Séminaire INFOLINGU du LIGM
14/05/2012 - Marne-la-Vallée

Nuages arborés et analyse textuelle

Philippe Gambette

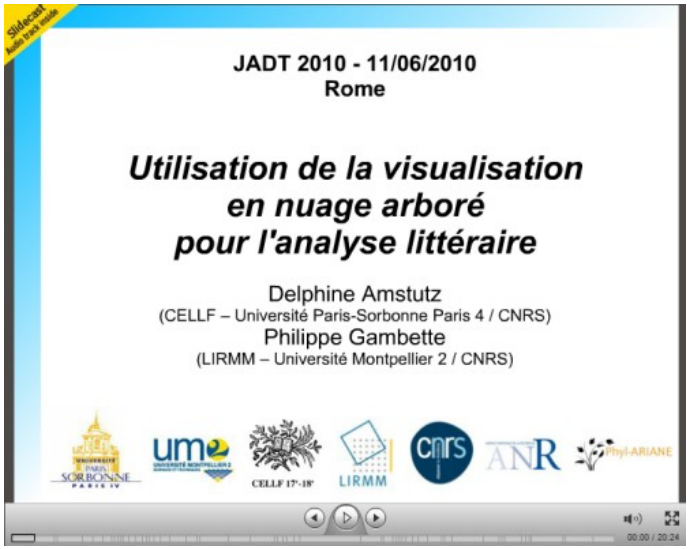
LIGM
Université Paris-Est
Marne-la-Vallée

Plan

- Nuages de mots et nuages arborés
- Caractéristiques du nuage arboré
- Utilisations du nuage arboré
- Construction d'un nuage arboré
- Évaluation de la robustesse de l'arbre
- Calcul des longueurs d'arêtes de l'arbre
- Perspectives

Plan

- Nuages de mots et nuages arborés
- Caractéristiques du nuage arboré
- Utilisations du nuage arboré
- Construction d'un nuage arboré
- Évaluation de la robustesse de l'arbre
- Calcul des longueurs d'arêtes de l'arbre
- Perspectives

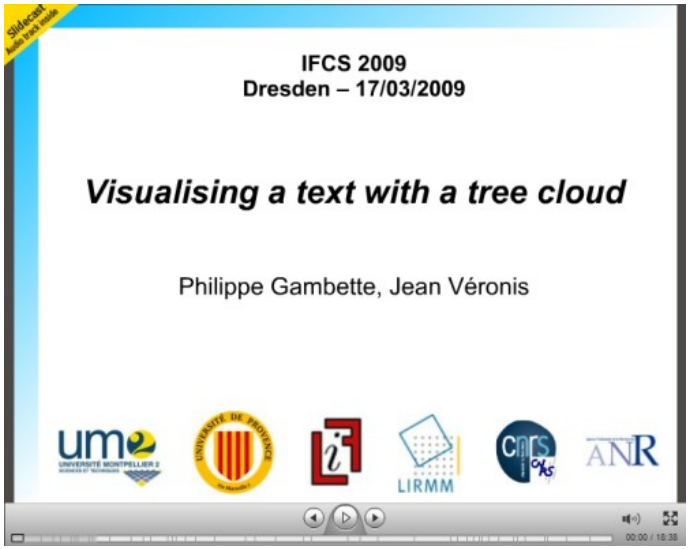


Delphine Amstutz, Philippe Gambette (2010)
Utilisation de la visualisation en nuage arboré pour l'analyse littéraire, JADT'10, *Statistical Analysis of Textual Data*, p. 227-238

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

Plan

- Nuages de mots et nuages arborés
- Caractéristiques du nuage arboré
- Utilisations du nuage arboré
- Construction d'un nuage arboré
- Évaluation de la robustesse de l'arbre
- Calcul des longueurs d'arêtes de l'arbre
- Perspectives



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

Plan

- Nuages de mots et nuages arborés
- Caractéristiques du nuage arboré
- Utilisations du nuage arboré
- Construction d'un nuage arboré
- Évaluation de la robustesse de l'arbre
- Calcul des longueurs d'arêtes de l'arbre
- Perspectives

Philippe Gambette, Alain Guénoche, Nuria Gala, Alexis Nasr (2012)
Longueur de branches et arbres de mots,
colloque *La cooccurrence, du fait statistique au fait textuel* (Besançon)

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

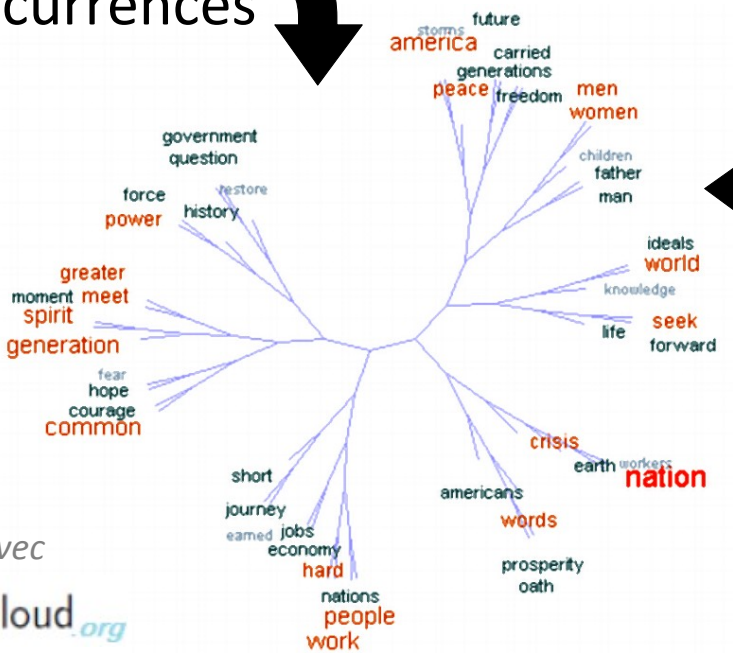
Plan

- Nuages de mots et nuages arborés
- Caractéristiques du nuage arboré
- Utilisations du nuage arboré
- Construction d'un nuage arboré
- Évaluation de la robustesse de l'arbre
- Calcul des longueurs d'arêtes de l'arbre
- Perspectives

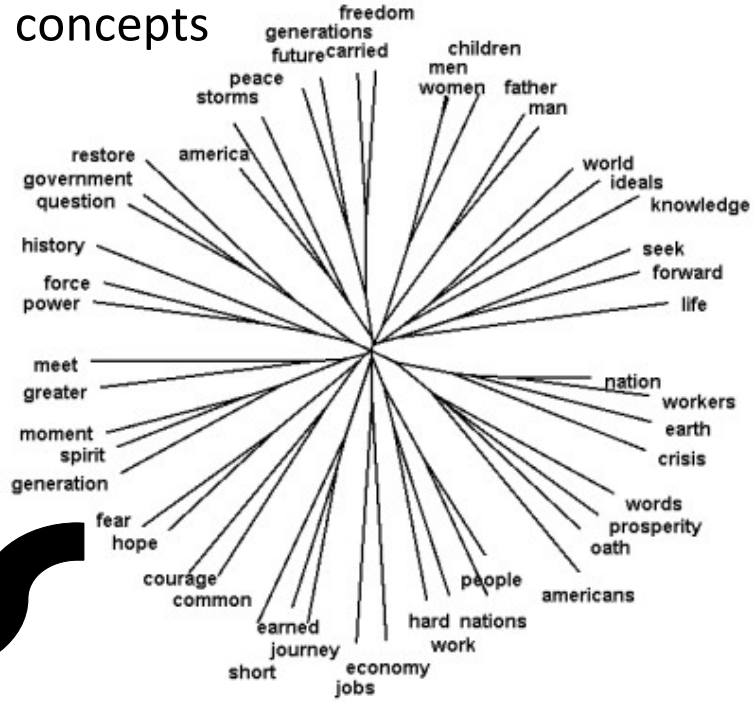
Nuage arboré, une information double

hiérarchie des mots

occurrences



hiérarchie des concepts



cooccurrences

Discours inaugural de Barack Obama

construit avec

TreeCloud.org

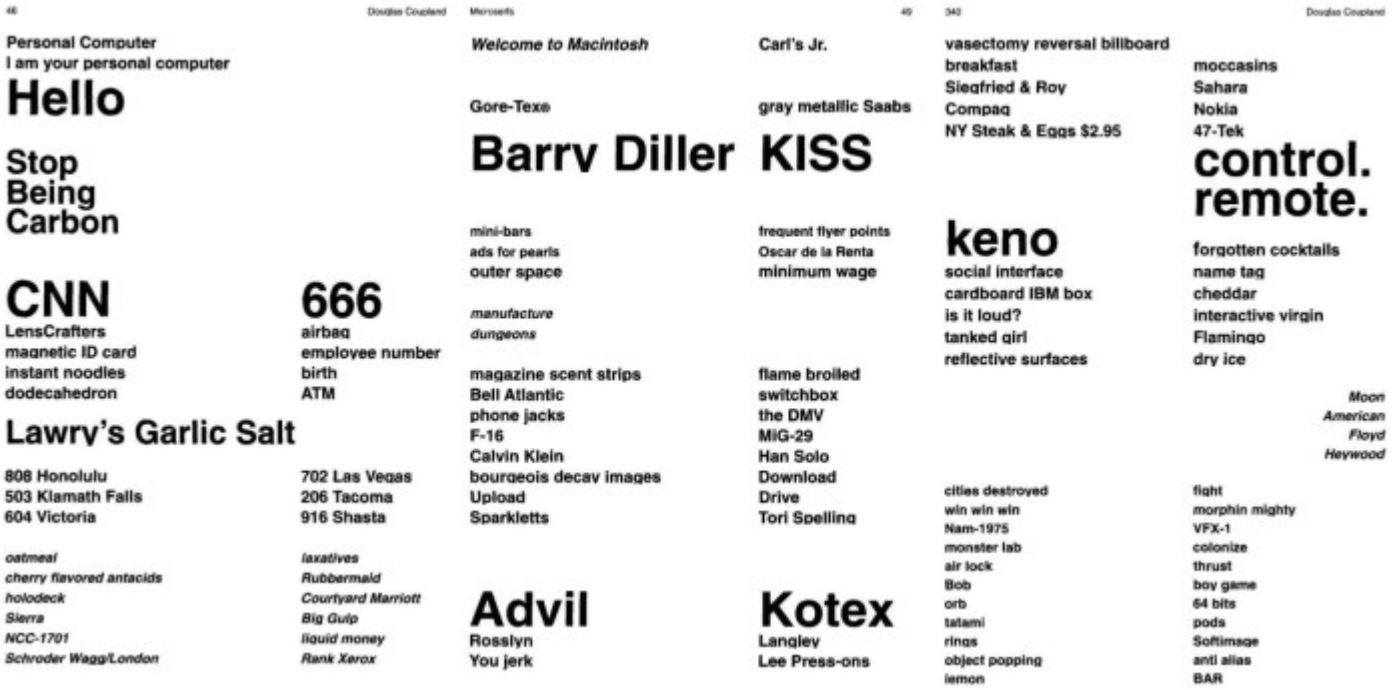
SplitsTree4

SplitsTree : Huson & Bryant, *Bioinformatics*, 2006

TreeCloud : Gambette & Véronis, *IFCS'09*

Nuages de tags

- Construits depuis un ensemble de tags
- Taille de police liée à la fréquence



Ce qui est habituellement cité comme le premier nuage de tags, dans *Microserfs* de D. Coupland, HarperCollins, Toronto, 1995

Nuages de mots

- Construits depuis l'ensemble des **mots d'un texte**
- Taille de police liée à la fréquence
- Se sont popularisés avec Wordle
- Donnent un bon aperçu d'un texte

Nuages Wordle des mots les plus utilisés début 2009 dans les blogs des Top 100 politique et high-tech de Wikio
http://aixtal.blogspot.com/2009/04/web-de-quoi-parlent-les-blogs.html

Extrait d'un nuage des mots des titres des 800 articles présentés à ce jour aux conférences JADT
http://www.ledonline.it/ledonline/jadt-2010.html

Nuages de tags/mots améliorés

Ajout d'informations provenant du texte :

- Couleur intense = tag récent dans Amazon
- Groupement sur une même ligne de tags cooccurrents

  Hassan-Montero & Herrero-Solana, InScit'06
- Optimisation des vides et proximité sémantique

  Kaser & Lemire, WWW'07
- “Topigraphy”: placement 2D en fonction de la cooccurrence

  Fujimura, Fujimura, Matsubayashi, Yamada & Okuda, WWW'08

Nuages de tags/mots améliorés

Ajout d'informations provenant du texte :

- Couleur intense = tag récent dans Amazon
- Groupement sur une même ligne de tags cooccurrents

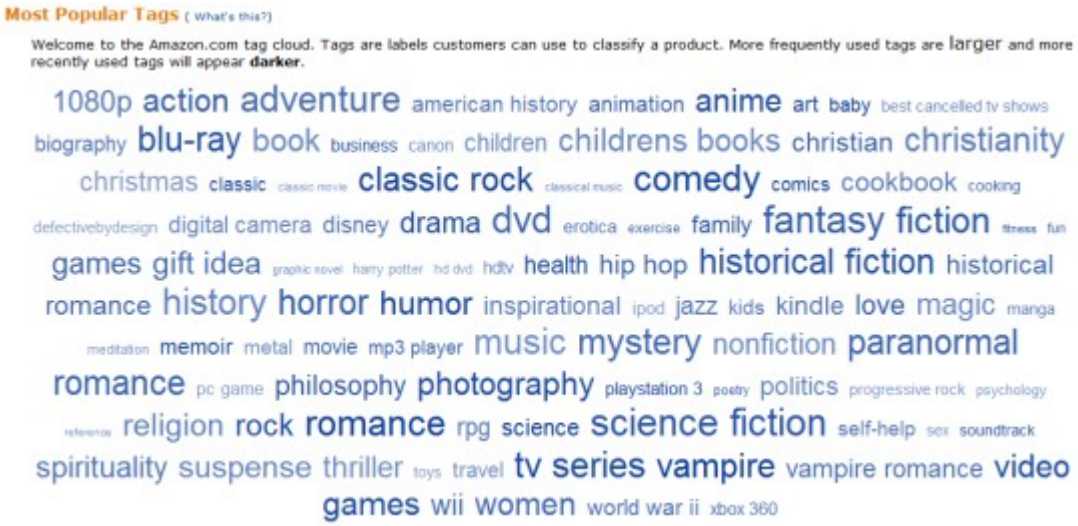
Hassan-Montero & Herrero-Solana, InScit'06

- Optimisation des vides et proximité sémantique

Kaser & Lemire, WWW'07

- “Topigraphy”: placement 2D en fonction de la cooccurrence

Fujimura, Fujimura, Matsubayashi, Yamada & Okuda, WWW'08



Nuages de tags/mots améliorés

Ajout d'informations provenant du texte :

- Couleur intense = tag récent dans Amazon
- Groupement sur une même ligne de tags cooccurrents

Hassan-Montero & Herrero-Solana, InScit'06

- Optimisation des vides et proximité sémantique

Kaser & Lemire, WWW'07

- “Topigraphy”: placement 2D en fonction de la cooccurrence

Fujimura, Fujimura, Matsubayashi, Yamada & Okuda, WWW'08

Nuages de tags/mots améliorés

Ajout d'informations provenant du texte :

- Couleur intense = tag récent dans Amazon
- Groupement sur une même ligne de tags cooccurrents

Hassan-Montero & Herrero-Solana, InScit'06

- Optimisation des vides et proximité sémantique

Kaser & Lemire, WWW'07

- “Topigraphy”: placement 2D en fonction de la cooccurrence

Fujimura, Fujimura, Matsubayashi, Yamada & Okuda, WWW'08

Nuages de tags/mots améliorés

Ajout d'informations provenant du texte :

- Couleur intense = tag récent dans Amazon
- Groupement sur une même ligne de tags cooccurrents

Hassan-Montero & Herrero-Solana, InScit'06

- Optimisation des vides et proximité sémantique

Kaser & Lemire, WWW'07

- “Topigraphy”: placement 2D en fonction de la cooccurrence

Fujimura, Fujimura, Matsubayashi, Yamada & Okuda, WWW'08

Extraire l'information sémantique d'un texte

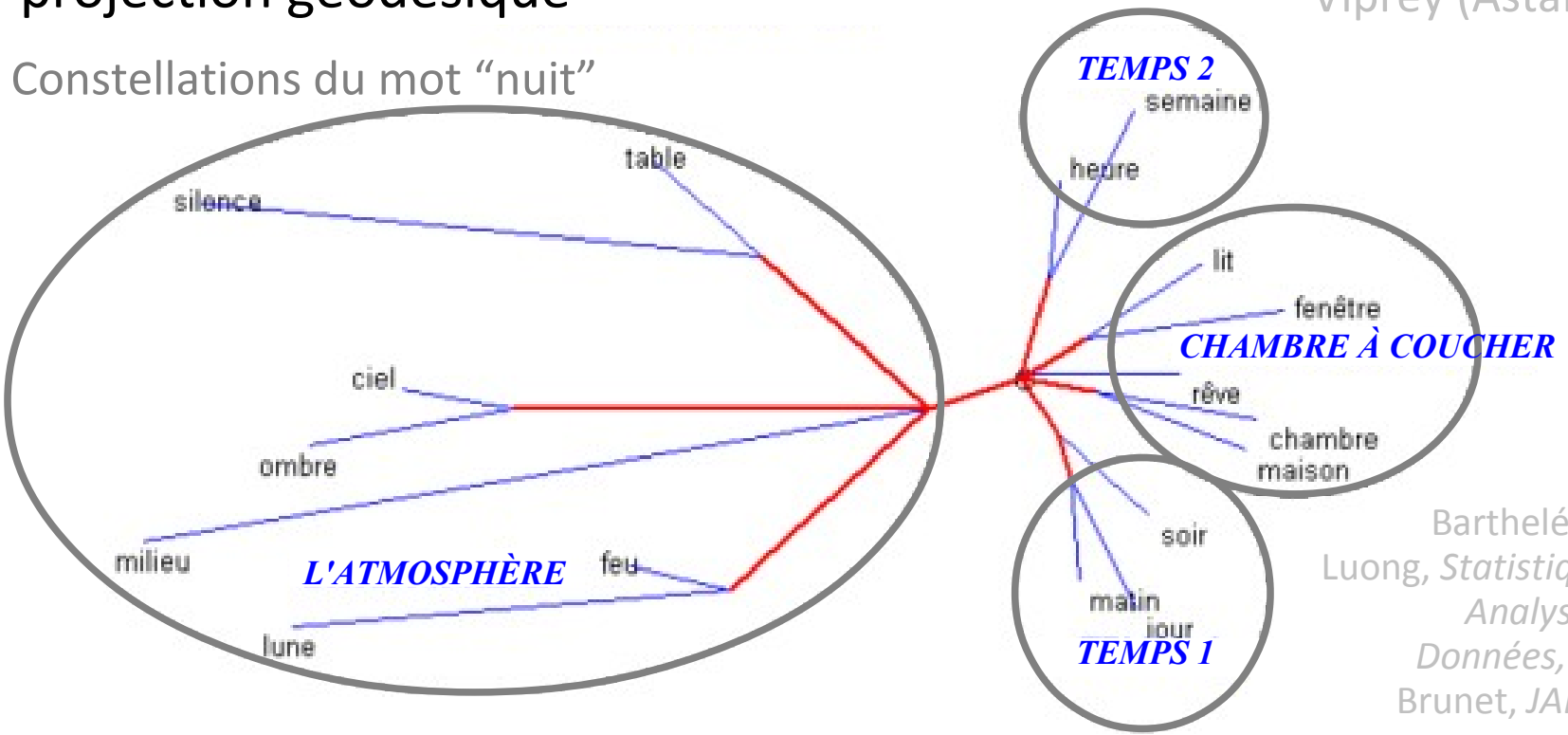
- analyse arborée — Brunet (Hyperbase)
- graphe de coocurrence — Brunet (Hyperbase)
- graphe sémantique — Grimmer (Wordmapper)
- lexicogramme récursif — Martinez (Coocs)
- désambiguïsation lexicale — Véronis (Hyperlex)
- réseau Phrasenet — Viegas et al. (IBM Many Eyes)
- projection géodésique — Viprey (Astartex)

Extraire l'information sémantique d'un texte

- analyse arborée
- graphe de coocurrence
- graphe sémantique
- lexicogramme récursif
- désambiguïsation lexicale
- réseau Phrasenet
- projection géodésique

Brunet (Hyperbase)
Brunet (Hyperbase)
Grimmer (Wordmapper)
Martinez (Coocs)
Véronis (Hyperlex)
Viegas et al. (IBM Many Eyes)
Viprey (Astartex)

Constellations du mot “nuit”

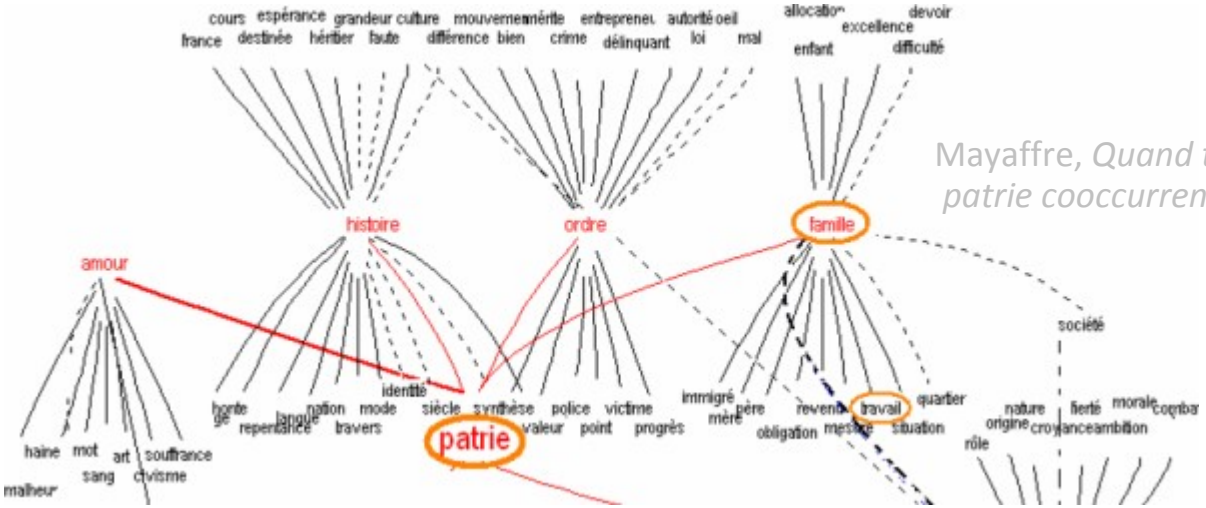


Barthelémy & Luong, *Statistique et Analyse des Données*, 1986
Brunet, *JADT'08*

Extraire l'information sémantique d'un texte

- analyse arborée
- graphe de coocurrence
- graphe sémantique
- lexicogramme récursif
- désambiguïsation lexicale
- réseau Phrasenet
- projection géodésique

Brunet (Hyperbase)
Brunet (Hyperbase)
Grimmer (Wordmapper)
Martinez (Coocs)
Véronis (Hyperlex)
Viegas et al. (IBM Many Eyes)
Viprey (Astartex)

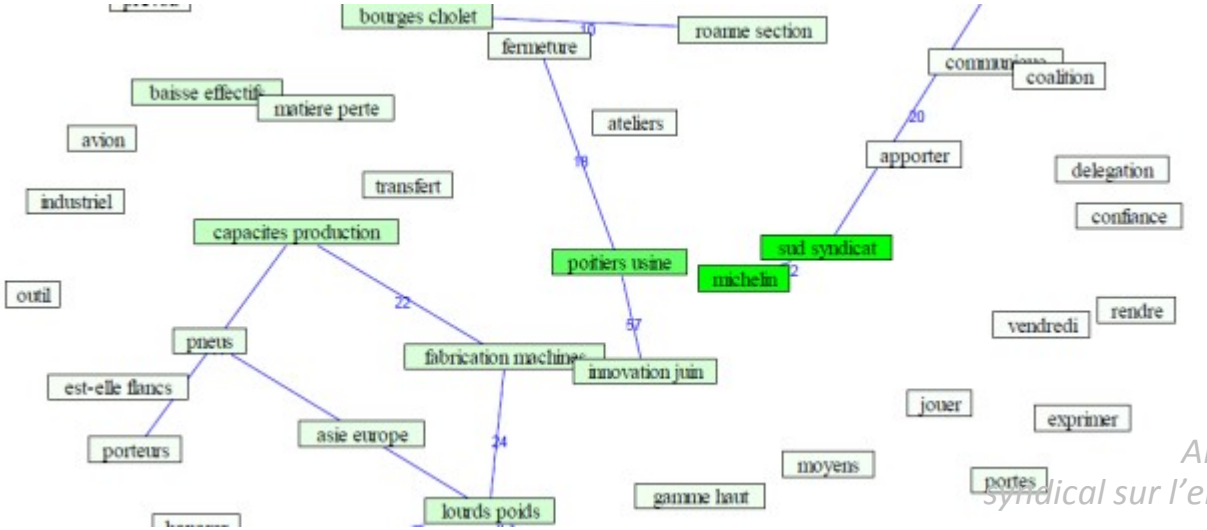


Mayaffre, *Quand travail, famille, et patrie cooccurrent dans le discours de Nicolas Sarkozy*, JADT'08

Extraire l'information sémantique d'un texte

- analyse arborée
- graphe de coocurrence
- graphe sémantique
- lexicogramme récursif
- désambiguïsation lexicale
- réseau Phrasenet
- projection géodésique

Brunet (Hyperbase)
Brunet (Hyperbase)
Grimmer (Wordmapper)
Martinez (Coocs)
Véronis (Hyperlex)
Viegas et al. (IBM Many Eyes)
Viprey (Astartex)

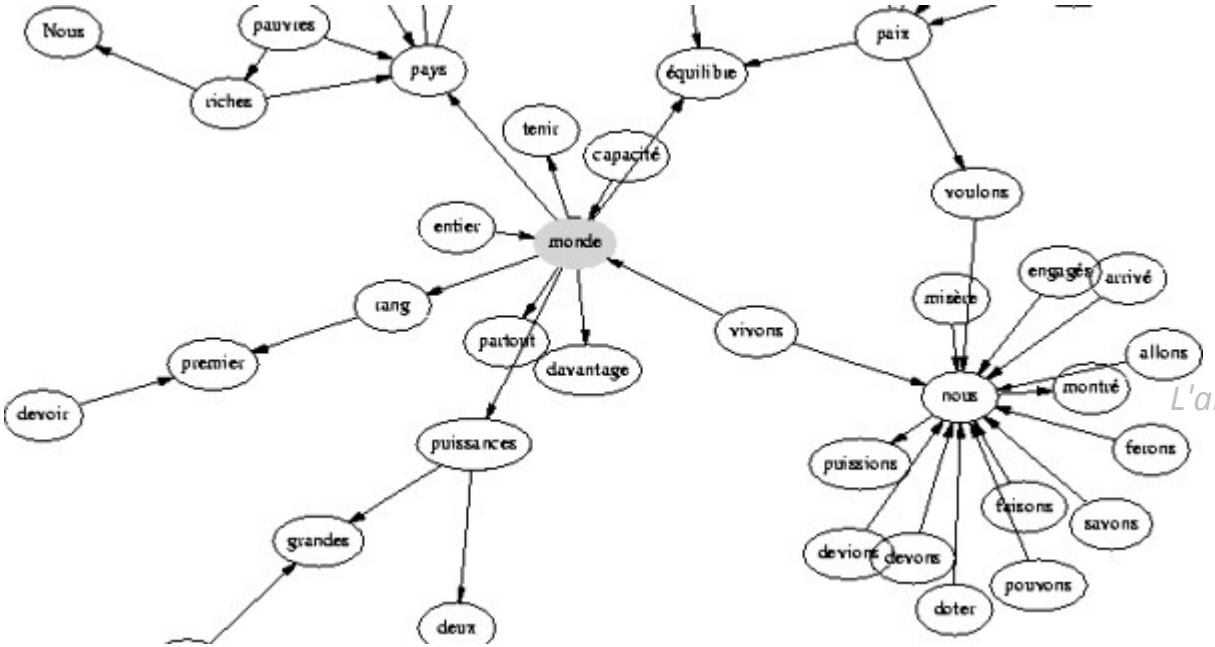


Peyrat-Guillard, *Analyse du discours syndical sur l'entreprise*, JADT'08

Extraire l'information sémantique d'un texte

- analyse arborée
- graphe de coocurrence
- graphe sémantique
- lexicogramme récursif
- désambiguïsation lexicale
- réseau Phrasenet
- projection géodésique

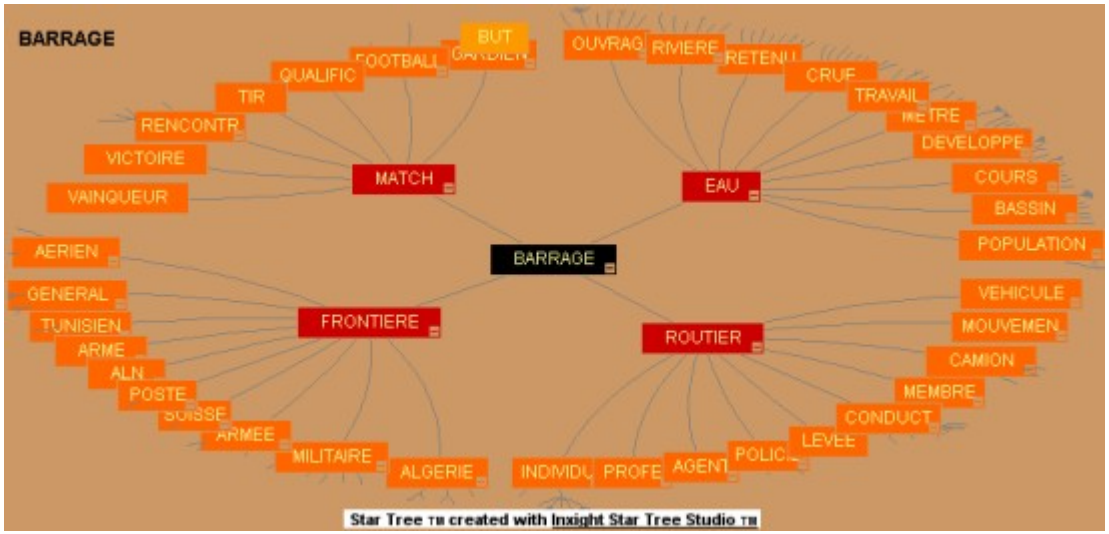
Brunet (Hyperbase)
Brunet (Hyperbase)
Grimmer (Wordmapper)
Martinez (Coocs)
Véronis (Hyperlex)
Viegas et al. (IBM Many Eyes)
Viprey (Astartex)



Leblanc, Martinez
L'analyse contrastive des réseaux de cooccurrence
JADT 2006.

Extraire l'information sémantique d'un texte

- analyse arborée — Brunet (Hyperbase)
- graphe de coocurrence — Brunet (Hyperbase)
- graphe sémantique — Grimmer (Wordmapper)
- lexicogramme récursif — Martinez (Coocs)
- désambiguïsation lexicale — Véronis (Hyperlex)
- réseau Phrasenet — Viegas et al. (IBM Many Eyes)
- projection géodésique — Viprey (Astartex)



Désambiguïsation du mot “barrage”.

