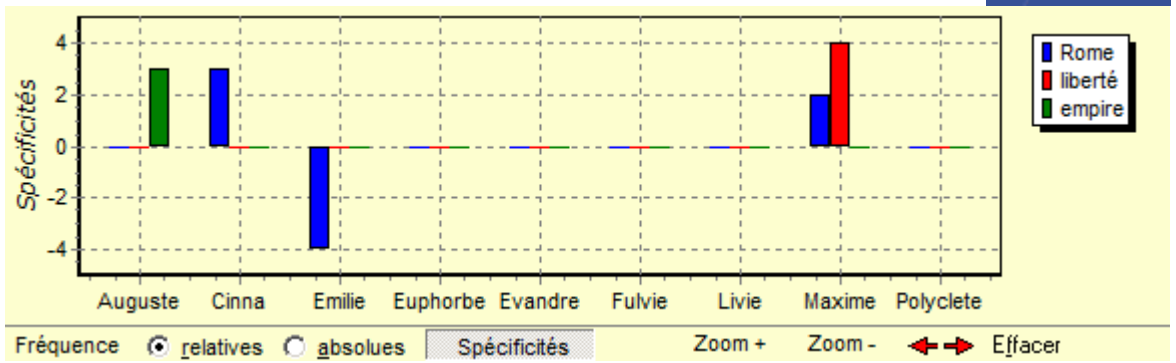
Nuages arborés globaux des 60 mots les plus fréquents dans Cinna de Corneille (distance Liddell, fenêtre de largeur 20), colorés chronologiquement (rouge au début, bleu à la fin)

Amstutz & Gambette, 2010

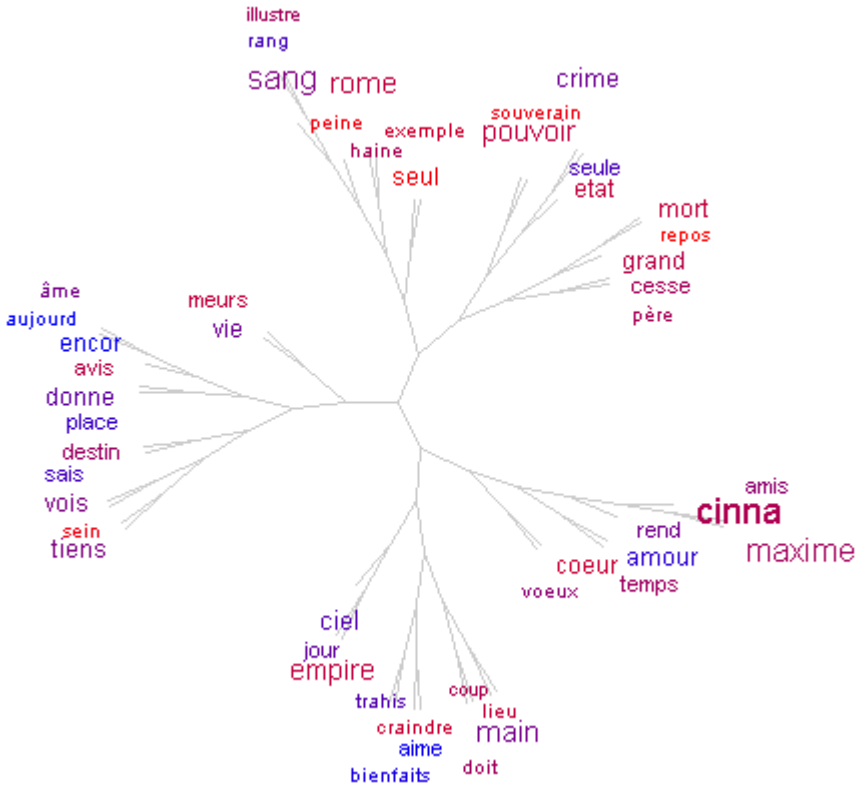
Analyse littéraire : illustration sur *Cinna*



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

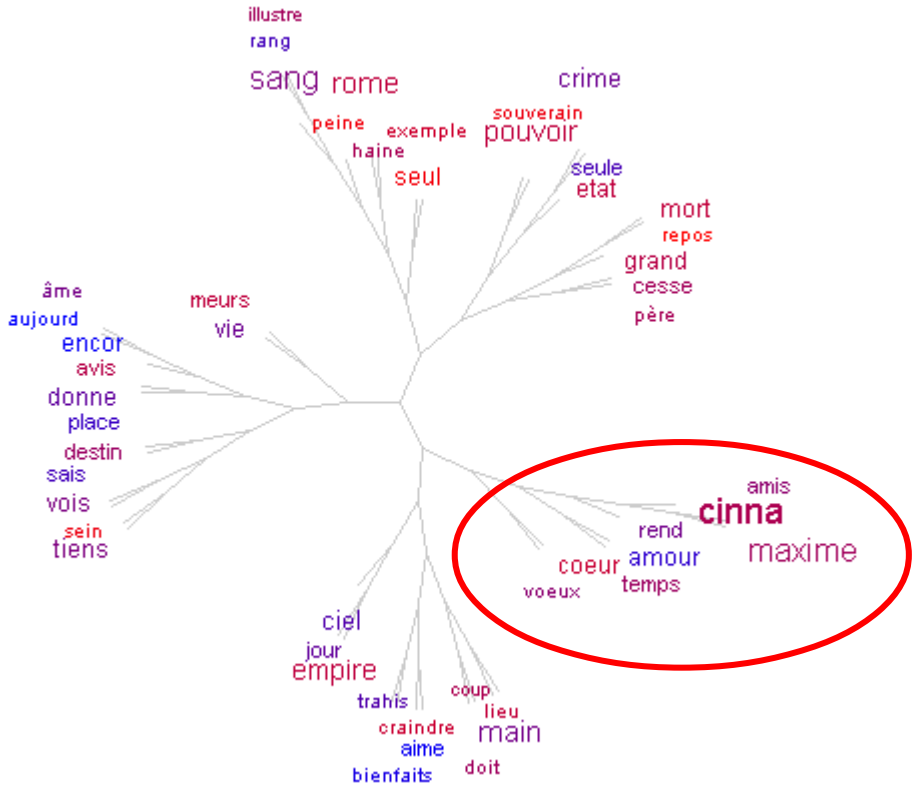


Analyse littéraire : illustration sur *Cinna*



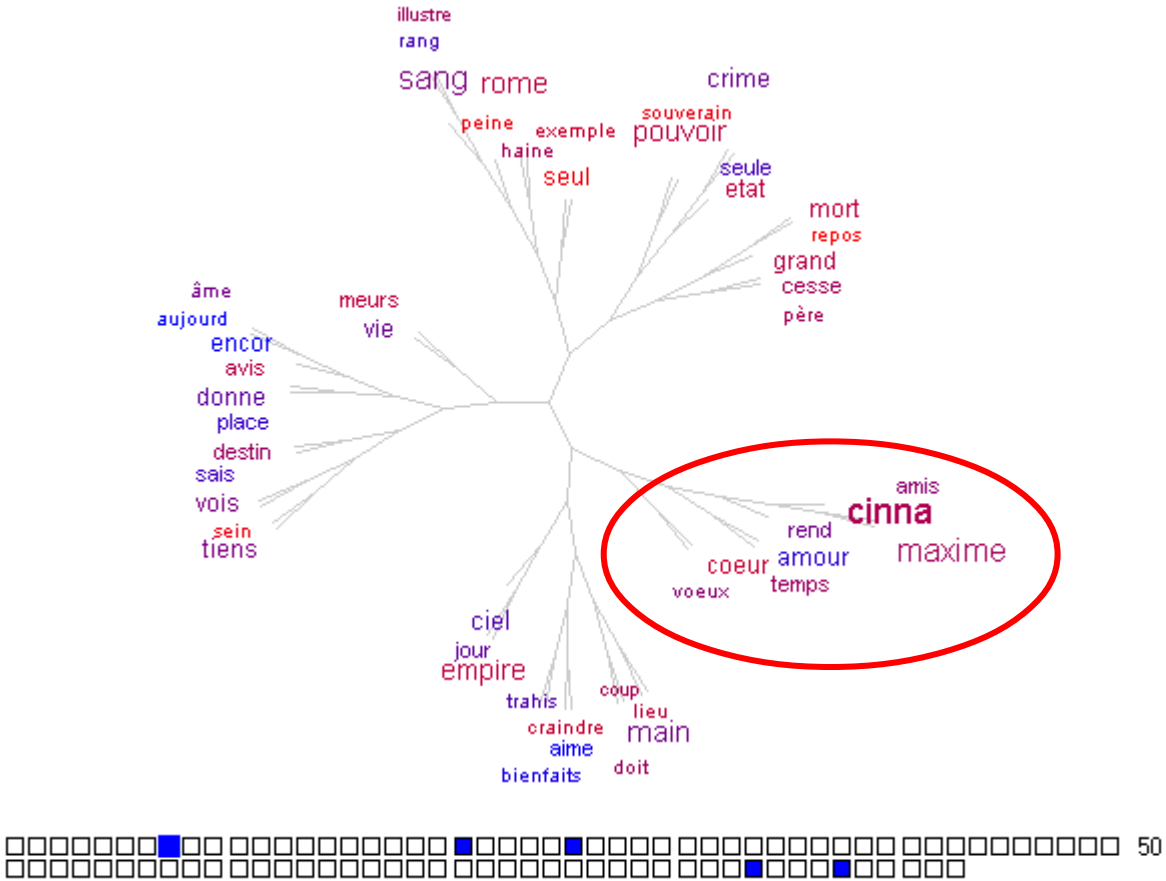
Nuage arboré des 50 mots les plus fréquents des paroles d'Auguste dans Cinna