Véronis, *HyperLex: Lexical Cartography for Information Retrieval*, 2004

Extraire l'information sémantique d'un texte

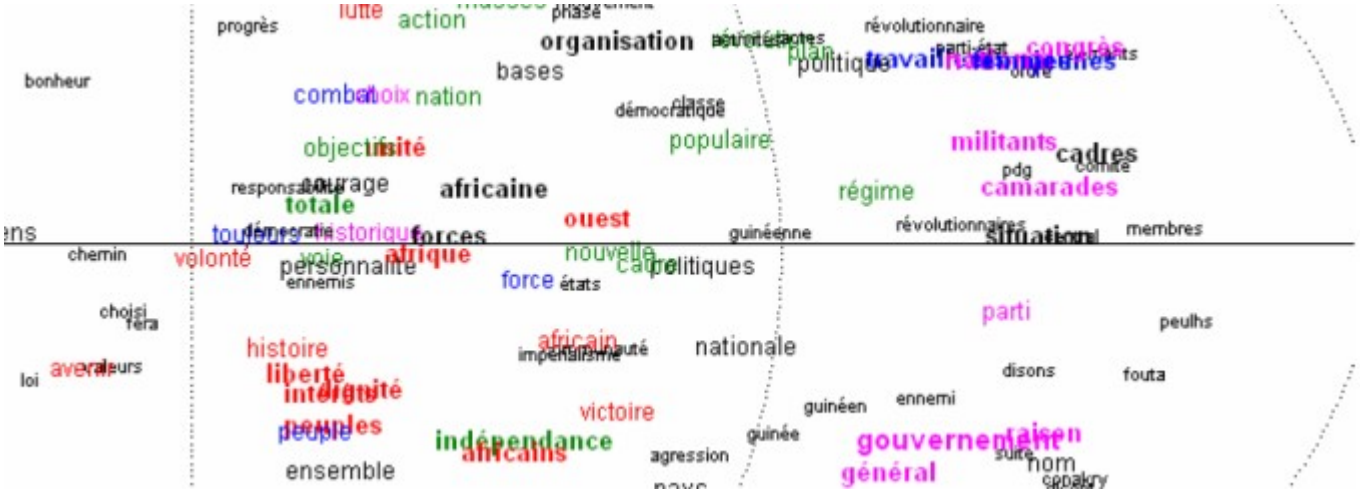
- analyse arborée — Brunet (Hyperbase)
- graphe de coocurrence — Brunet (Hyperbase)
- graphe sémantique — Grimmer (Wordmapper)
- lexicogramme récursif — Martinez (Coocs)
- désambiguïsation lexicale — Véronis (Hyperlex)
- réseau Phrasenet — Viegas et al. (IBM Many Eyes)
- projection géodésique — Viprey (Astartex)

Visualisation PhraseNet de paroles des Beatles créé avec Many Eyes (IBM) http://many-eyes.com

http://visualthinkmap.ning.com/photo/phrasenet_beatles-many-eyes

Extraire l'information sémantique d'un texte

- analyse arborée — Brunet (Hyperbase)
- graphe de coocurrence — Brunet (Hyperbase)
- graphe sémantique — Grimmer (Wordmapper)
- lexicogramme récursif — Martinez (Coocs)
- désambiguïsation lexicale — Véronis (Hyperlex)
- réseau Phrasenet — Viegas et al. (IBM Many Eyes)
- projection géodésique — Viprey (Astartex)



Barry, Viprey, *Approche comparative des résultats d'exploration textuelle des discours de deux leaders africains Keita et Touré*, JADT'08

Plan

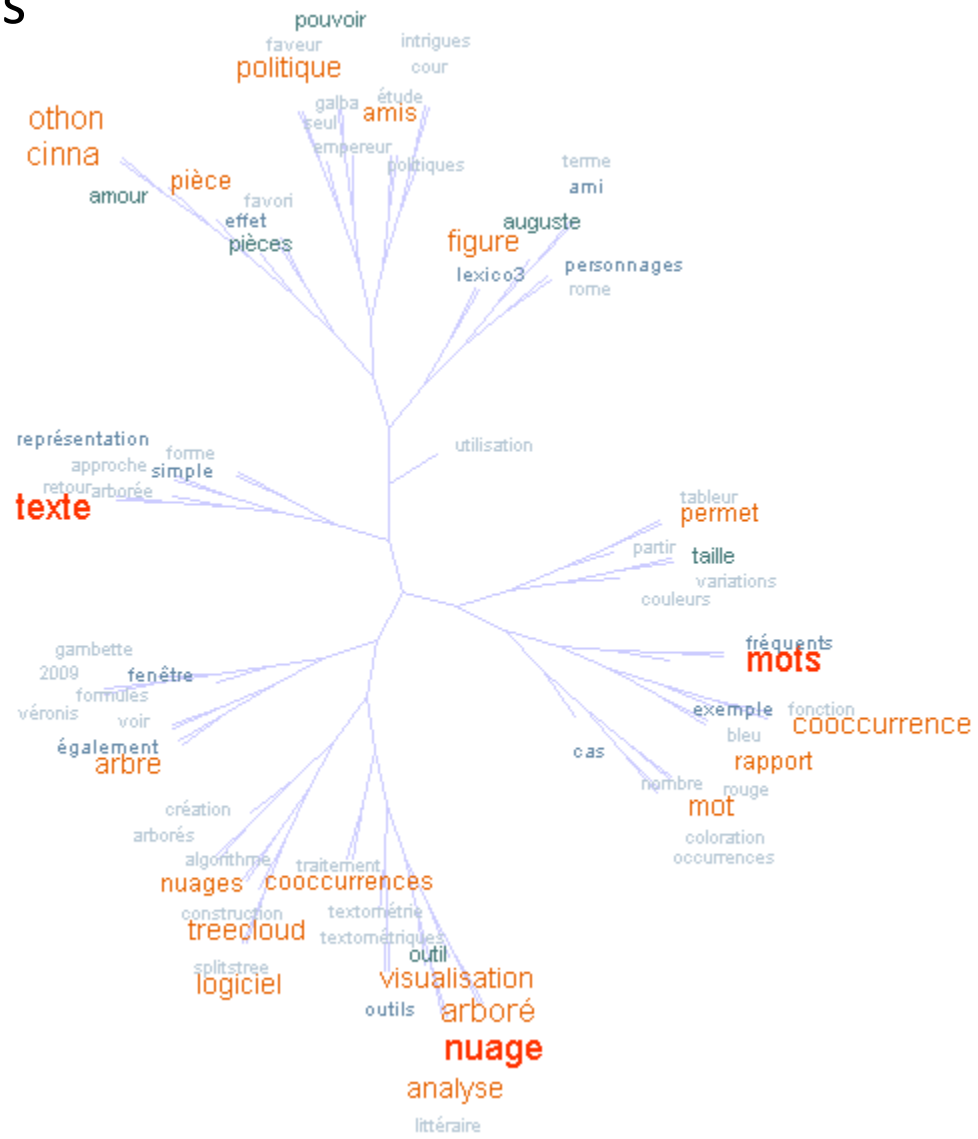
- Nuages de mots et nuages arborés
- Caractéristiques du nuage arboré
- Utilisations du nuage arboré
- Construction d'un nuage arboré
- Évaluation de la robustesse de l'arbre
- Calcul des longueurs d'arêtes de l'arbre
- Perspectives

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

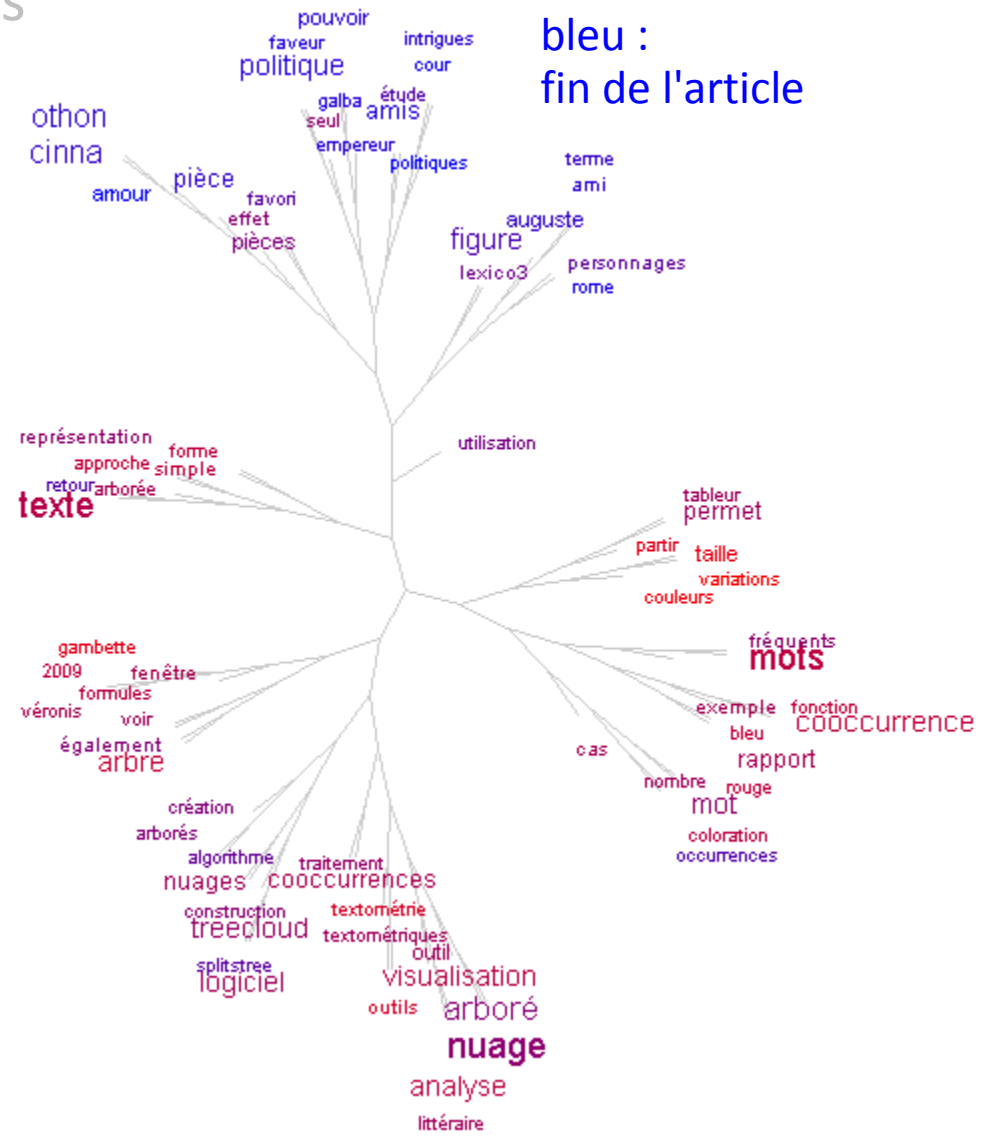
- coloration selon les fréquences
- coloration chronologique
- coloration de la dispersion
- coloration ciblée sur un mot
- coloration grammaticale



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

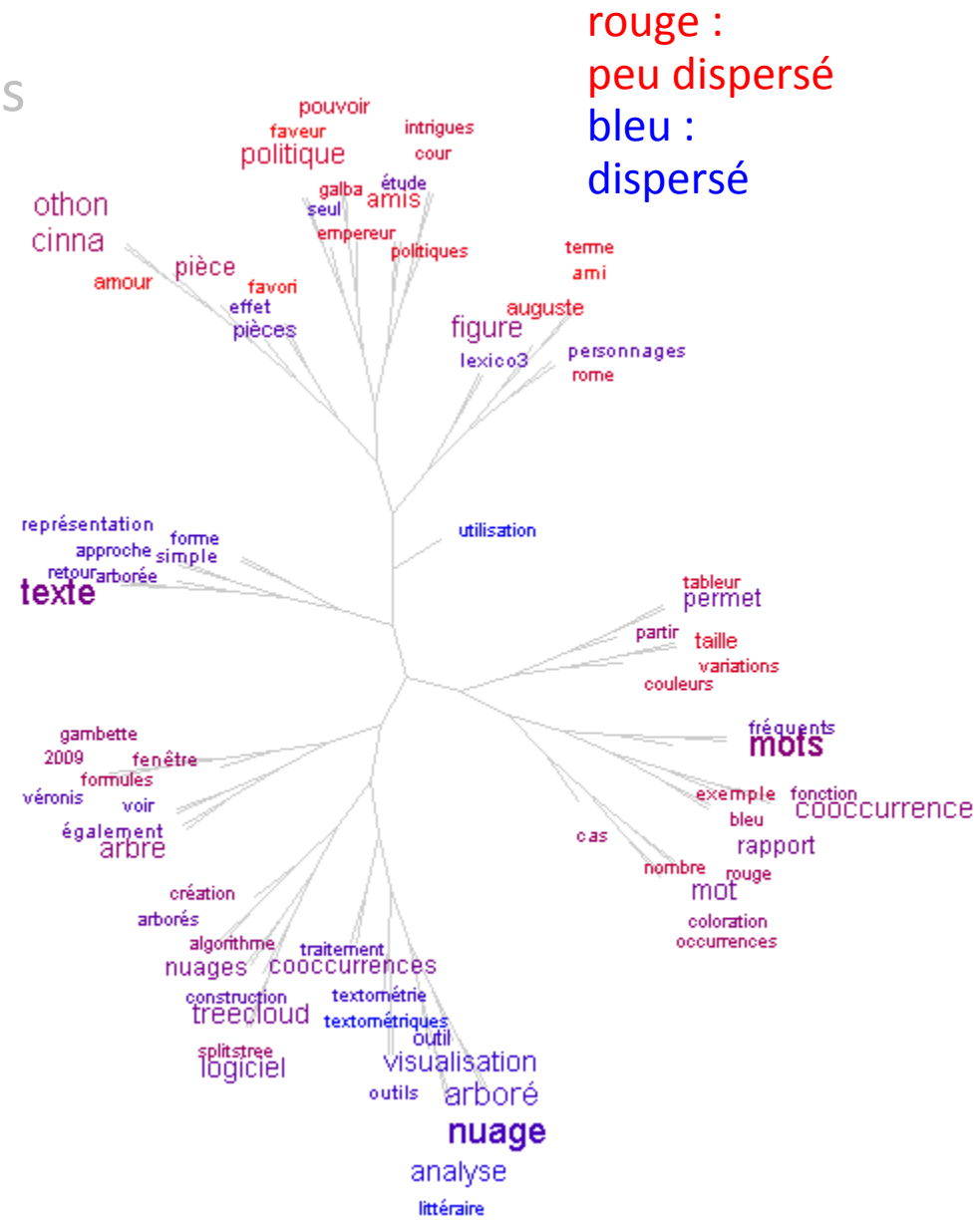
- coloration selon les fréquences
- coloration chronologique
- coloration de la dispersion
- coloration ciblée sur un mot
- coloration grammaticale



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

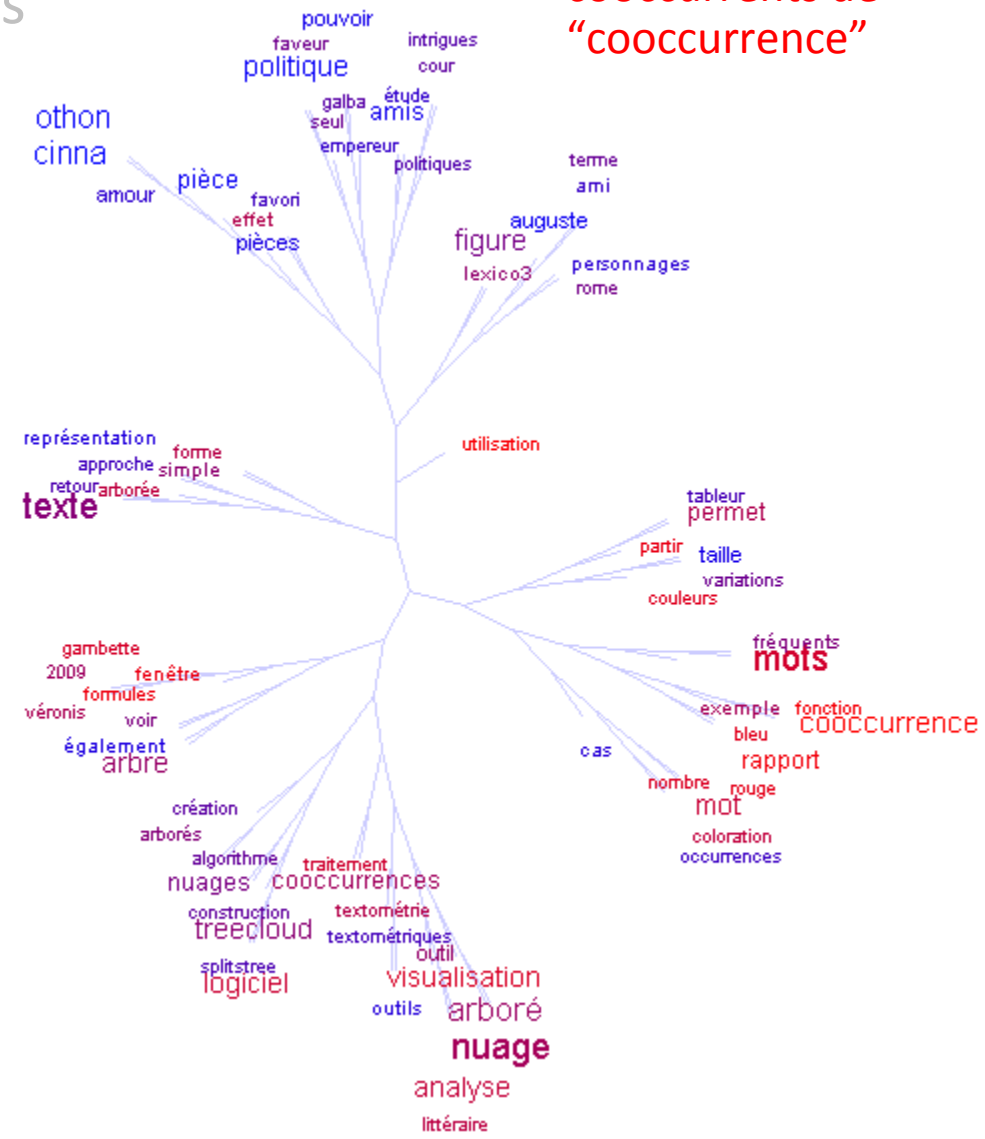
- coloration selon les fréquences
- coloration chronologique
- coloration de la dispersion
- coloration ciblée sur un mot
- coloration grammaticale



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

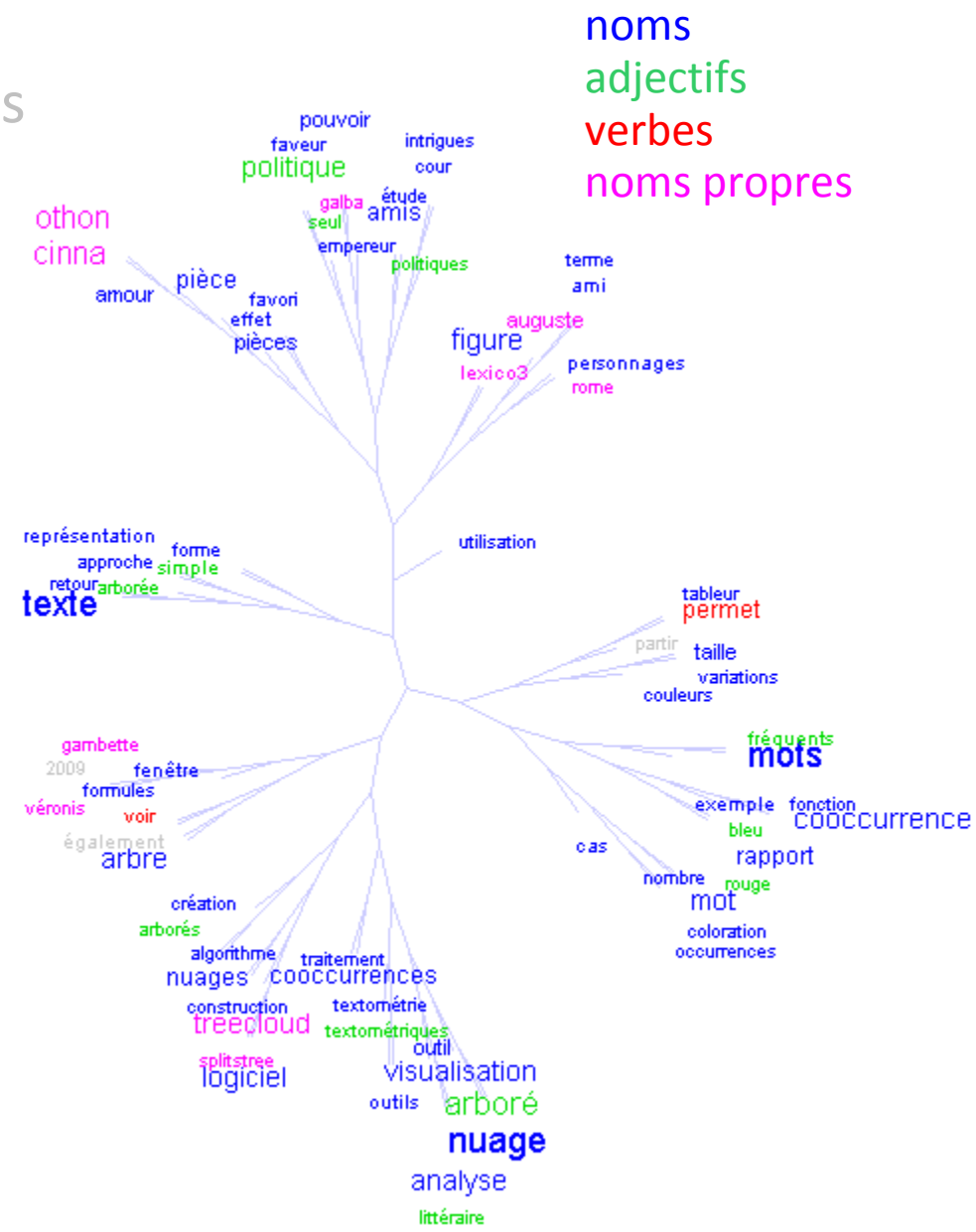
- coloration selon les fréquences
- coloration chronologique
- coloration de la dispersion
- coloration ciblée sur un mot
- coloration grammaticale



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

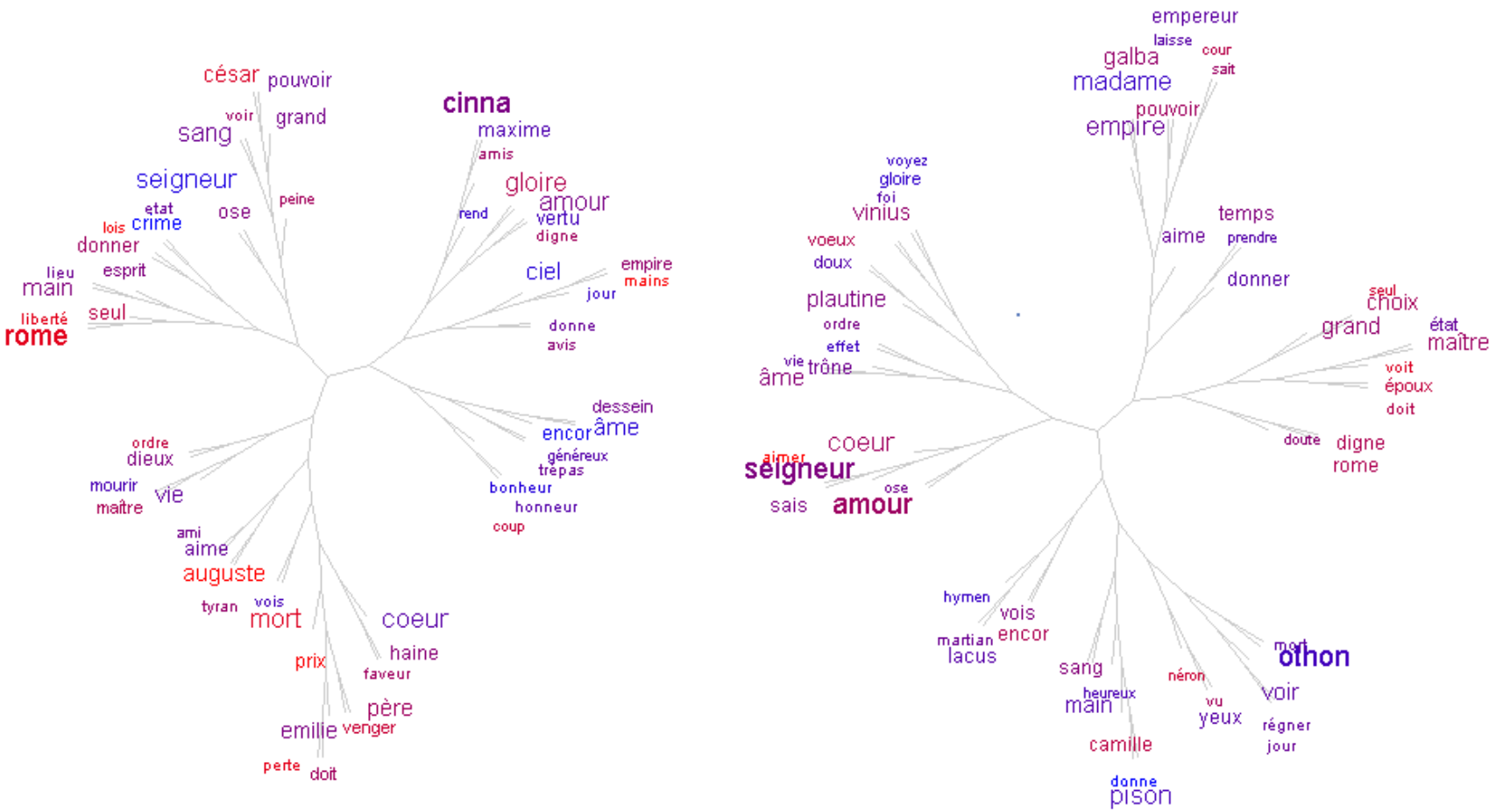
Plan

- Nuages de mots et nuages arborés
- Caractéristiques du nuage arboré
- Utilisations du nuage arboré
- Construction d'un nuage arboré
- Évaluation de la robustesse de l'arbre
- Calcul des longueurs d'arêtes de l'arbre
- Perspectives

Utilisations du nuage arboré

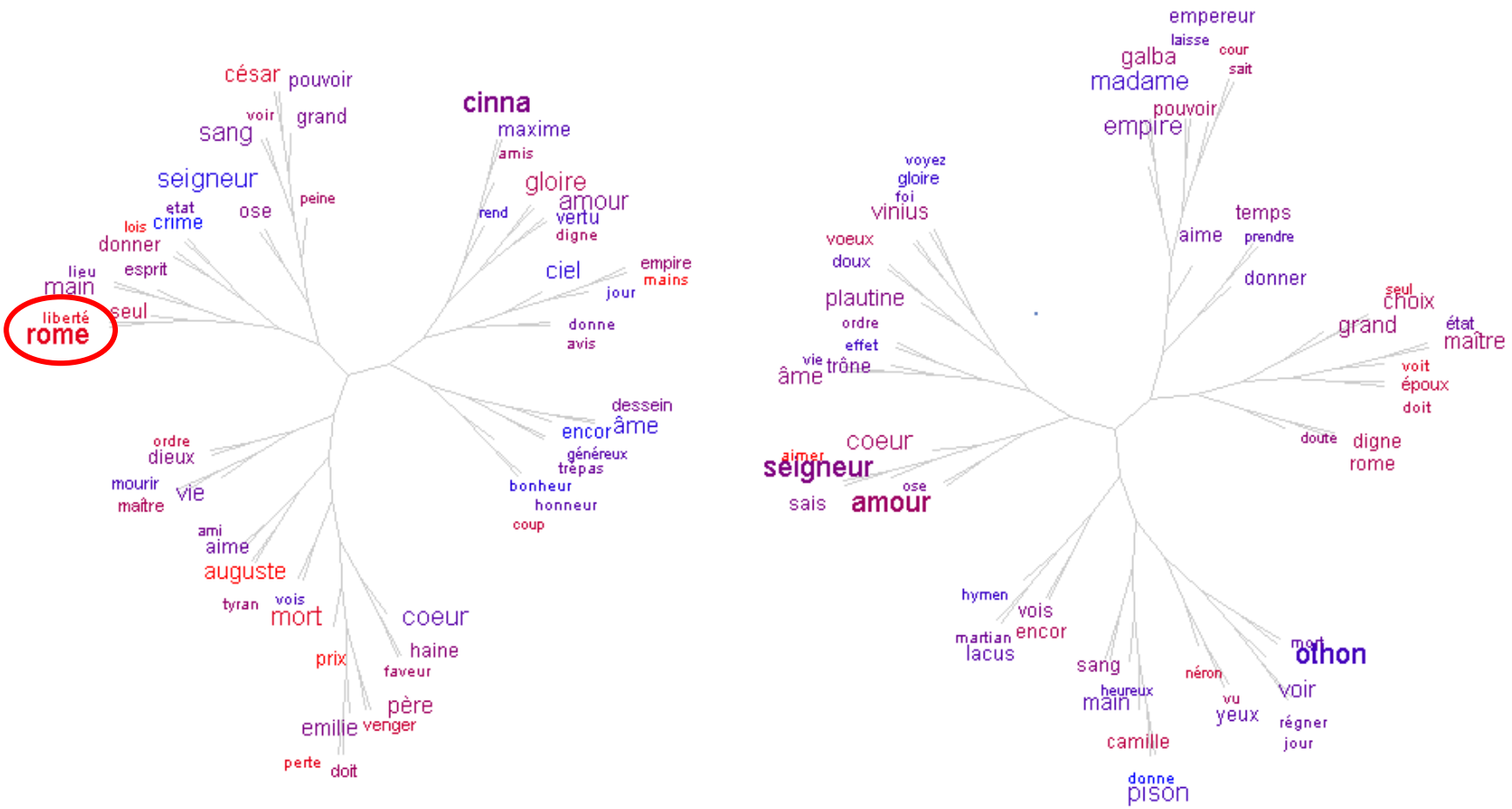
- **Résumé visuel** des thématiques d'un texte
- **Clarification** de rapports ou discours
- Analyse des **réponses aux questions ouvertes** ?
- En **analyse littéraire** :
  - susciter, formaliser et étayer des **hypothèses de travail**
  - **comparer des textes** selon leur représentation arborée
  - hiérarchiser l'utilisation d'**autres outils textométriques**
  - représenter les **résultats de l'analyse**

Illustration sur *Cinna* et *Othon*



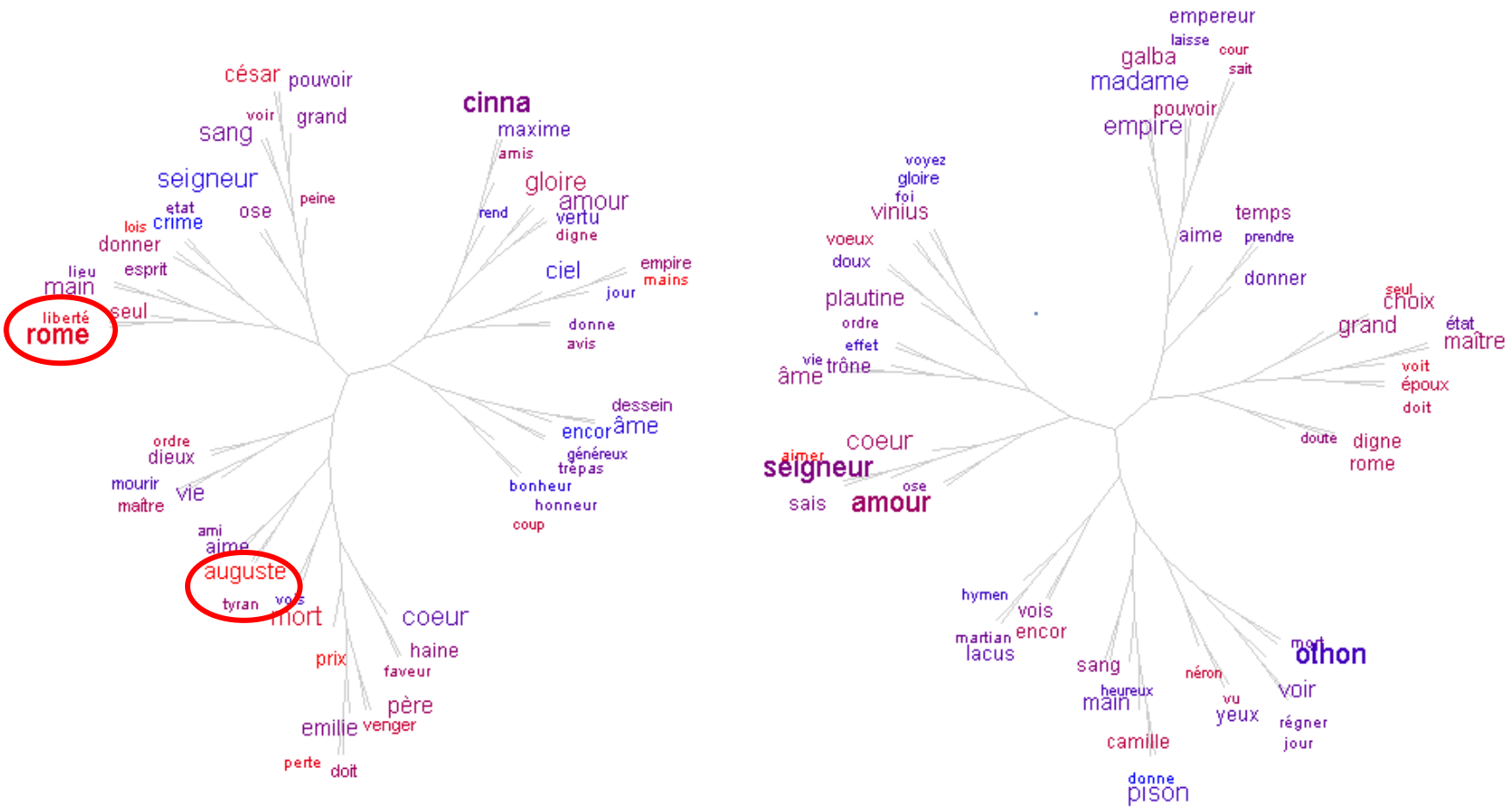
Nuages arborés globaux des 60 mots les plus fréquents dans Cinna et Othon (distance Liddell, fenêtre de largeur 20), colorés chronologiquement (rouge au début, bleu à la fin)

Illustration sur *Cinna* et *Othon*



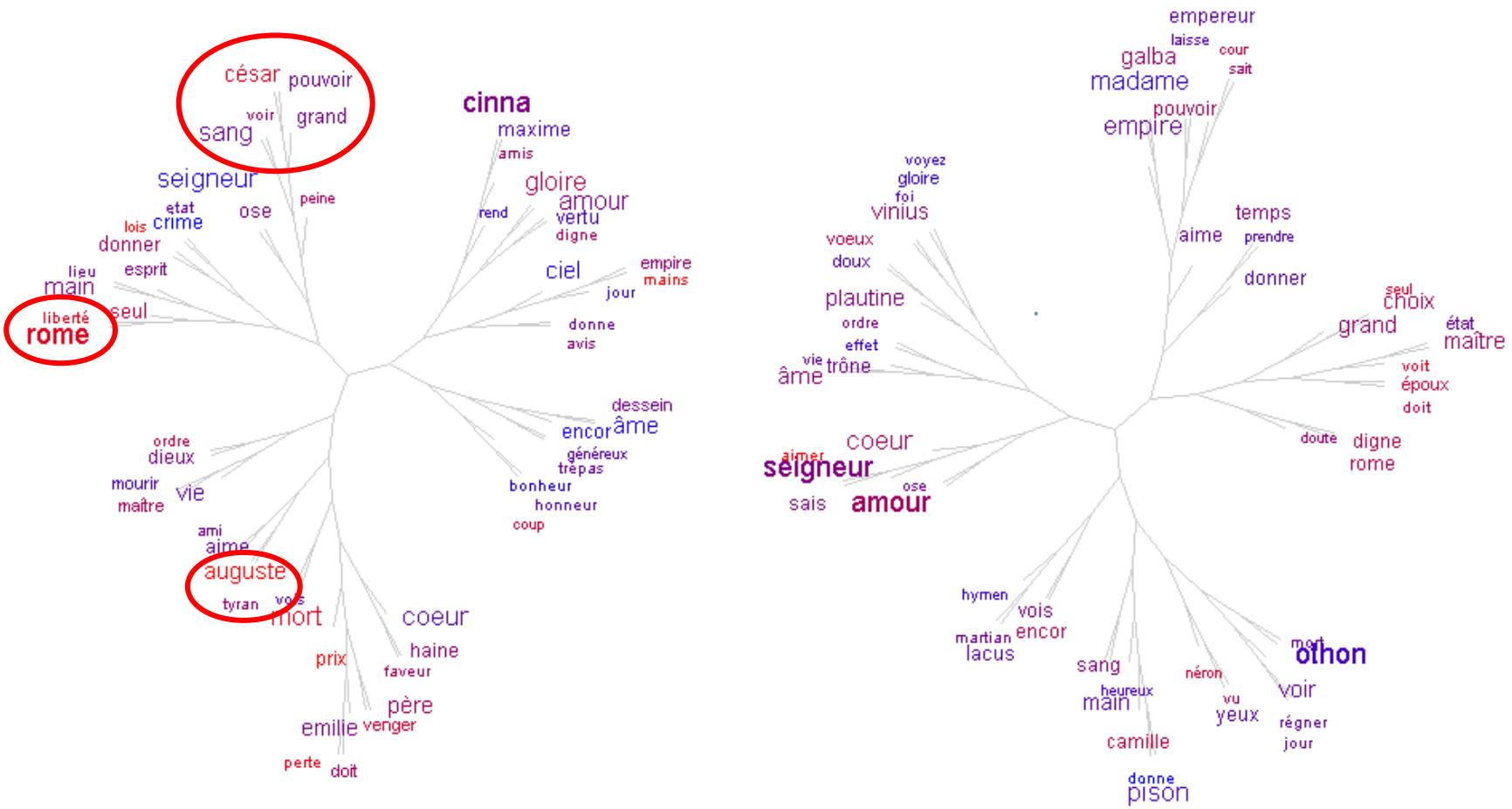
Nuages arborés globaux des 60 mots les plus fréquents dans Cinna et Othon (distance Liddell, fenêtre de largeur 20), colorés chronologiquement (rouge au début, bleu à la fin)

Illustration sur *Cinna* et *Othon*



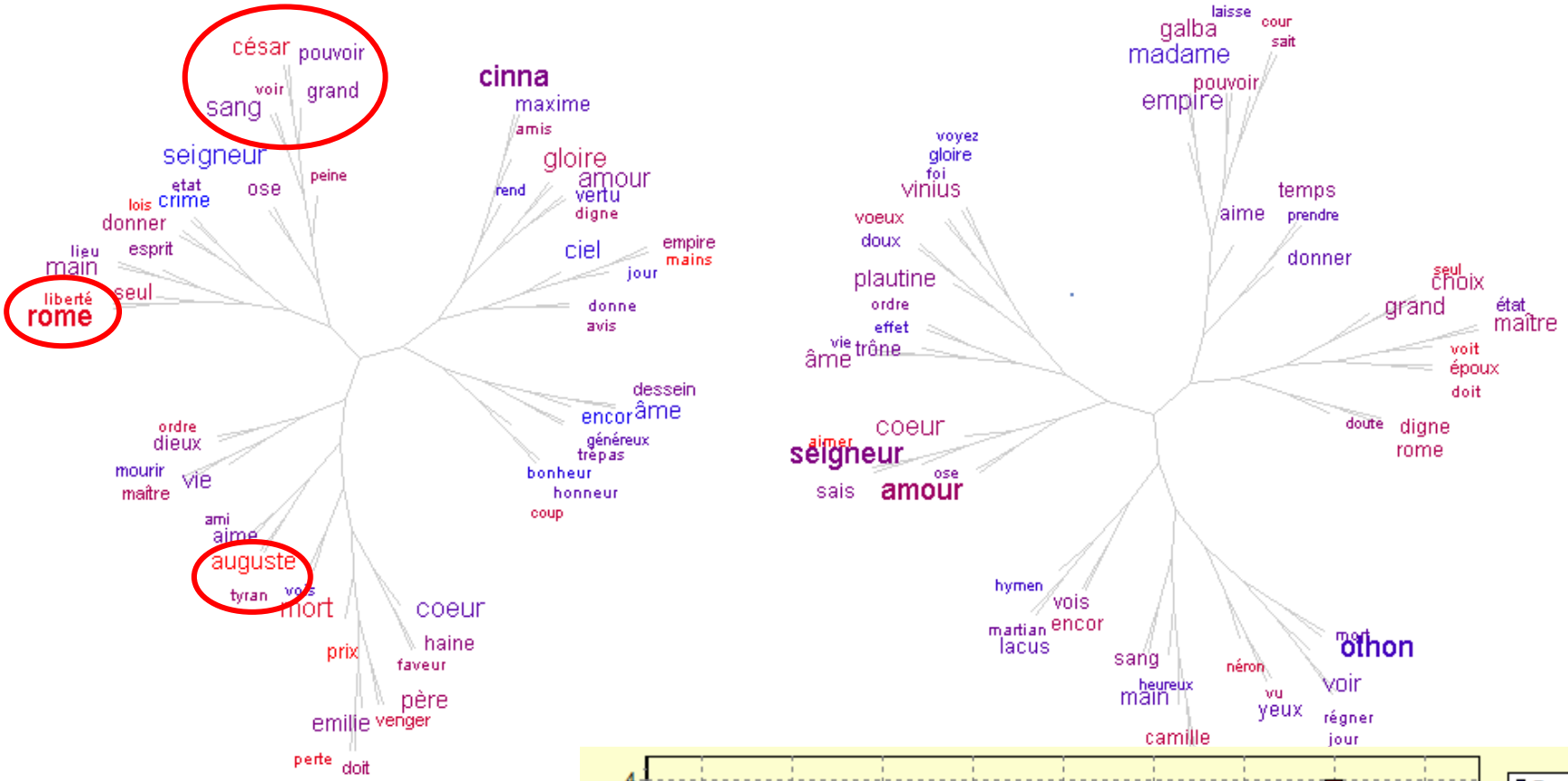
Nuages arborés globaux des 60 mots les plus fréquents dans Cinna et Othon (distance Liddell, fenêtre de largeur 20), colorés chronologiquement (rouge au début, bleu à la fin)

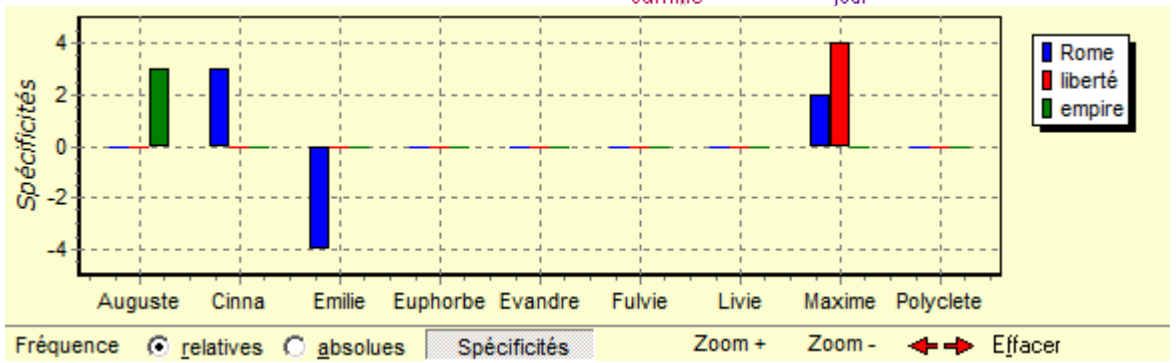
Illustration sur *Cinna* et *Othon*



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

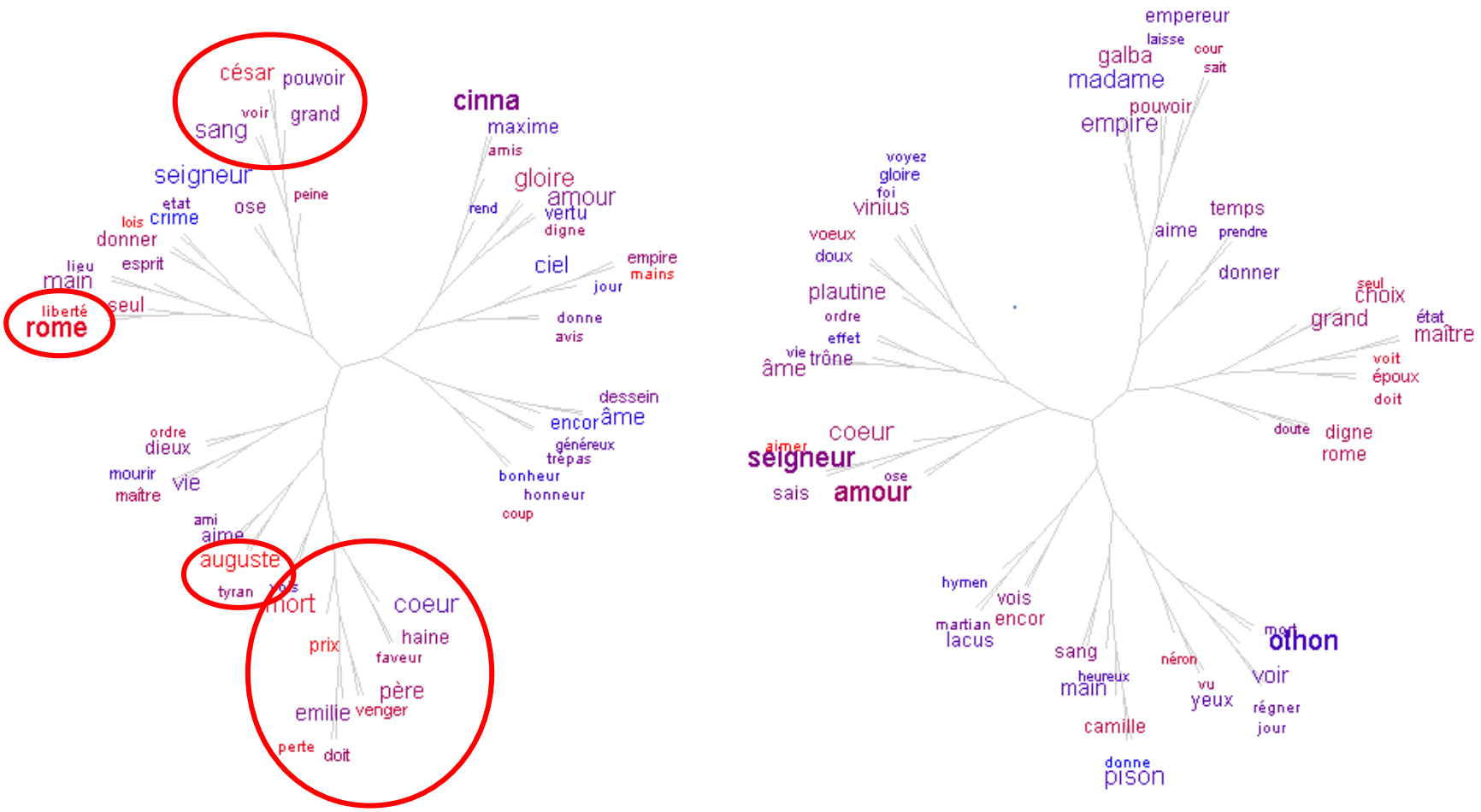
Illustration sur *Cinna* et *Othon*





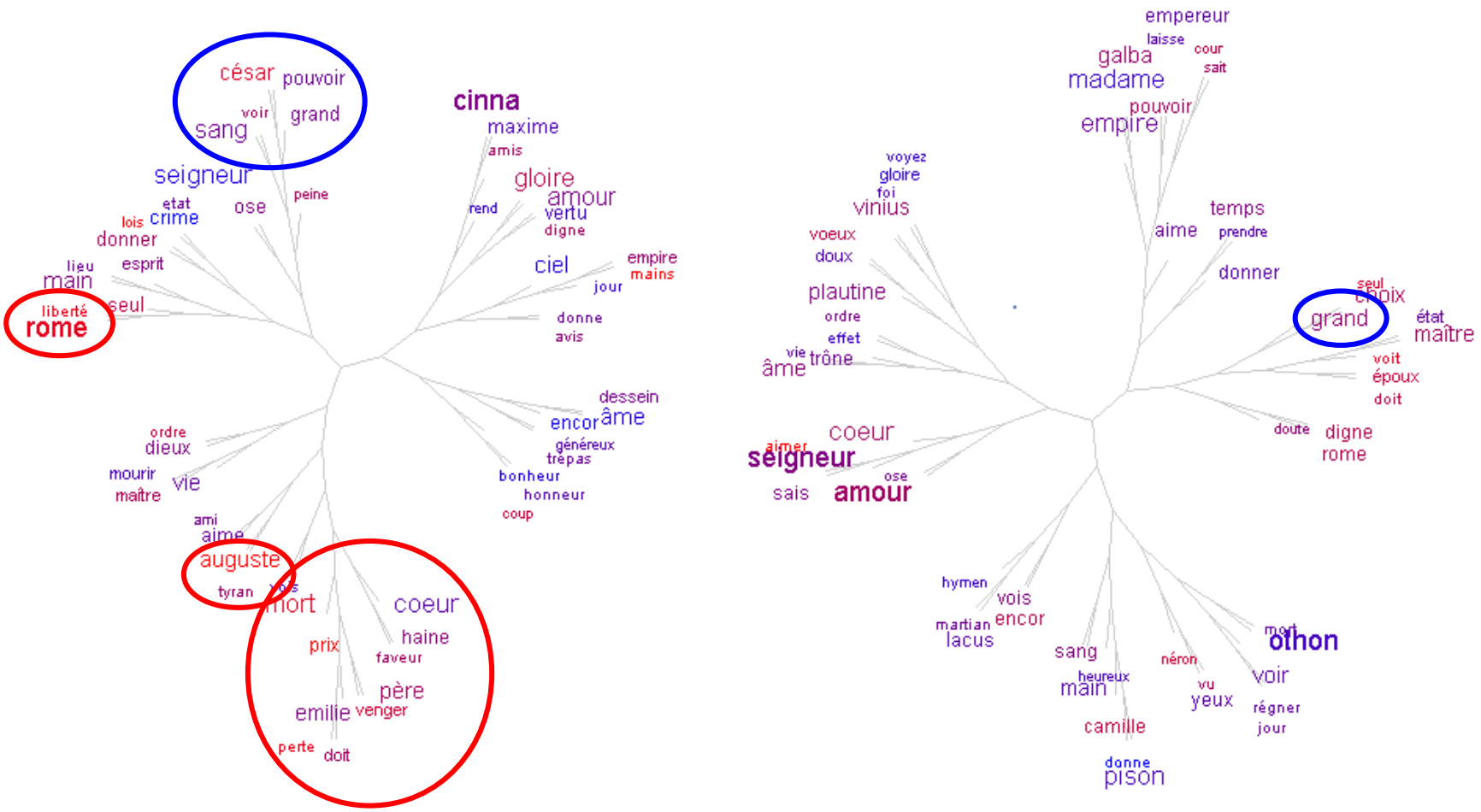
Spécificités d'emploi de « Rome », « liberté » et « empire » chez les différents personnages de Cinna dans Lexico3.

Illustration sur *Cinna* et *Othon*



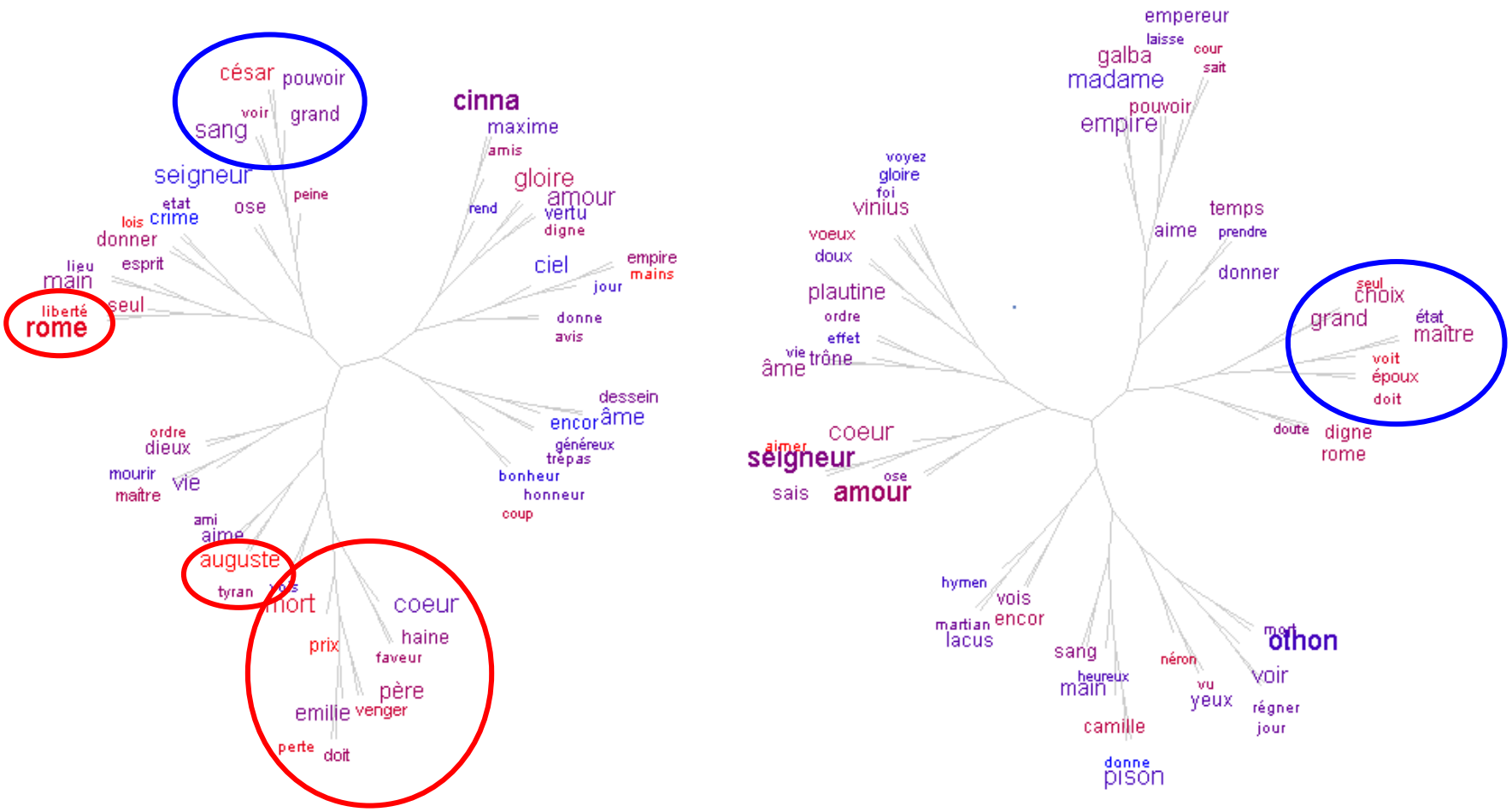
Nuages arborés globaux des 60 mots les plus fréquents dans Cinna et Othon (distance Liddell, fenêtre de largeur 20), colorés chronologiquement (rouge au début, bleu à la fin)

Illustration sur *Cinna* et *Othon*



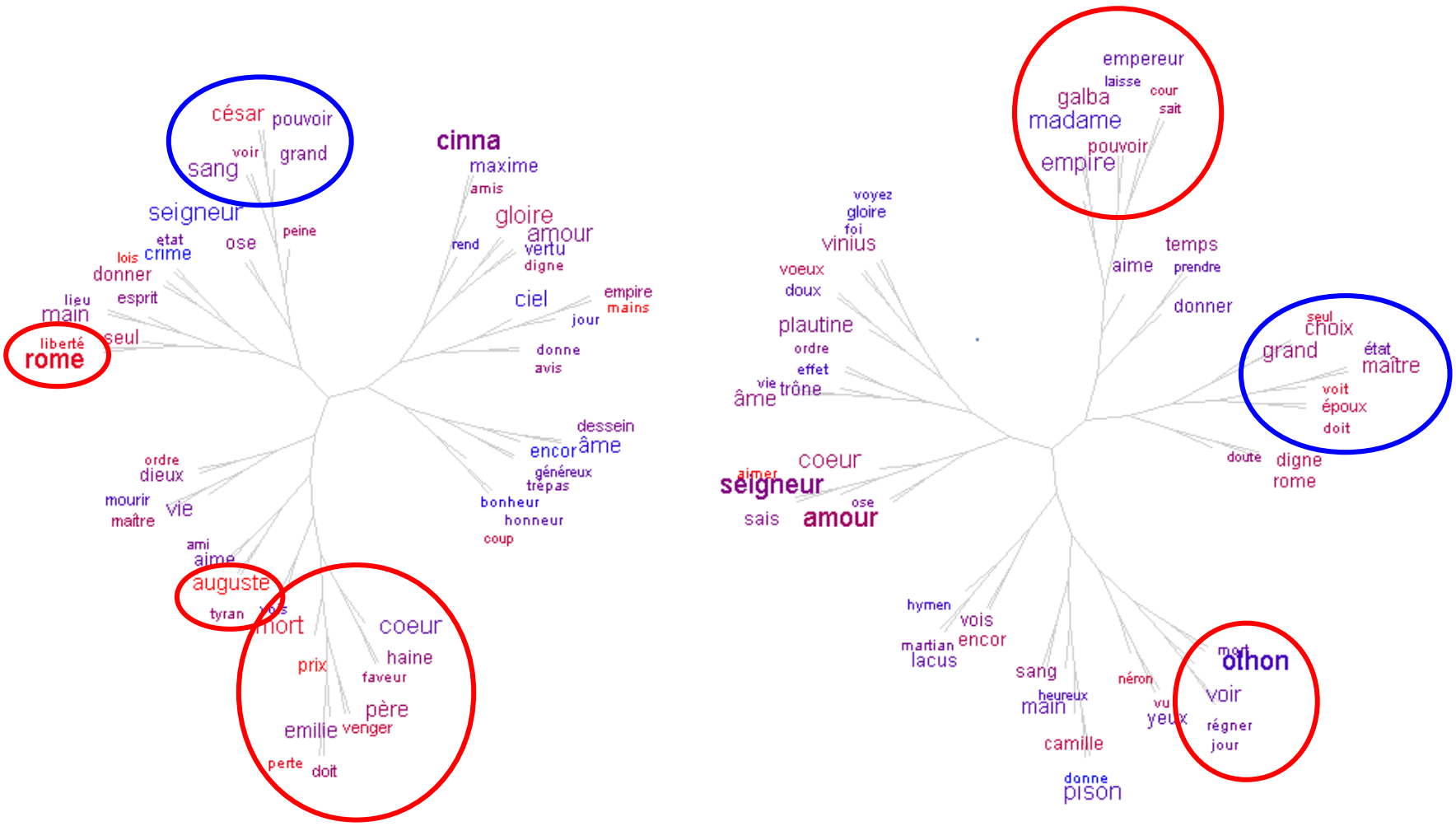
Nuages arborés globaux des 60 mots les plus fréquents dans Cinna et Othon (distance Liddell, fenêtre de largeur 20), colorés chronologiquement (rouge au début, bleu à la fin)

Illustration sur *Cinna* et *Othon*



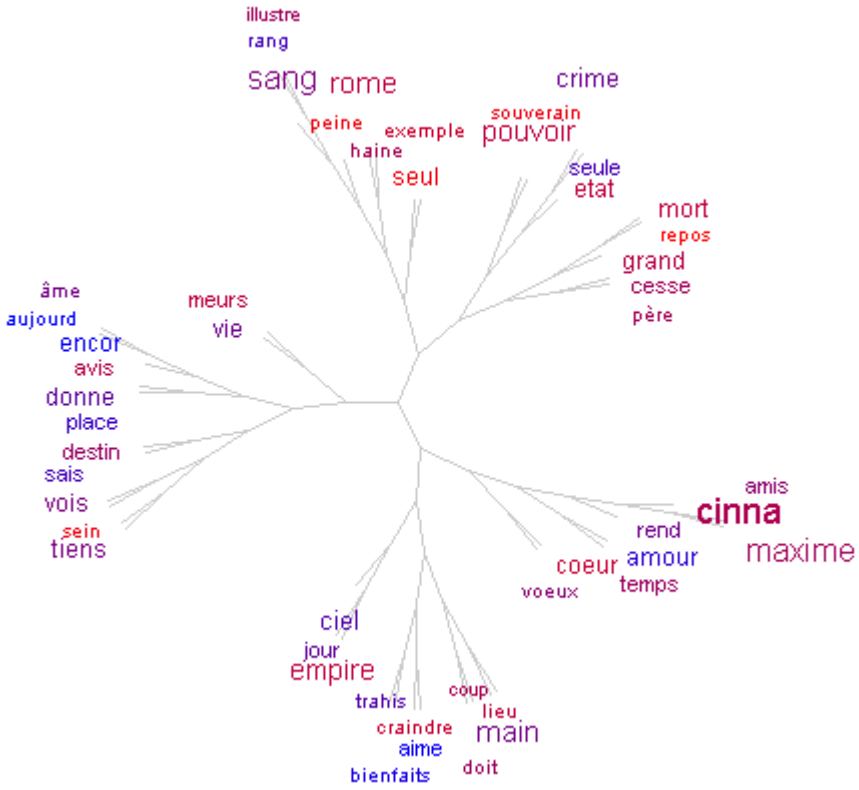
Nuages arborés globaux des 60 mots les plus fréquents dans Cinna et Othon (distance Liddell, fenêtre de largeur 20), colorés chronologiquement (rouge au début, bleu à la fin)

Illustration sur *Cinna* et *Othon*



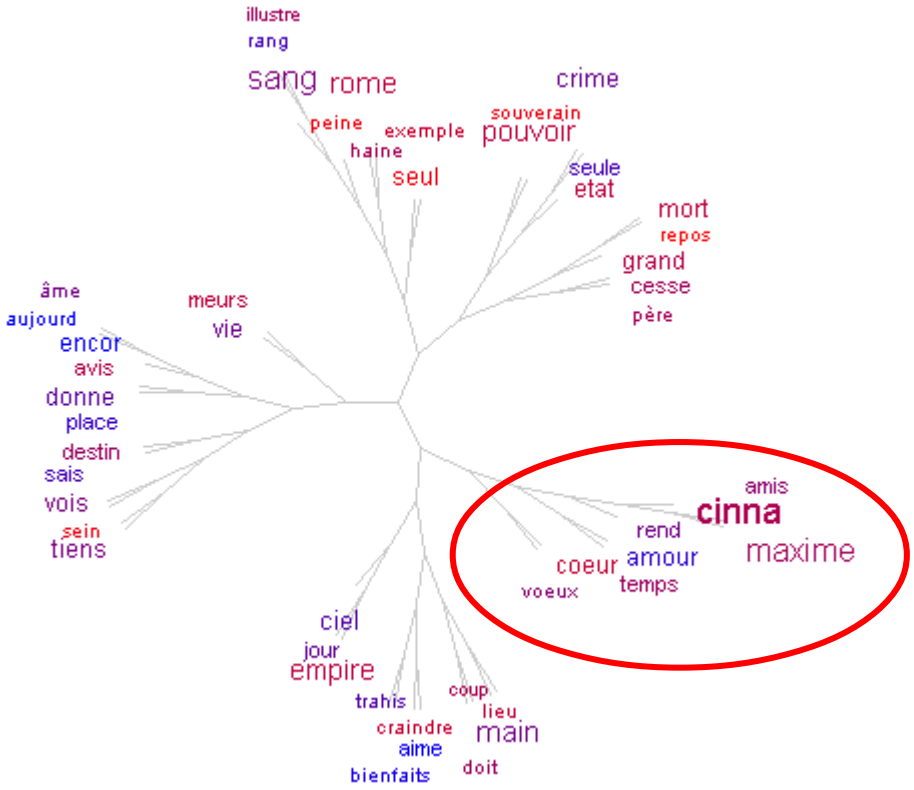
Nuages arborés globaux des 60 mots les plus fréquents dans Cinna et Othon (distance Liddell, fenêtre de largeur 20), colorés chronologiquement (rouge au début, bleu à la fin)

Illustration sur *Cinna* et *Othon*



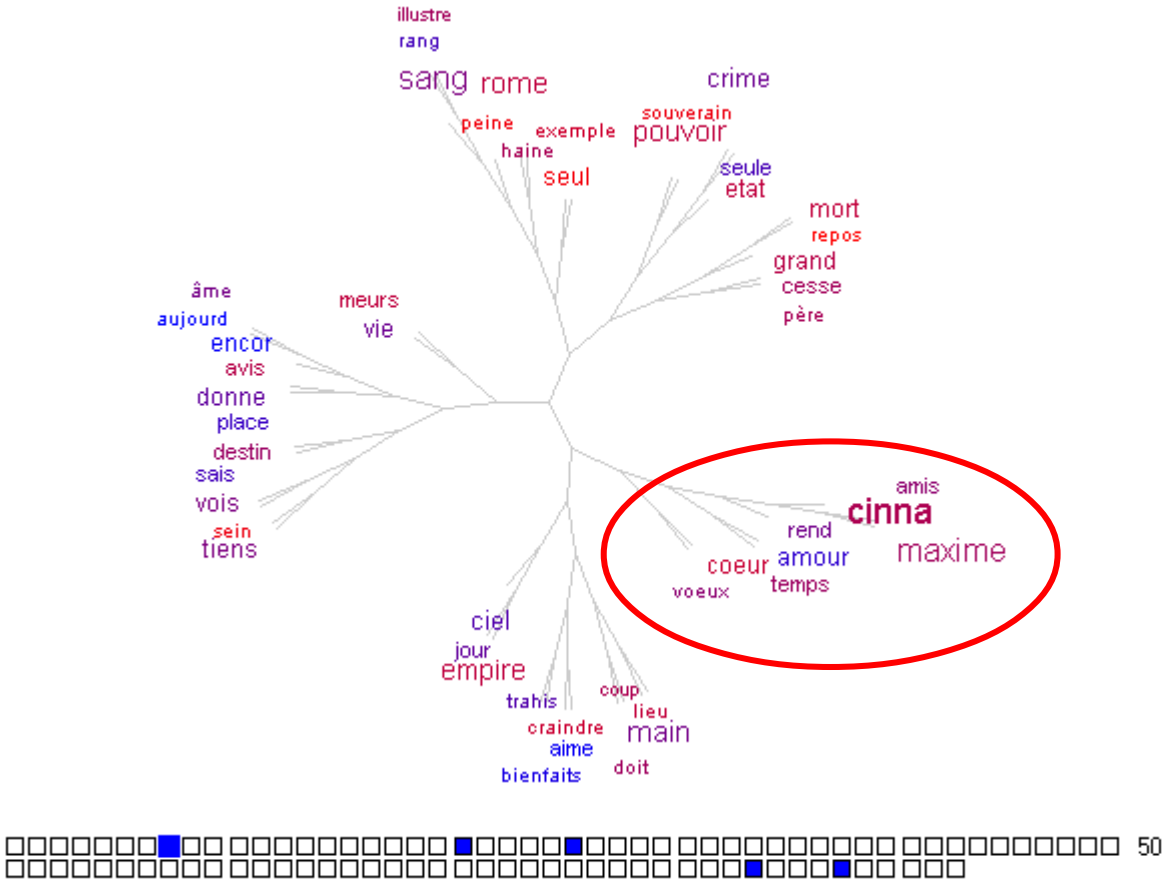
Nuage arboré des 50 mots les plus fréquents des paroles d'Auguste dans Cinna

Illustration sur *Cinna* et *Othon*



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

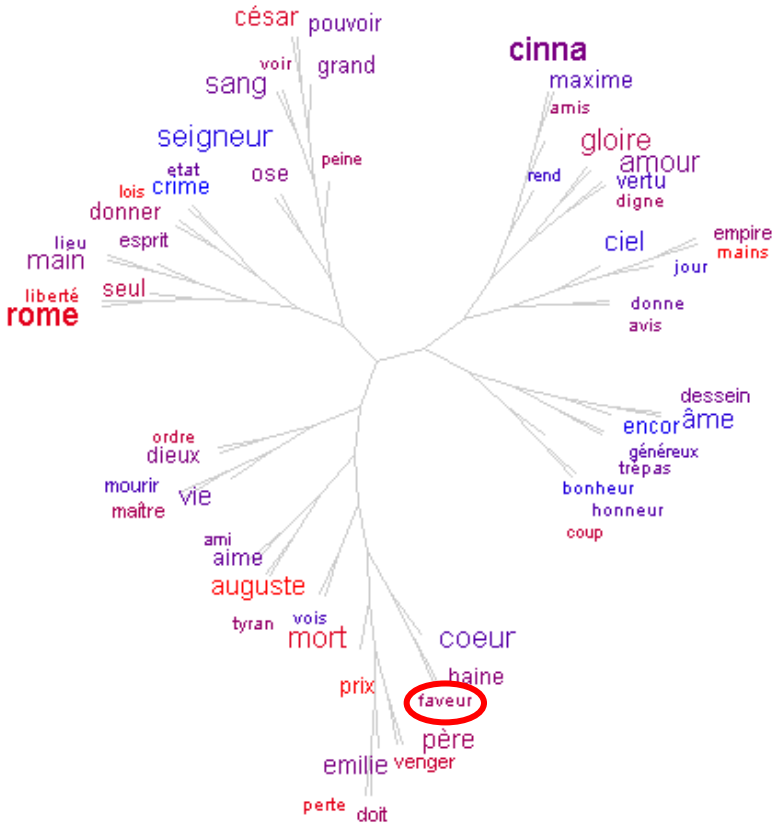
Illustration sur *Cinna* et *Othon*



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

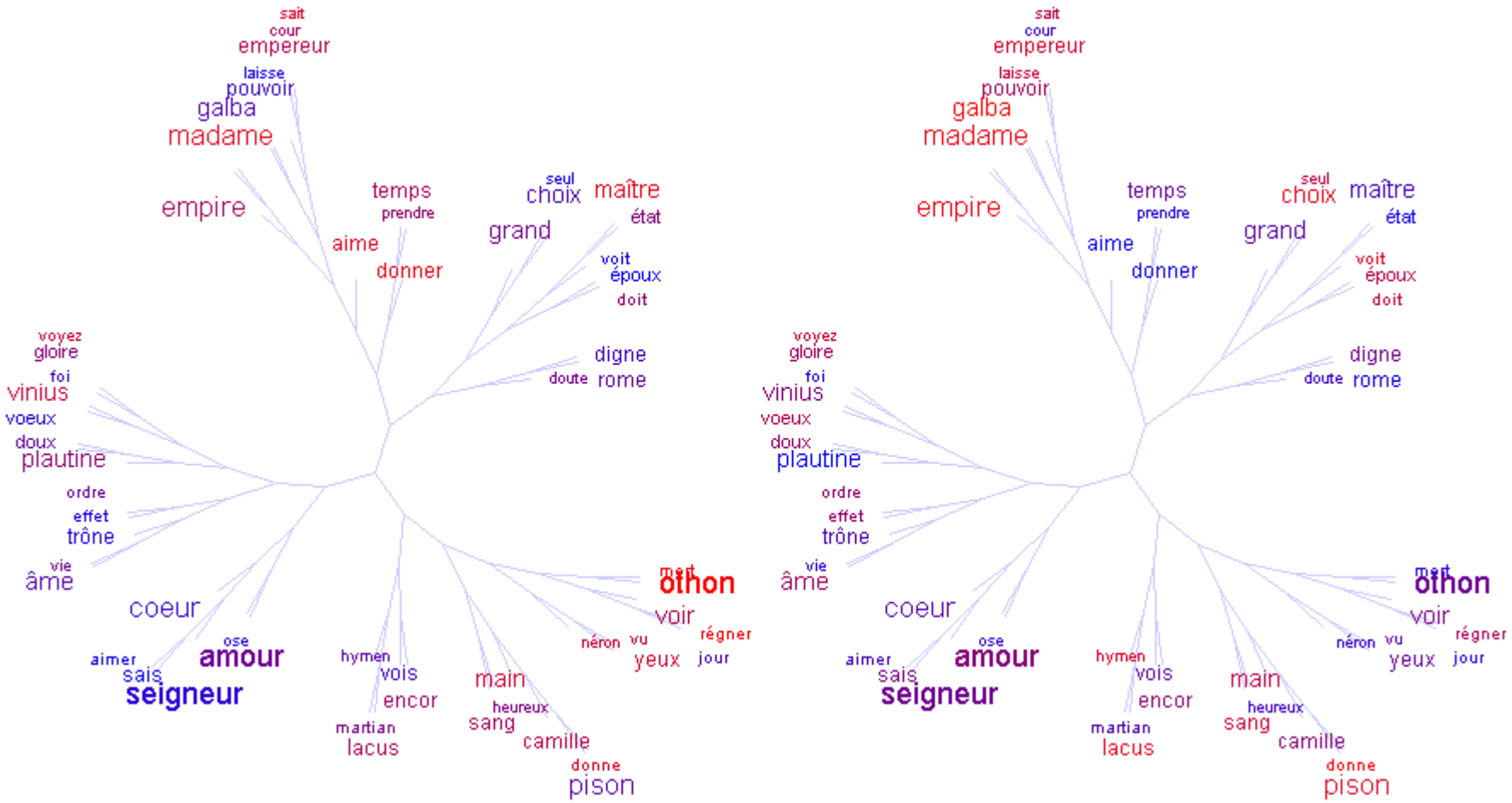
1. *Voilà, mes chers **amis**, ce qui me met en peine.*
2. *Quoi ! mes plus chers **amis** ! quoi ! Cinna ! quoi ! Maxime !*
3. *Reprenez le pouvoir que vous m'avez commis, Si donnant des sujets il ôte les **amis***
4. *Soyons **amis**, Cinna, c'est moi qui t'en convie*
5. *Il nous a trahis tous ; mais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