Analyse littéraire : illustration sur *Cinna*



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

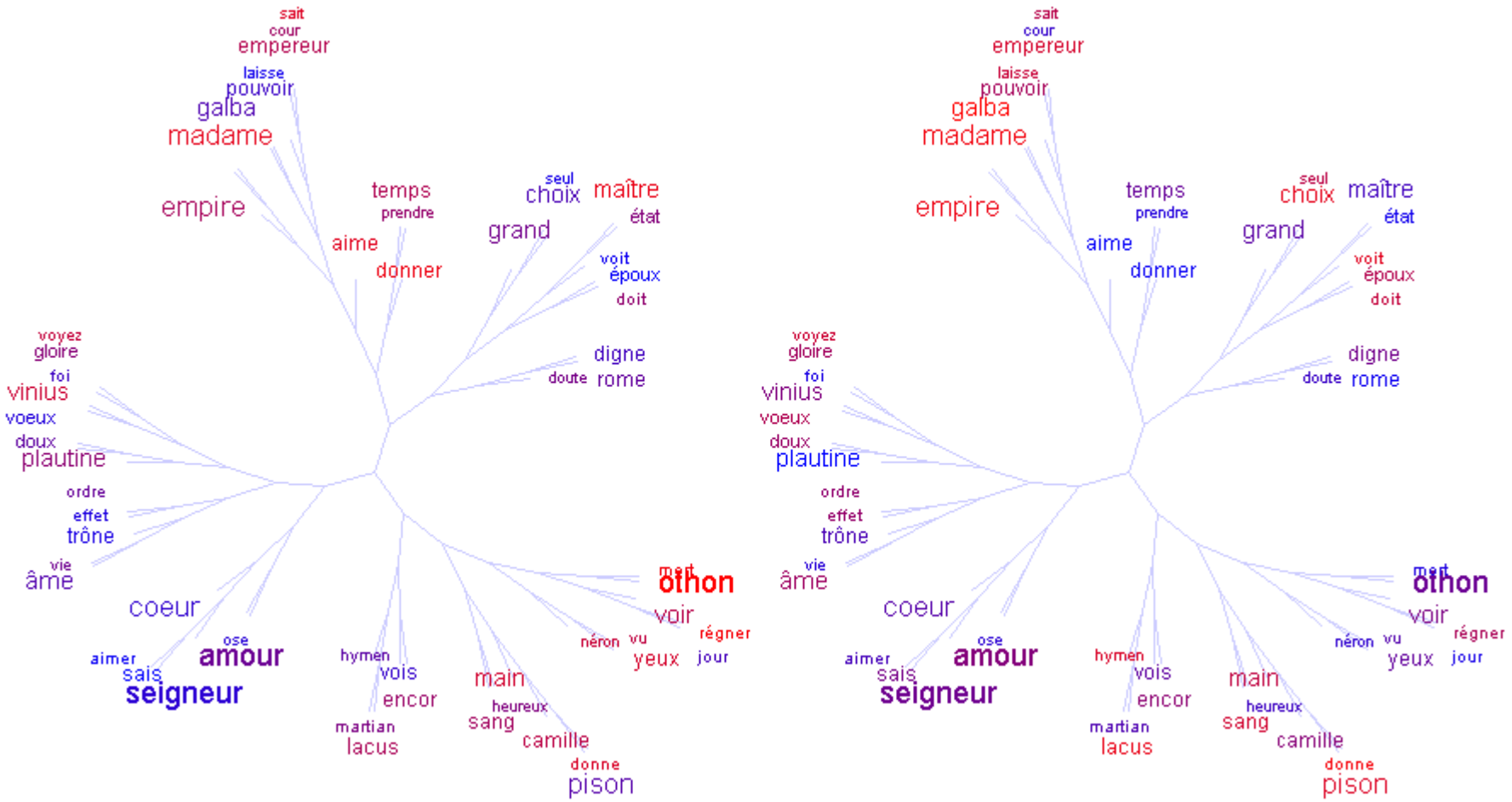
Analyse littéraire : illustration sur *Cinna*



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

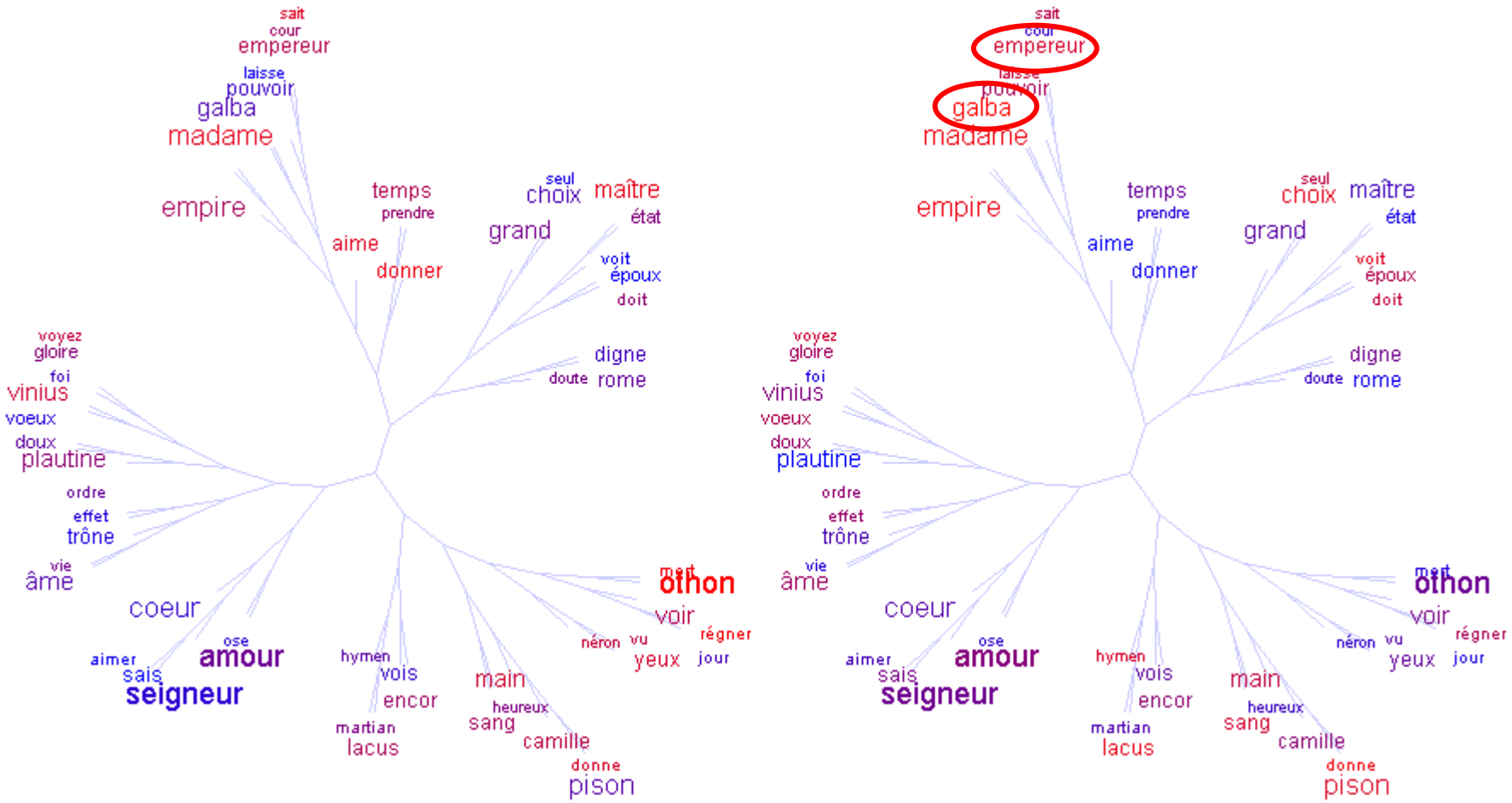
1. *Voilà, mes chers **amis**, ce qui me met en peine.*
2. *Quoi ! mes plus chers **amis** ! quoi ! Cinna ! quoi ! Maxime !*
3. *Reprenez le pouvoir que vous m'avez commis, Si donnant des sujets il ôte les **amis***
4. *Soyons **amis**, Cinna, c'est moi qui t'en convie*
5. *Il nous a trahis tous ; mais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