Illustration sur *Cinna* et *Othon*



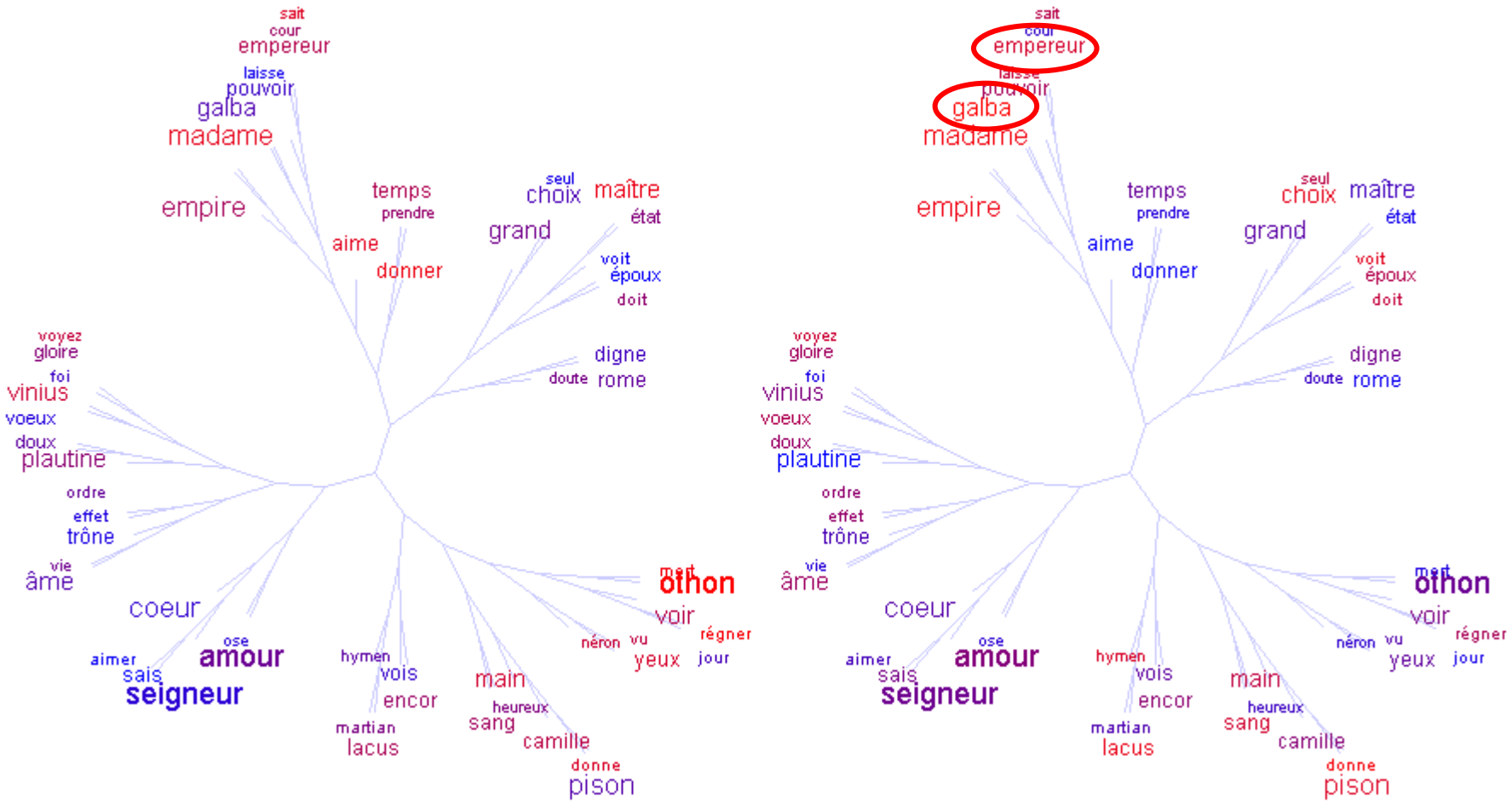
Nuages arborés globaux des 60 mots les plus fréquents dans Cinna et Othon (distance Liddell, fenêtre de largeur 20), colorés chronologiquement (rouge au début, bleu à la fin)

Illustration sur *Cinna* et *Othon*



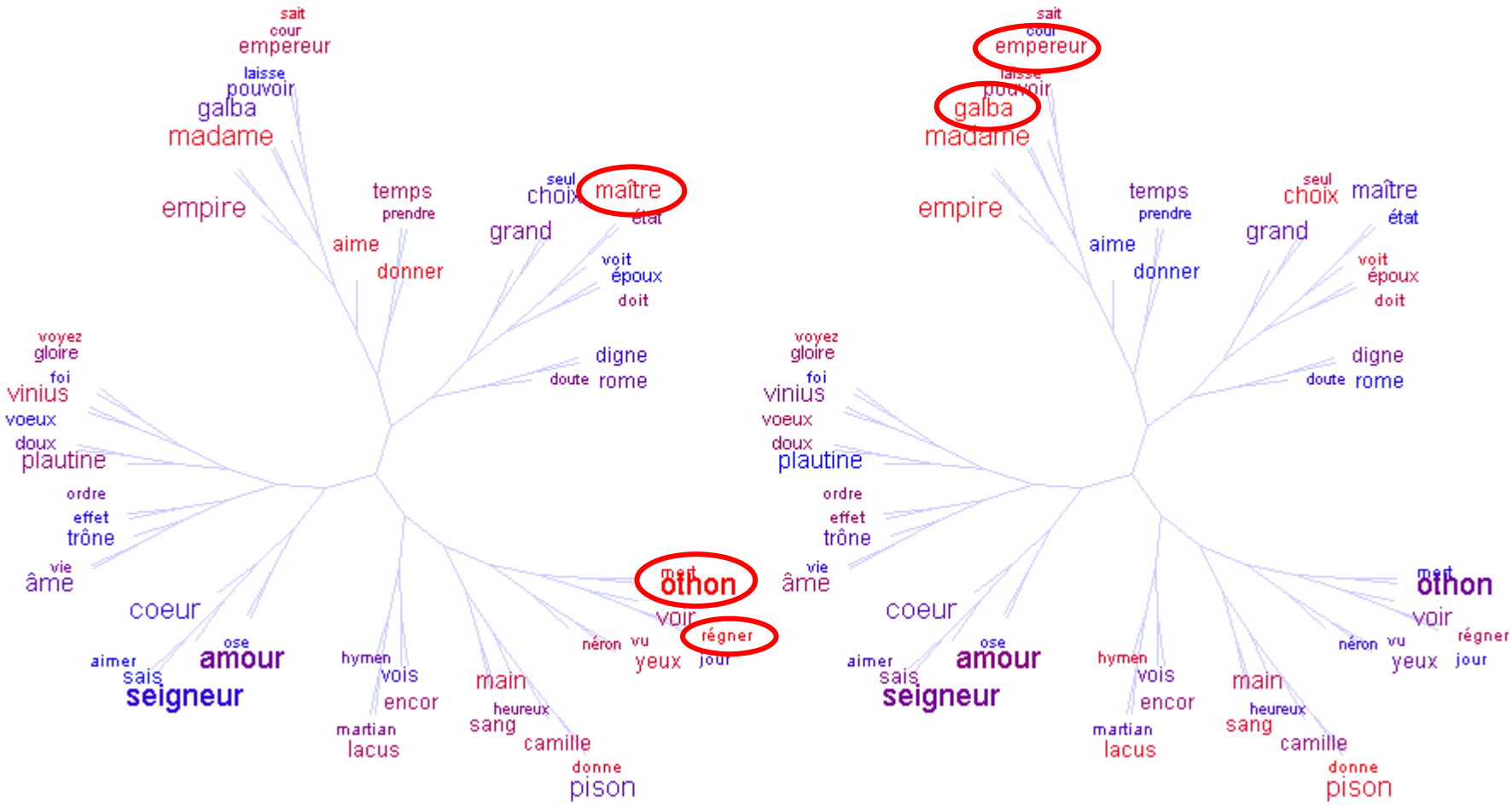
Nuage arboré des 30 mots les plus fréquents de la pièce Othon, coloré à gauche par rapport aux cooccurrences avec « Othon », à droite par rapport à celles avec « Galba »

Illustration sur *Cinna* et *Othon*



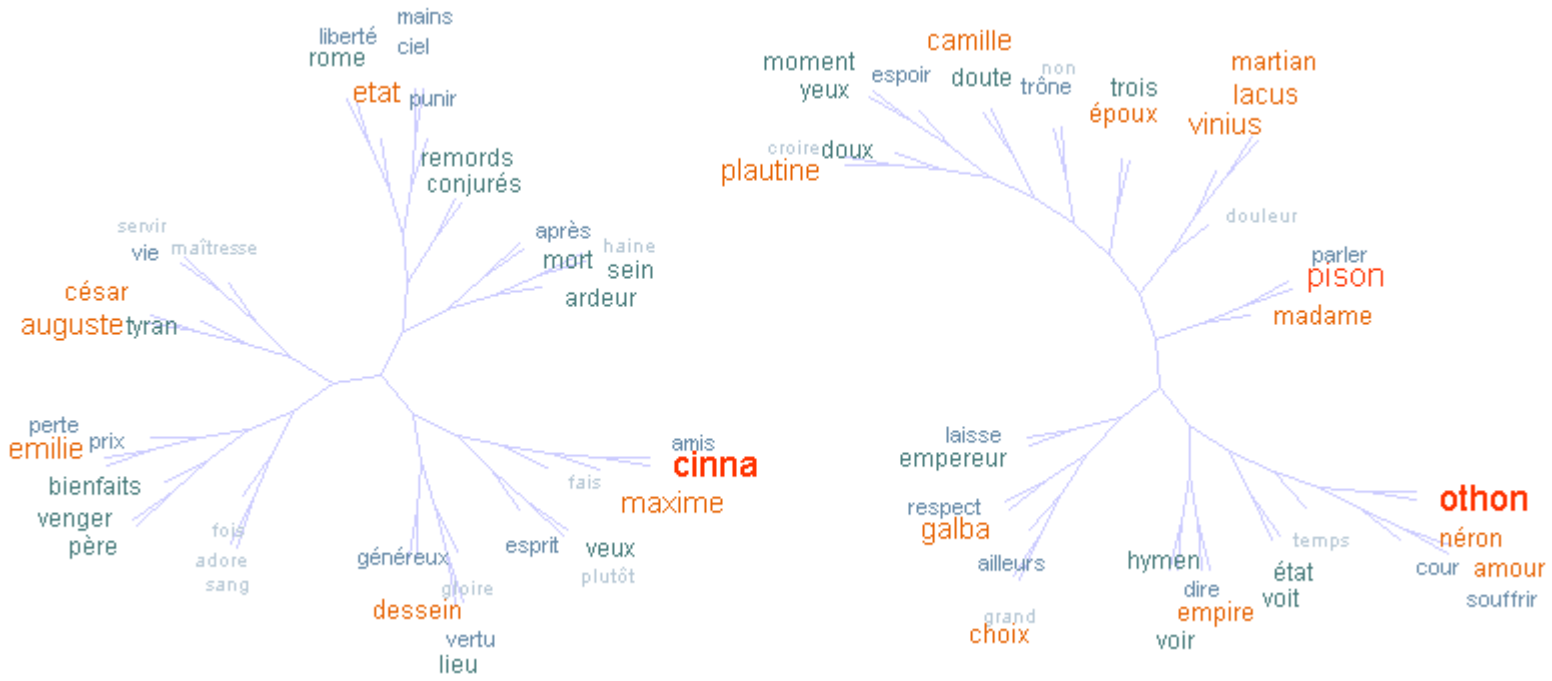
Nuage arboré des 30 mots les plus fréquents de la pièce Othon, coloré à gauche par rapport aux cooccurrences avec « Othon », à droite par rapport à celles avec « Galba »

Illustration sur *Cinna* et *Othon*



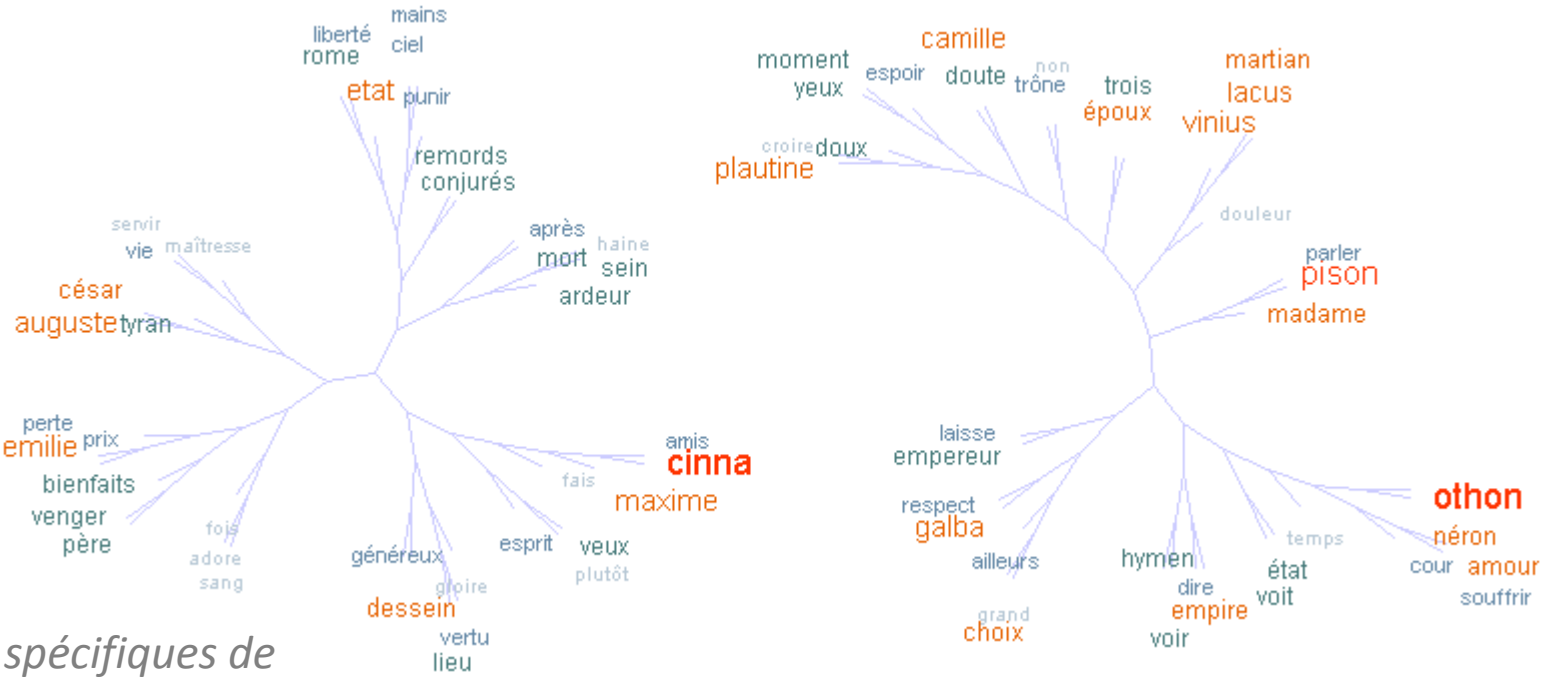
Nuage arboré des 30 mots les plus fréquents de la pièce Othon, coloré à gauche par rapport aux cooccurrences avec « Othon », à droite par rapport à celles avec « Galba »

Illustration sur *Cinna* et *Othon*



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

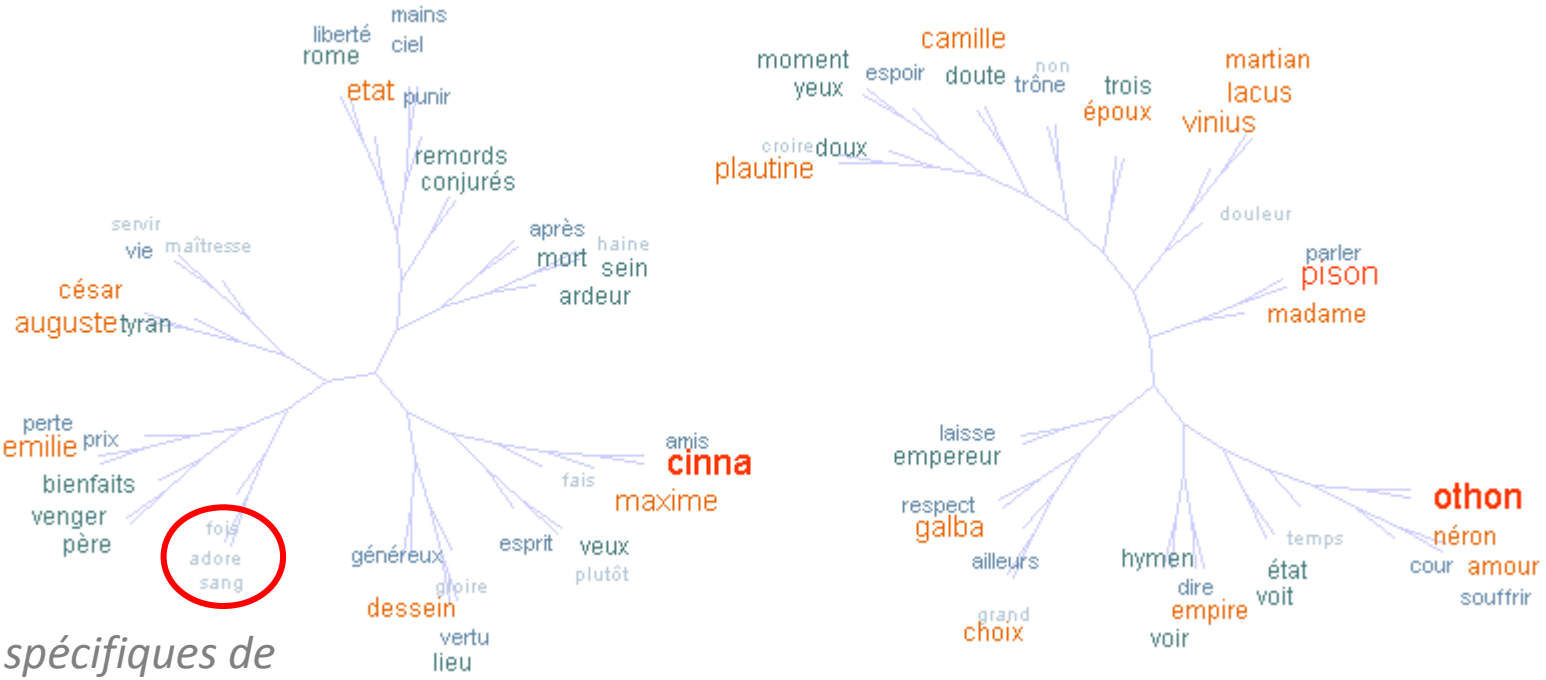
Illustration sur *Cinna* et *Othon*



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

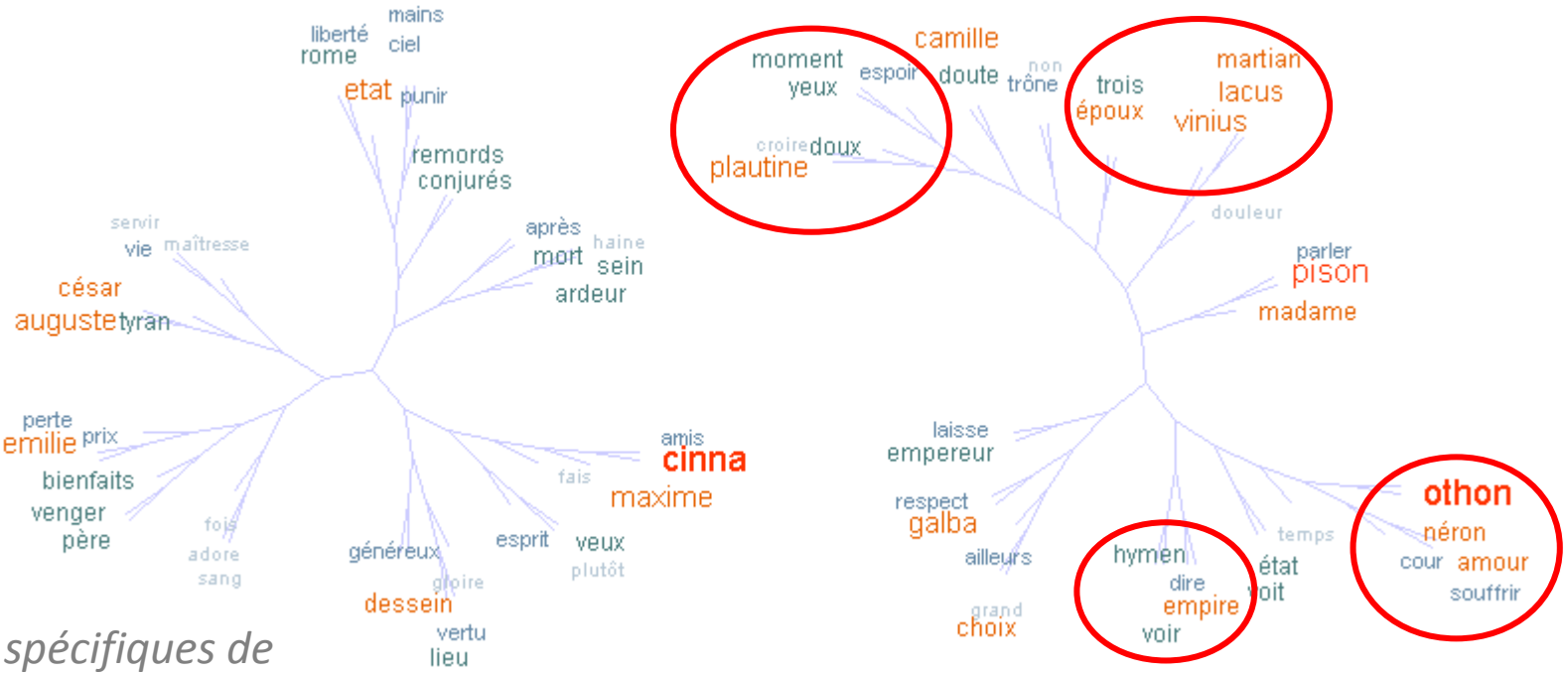
Illustration sur *Cinna* et *Othon*



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 *Cinna* et *Othon*



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

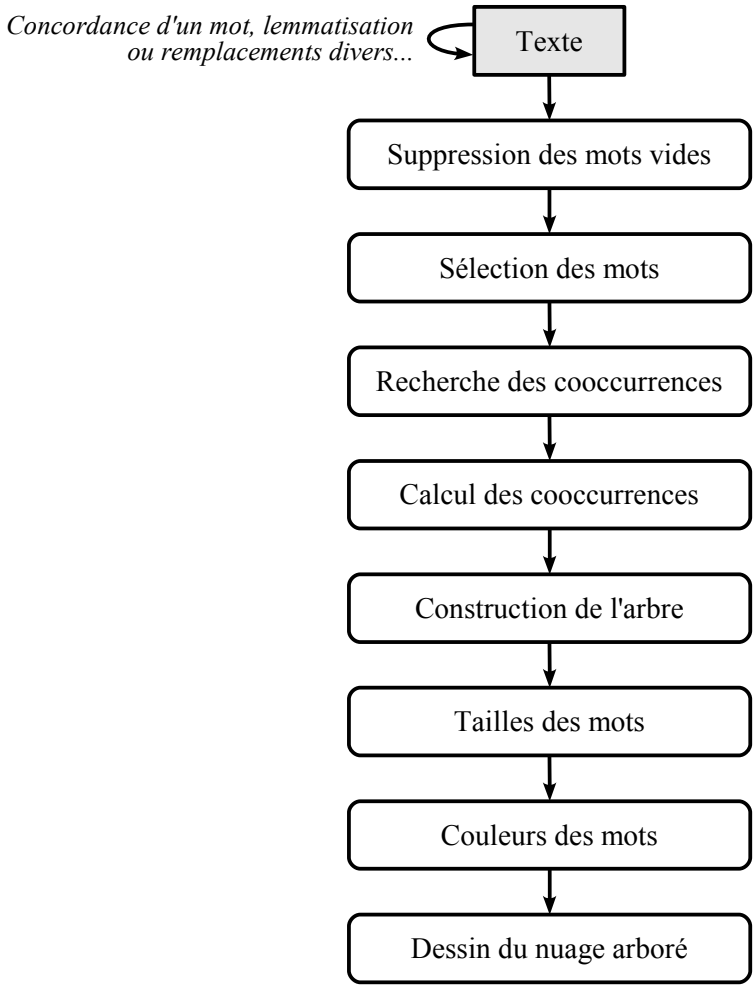
	Cinna	*Othon*
Lieu du pouvoir et objet de la confrontation entre les personnages	Rome (« liberté »)	Empire (« trône »)
Souverain en place	tyran	Empereur
Membres du corps politique	amis	maîtres / seigneurs
Moyens au service de la cause politique	gloire	amour matrimonial (« amour », « hymen », « choix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

Plan

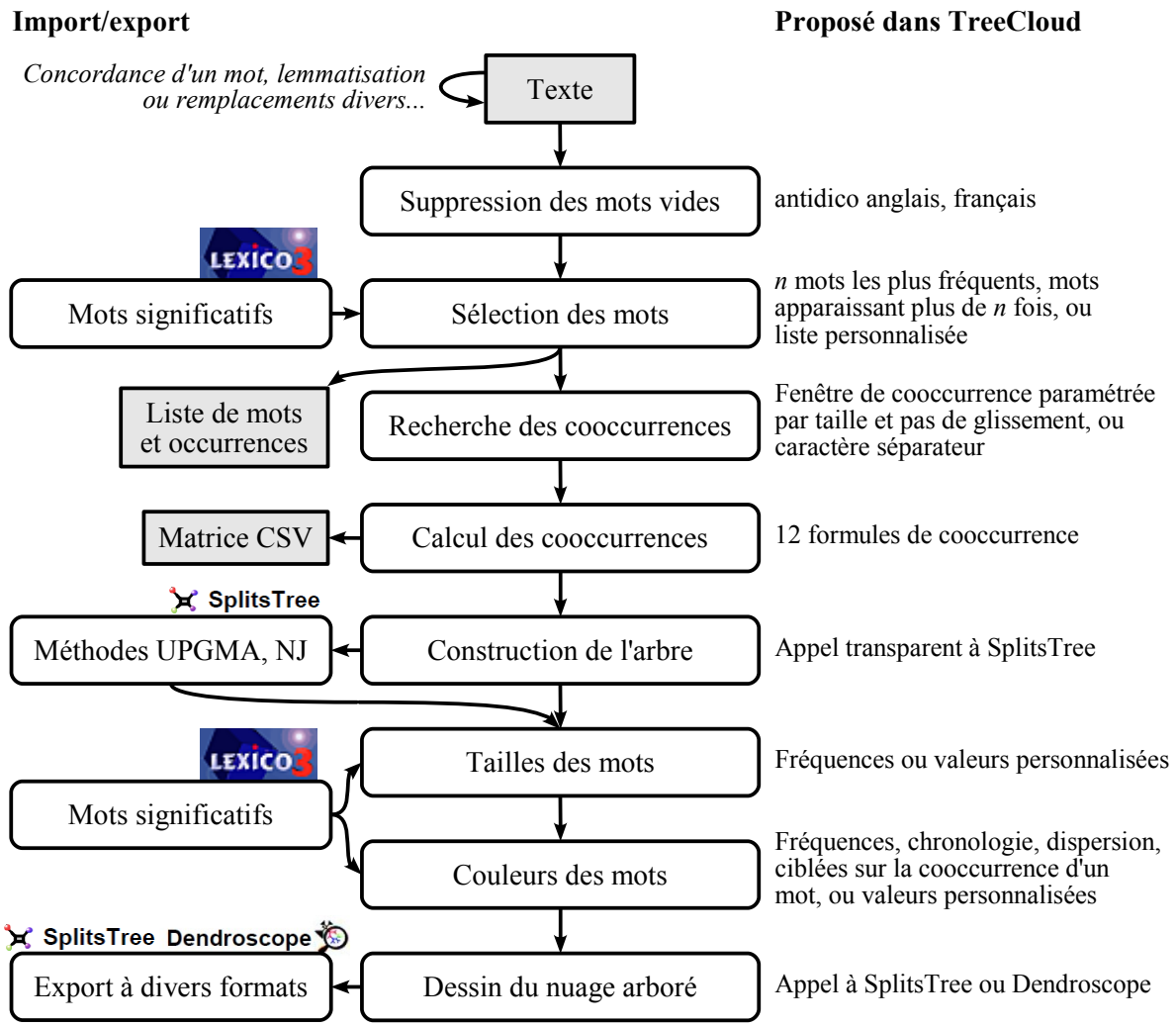
- Nuages de mots et nuages arborés
- Caractéristiques du nuage arboré
- Utilisations du nuage arboré
- Construction d'un nuage arboré
- Évaluation de la robustesse de l'arbre
- Calcul des longueurs d'arêtes de l'arbre
- Perspectives

Processus de construction

Import/export

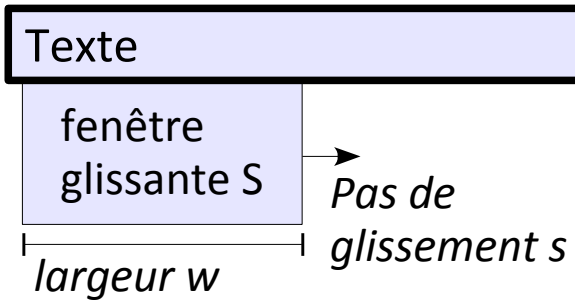


Processus de construction



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 fenêtre *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] ?

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]

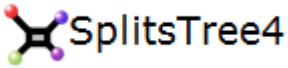
Normalisation des scores de similarité sur [0,1] :
- normalisation linéaire pour les matrices positives
- normalisation affines pour les matrices contenant des valeurs négatives, afin d'obtenir des distances dans l'intervalle [a,1] (a=0.1)

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!)

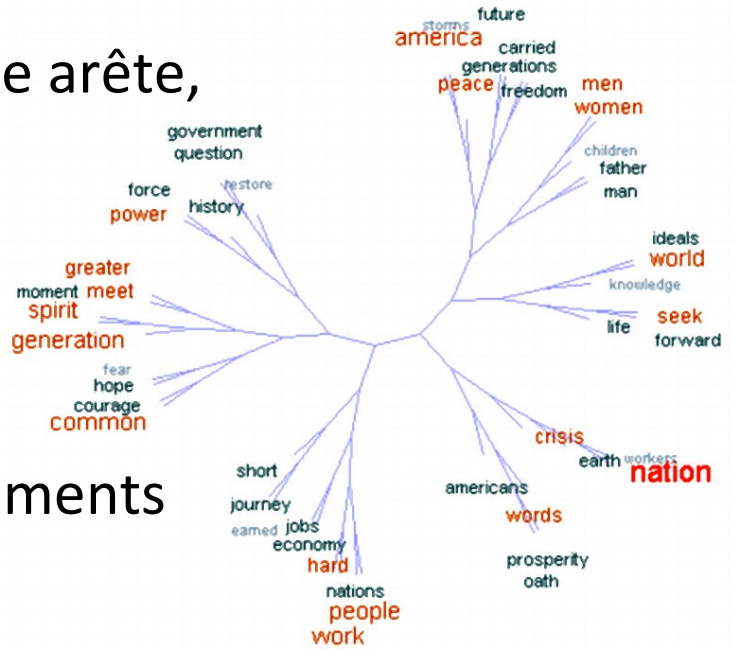
- calculées à partir des **rangs des fréquences**
  (distribution exponentielle)

- **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

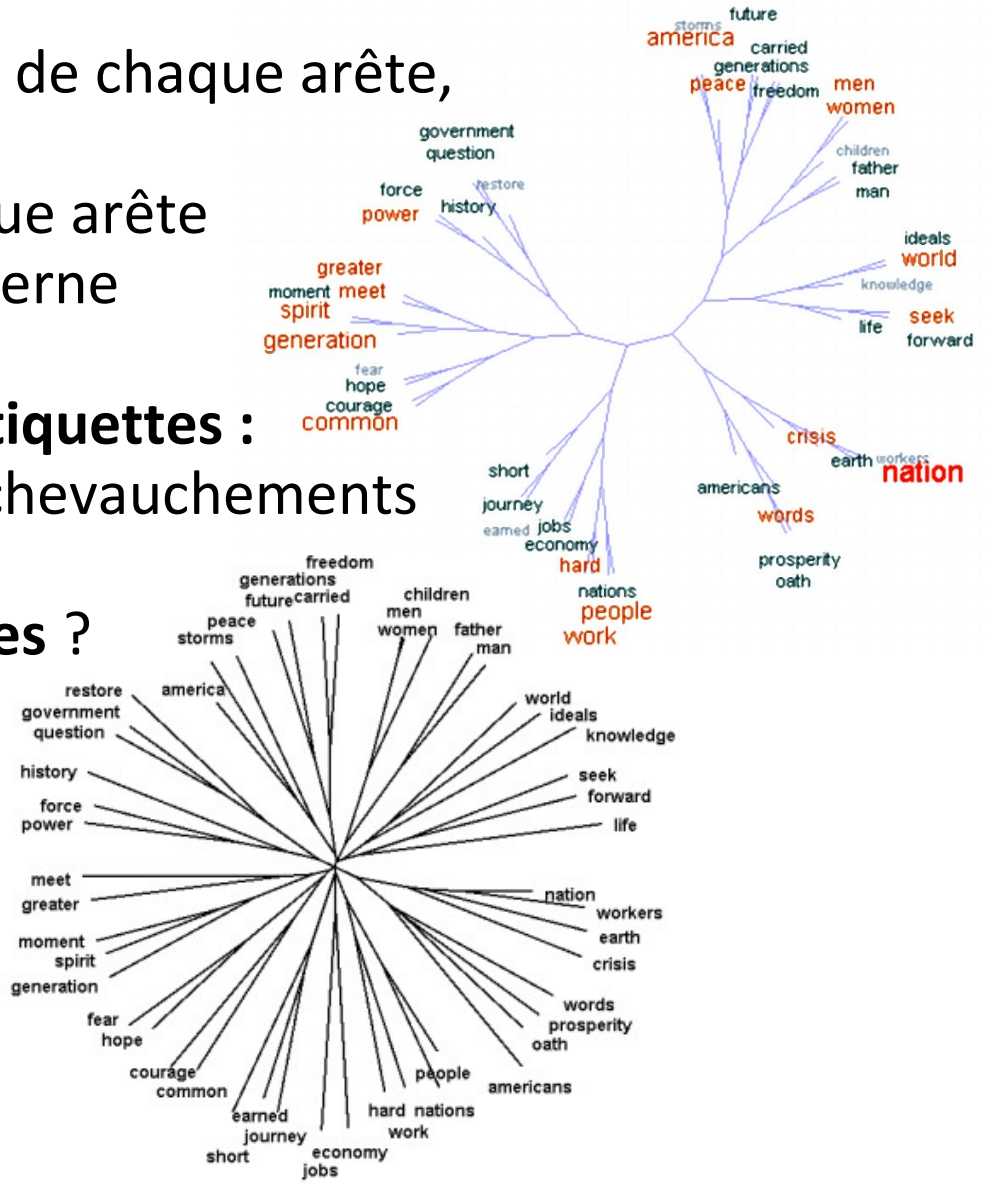
Algorithme “equal angle” :

- 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 :

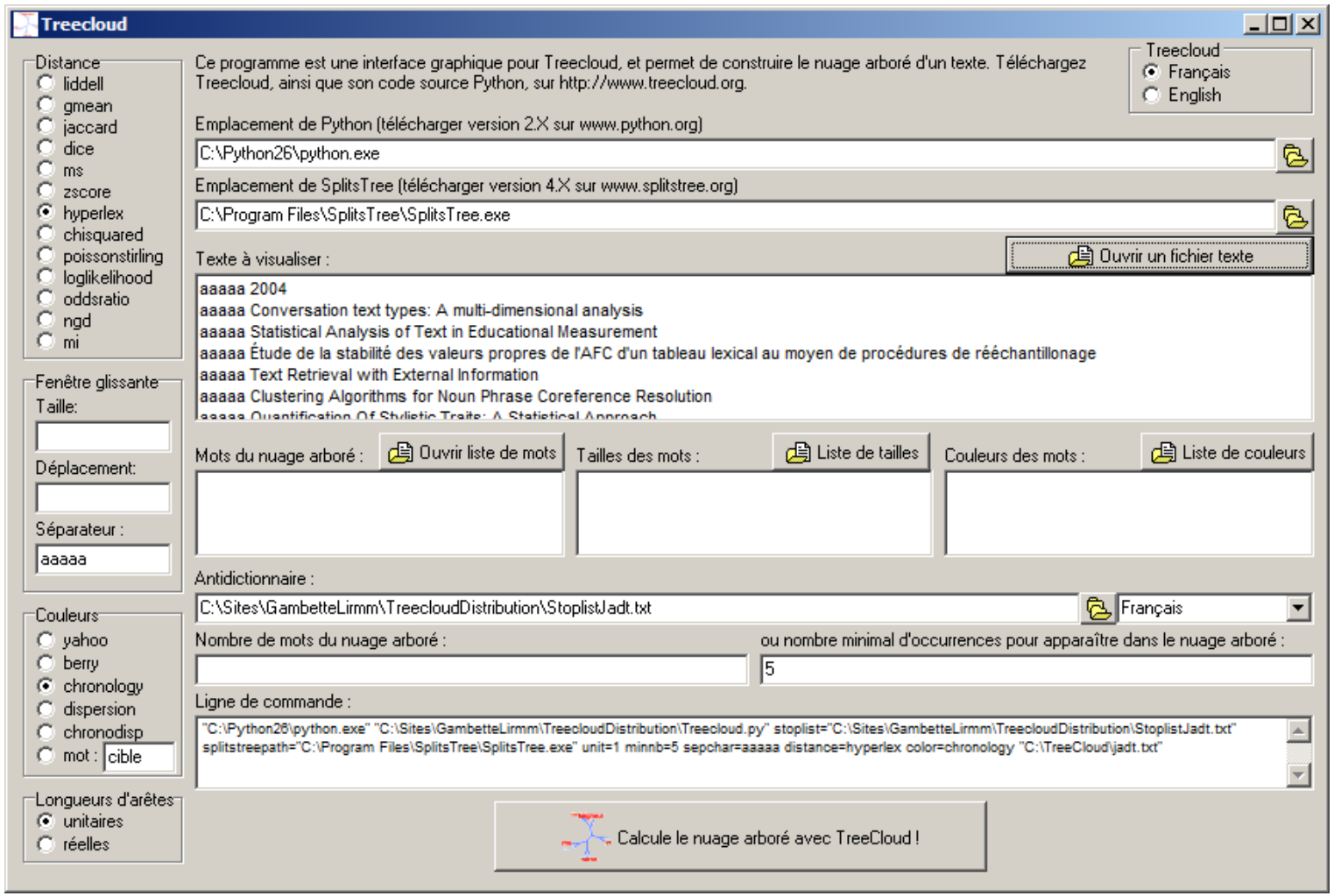
→ heuristique pour éviter les chevauchements

Question des longueurs d'arêtes ?