Analyse littéraire : illustration sur *Othon*



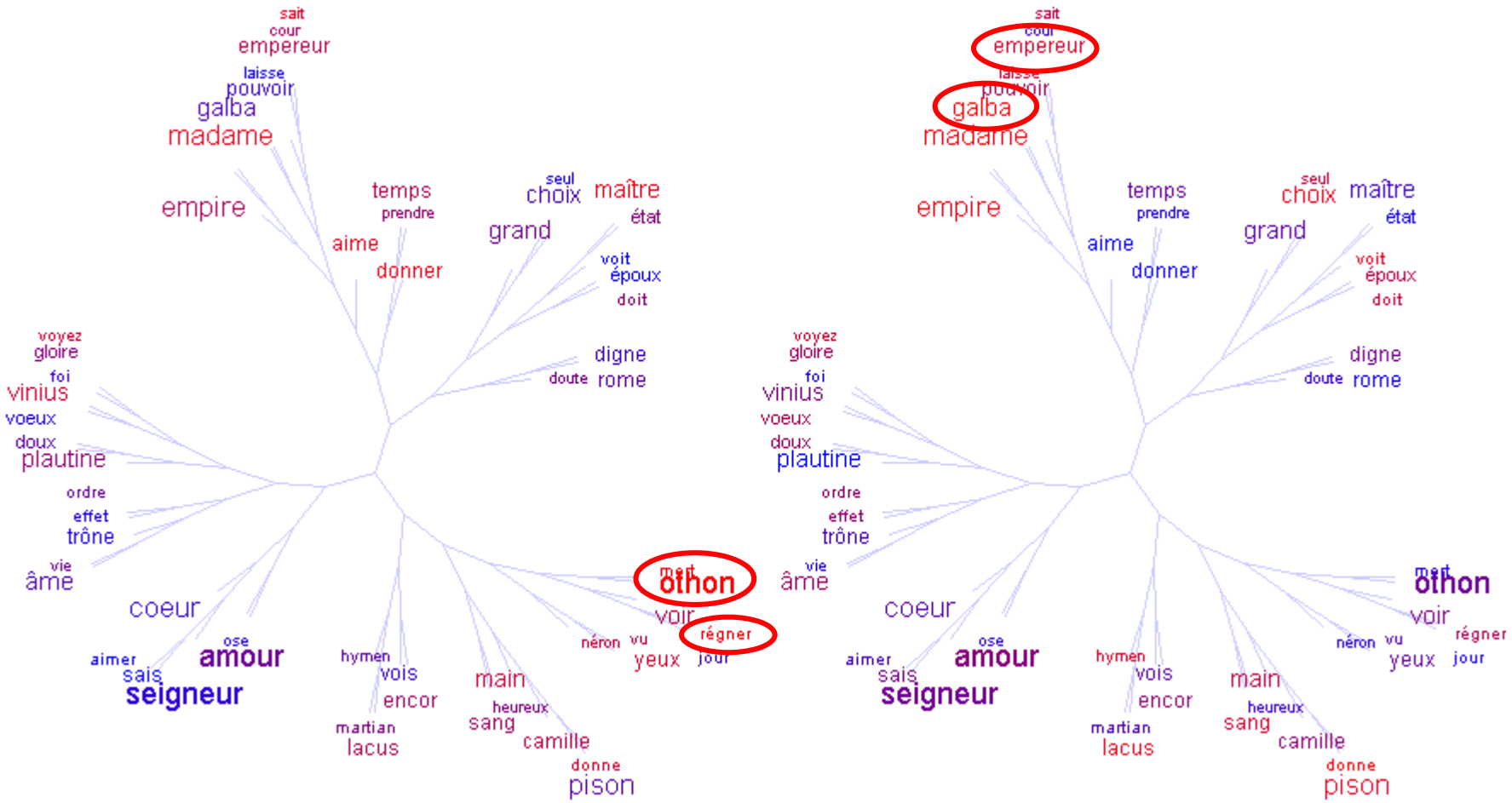
Nuage arboré des 30 mots les plus fréquents de la pièce Othon, coloré à gauche par rapport aux cooccurrences avec « Othon », à droite par rapport à celles avec « Galba »

Analyse littéraire : illustration sur *Othon*



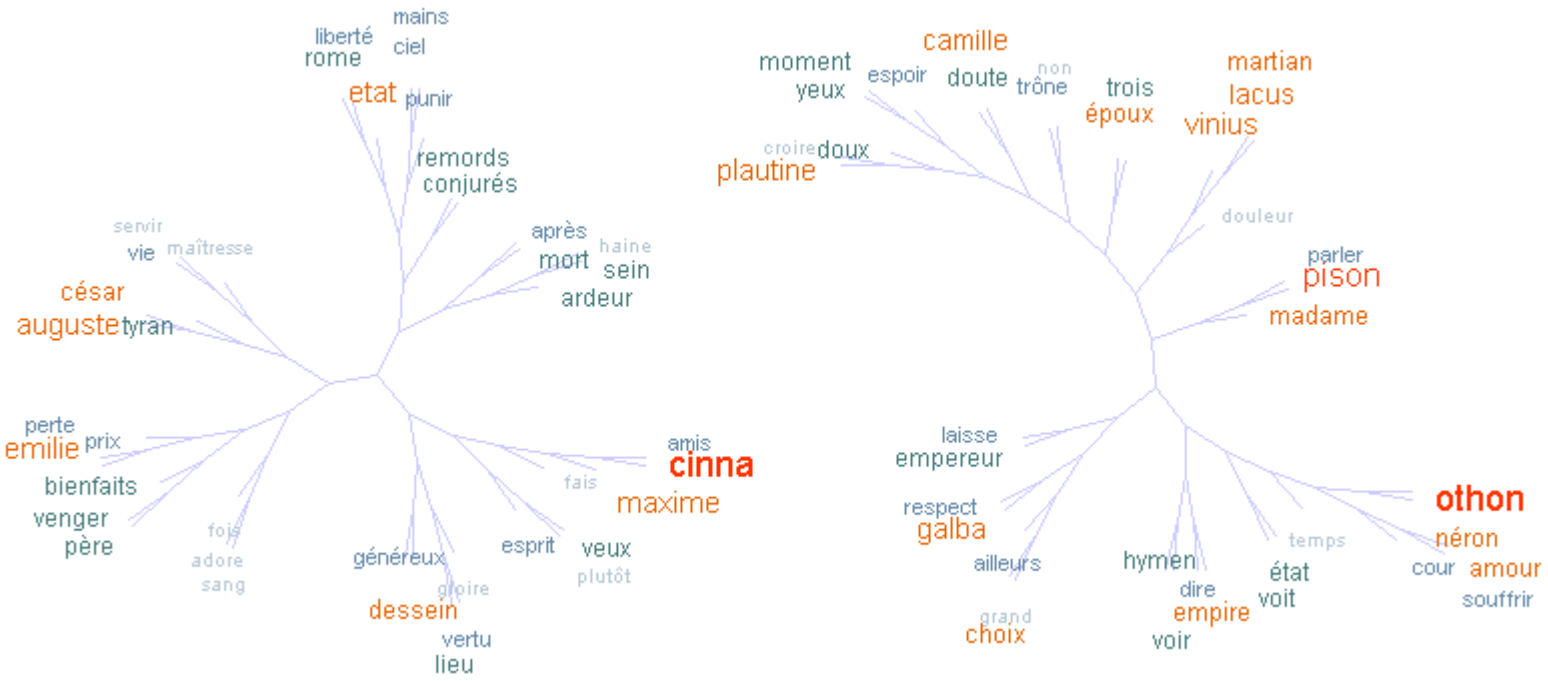
Nuage arboré des 30 mots les plus fréquents de la pièce Othon, coloré à gauche par rapport aux cooccurrences avec « Othon », à droite par rapport à celles avec « Galba »

Analyse littéraire : illustration sur *Othon*



Nuage arboré des 30 mots les plus fréquents de la pièce Othon, coloré à gauche par rapport aux cooccurrences avec « Othon », à droite par rapport à celles avec « Galba »

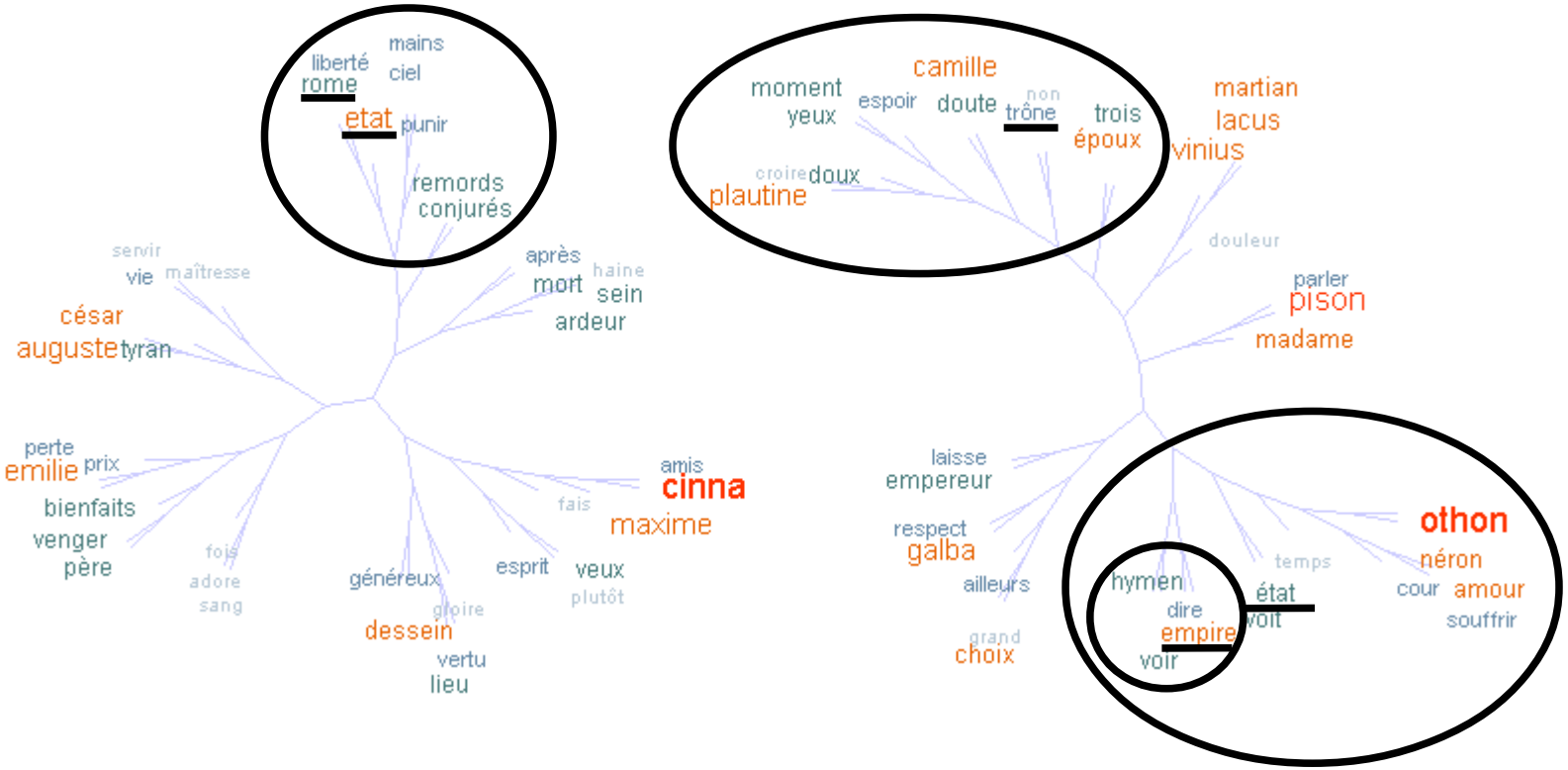
Analyse littéraire : comparaison de *Cinna* et *Othon*