Implémentations

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

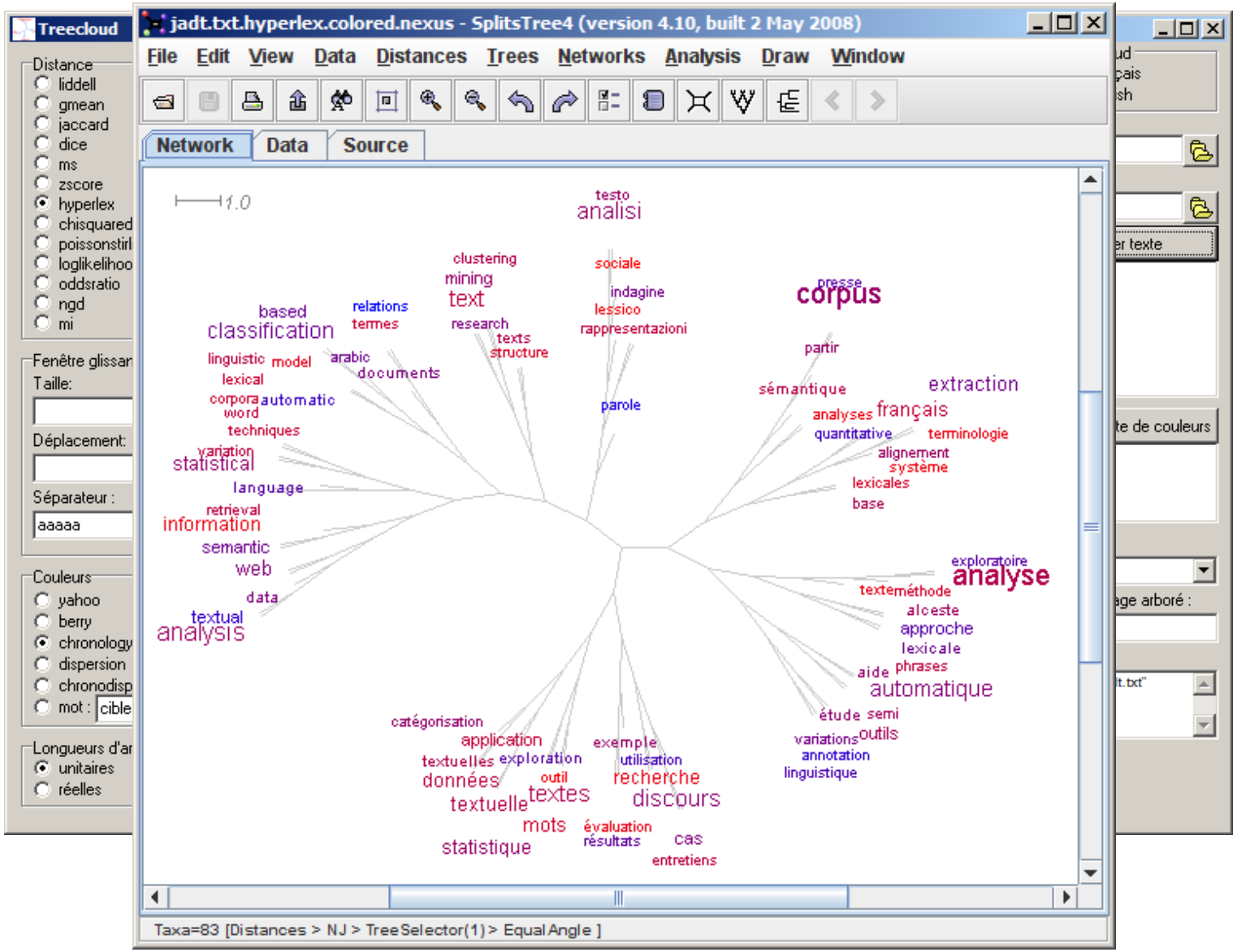


www.treecloud.org

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 !

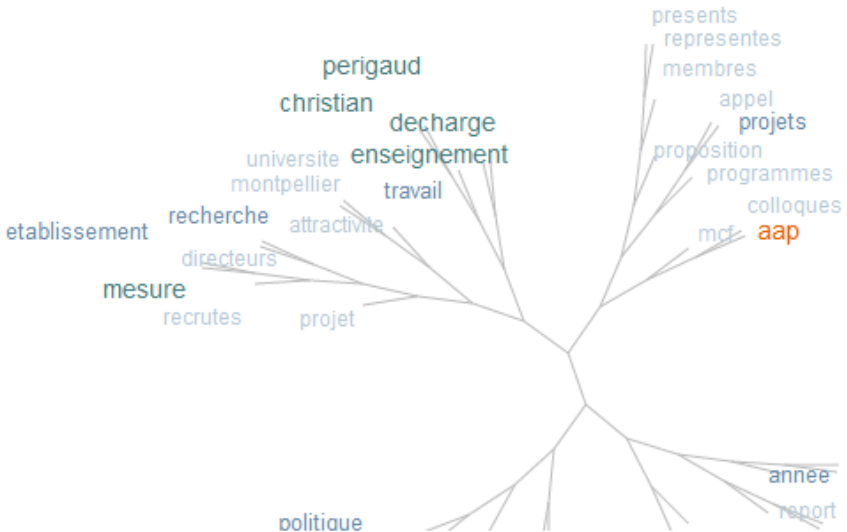
Collez votre texte dans le cadre ci-dessous, puis cliquez sur *Envoyer* ! Attention, l'utilisateur suivant verra votre texte quand il se connectera au site, si vous ne voulez pas faire apparaître vos textes, installez plutôt TreeCloud sur votre machine.

Texte :

conseil
scientifique
projet
proces
verbal

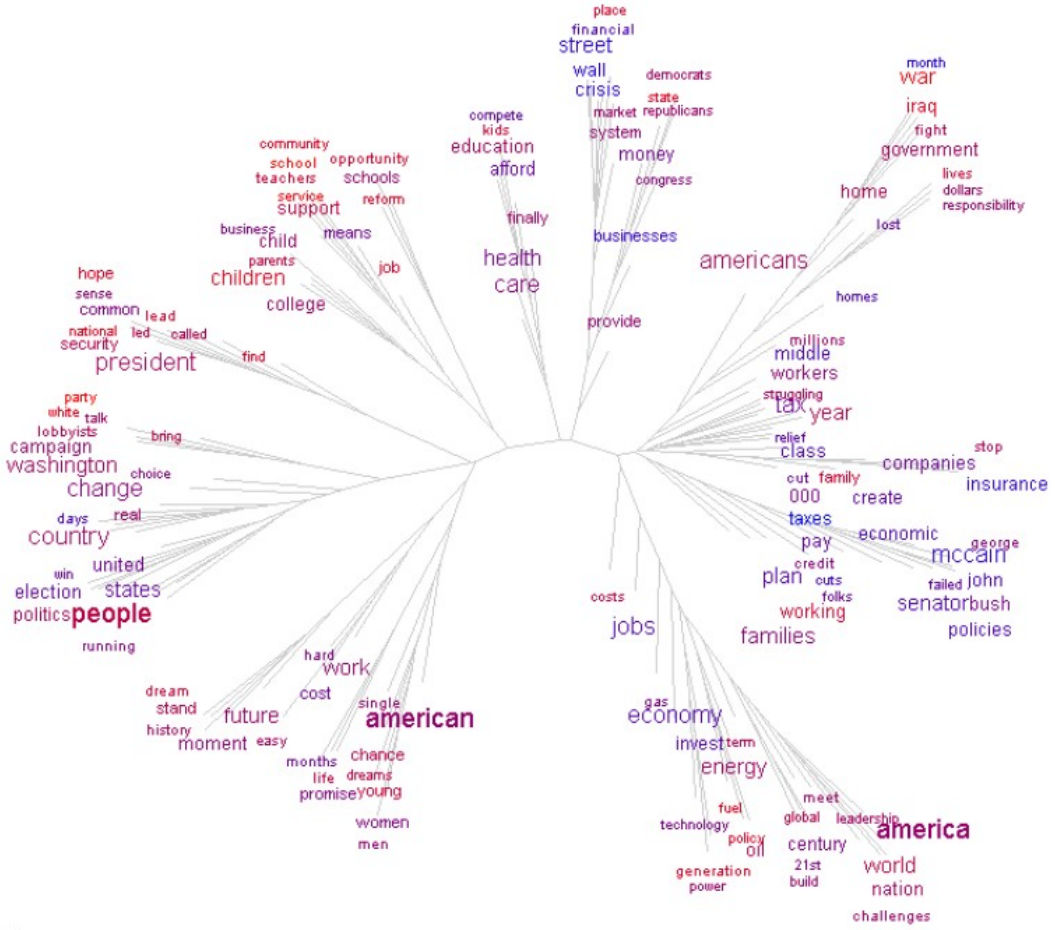
Envoyer

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)

Plan

- Nuages de mots et nuages arborés
- Caractéristiques du nuage arboré
- Utilisations du nuage arboré
- Construction d'un nuage arboré
- Évaluation de la robustesse de l'arbre
- Calcul des longueurs d'arêtes de l'arbre
- Perspectives

Contrôle qualité

Distances dans l'arbre = approximation des distances entre mots

Mesure objective de la qualité de l'arbre ?

Contrôle qualité

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

Contrôle qualité

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 ?

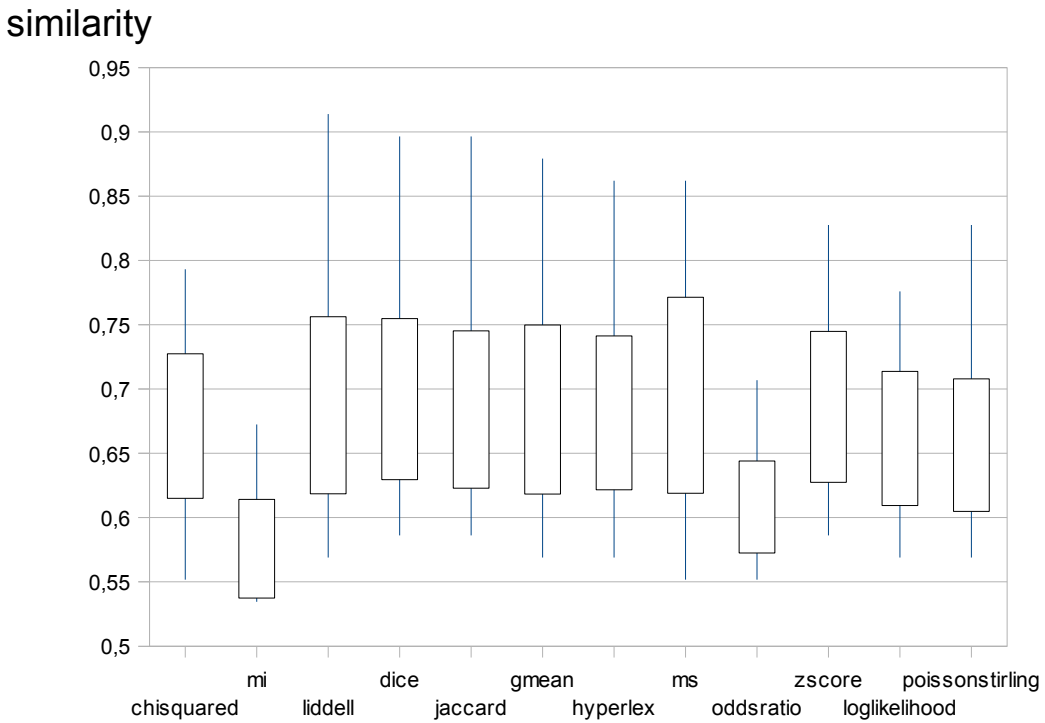
➡ **arboricité** pour mesurer à quel point la matrice de distance correspond à une distance d'arbre

Guénoche & Garreta, 2001

Guénoche & Darlu, 2009

Contrôle qualité - Bootstrap

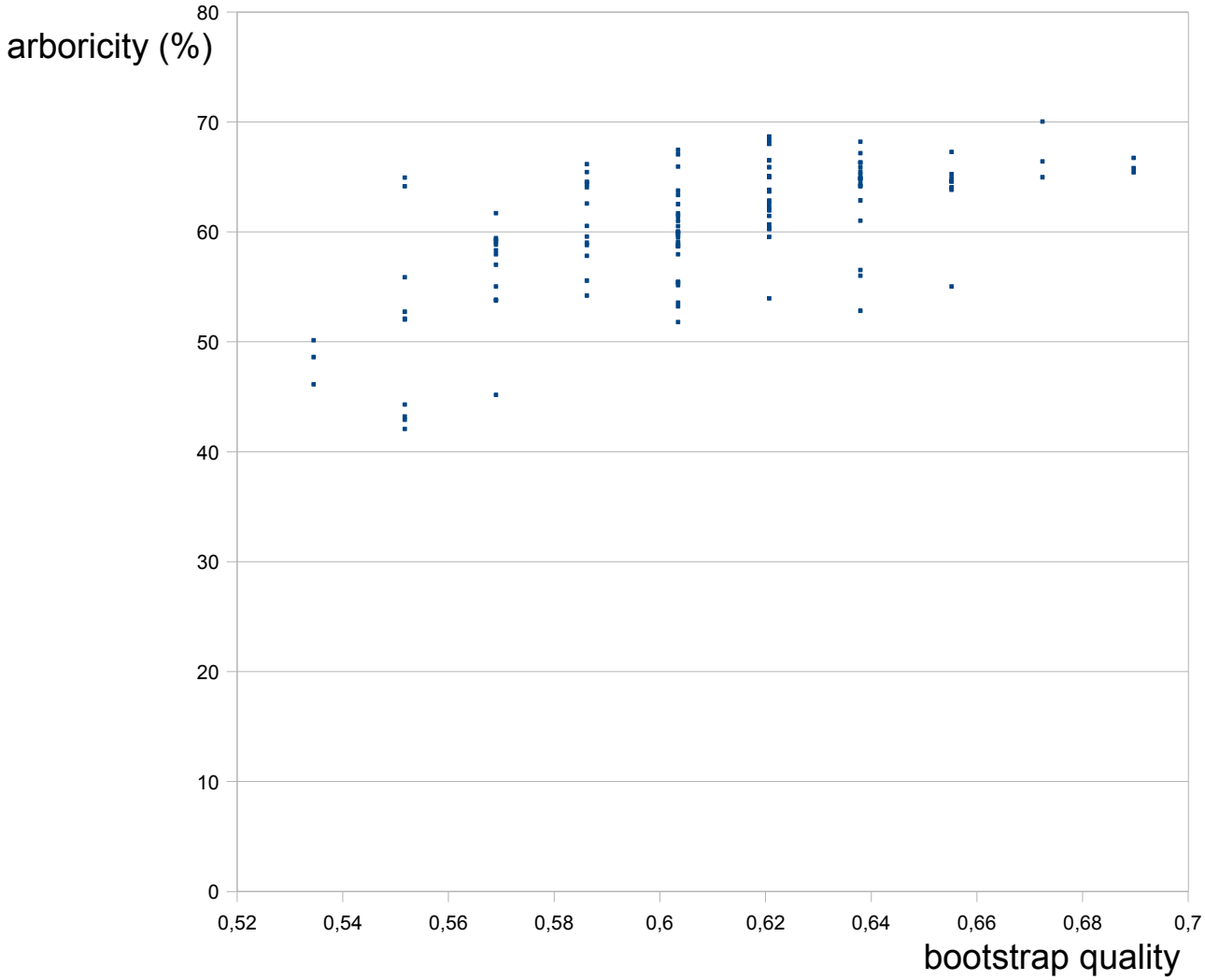
- Suppression des mots avec probabilité 50%.
- Construction du nuage arboré des textes original et altéré.
- Similarité des deux arbres (1-normalized RobinsonFoulds)



4 versions altérées de 10 discours d'Obama's, 3000 mots en moyenne, w=30, arbre NJ

Contrôle qualité - Arboricité

Relation entre “qualité bootstrap” et arboricité :

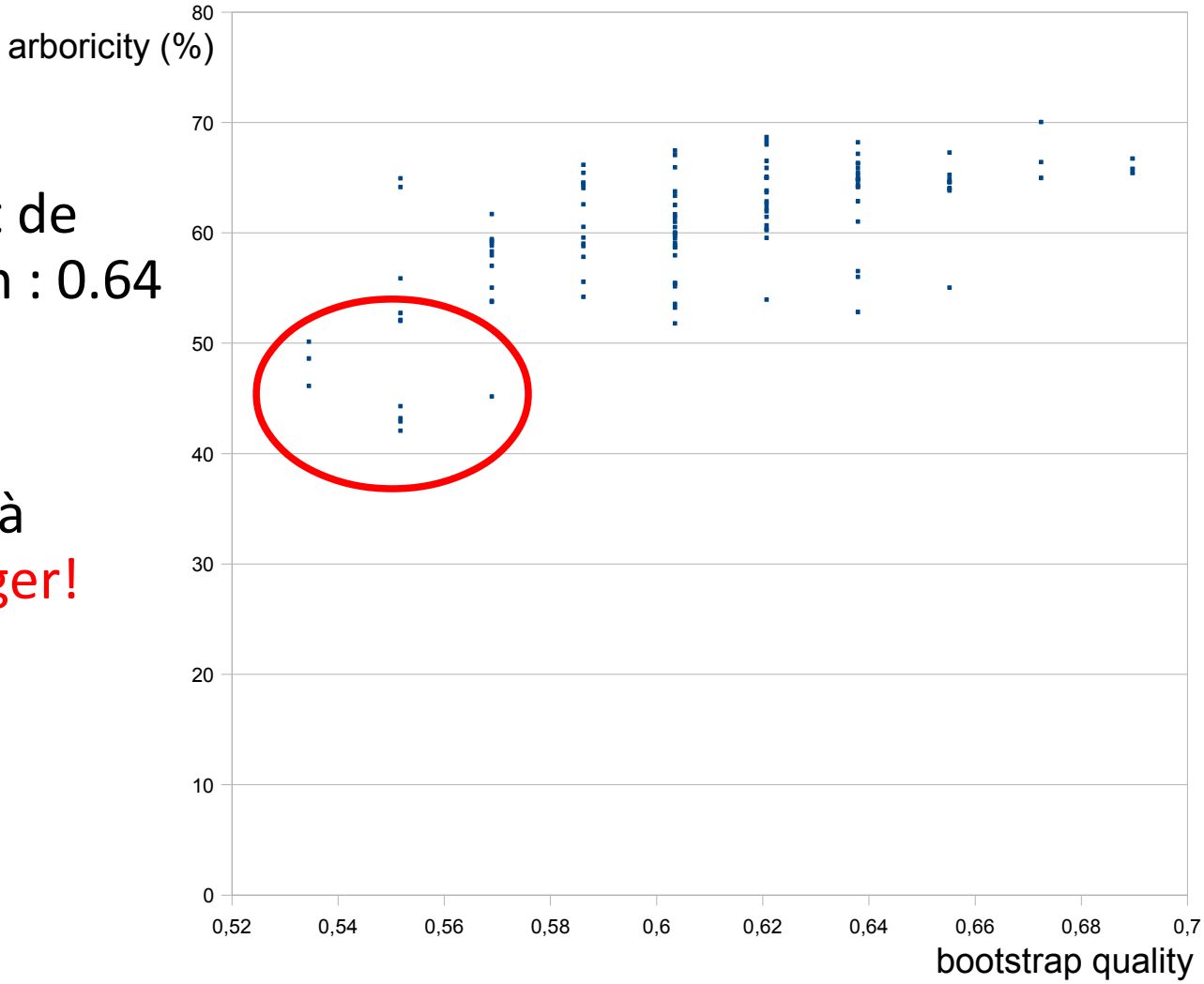


Contrôle qualité - Arboricité

Relation entre “qualité bootstrap” et arboricité :

coefficient de corrélation : 0.64

arboricité inférieure à 50% : danger!



Plan

- Nuages de mots et nuages arborés
- Caractéristiques du nuage arboré
- Utilisations du nuage arboré
- Construction d'un nuage arboré
- Évaluation de la robustesse de l'arbre
- Calcul des longueurs d'arêtes de l'arbre
- Perspectives

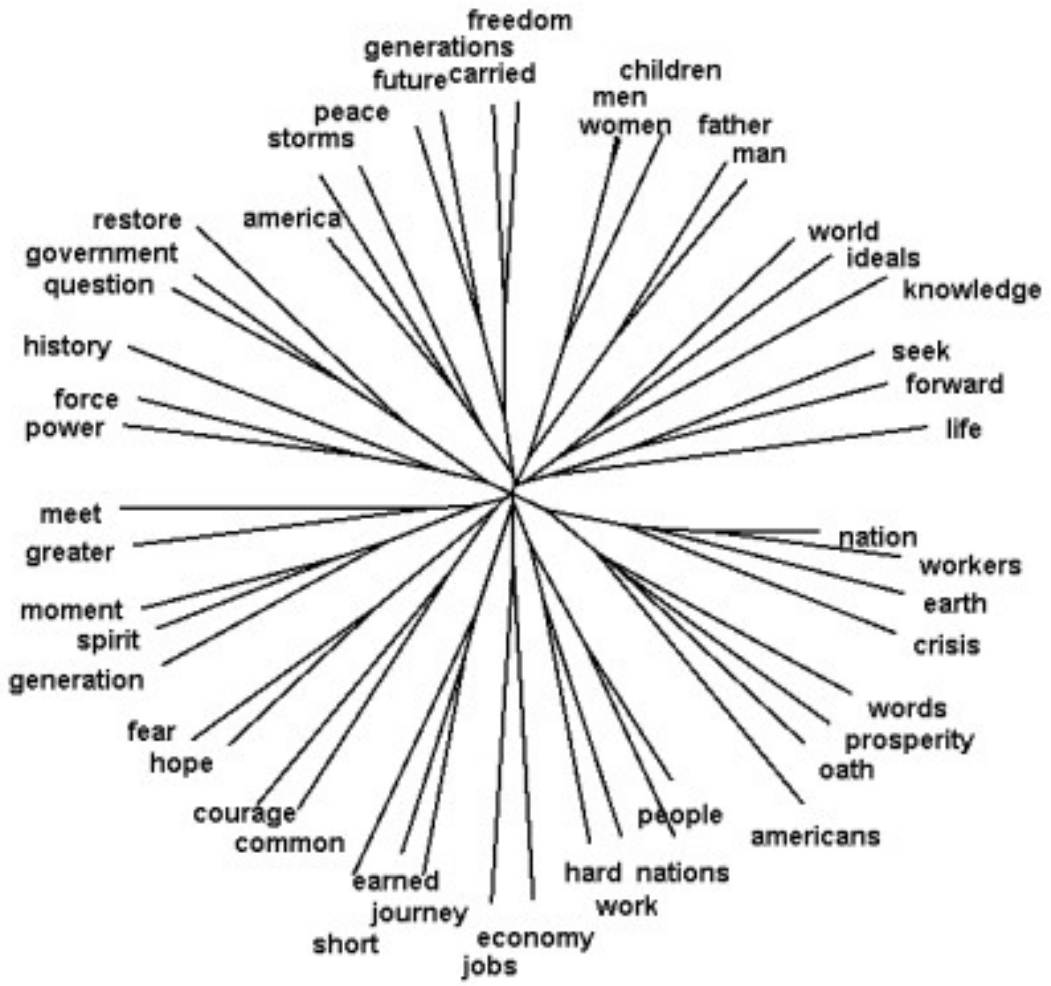
Calcul des longueurs d'arêtes

- Interprétation visuelle des longueurs d'arêtes
- Formules de longueurs d'arêtes
- Protocole d'évaluation
- Résultats
- Visualisations

Calcul des longueurs d'arêtes

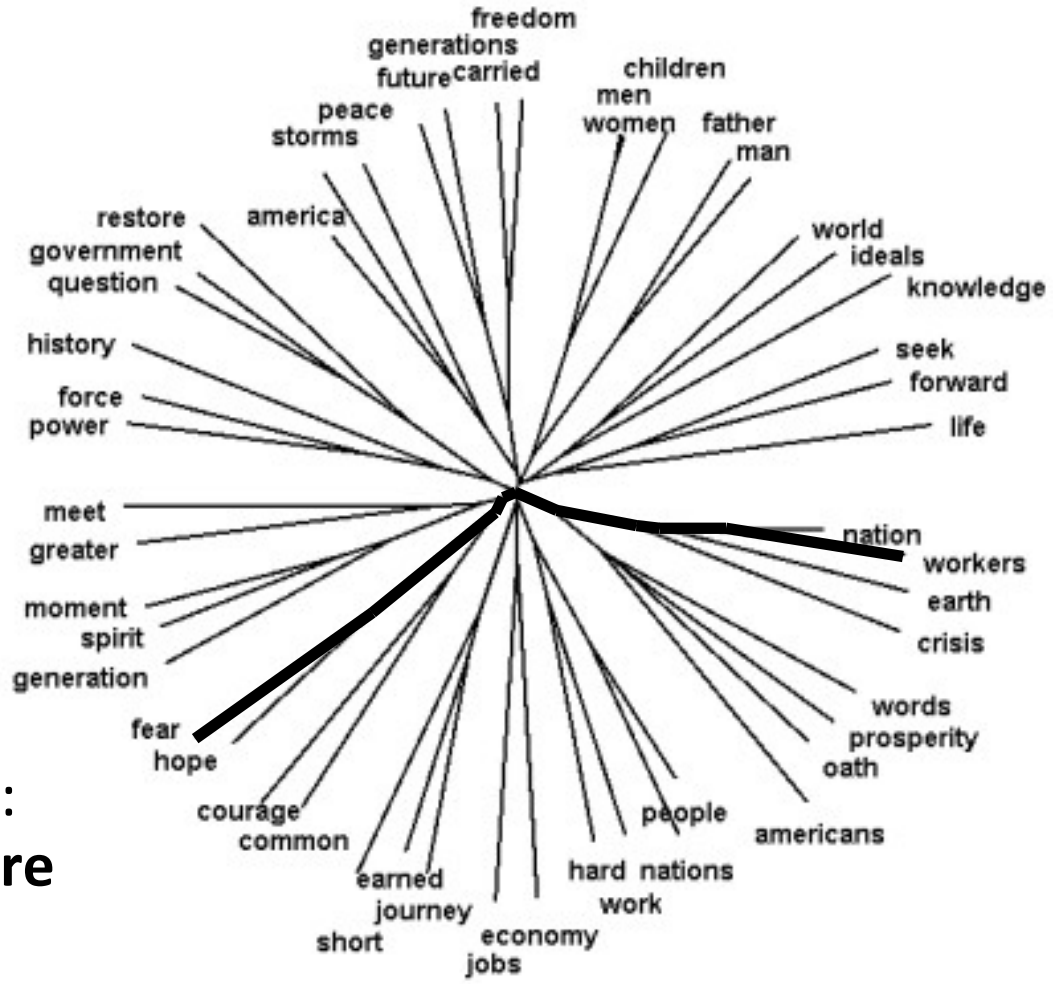
- Interprétation visuelle des longueurs d'arêtes
- Formules de longueurs d'arêtes
- Protocole d'évaluation
- Résultats
- Visualisations

Interprétation réelle



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

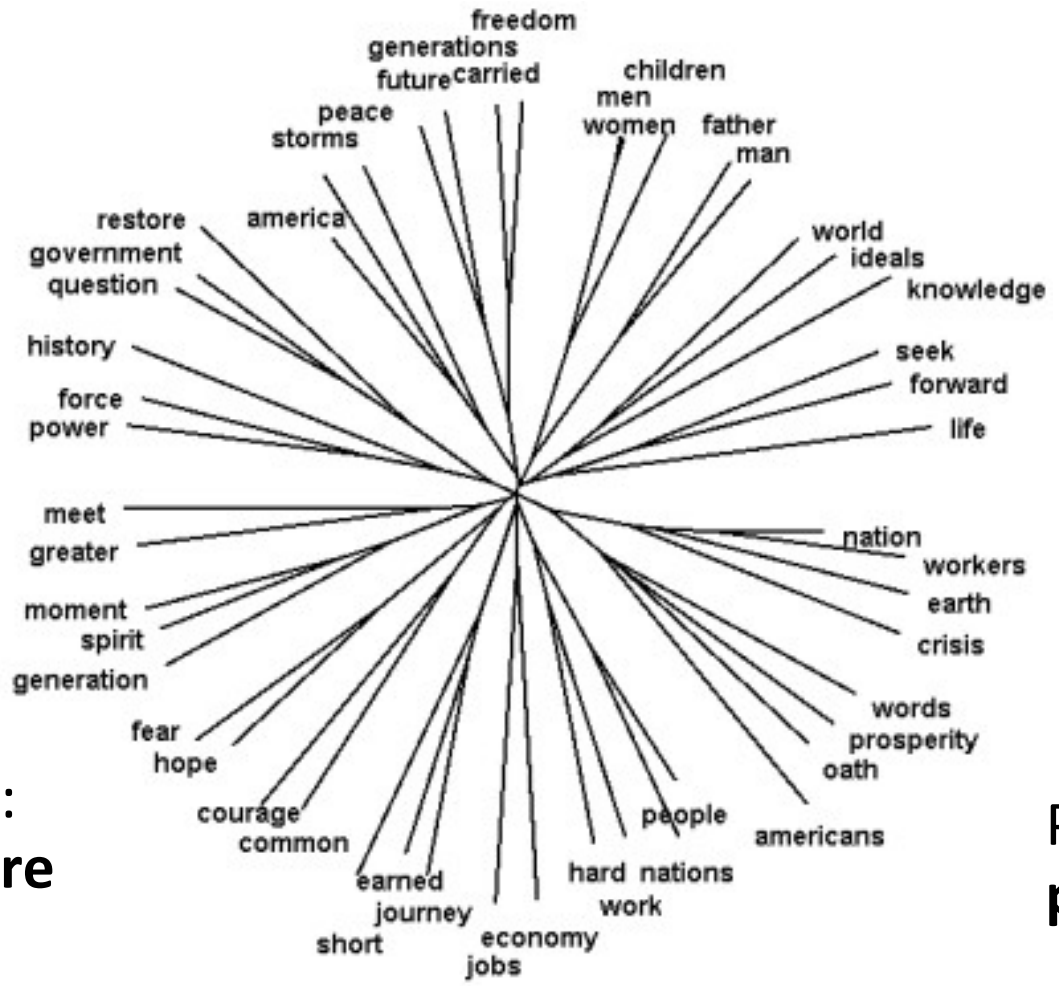
Interprétation réelle



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

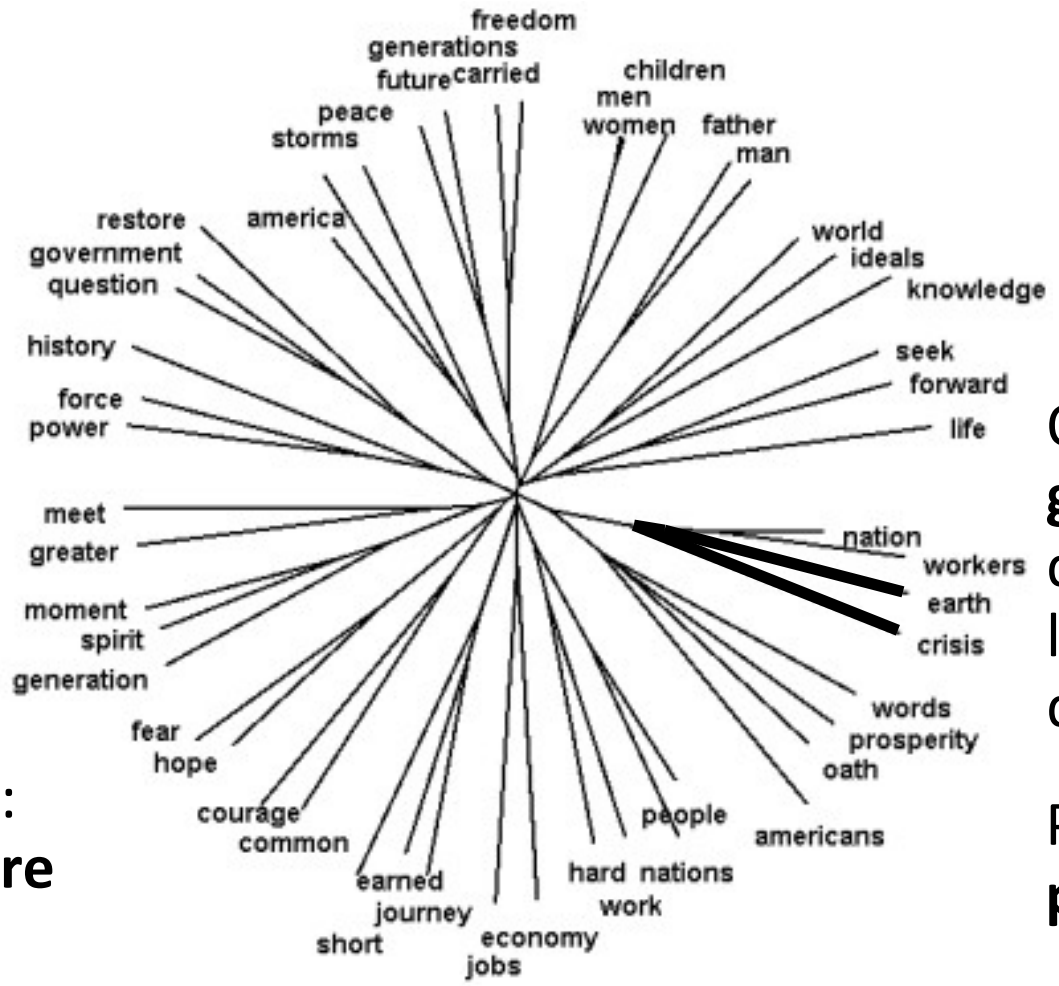


Problème 1 :
difficiles à lire

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



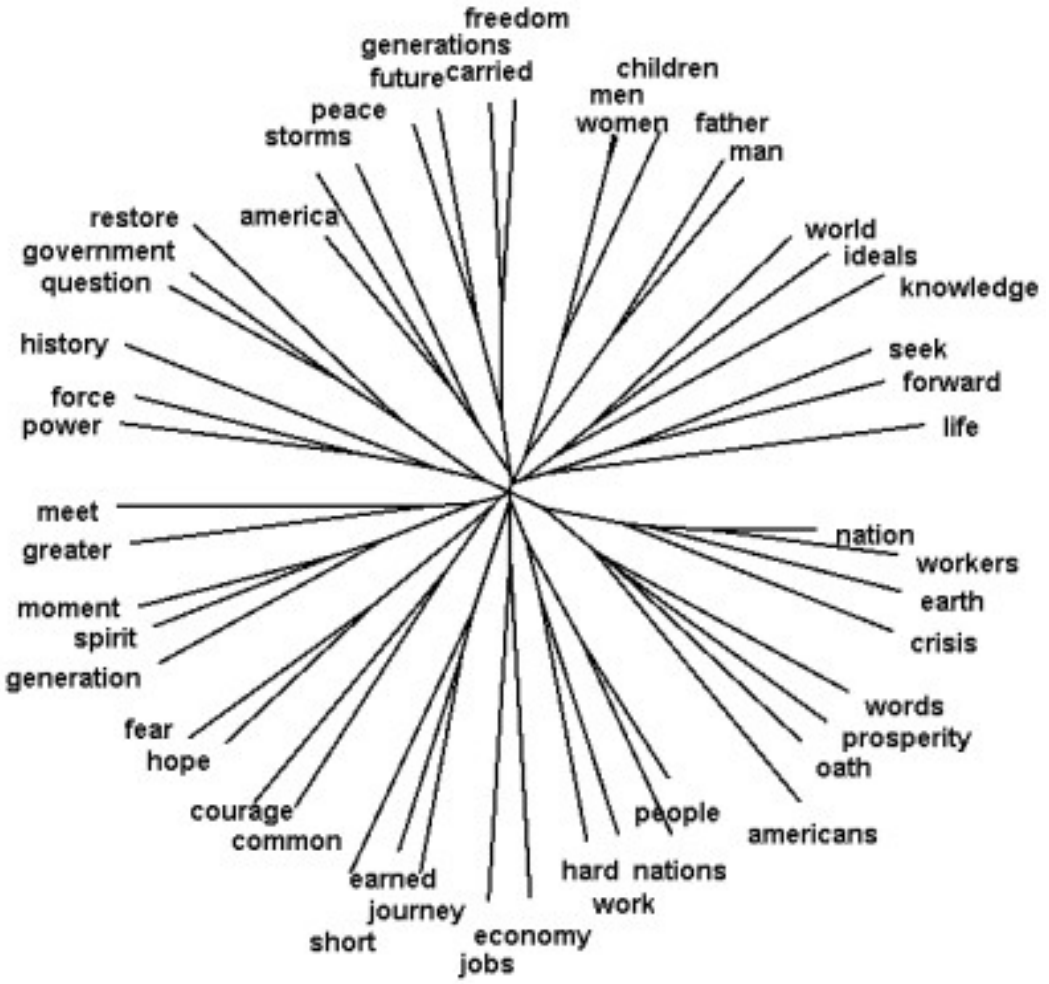
Optimisation **globale**, pas de garanties locales de qualité