*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 ?

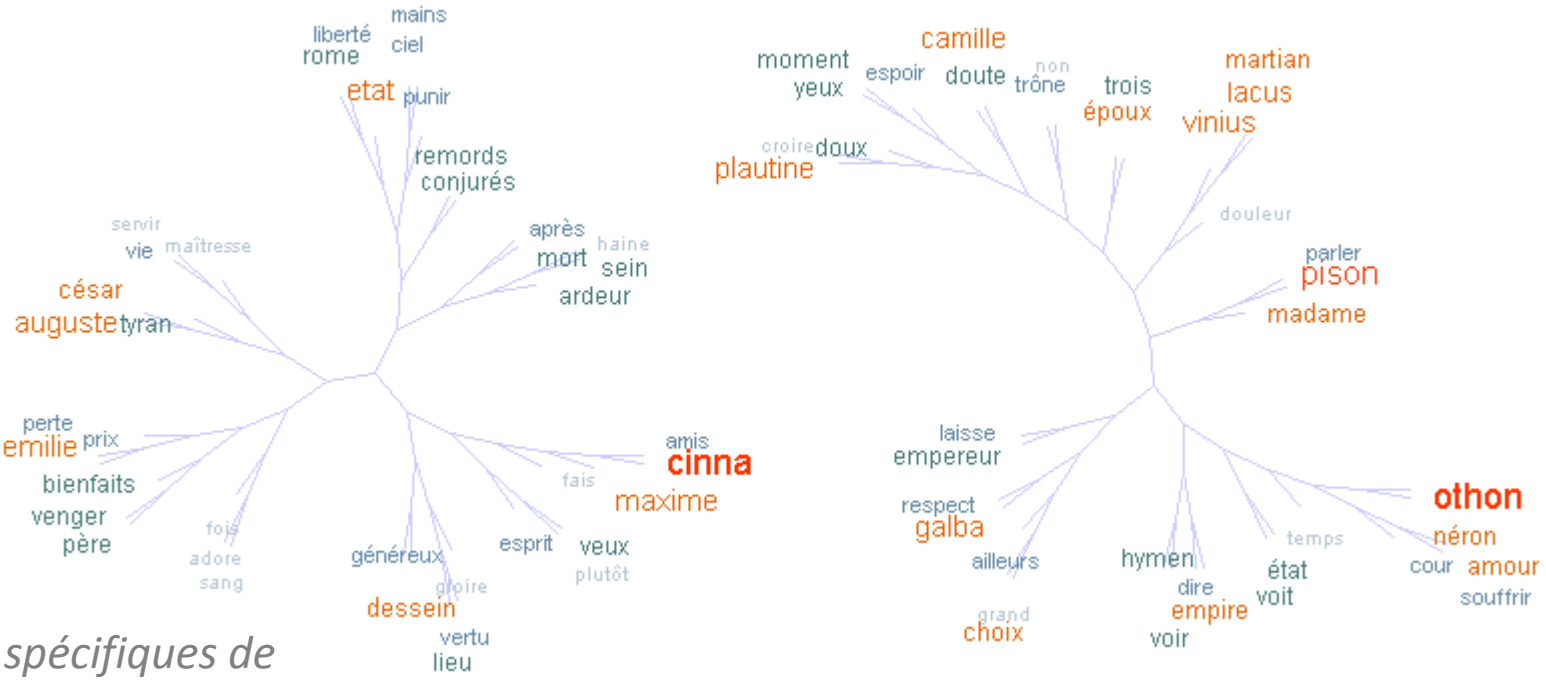
Analyse littéraire : comparaison de *Cinna* et *Othon*



*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 ?

Analyse littéraire : comparaison de *Cinna* et *Othon*



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

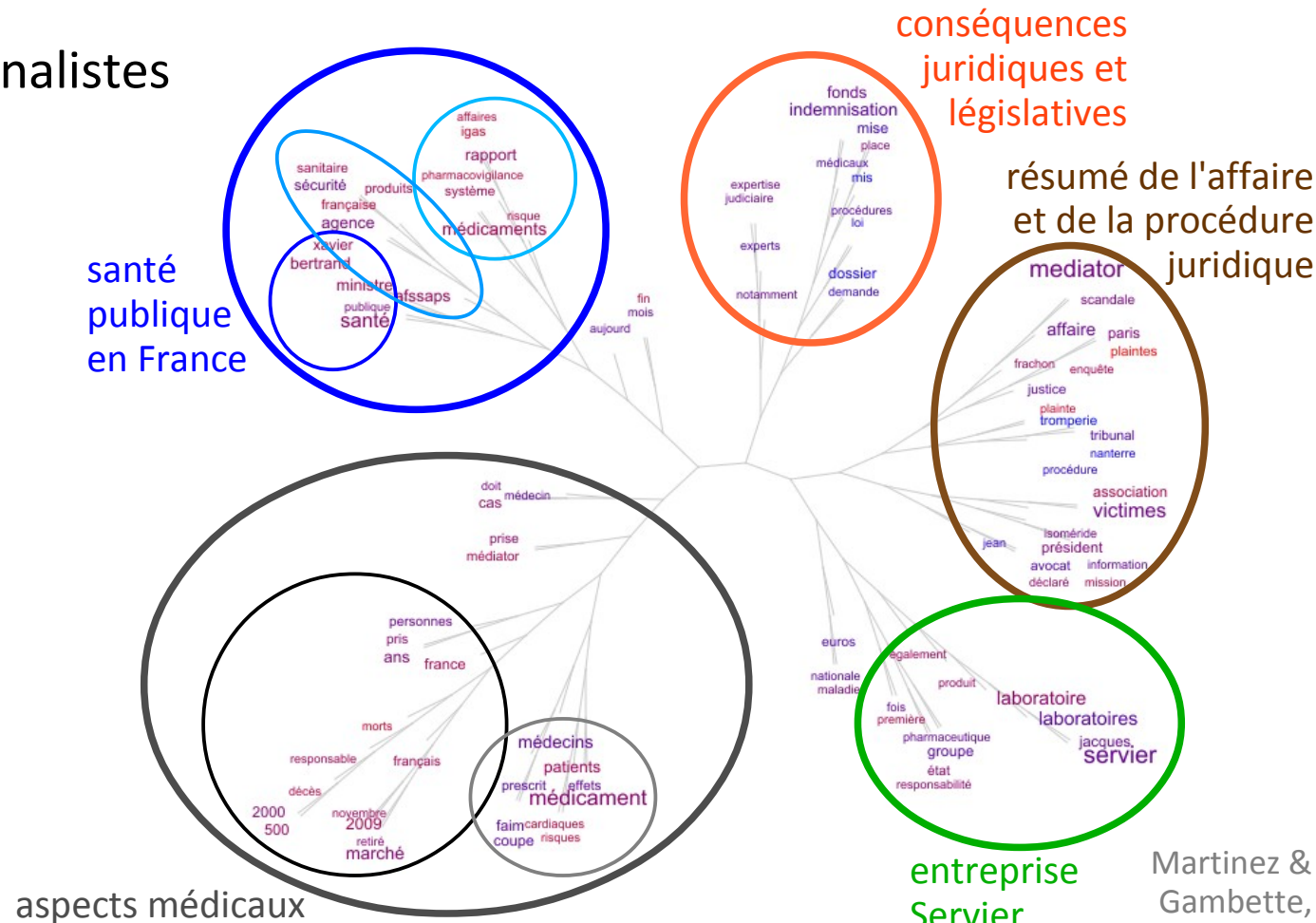
	Cinna	*Othon*
Lieu du pouvoir et objet de la confrontation entre les personnages	Rome (« liberté »)	Empire (« trône »)
Souverain en place	tyran	Empereur
Membres du corps politique	amis	maîtres / seigneurs
Moyens au service de la cause politique	gloire	amour matrimonial (« amour », « hymen », « choix »)
Caractérisation de la pièce	Pièce de FONDATION	Pièce de SUCCESSION DYNASTIQUE