Problème 1 : **difficiles à lire**

Problème 2 : **peu fiables**

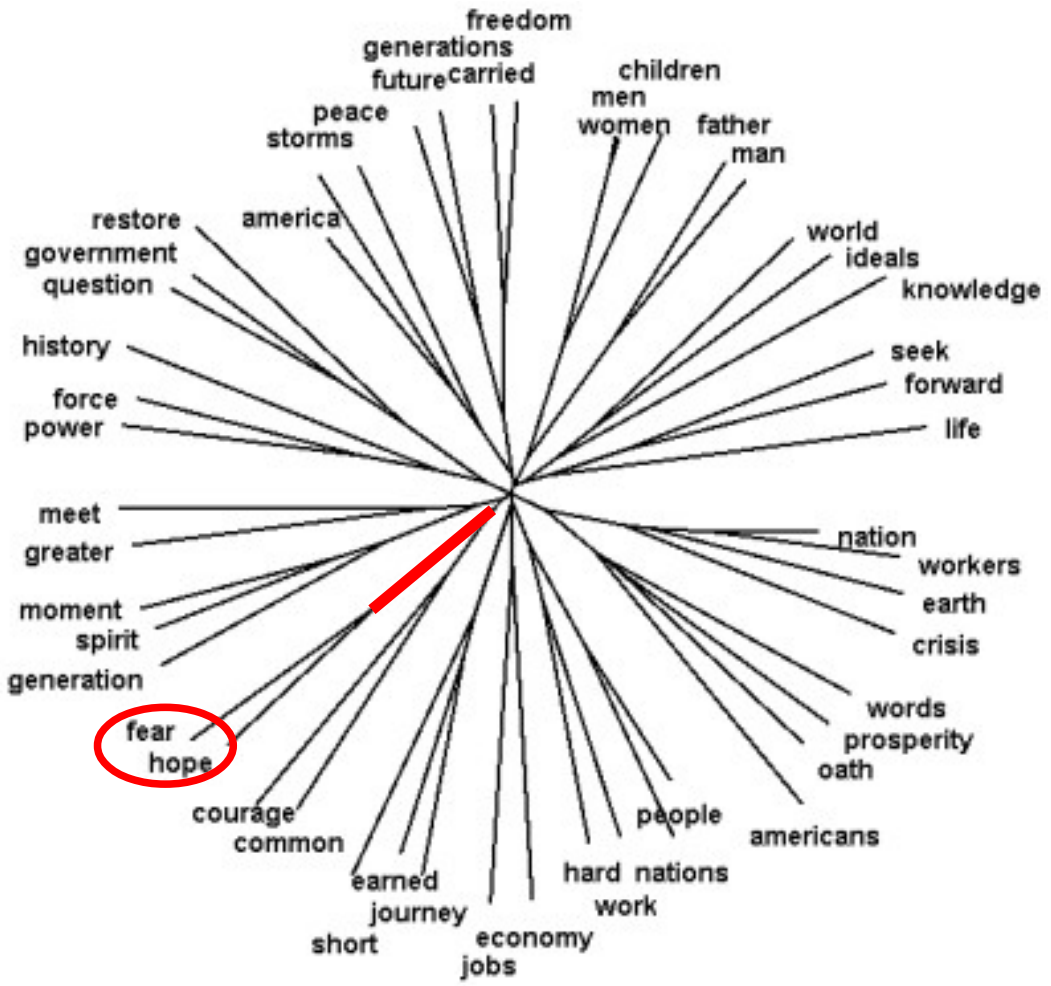
Les **distances dans l'arbre** entre deux mots reflètent *au mieux* le **degré de cooccurrence** entre ces deux mots

Interprétation pratique



arbre de distances utilisé comme **classification**

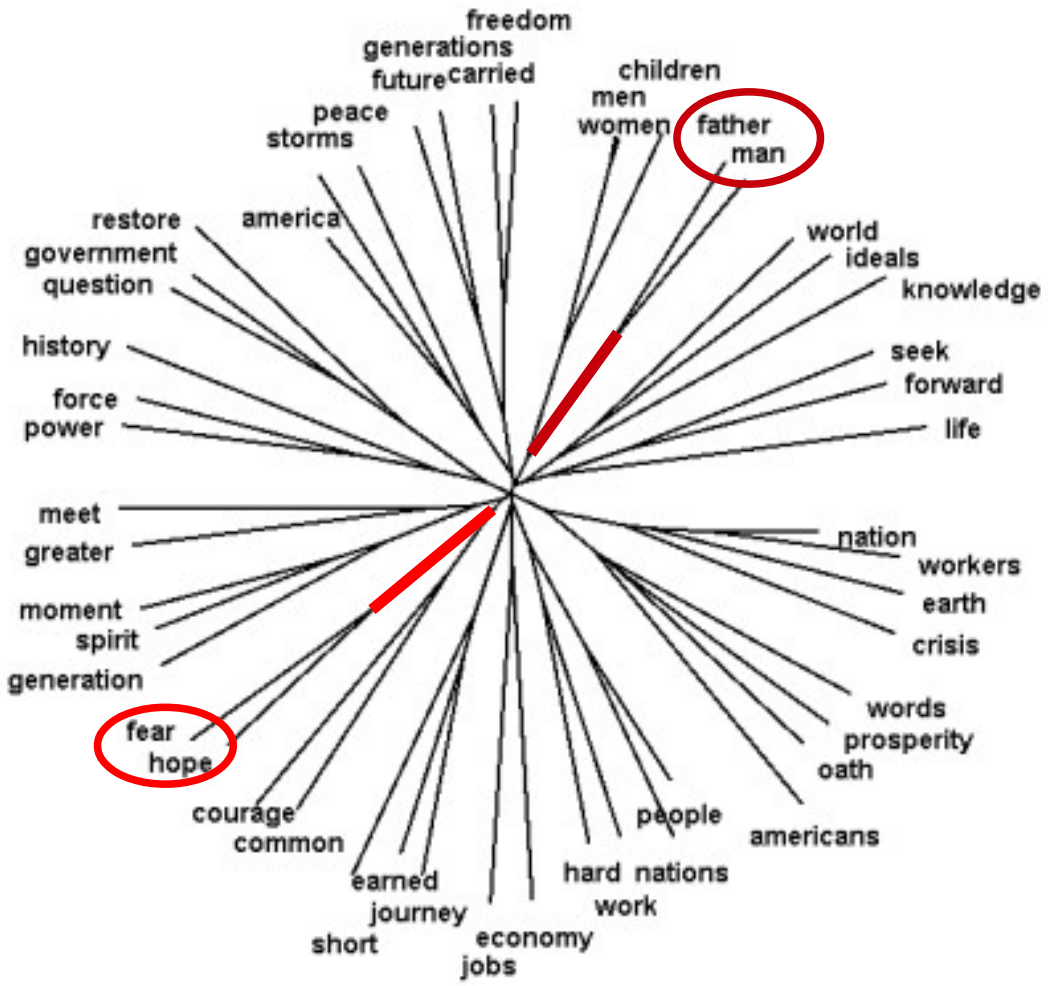
Interprétation pratique



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

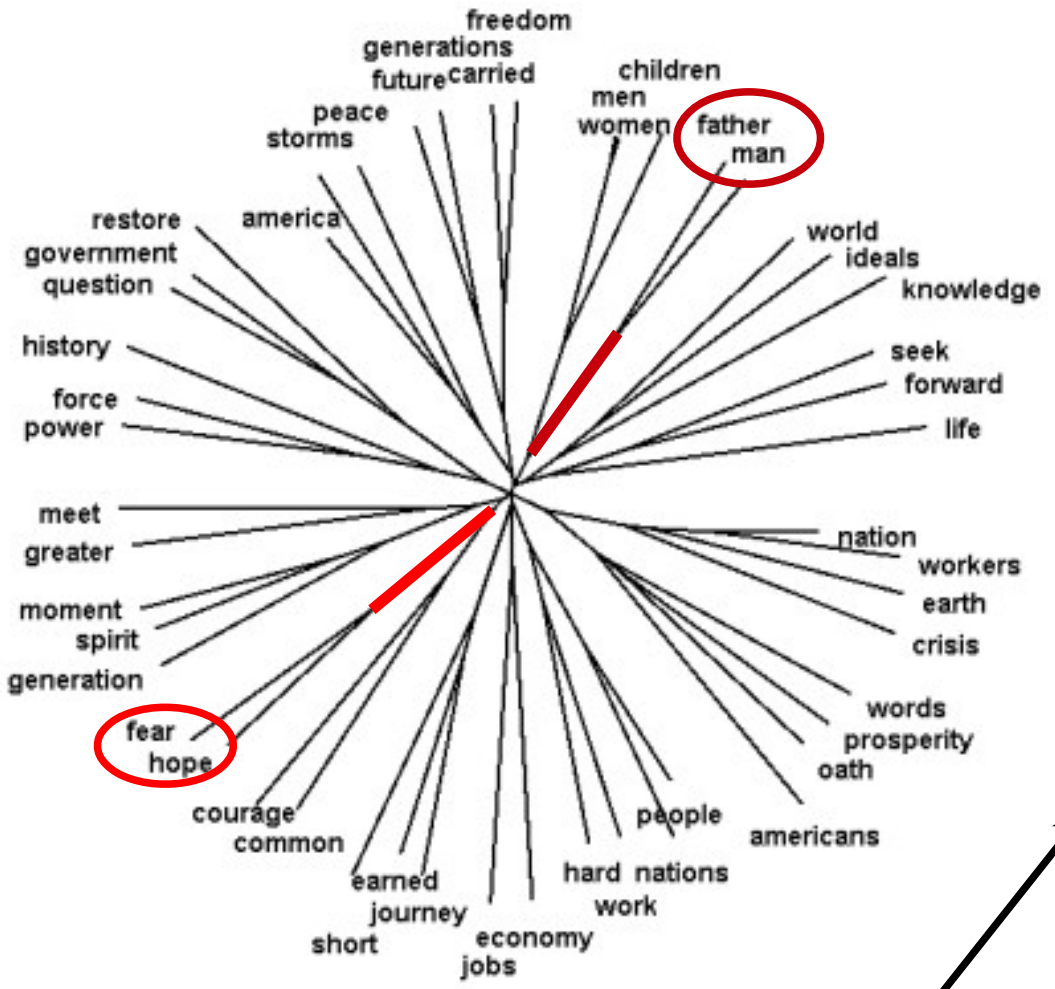
Interprétation pratique



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



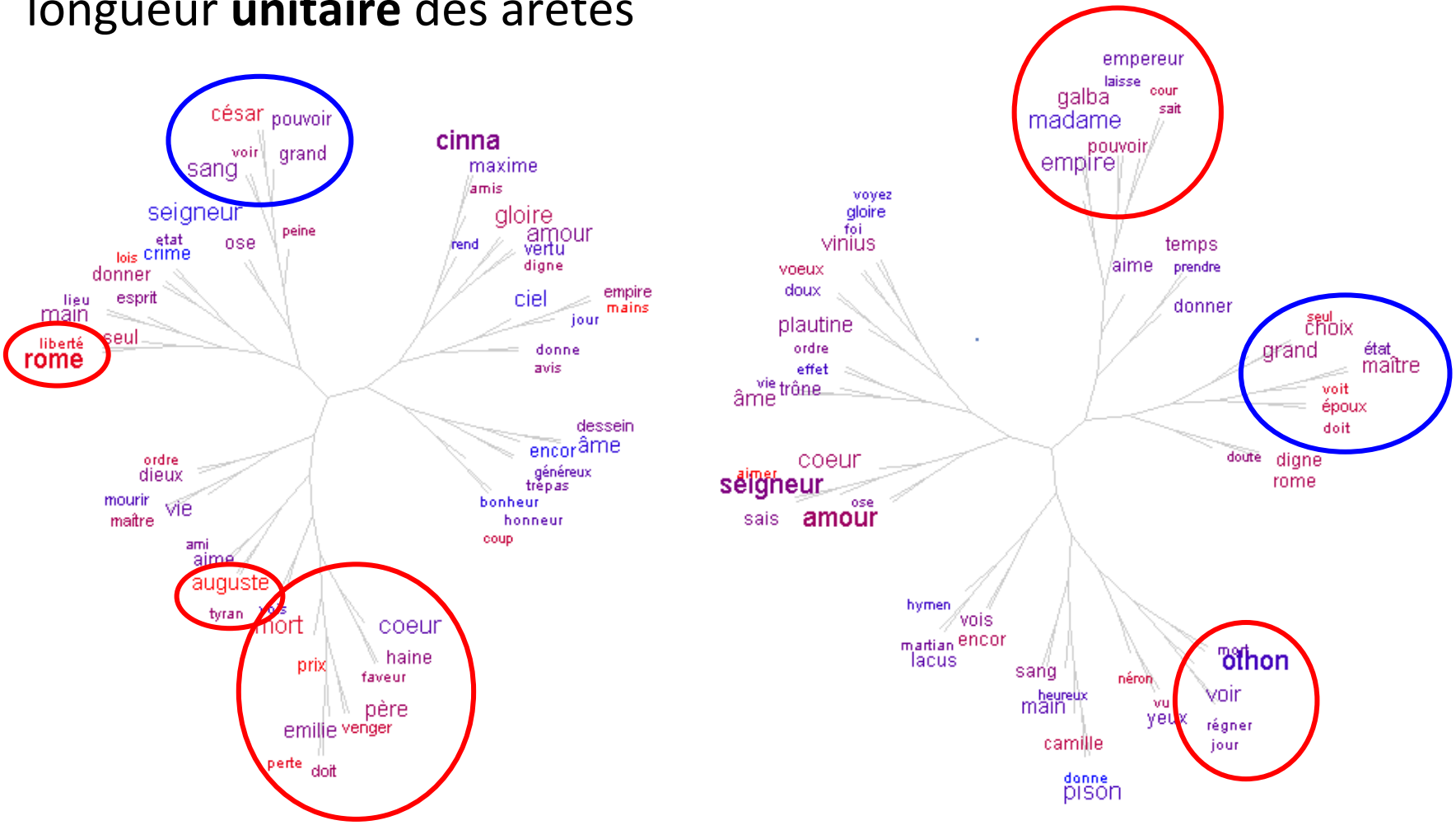
arbre de distances utilisé comme **classification**

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

Interprétation pratique

Astuce de visualisation pour améliorer la lisibilité :
longueur **unitaire** des arêtes



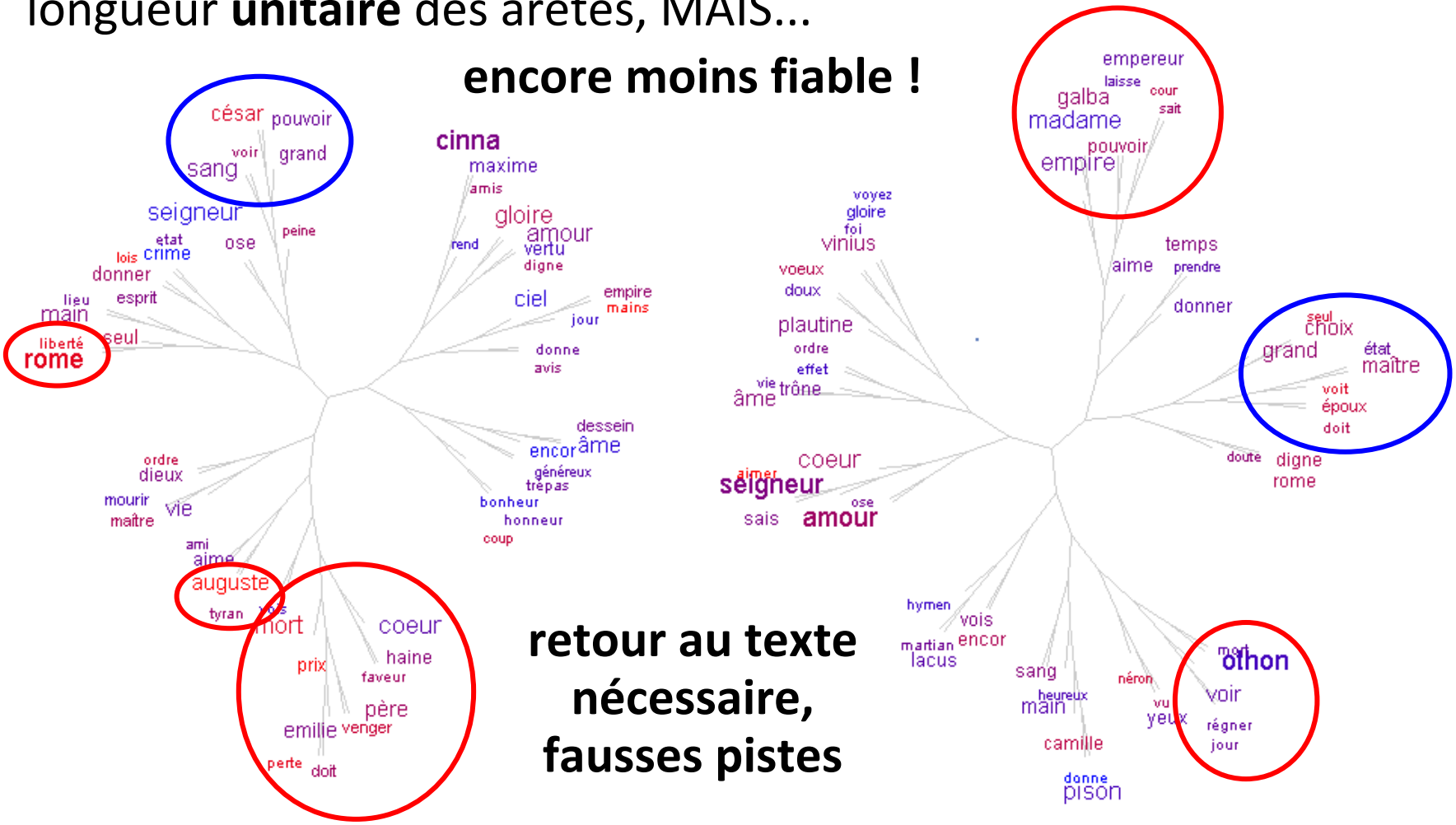
Nuages arborés globaux des 60 mots les plus fréquents dans Cinna et Othon (distance Liddell, fenêtre de largeur 20), colorés chronologiquement (rouge au début, bleu à la fin)

Amstutz & Gambette, *JADT'10*

Interprétation pratique

Astuce de visualisation pour améliorer la lisibilité :
longueur **unitaire** des arêtes, MAIS...

encore moins fiable !



retour au texte nécessaire, fausses pistes

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

Amstutz & Gambette, *JADT'10*

Interprétation pratique

Problème :

Comment calculer les longueurs des arêtes de l'arbre pour une **interprétation fiable des classes** ?

Arête longue = classe de mots significative (proches les uns des autres, bien séparés du reste)

Arête courte = classe de mots peu significative

Calcul des longueurs d'arêtes

- Interprétation visuelle
- Formules de longueurs d'arêtes
- Protocole d'évaluation
- Résultats
- Visualisations
- Perspectives

Formules de longueurs d'arêtes

Post-calcul des longueurs d'arêtes après la construction de l'arbre, pour que :

arêtes les plus longues ↔ classes de mots **les plus significatives**
↔ classes de mots **bien séparées**
d'après la **distance de cooccurrence**

Formules de longueurs d'arêtes

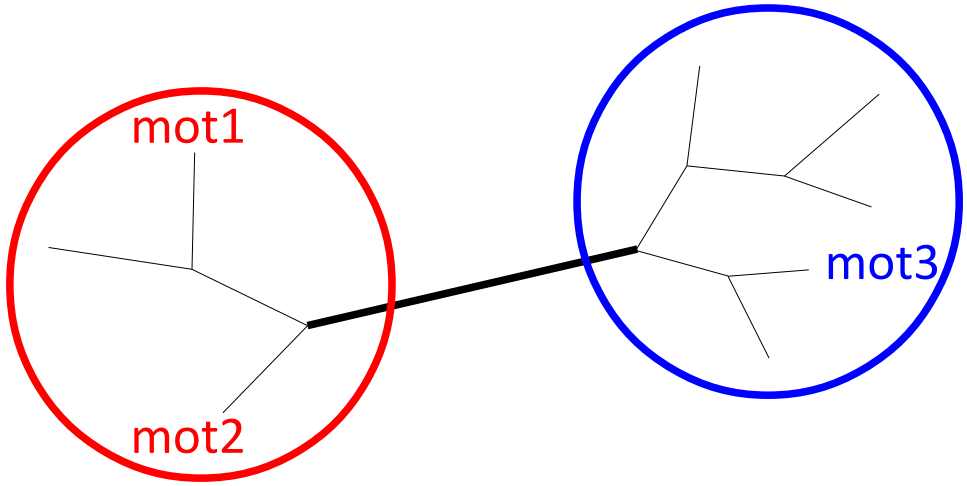
Post-calcul des longueurs d'arêtes après la construction de l'arbre, pour que :

arêtes les plus longues ↔ classes de mots **les plus significatives**
↔ classes de mots **bien séparées**
d'après la **distance de cooccurrence**

Formule du **ratio des bons triplets** (“triples”) :

Si mot1 et mot2 d'un côté de l'arête, mot3 de l'autre côté,

“**bon triplet**” si
distance(mot1,mot2) <
min**(**distance(mot1,mot3),
distance(mot2,mot3)**)**



ratio espéré proche de 1

Formules de longueurs d'arêtes

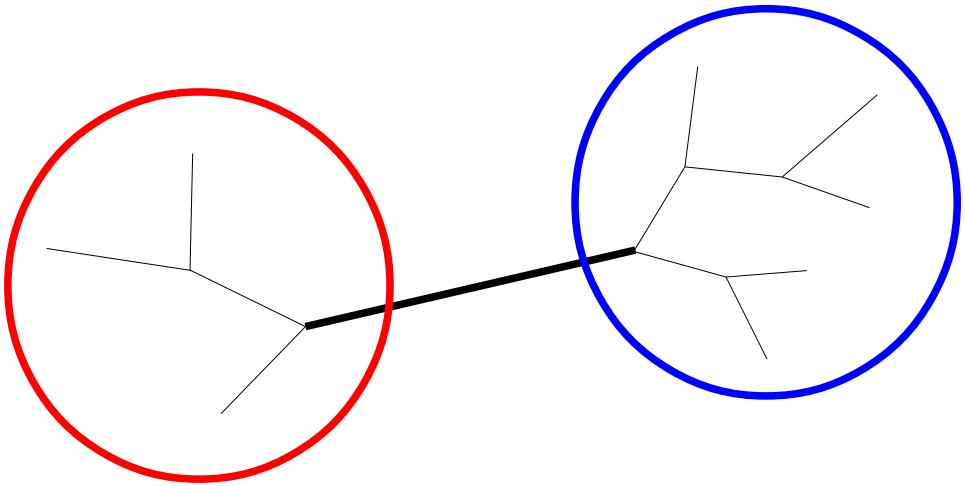
Post-calcul des longueurs d'arêtes après la construction de l'arbre, pour que :

arêtes les plus longues ↔ classes de mots **les plus significatives**
↔ classes de mots **bien séparées**
d'après la **distance de cooccurrence**

Formule du **ratio des distances moyennes** (“distanceRatio”) :

$$\frac{\text{moyenne(distances inter-classes)}}{\text{moyenne(distances intra-classes)}}$$

ratio espéré supérieur à 1



Guénoche & Garreta, *IFCS'02*

Formules de longueurs d'arêtes

Post-calcul des longueurs d'arêtes après la construction de l'arbre, pour que :

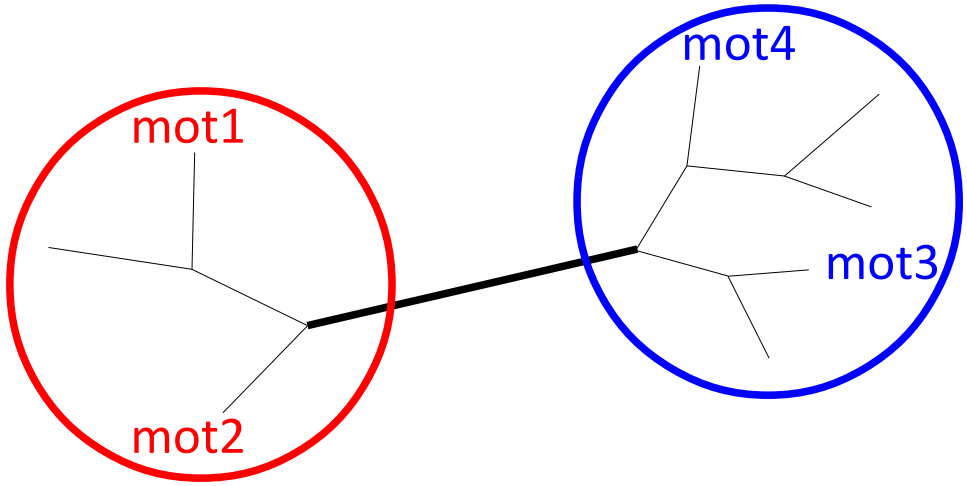
arêtes les plus longues ↔ classes de mots **les plus significatives**
↔ classes de mots **bien séparées**
d'après la **distance de cooccurrence**

Formule du **ratio des bons quadruplets** (“quartets”) :

Si mot1 et mot2 d'un côté de l'arête, mot3 et mot4 de l'autre côté,

“**bon quadruplet**” si
distance(mot1,mot2) +
distance(mot2,mot3) <
min**(**distance(mot1,mot3) +
distance(mot2,mot4) +,
distance(mot1,mot4) +
distance(mot2,mot3)**)**

ratio espéré proche de 1



Guénoche & Garreta, *IFCS'02*

Calcul des longueurs d'arêtes

- Interprétation visuelle
- Formules de longueurs d'arêtes
- Protocole d'évaluation
- Résultats
- Visualisations
- Perspectives

Protocole d'évaluation

Post-calcul des longueurs d'arêtes après la construction de l'arbre, puis :

Vérification que les classes de mots **les mieux séparées**
(d'après ces longueurs) sont **significatives**

Protocole d'évaluation

Post-calcul des longueurs d'arêtes après la construction de l'arbre, puis :

Vérification que les classes de mots **les mieux séparées** (d'après ces longueurs) sont **significatives**

Partition obtenue en **découpant les arêtes les plus longues** comparée avec une partition de référence

Protocole d'évaluation

Post-calcul des longueurs d'arêtes après la construction de l'arbre, puis :

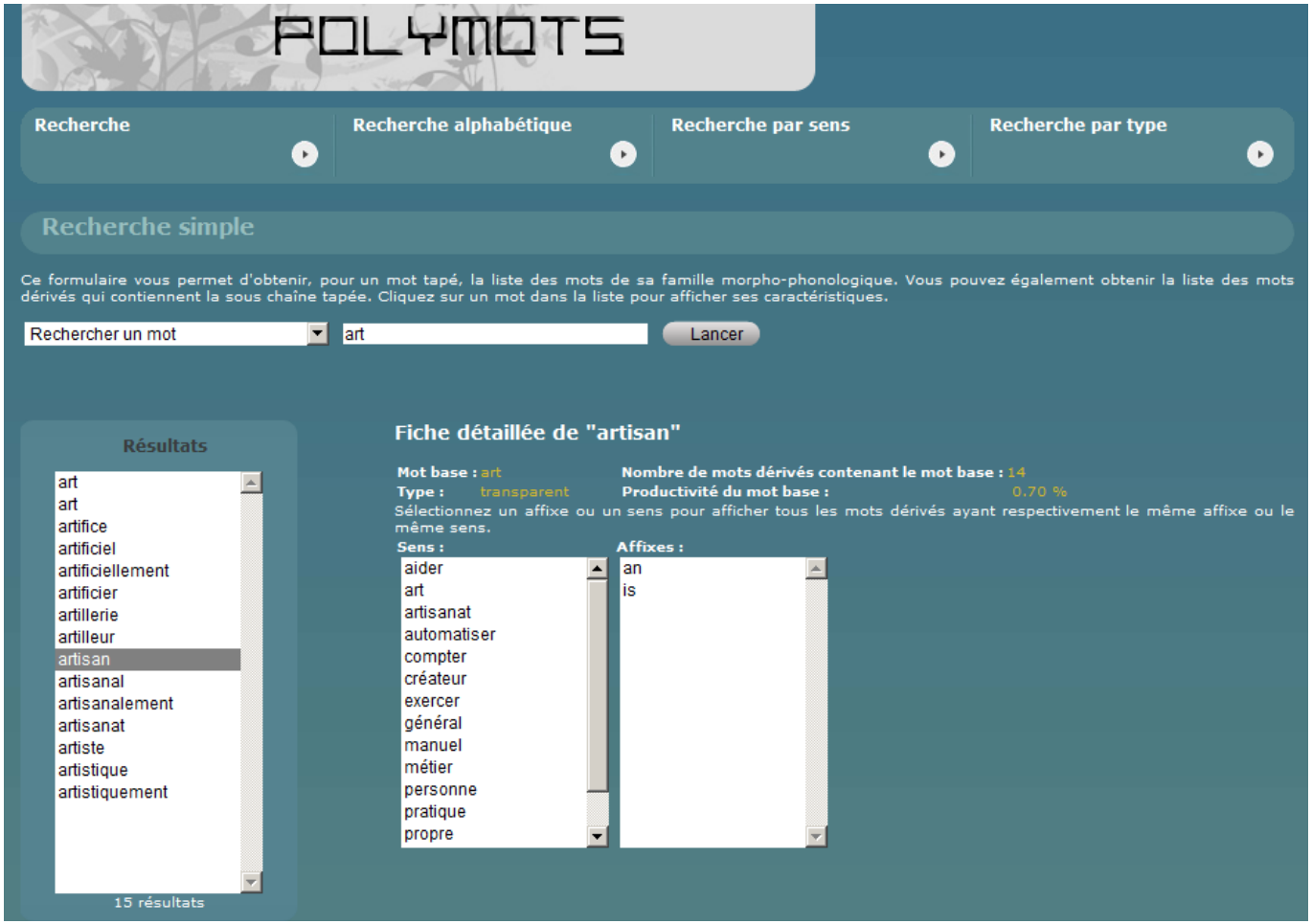
Vérification que les classes de mots **les mieux séparées** (d'après ces longueurs) sont **significatives**

Partition obtenue en **découpant les arêtes les plus longues** comparée avec une partition de référence

quelles données ?

Protocole d'évaluation

Base de données Polymots
Base lexicale de familles morpho-phonologiques
20 000 mots, 2000 familles



Protocole d'évaluation

Base de données Polymots
Base lexicale de familles morpho-phonologiques
20 000 mots, 2000 familles

+ **partitions sémantiques** des familles de 20 mots
(arbre, art, boule, carte, corde, dent, dict, fil, fusée, lune, meuble, mode, onde, paille, penser, pot, presse, tenir, terre, val).