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



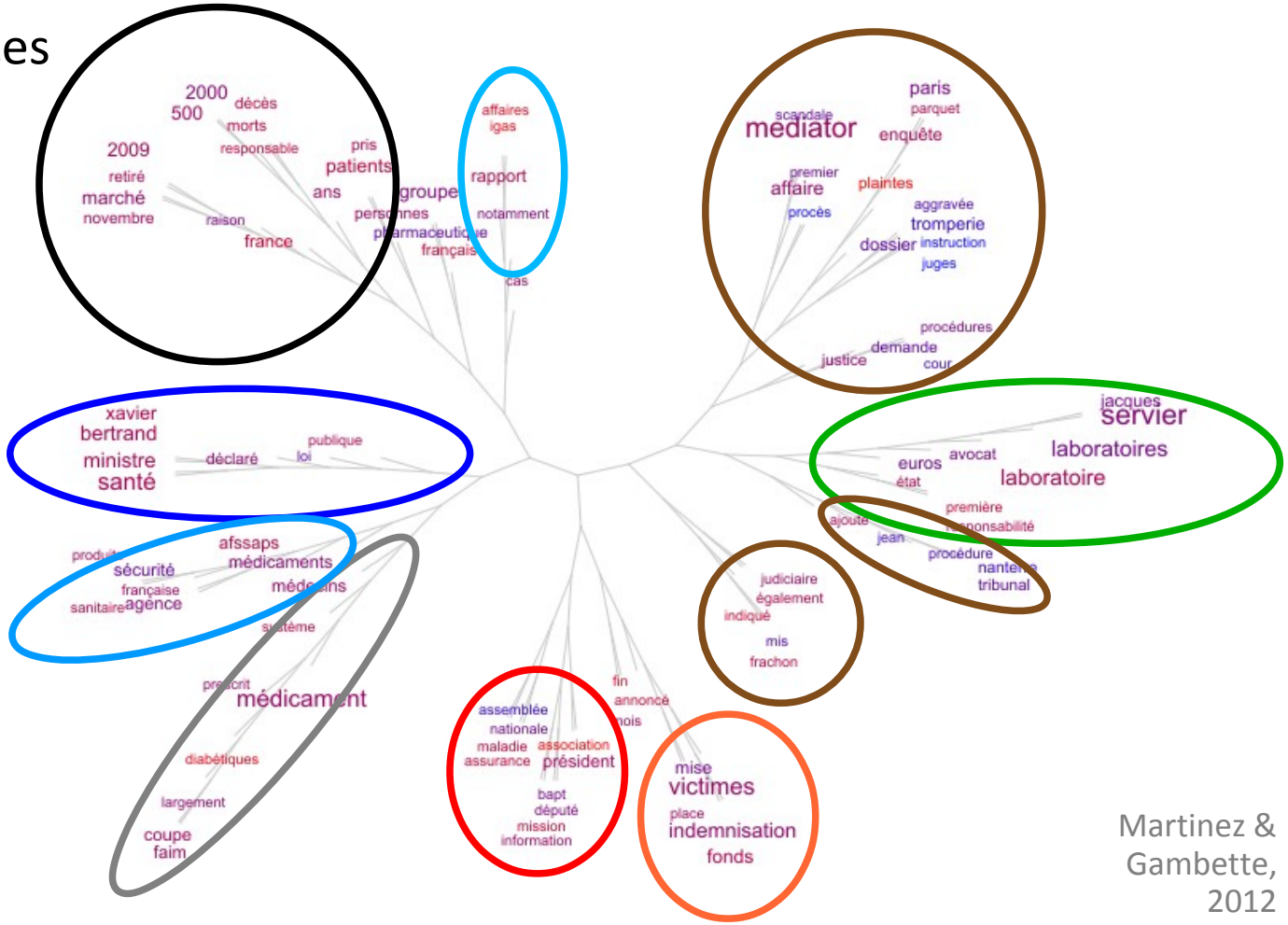
Martinez & Gambette, 2012

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



Martinez & Gambette, 2012

Plan

- Nuages arborés, intérêts et limites
- Construction des nuages arborés
- Options de coloration
- Utilisation des nuages arborés
- Évaluation de qualité
- Perspectives

Évaluation de qualité

1. Qualité de la **méthode de construction** de l'arbre :
   → l'arbre correspond-il aux données ?

2. Qualité de l'arbre en vue de l'**interprétation** :
   → Les classes lisibles dans l'arbre correspondent-elles aux données ?

3. Qualité du nuage arboré comme **outil d'analyse** :
   → La visualisation facilite-t-elle la lecture et l'interprétation ?

Correspondance de l’arbre avec les données

Distances dans l'arbre = approximation des distances entre mots

Mesure objective de la qualité de l'arbre ?

Correspondance de l’arbre avec les données

Distances dans l'arbre = approximation des distances entre mots

Mesure objective de la qualité de l'arbre ?

Variations du nuage de mots suite à l'altération du texte?

➡ **bootstrap** pour évaluer :

- stabilité du résultat

- robustesse de la méthode

Correspondance de l'arbre avec les données

Distances dans l'arbre = approximation des distances entre mots

Mesure objective de la qualité de l'arbre ?

Variations du nuage de mots suite à l'altération du texte?

➡ **bootstrap** pour évaluer :

- stabilité du résultat

- robustesse de la méthode

Méthode directe ?

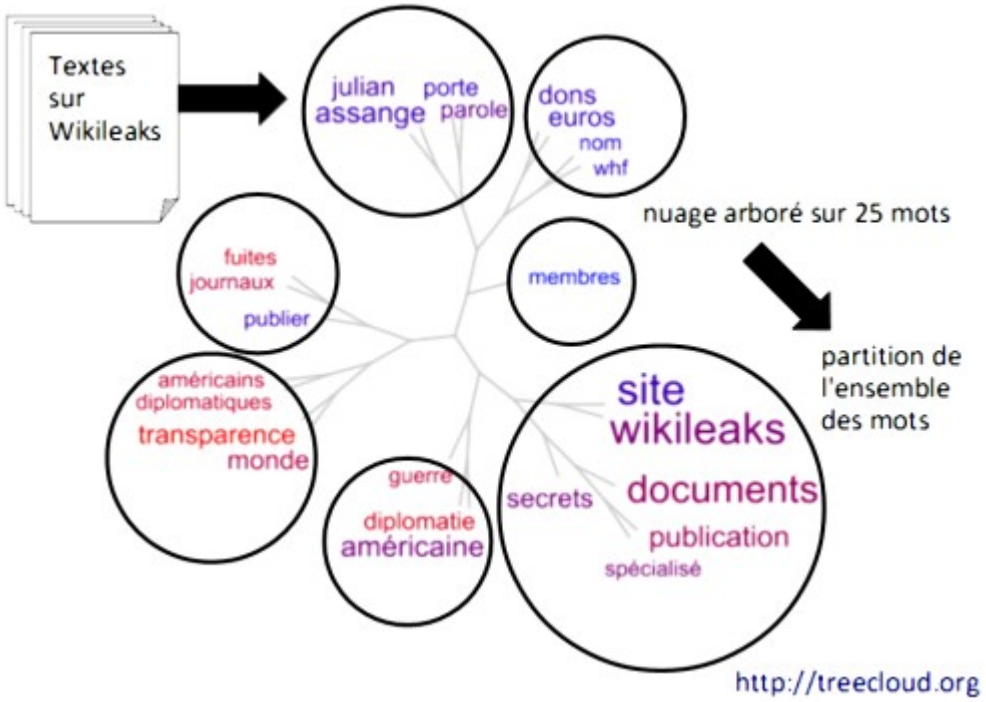
➡ **arboricité** pour mesurer à quel point la matrice de distance correspond à une distance d'arbre

Guénoche & Garreta, 2001

Guénoche & Darlu, 2009

Lisibilité des classes dans l'arbre

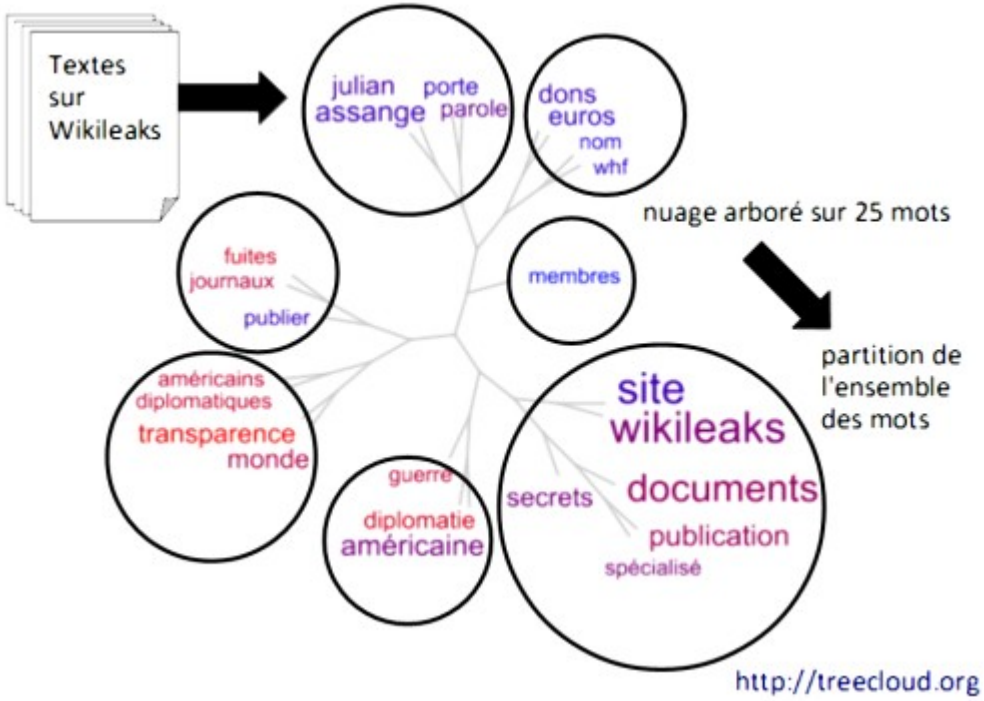
Protocole d'évaluation à partir d'un texte généré depuis le résultat attendu :



• Rédaction, par les étudiants de M1 de l'UPEMLV, de 10 textes respectant cette partition des mots en 7 classes

• Calcul des classes selon les différentes formules de longueurs d'arêtes, après découpage des 6 arêtes les plus longues

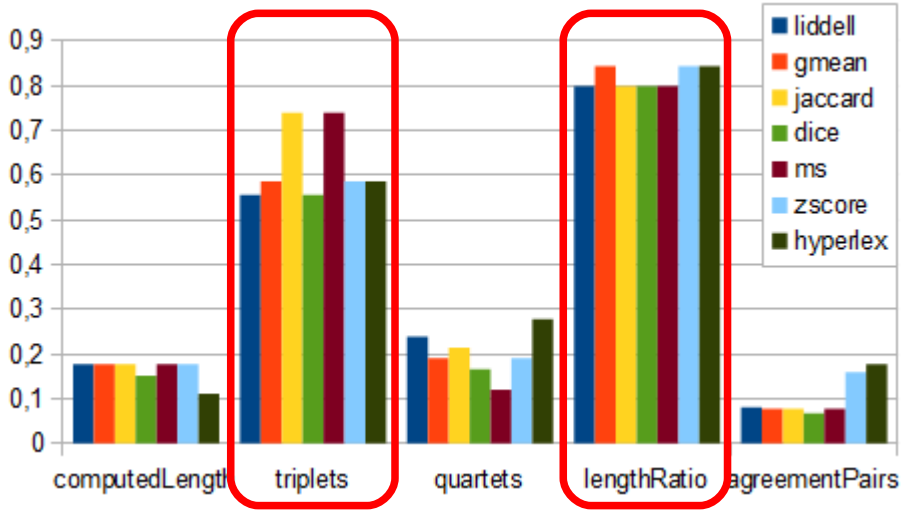
Lisibilité des classes dans l'arbre



- Rédaction, par les étudiants de M1 de l'UPEMLV, de 10 textes respectant cette partition des mots en 7 classes

- Calcul des classes selon les différentes formules de longueurs d'arêtes, après découpage des 6 arêtes les plus longues

score Rand corrigé



Plan

- Nuages arborés, intérêts et limites
- Construction des nuages arborés
- Options de coloration
- Utilisation des nuages arborés
- Évaluation de qualité
- Perspectives

Perspectives

- intégration de la **visualisation en nuages arborés** avec longueurs de branches post-calculées :
  - avec des outils de prétraitement linguistique :
    - → traitement des expressions composées (Unitex)
    - → détection des entités nommées
    - → étiquetage morphosyntaxique pour coloration (TreeTagger)
    - → sélection de groupes syntaxiques dans les concordances (Unitex)
  - dans les outils de textométrie existants
  - par des interfaces d'import/export adaptées
  - pour faciliter le retour au texte

- amélioration des méthodes de **construction de l'arbre**
  - transformée de Farris pour le calcul des distances
  - algorithme de Luong pour le calcul de l'arbre

Références

Disponibles sur 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

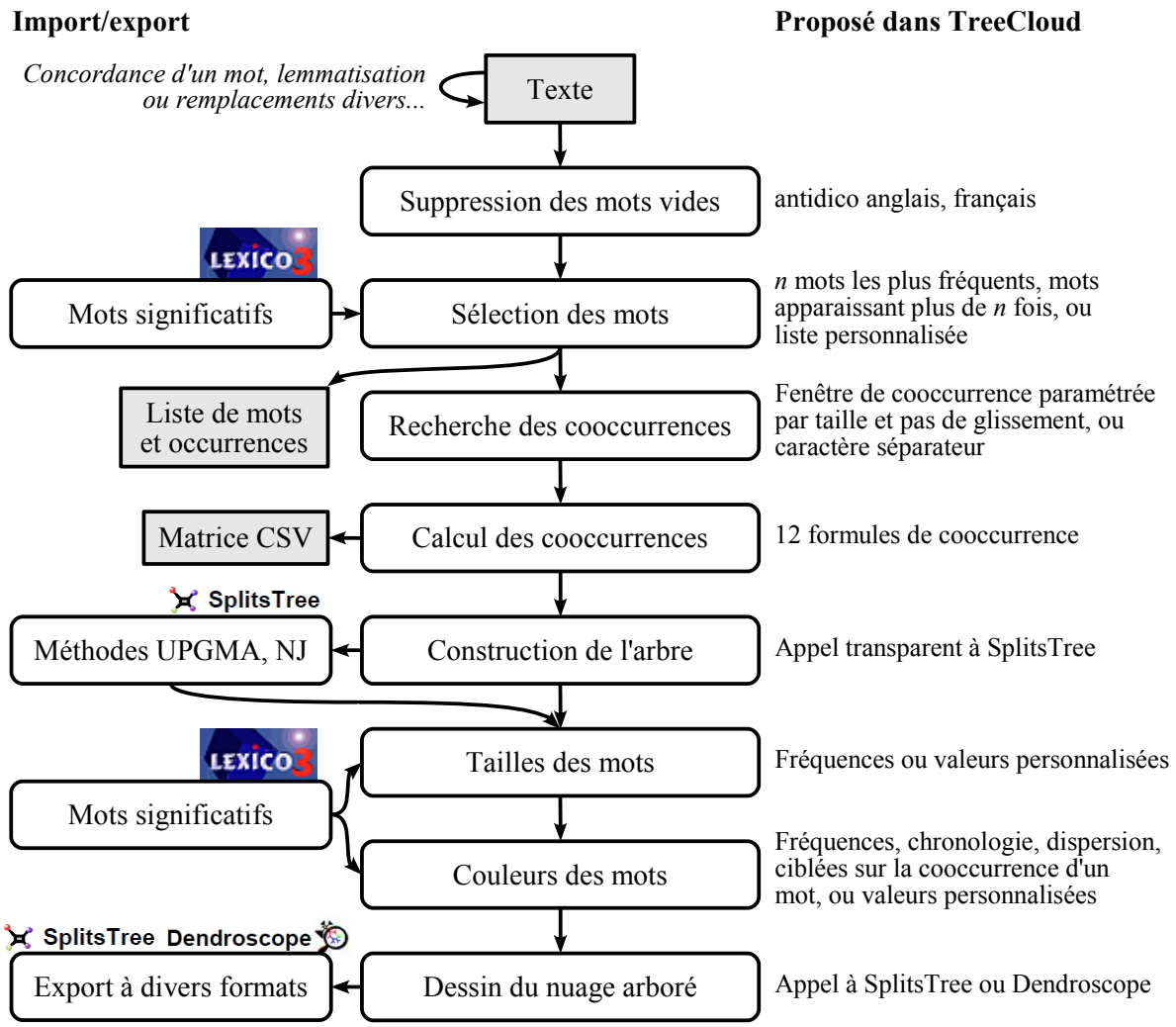
http://www.slideshare.net/PhilippeGambette/utilisation-de-la-visualisation-en-nuage-arbor-pour-lanalyse-littraire

Philippe Gambette, Nuria Gala & Alexis Nasr (2012)
Longueur de branches et arbres de mots,
Corpus 11:129-146