Protocole d'évaluation

Base de données Polymots
Base lexicale de familles morpho-phonologiques
20 000 mots, 2000 familles

+ **partitions sémantiques** des familles de 20 mots
(arbre, art, boule, carte, corde, dent, dict, fil, fusée, lune,
meuble, mode, onde, paille, penser, pot, presse, tenir, terre, val).

Exemple pour la famille de **art** :

{ {artifice, artificiel, artificiellement, artificier},
{artillerie, artilleur},
{artisan, artisanal, artisanalement, artisanat},
{artiste, artistique, artistiquement, art} **}**

Protocole d'évaluation

Idée :

- Construire une représentation arborée des mots de la famille
- Découper les *k* arêtes les plus longues de l'arbre pour obtenir *Pk*
- La partition obtenue est-elle proche de la partition “manuelle” ?

Protocole d'évaluation

Idée :

- Construire une représentation arborée des mots de la famille
- Découper les *k* arêtes les plus longues de l'arbre pour obtenir *Pk*
- La partition obtenue est-elle proche de la partition “manuelle” ?

Distance utilisée pour le calcul de la représentation arborée ?

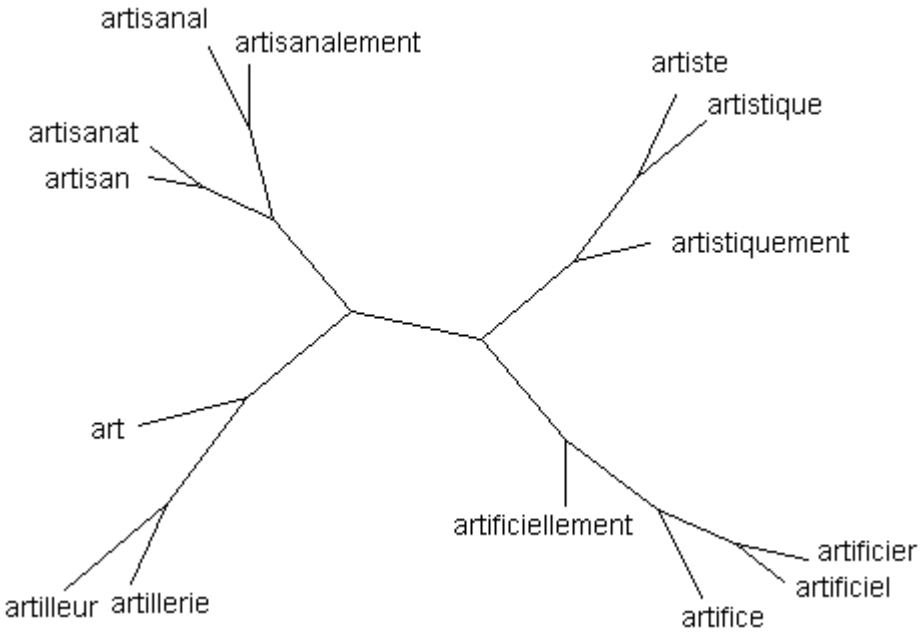
Distance composite entre :

- nombre d'affixes communs
- degré de cooccurrence dans Le Trésor de la Langue Française informatisé

Protocole d'évaluation

Idée :

- Construire une représentation arborée des mots de la famille
- Découper les *k* arêtes les plus longues de l'arbre pour obtenir *Pk*
- La partition obtenue est-elle proche de la partition “manuelle” ?



Partition automatique :
P0 = {{artisan, artisanat, artisanal, artisanalement, artiste, artistique, artistiquement, artificier, artificiel, artifice, artificiellement, artillerie, artilleur, art}**}**

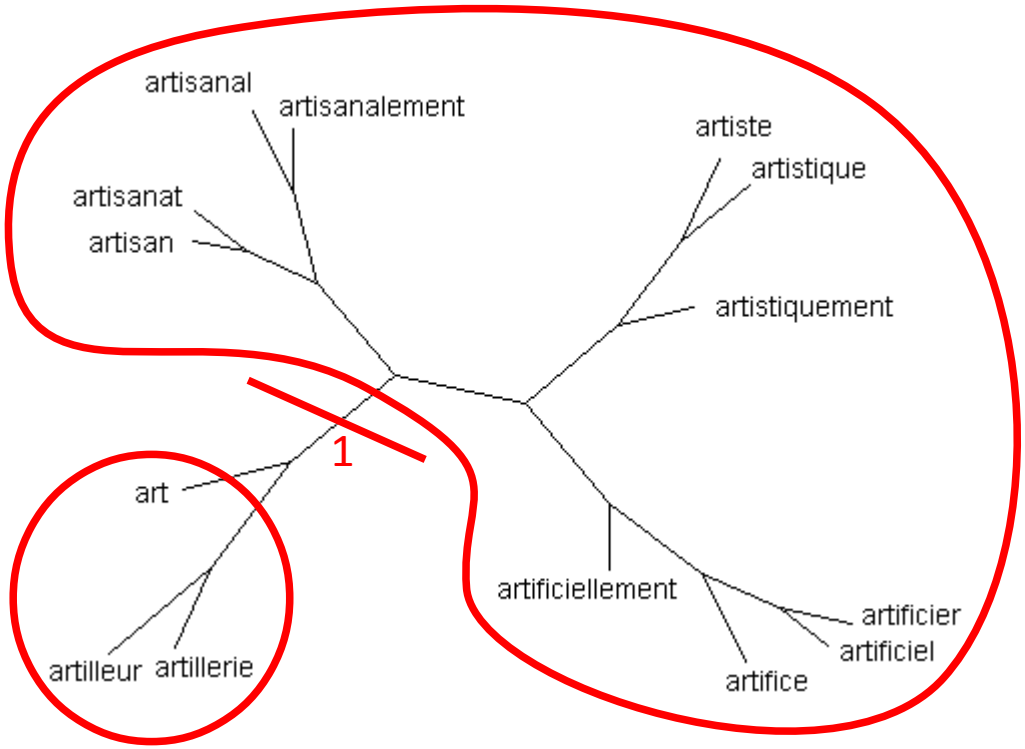
Partition manuelle : **Pm = {**{artificier, artifice, artificiel, artificiellement}, {artillerie, artilleur}, {artisan, artisanal, artisanalement, artisanat}, {artiste, artistique, artistiquement, art}**}**

Gala, Hathout, Nasr, Rey, Seppälä, *TALN'11*

Protocole d'évaluation

Idée :

- Construire une représentation arborée des mots de la famille
- Découper les *k* arêtes les plus longues de l'arbre pour obtenir *Pk*
- La partition obtenue est-elle proche de la partition “manuelle” ?



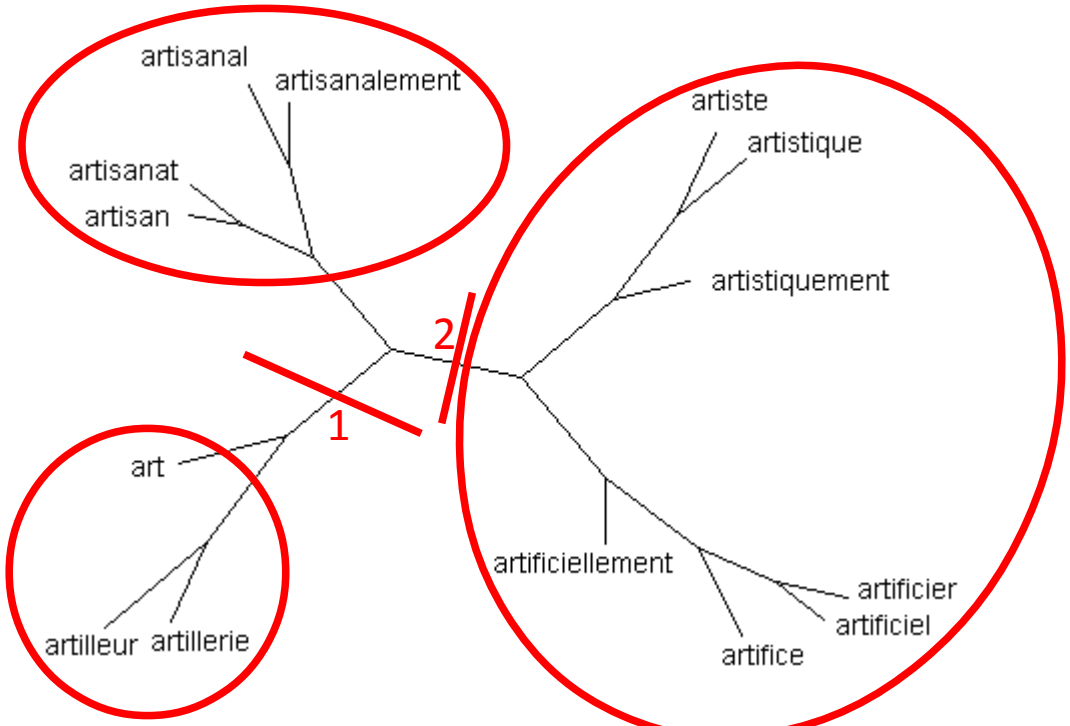
Partition automatique :
P1 = {{artisan, artisanat, artisanal, artisanalement, artiste, artistique, artistiquement, artificier, artificiel, artifice, artificiellement}, {artillerie, artilleur, art}**}**

Partition manuelle : **Pm = {**{artificier, artifice, artificiel, artificiellement}, {artillerie, artilleur}, {artisan, artisanal, artisanalement, artisanat}, {artiste, artistique, artistiquement, art}**}**

Protocole d'évaluation

Idée :

- Construire une représentation arborée des mots de la famille
- Découper les *k* arêtes les plus longues de l'arbre pour obtenir *Pk*
- La partition obtenue est-elle proche de la partition “manuelle” ?



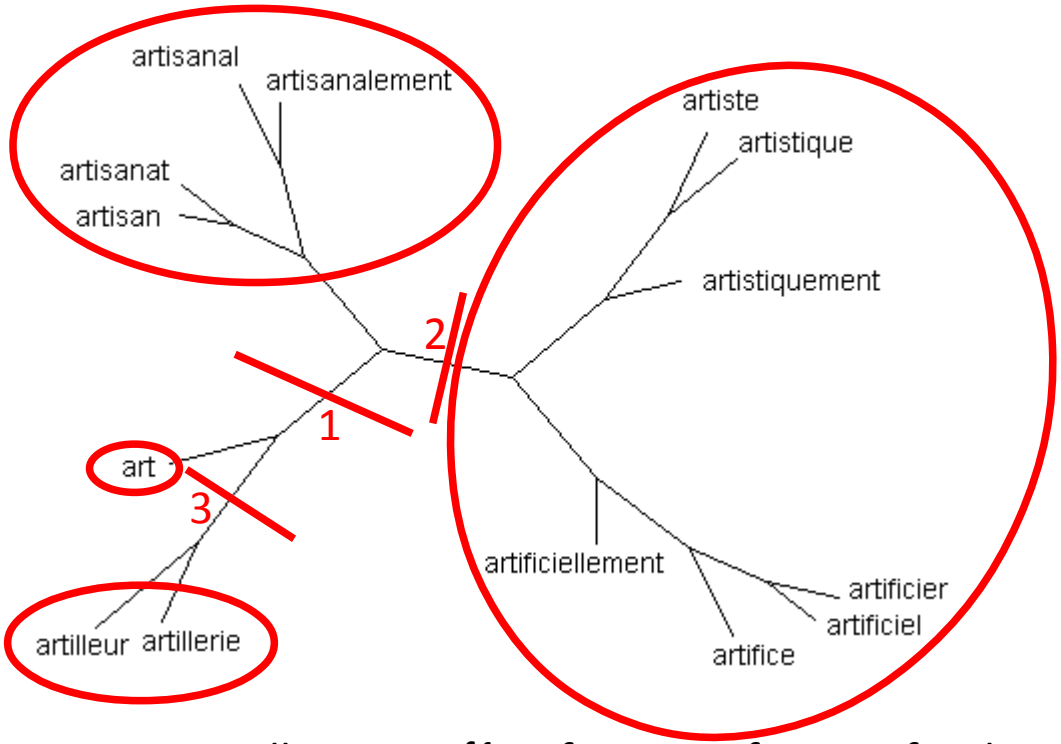
Partition automatique :
P2 = {{artisan, artisanat, artisanal, artisanalement}, {artiste, artistique, artistiquement, artificier, artificiel, artifice, artificiellement}, {artillerie, artilleur, art}**}**

Partition manuelle : **Pm = {**{artificier, artifice, artificiel, artificiellement}, {artillerie, artilleur}, {artisan, artisanal, artisanalement, artisanat}, {artiste, artistique, artistiquement, art}**}**

Protocole d'évaluation

Idée :

- Construire une représentation arborée des mots de la famille
- Découper les *k* arêtes les plus longues de l'arbre pour obtenir *Pk*
- La partition obtenue est-elle proche de la partition “manuelle” ?



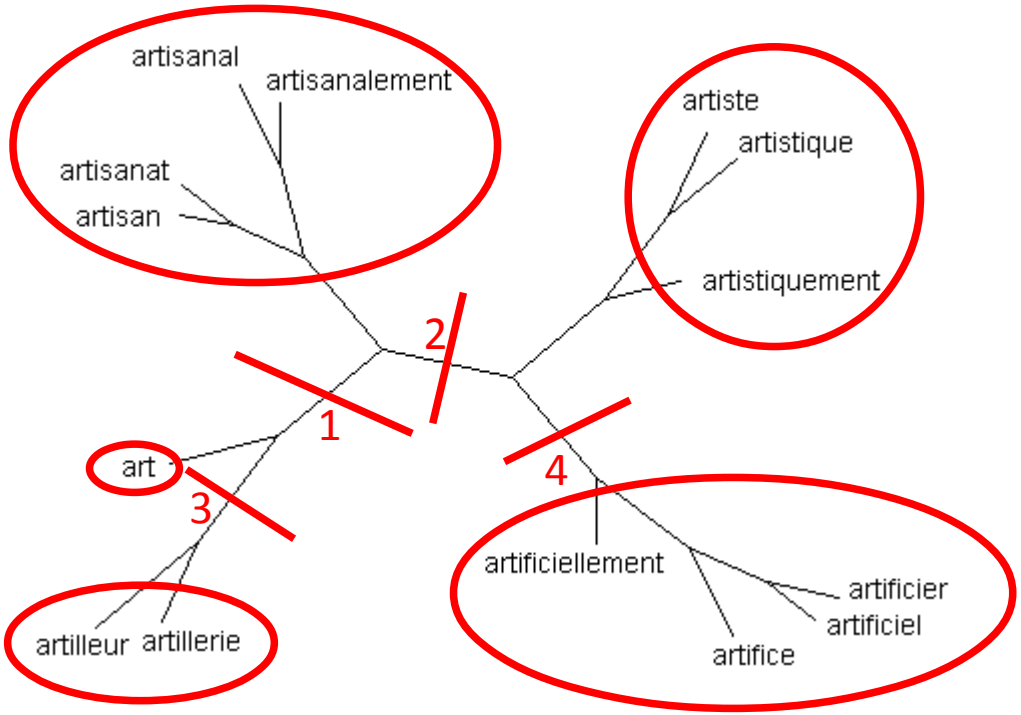
Partition automatique :
P3 = {{artisan, artisanat, artisanal, artisanalement}, {artiste, artistique, artistiquement, artificier, artificiel, artifice, artificiellement}, {artillerie, artilleur}, {art}**}**

Partition manuelle : **Pm = {**{artificier, artifice, artificiel, artificiellement}, {artillerie, artilleur}, {artisan, artisanal, artisanalement, artisanat}, {artiste, artistique, artistiquement, art}**}**

Protocole d'évaluation

Idée :

- Construire une représentation arborée des mots de la famille
- Découper les *k* arêtes les plus longues de l'arbre pour obtenir *Pk*
- La partition obtenue est-elle proche de la partition “manuelle” ?



Partition automatique :
P4 = {{artisan, artisanat, artisanal, artisanalement}, {artiste, artistique, artistiquement}, {artificier, artificiel, artifice, artificiellement}, {artillerie, artilleur}, {art}**}**

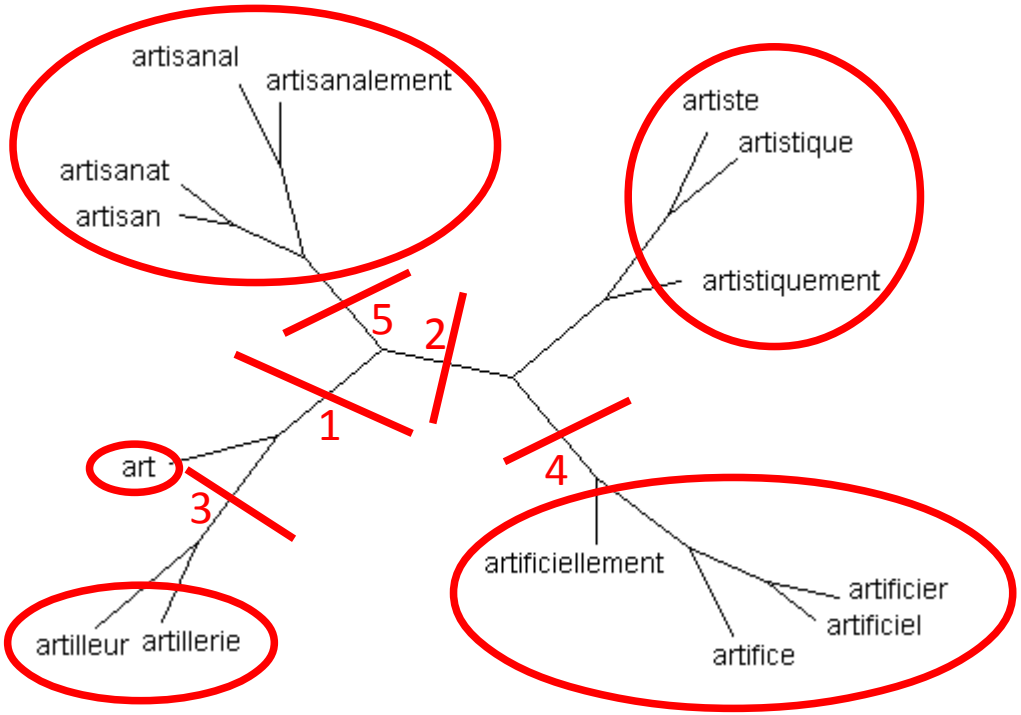
Partition manuelle : **Pm = {**{artificier, artifice, artificiel, artificiellement}, {artillerie, artilleur}, {artisan, artisanal, artisanalement, artisanat}, {artiste, artistique, artistiquement, art}**}**

Gala, Hathout, Nasr, Rey, Seppälä, *TALN'11*

Protocole d'évaluation

Idée :

- Construire une représentation arborée des mots de la famille
- Découper les *k* arêtes les plus longues de l'arbre pour obtenir *Pk*
- La partition obtenue est-elle proche de la partition “manuelle” ?



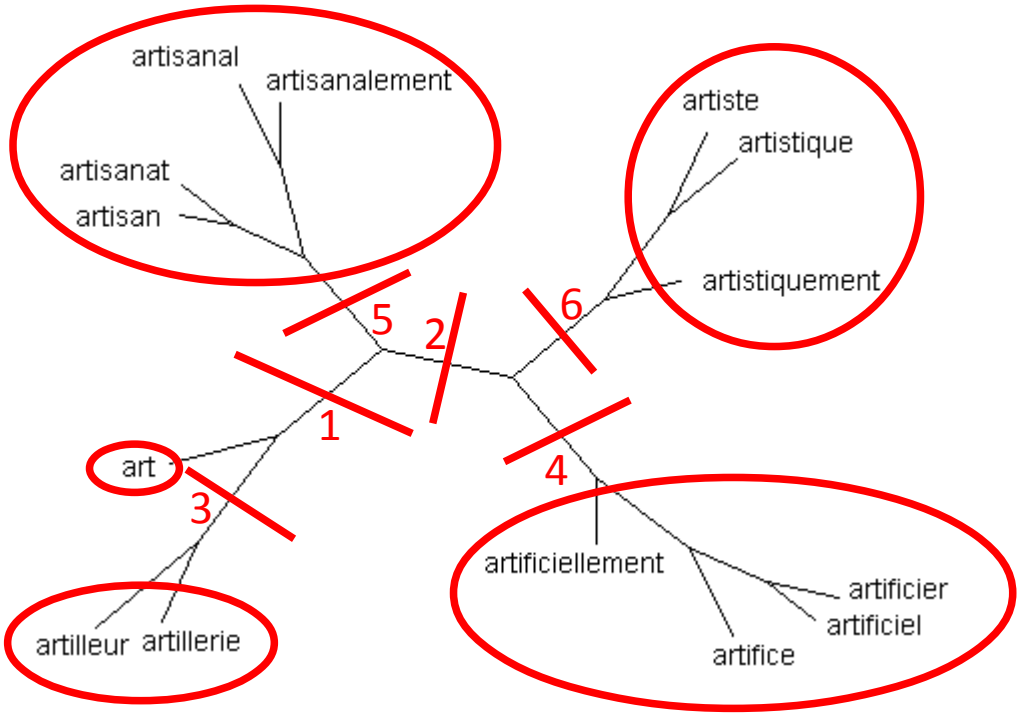
Partition automatique :
P5 = {{artisan, artisanat, artisanal, artisanalement}, {artiste, artistique, artistiquement}, {artificier, artificiel, artifice, artificiellement}, {artillerie, artilleur}, {art}**}**

Partition manuelle : **Pm = {**{artificier, artifice, artificiel, artificiellement}, {artillerie, artilleur}, {artisan, artisanal, artisanalement, artisanat}, {artiste, artistique, artistiquement, art}**}**

Protocole d'évaluation

Idée :

- Construire une représentation arborée des mots de la famille
- Découper les *k* arêtes les plus longues de l'arbre pour obtenir *Pk*
- La partition obtenue est-elle proche de la partition “manuelle” ?



Partition automatique :
P6 = {{artisan, artisanat, artisanal, artisanalement}, {artiste, artistique, artistiquement}, {artificier, artificiel, artifice, artificiellement}, {artillerie, artilleur}, {art}**}**

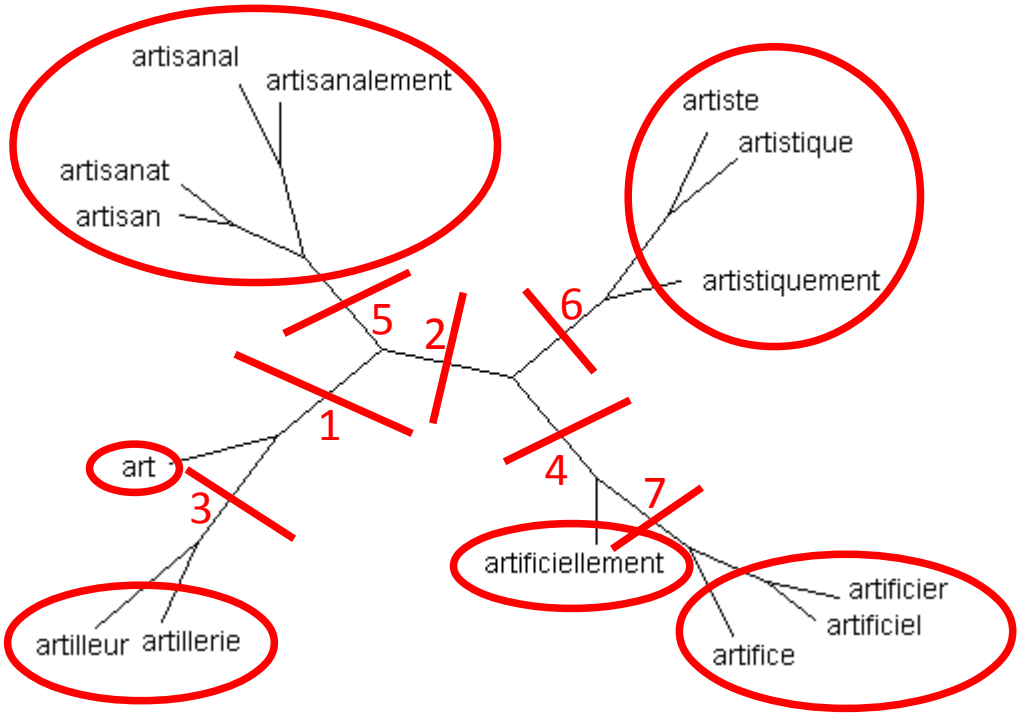
Partition manuelle : **Pm = {**{artificier, artifice, artificiel, artificiellement}, {artillerie, artilleur}, {artisan, artisanal, artisanalement, artisanat}, {artiste, artistique, artistiquement, art}**}**

Gala, Hathout, Nasr, Rey, Seppälä, *TALN'11*

Protocole d'évaluation

Idée :

- Construire une représentation arborée des mots de la famille
- Découper les *k* arêtes les plus longues de l'arbre pour obtenir *Pk*
- La partition obtenue est-elle proche de la partition “manuelle” ?



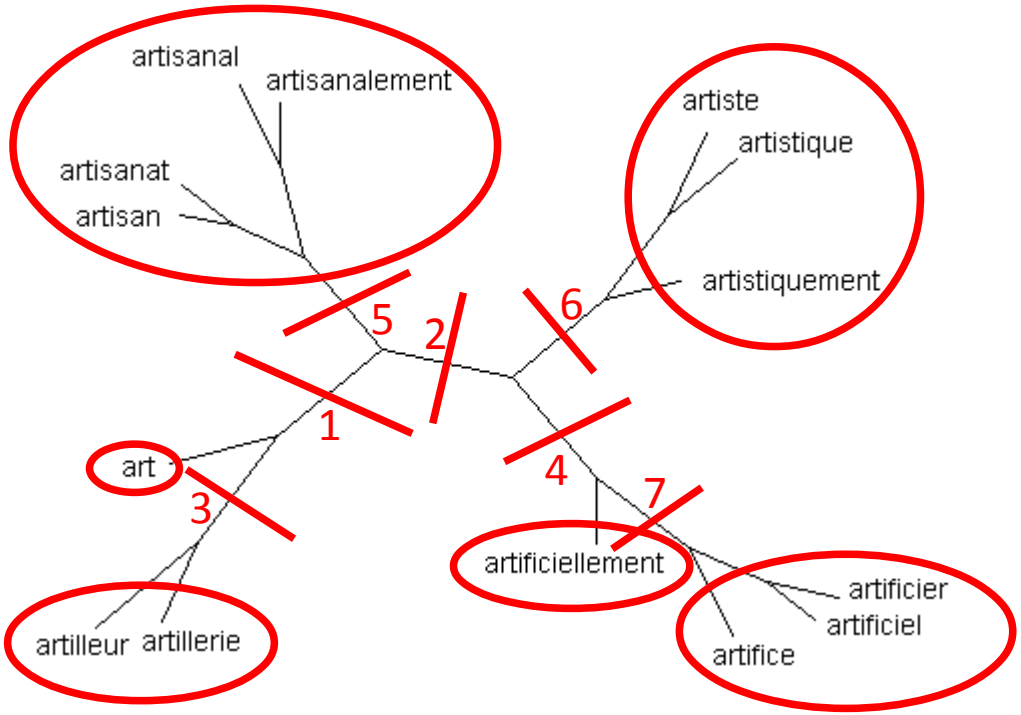
Partition automatique :
P7 = {{artisan, artisanat, artisanal, artisanalement}, {artiste, artistique, artistiquement}, {artificier, artificiel, artifice}, {artificiellement}, {artillerie, artilleur}, {art}**}**

Partition manuelle : **Pm = {**{artificier, artifice, artificiel, artificiellement}, {artillerie, artilleur}, {artisan, artisanal, artisanalement, artisanat}, {artiste, artistique, artistiquement, art}**}**

Protocole d'évaluation

Idée :

- Construire une représentation arborée des mots de la famille
- Découper les *k* arêtes les plus longues de l'arbre pour obtenir *Pk*
- La partition obtenue est-elle proche de la partition “manuelle” ?



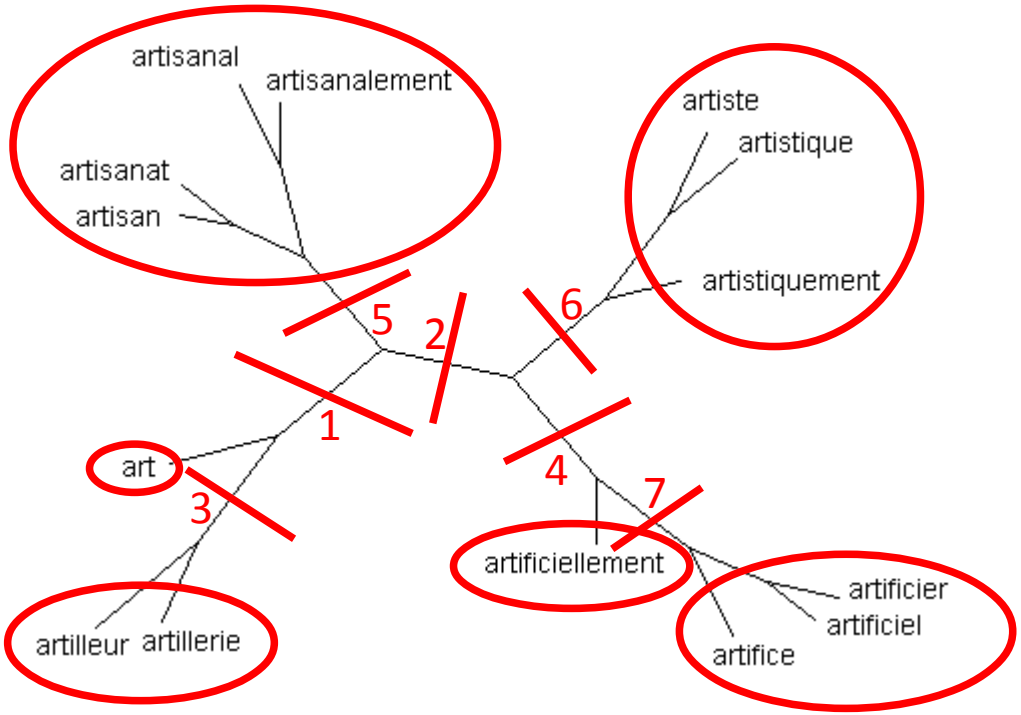
Partition automatique :
P7 = {{artisan, artisanat, artisanal, artisanalement}, {artiste, artistique, artistiquement}, {artificier, artificiel, artifice}, {artificiellement}, {artillerie, artilleur}, {art}**}**

Partition manuelle : **Pm = {**{artificier, artifice, artificiel, artificiellement}, {artillerie, artilleur}, {artisan, artisanal, artisanalement, artisanat}, {artiste, artistique, artistiquement, art}**}**

Protocole d'évaluation

Idée :

- Construire une représentation arborée des mots de la famille
- Découper les *k* arêtes les plus longues de l'arbre pour obtenir *Pk*
- La partition obtenue est-elle proche de la partition “manuelle” ?



Partition automatique :
P7 = {{artisan, artisanat, artisanal, artisanalement}, {artiste, artistique, artistiquement}, {artificier, artificiel, artifice}, {artificiellement}, {artillerie, artilleur}, {art}**}**

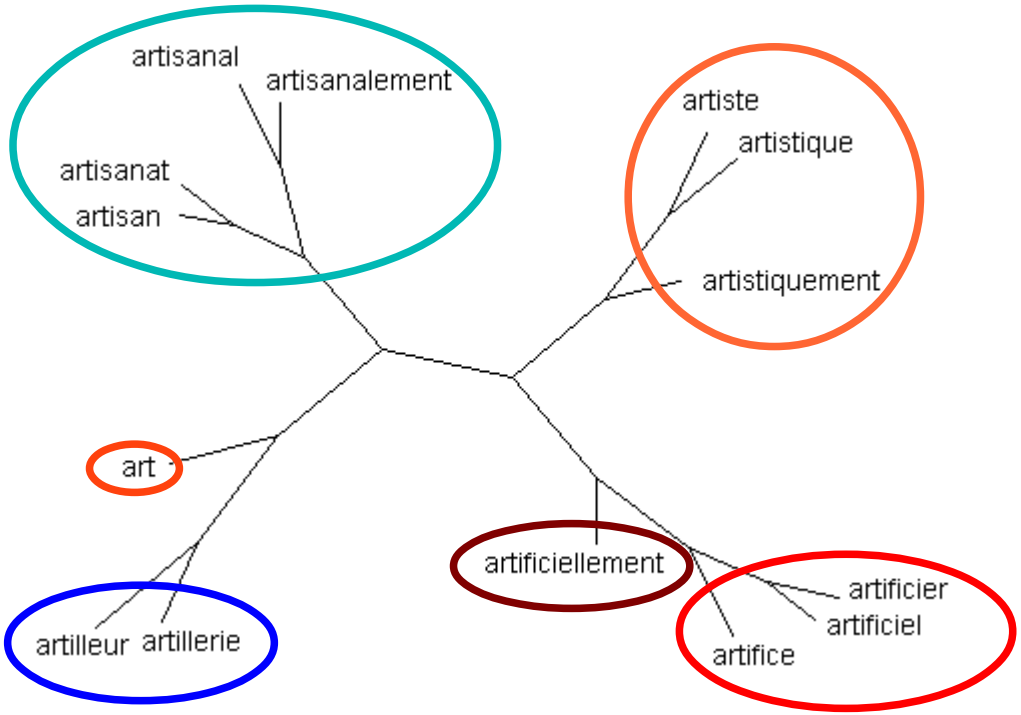
Comparer les partitions !
(indice de Rand, Rand corrigé)

Partition manuelle : **Pm = {**{artificier, artifice, artificiel, artificiellement}, {artillerie, artilleur}, {artisan, artisanal, artisanalement, artisanat}, {artiste, artistique, artistiquement, art}**}**

Protocole d'évaluation

Idée :

- Construire une représentation arborée des mots de la famille
- Découper les *k* arêtes les plus longues de l'arbre pour obtenir *Pk*
- La partition obtenue est-elle proche de la partition “manuelle” ?