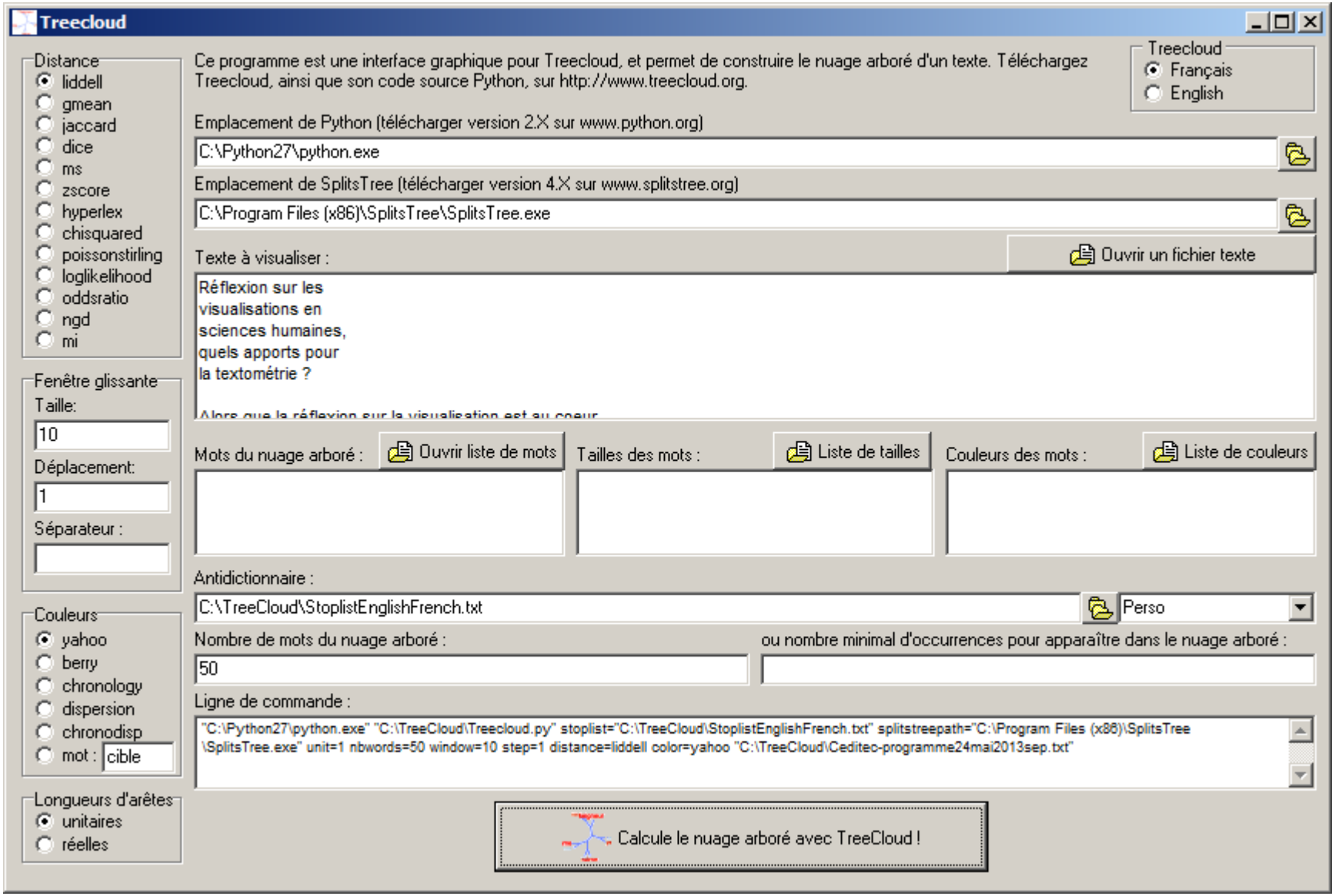
http://www.slideshare.net/PhilippeGambette/longueur-de-branches-et-arbres-de-mots

William Martinez & Philippe Gambette (2012)
L'affaire du Médiator au prisme de la textométrie,
Colloque Médias / Santé Publique

Processus de construction

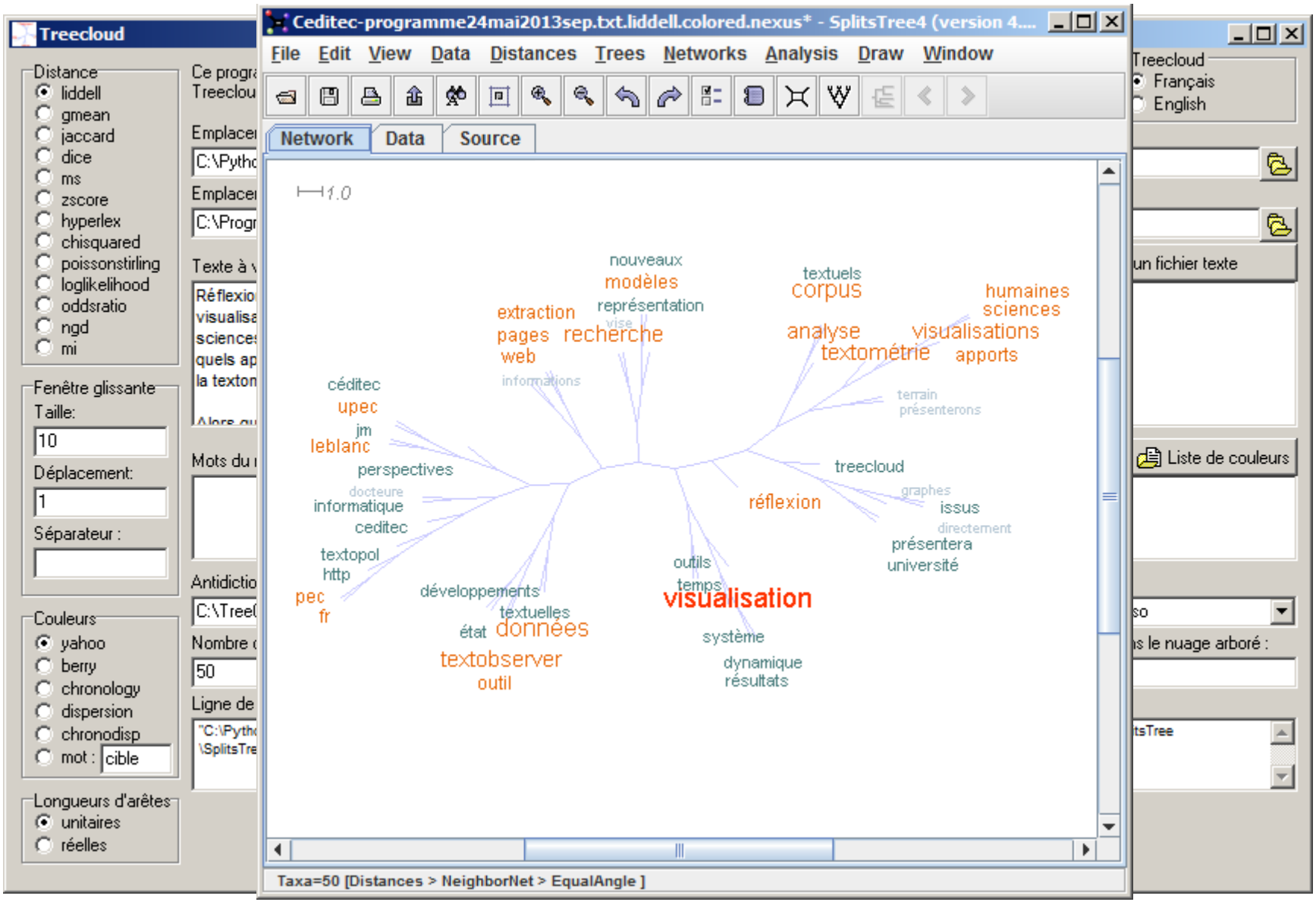


Implémentations



www.treecloud.org www.splitstree.org

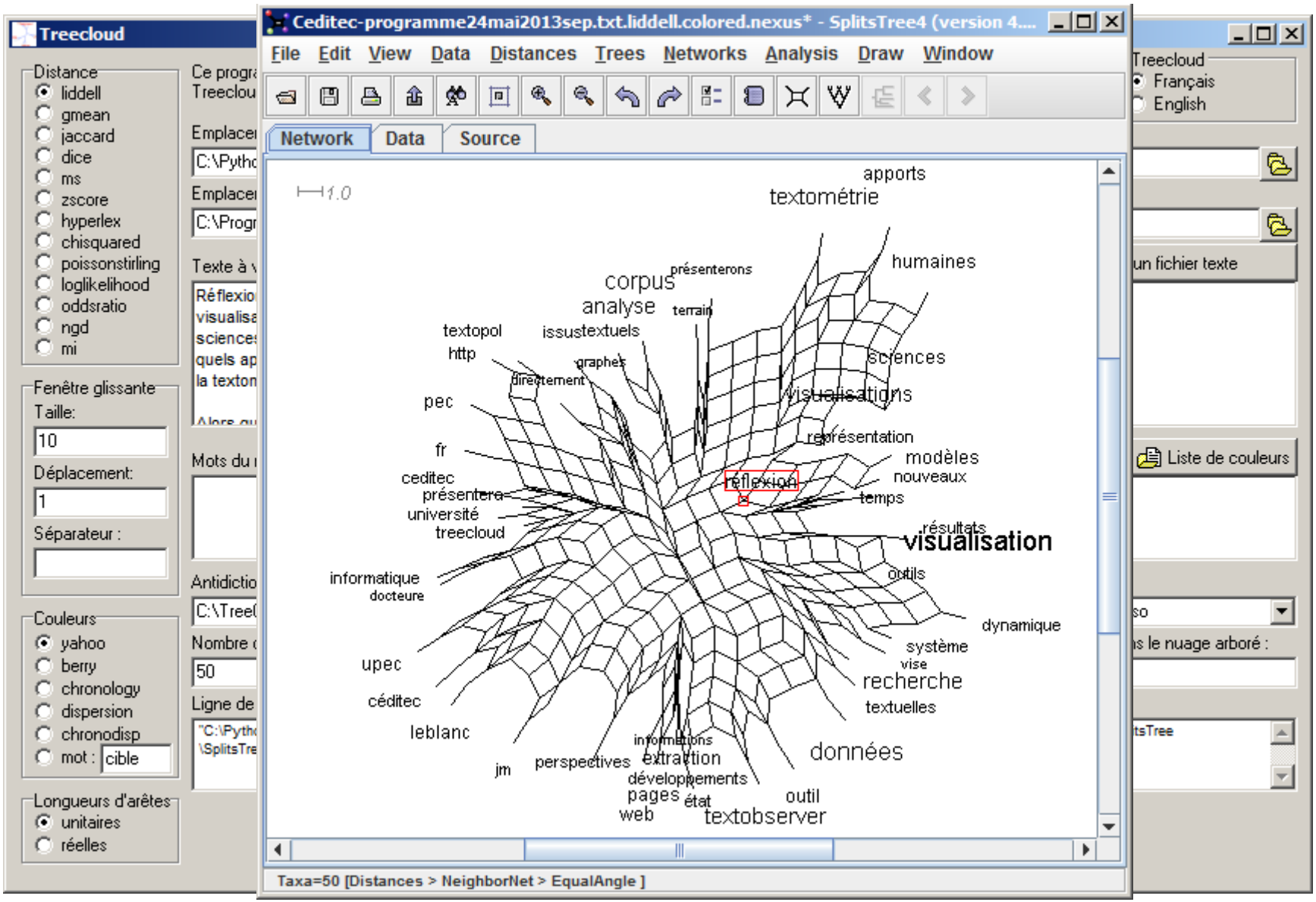
Implémentations



www.treecloud.org

www.splitstree.org

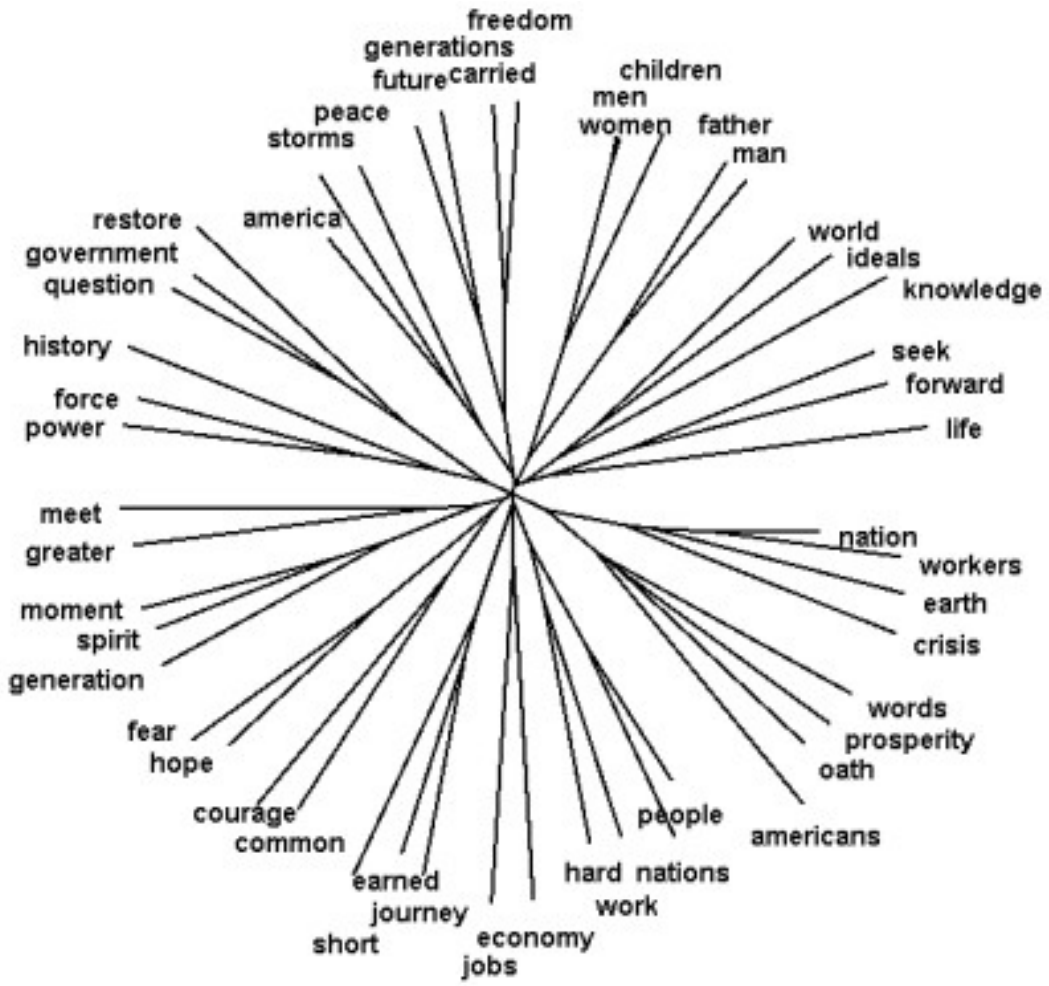
Implémentations



www.treecloud.org

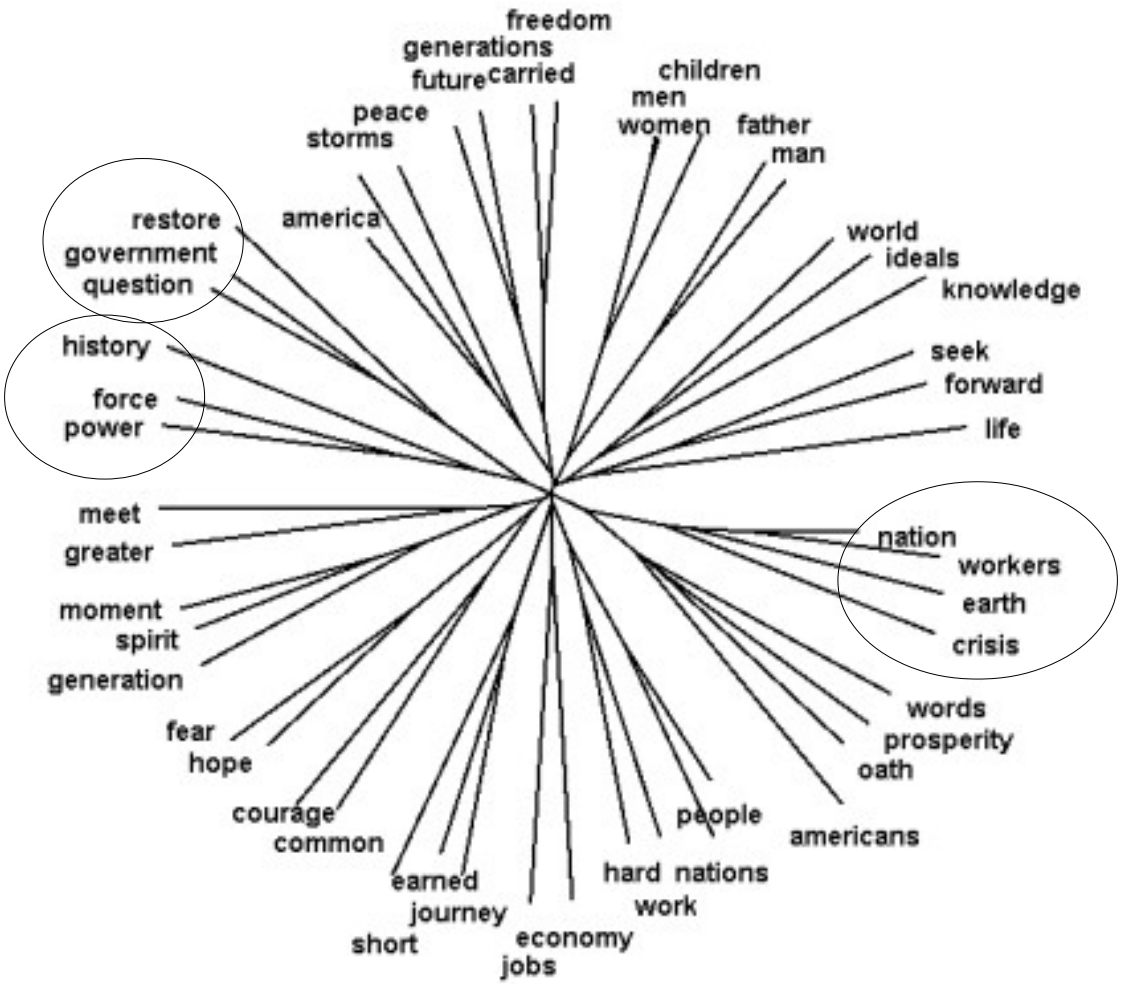
www.splitstree.org

Interprétation réelle



Principe de construction de l'arbre :
Les **distances dans l'arbre** entre deux mots reflètent *au mieux* le **degré de cooccurrence** entre ces deux mots

Interprétation réelle

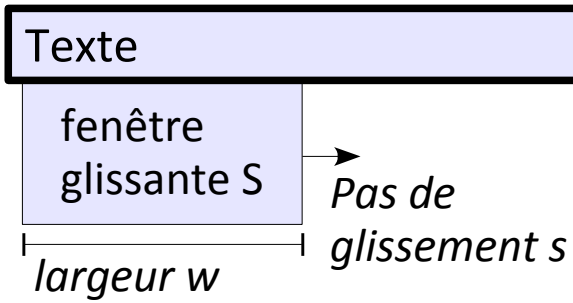


Principe de construction de l'arbre :

Les **distances dans l'arbre** entre deux mots reflètent *au mieux* le **degré de cooccurrence** entre ces deux mots

Calcul des scores de cooccurrence

Calcul de la matrice de distance entre mots



La dissimilarité sémantique entre deux mots *u* et *v* dépend du nombre de fenêtres *S* où ils apparaissent ensemble.

matrices de cooccurrence
O_{11}, O_{12}, O_{21}, O_{22}

Pour 2 mots *u* et *v*	$v \in S$	$v \notin S$
$u \in S$	O_{11}	O_{12}
$u \notin S$	O_{21}	O_{22}

matrice de dissimilarité sémantique
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
Gambette, *User manual for TreeCloud,* 2009

Références

Disponibles sur 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 (2012)
L'affaire du Médiator au prisme de la textométrie,
Colloque Médias / Santé Publique