Partition automatique :
P7 = {{artisan, artisanat, artisanal, artisanalement}, {artiste, artistique, artistiquement}, {artificier, artificiel, artifice}, {artificiellement}, {artillerie, artilleur}, {art}**}**

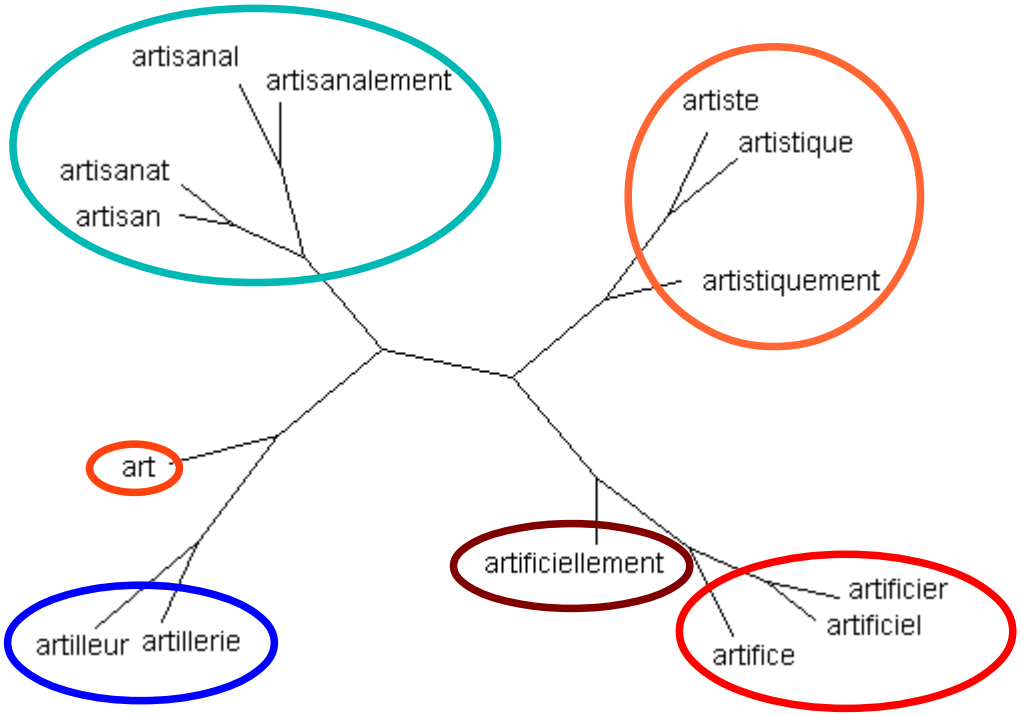
Comparer les partitions !
(indice de Rand, Rand corrigé)

Partition manuelle : **Pm = {**{artificier, artifice, artificiel, artificiellement}, {artillerie, artilleur}, {artisan, artisanal, artisanalement, artisanat}, {artiste, artistique, artistiquement, art}**}**

Protocole d'évaluation

Idée :

- Construire une représentation arborée des mots de la famille
- Découper les *k* arêtes les plus longues de l'arbre pour obtenir *Pk*
- La partition obtenue est-elle proche de la partition “manuelle” ?



Partition automatique :
P7 = {{artisan, artisanat, artisanal, artisanalement}, {artiste, artistique, artistiquement}, {artificier, artificiel, artifice}, {artificiellement}, {artillerie, artilleur}, {art}**}**

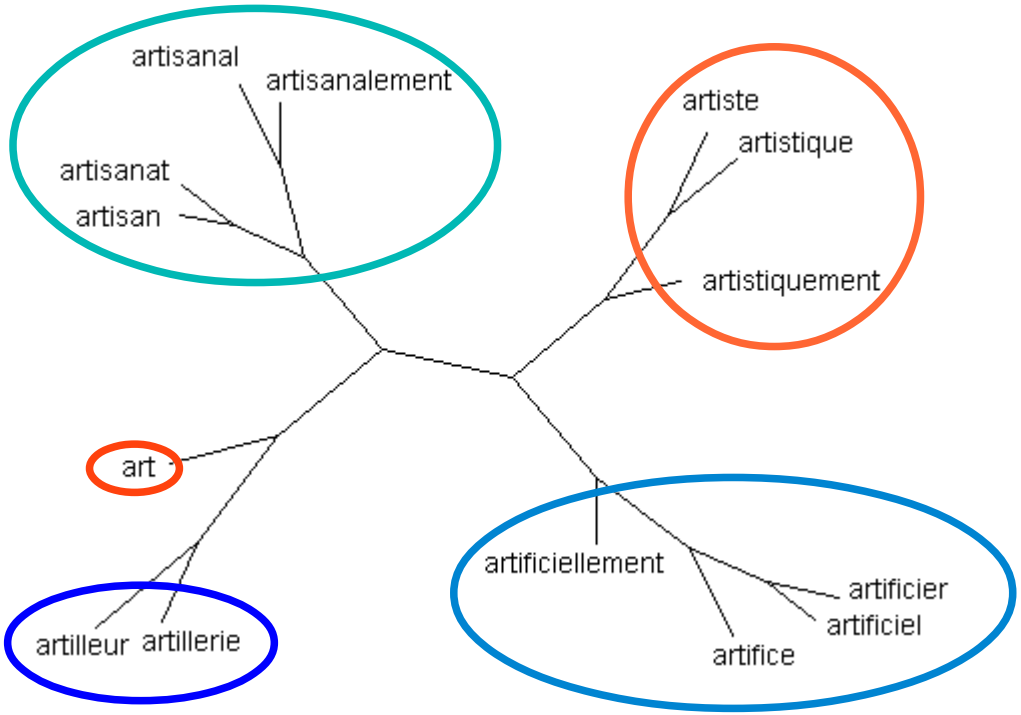
rand(Pm,P7) = 0.934
aRand(Pm,P7) = 0.774

Partition manuelle : **Pm = {**{artificier, artifice, artificiel, artificiellement}, {artillerie, artilleur}, {artisan, artisanal, artisanalement, artisanat}, {artiste, artistique, artistiquement, art}**}**

Protocole d'évaluation

Idée :

- Construire une représentation arborée des mots de la famille
- Découper les *k* arêtes les plus longues de l'arbre pour obtenir *Pk*
- La partition obtenue est-elle proche de la partition “manuelle” ?

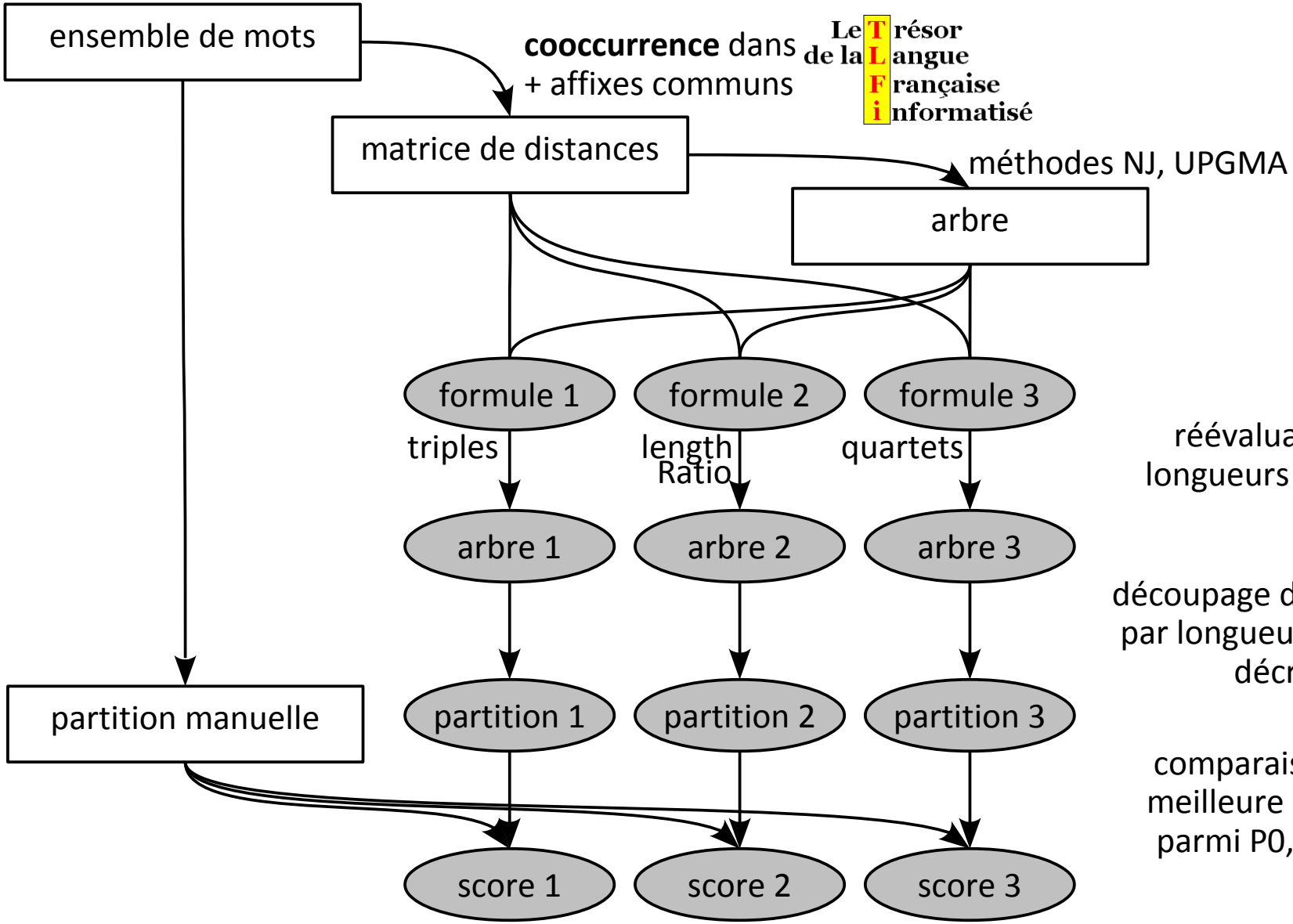


Partition automatique :
P4 = {{artisan, artisanat, artisanal, artisanalement}, {artiste, artistique, artistiquement}, {artificier, artificiel, artifice, artificiellement}, {artillerie, artilleur}, {art}**}**

rand(Pm,P4) = 0.967
aRand(Pm,P4) = 0.894

Partition manuelle : **Pm = {**{artificier, artifice, artificiel, artificiellement}, {artillerie, artilleur}, {artisan, artisanal, artisanalement, artisanat}, {artiste, artistique, artistiquement, art}**}**

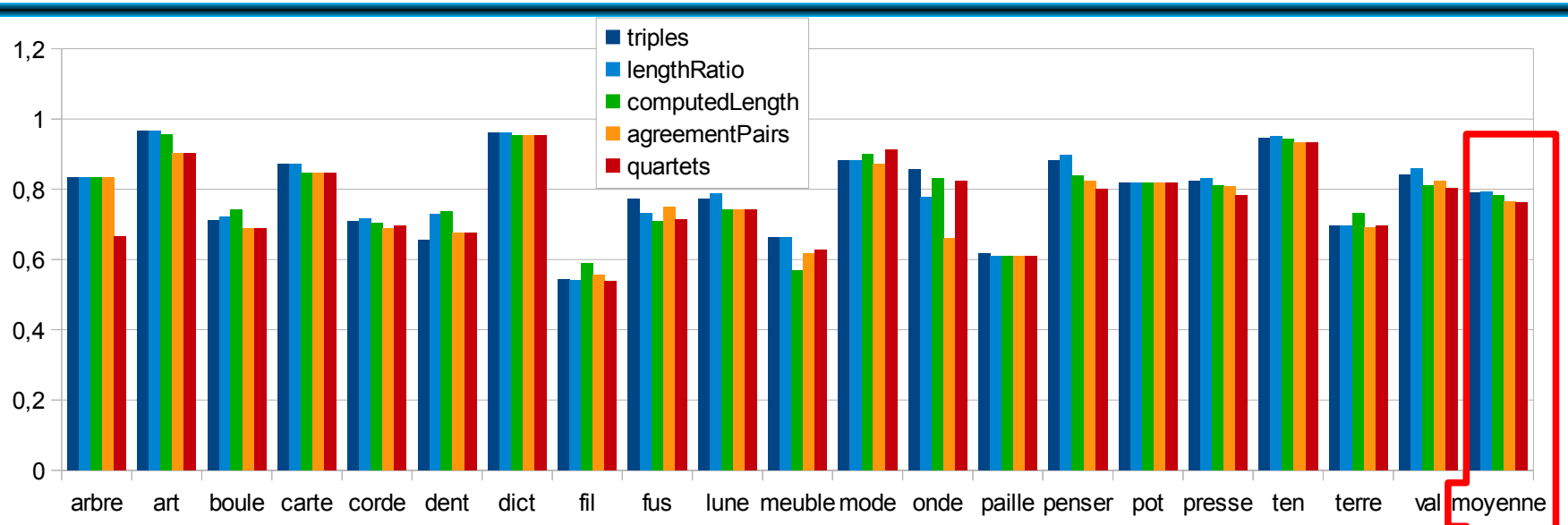
Protocole d'évaluation



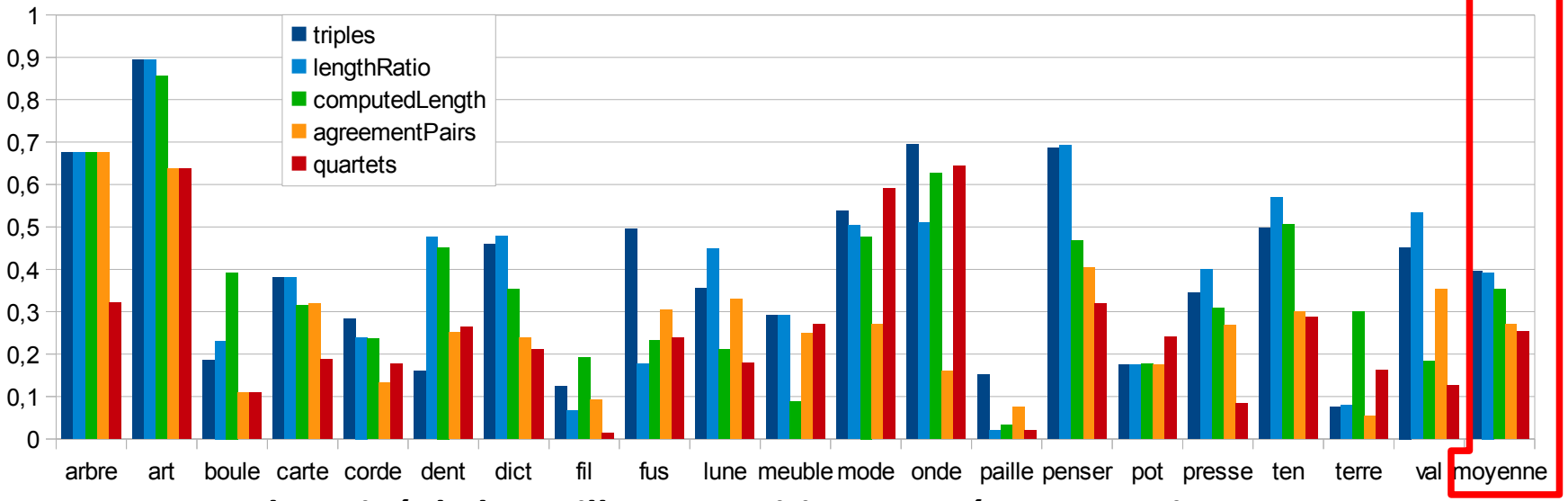
Calcul des longueurs d'arêtes

- Interprétation visuelle
- Formules de longueurs d'arêtes
- Protocole d'évaluation
- Résultats
- Visualisations
- Perspectives

Scores de chaque formule

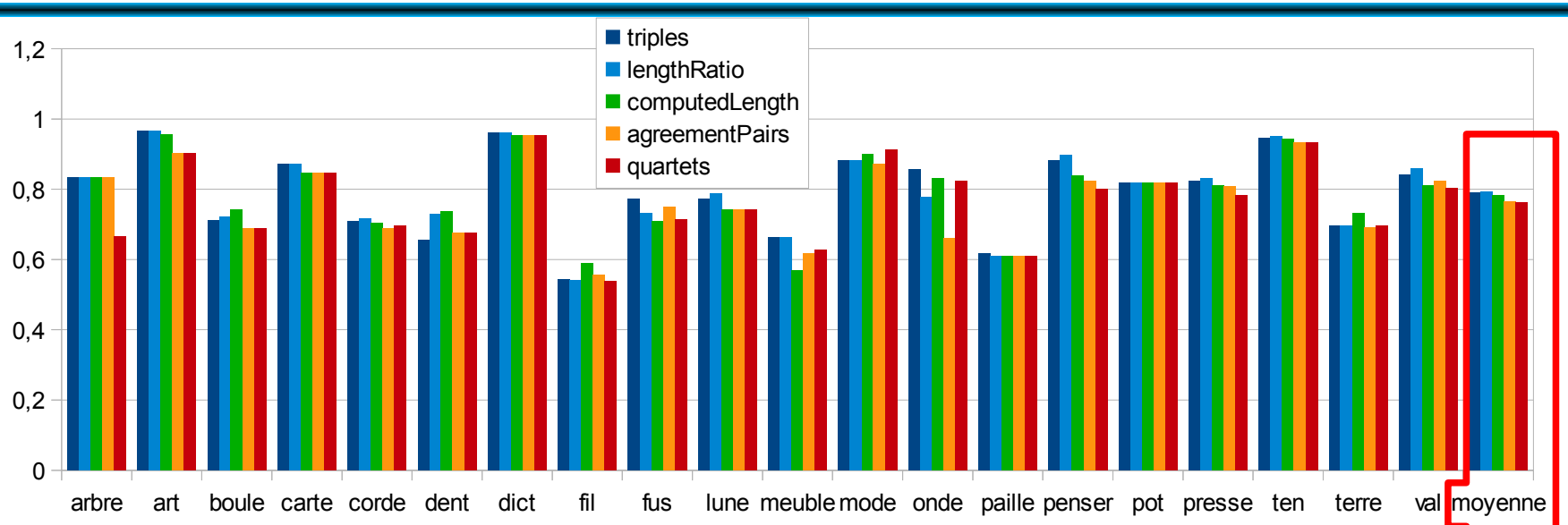


Score Rand de la meilleure partition trouvée automatiquement

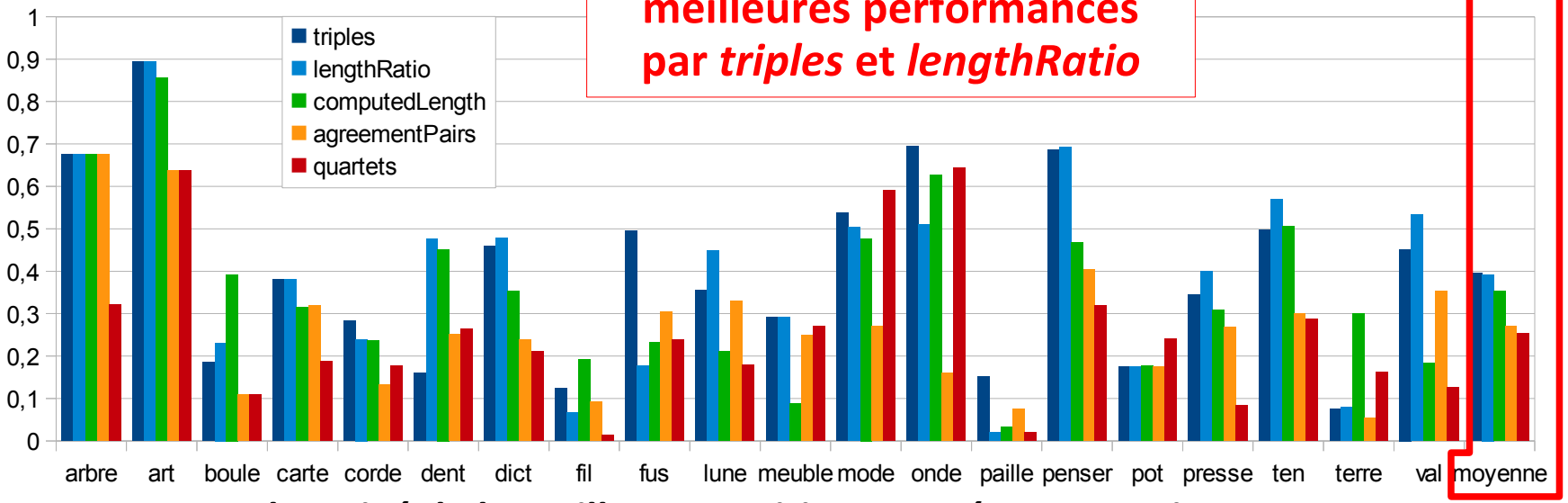


Score Rand corrigé de la meilleure partition trouvée automatiquement

Scores de chaque formule



Score Rand de la meilleure partition trouvée automatiquement



meilleures performances par *triples* et *lengthRatio*

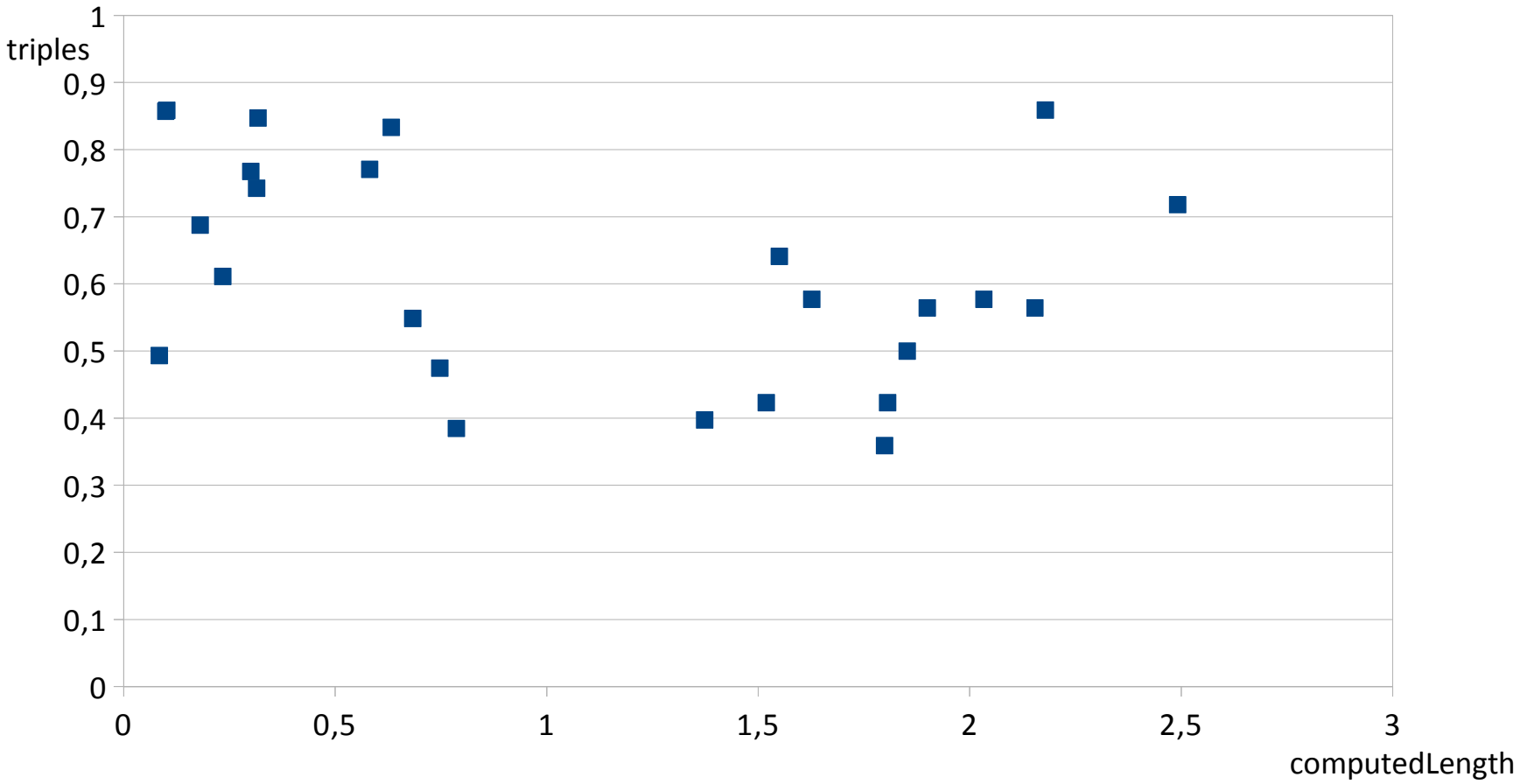
Score Rand corrigé de la meilleure partition trouvée automatiquement

Scores de chaque formule

Les formules de longueur d'arête sont-elles **cohérentes** ?

Scores de chaque formule

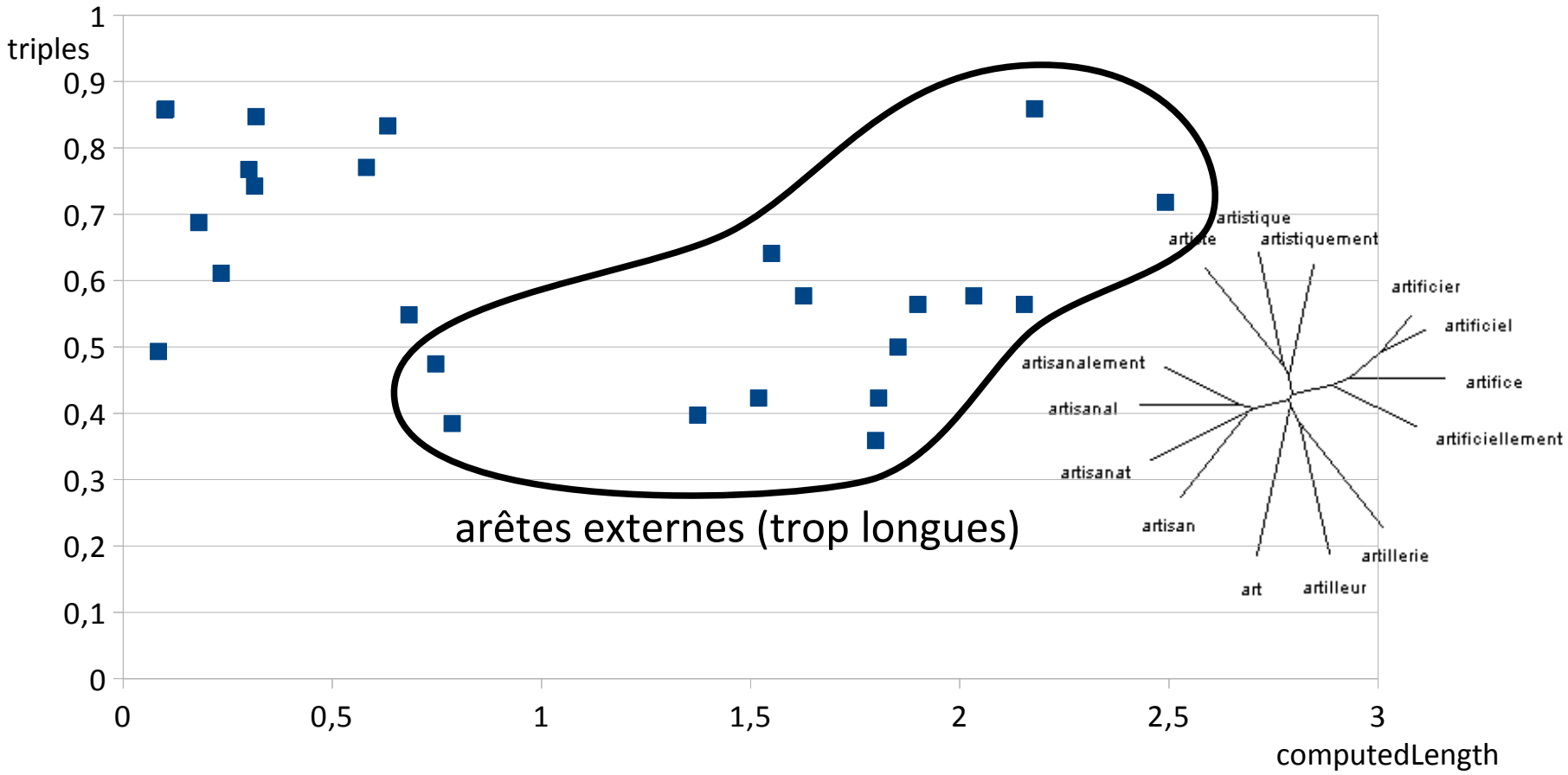
Les formules de longueur d'arête sont-elles **cohérentes** ?



Longueur selon la formule *triples* en fonction de la longueur originale de l'arête pour l'arbre de la famille de *art*

Scores de chaque formule

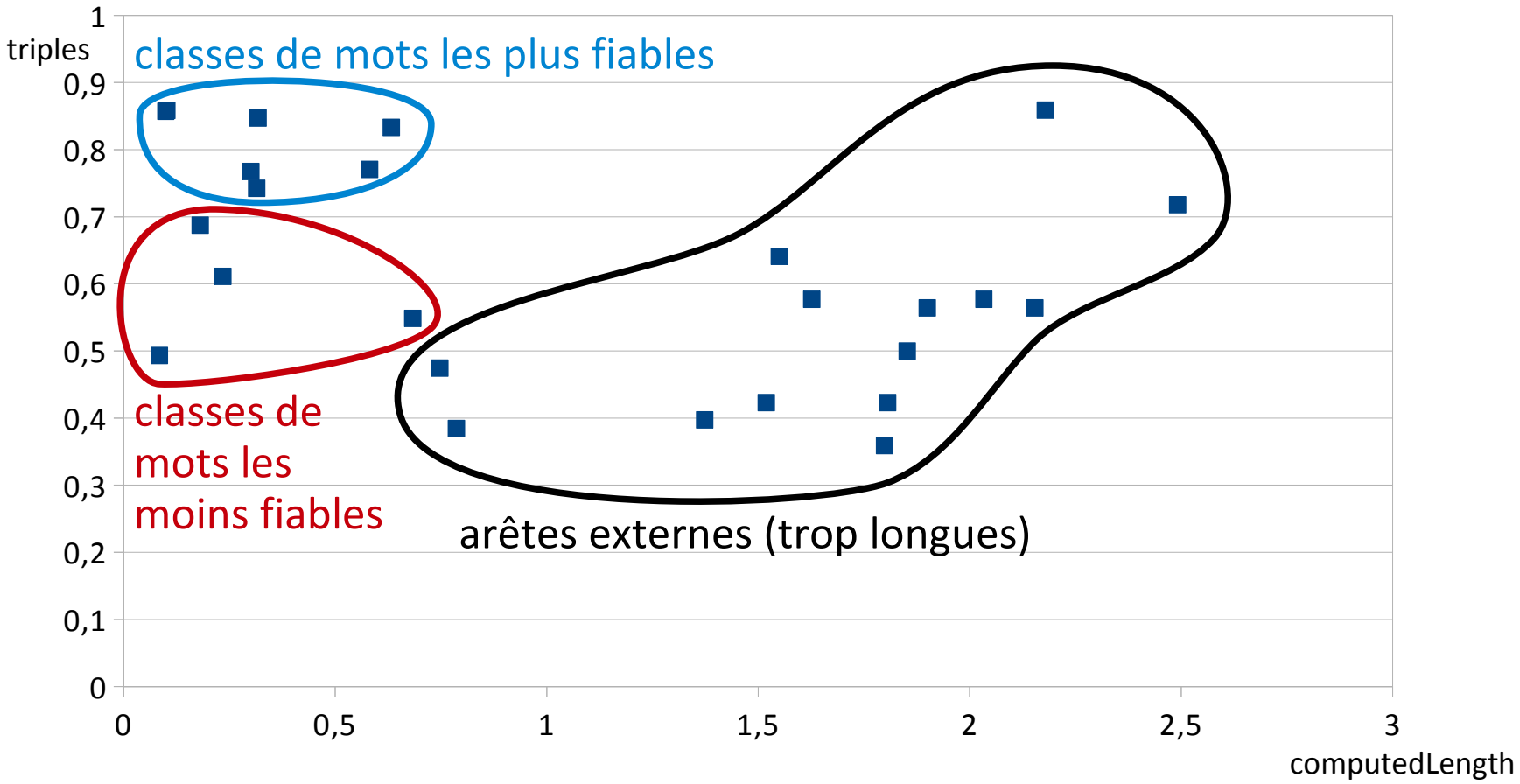
Les formules de longueur d'arête sont-elles **cohérentes** ?



Longueur selon la formule *triples* en fonction de la longueur originale de l'arête pour l'arbre de la famille de *art*

Scores de chaque formule

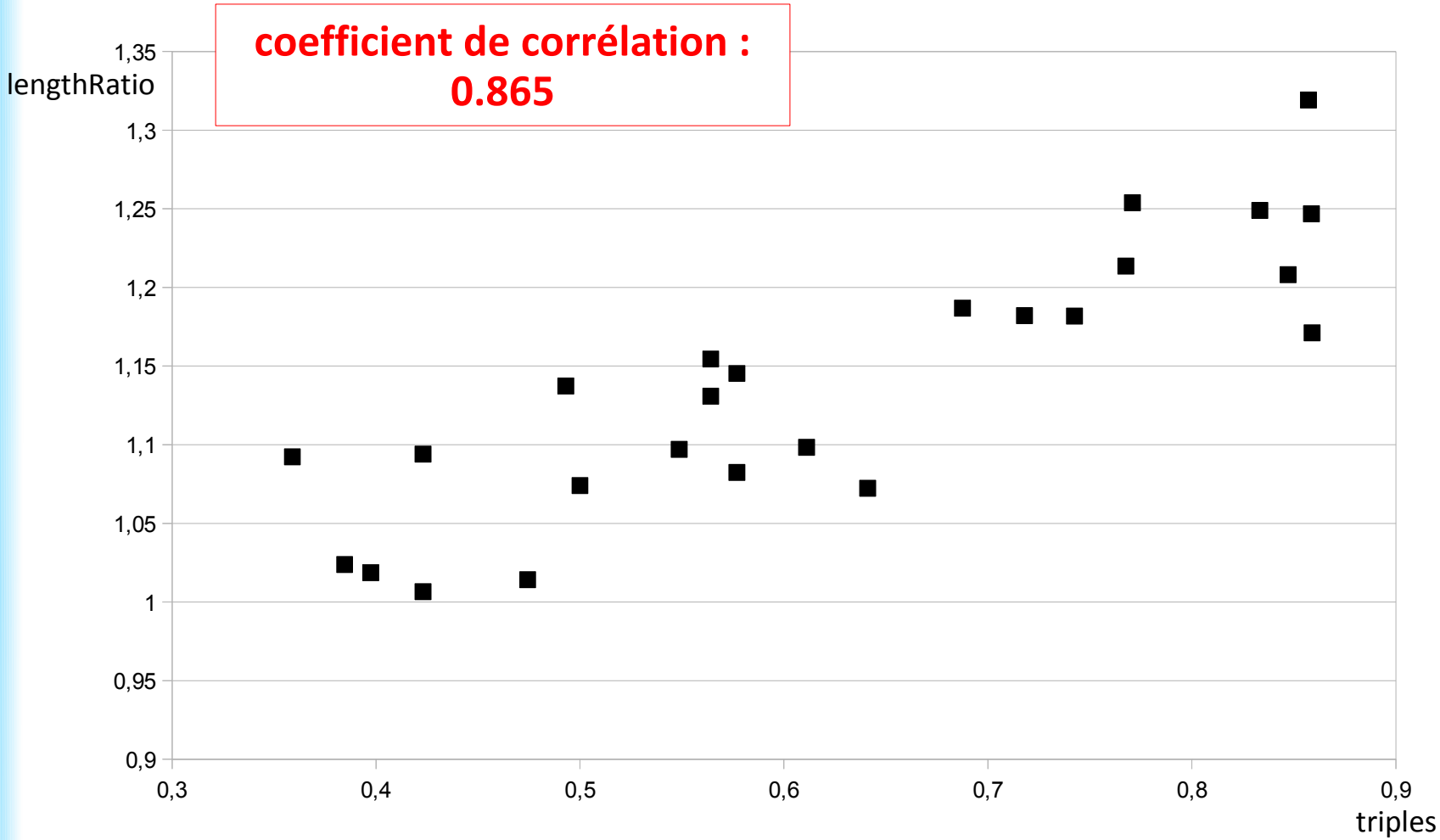
Les formules de longueur d'arête sont-elles **cohérentes** ?



Longueur selon la formule *triples* en fonction de la longueur originale de l'arête pour l'arbre de la famille de *art*

Scores de chaque formule

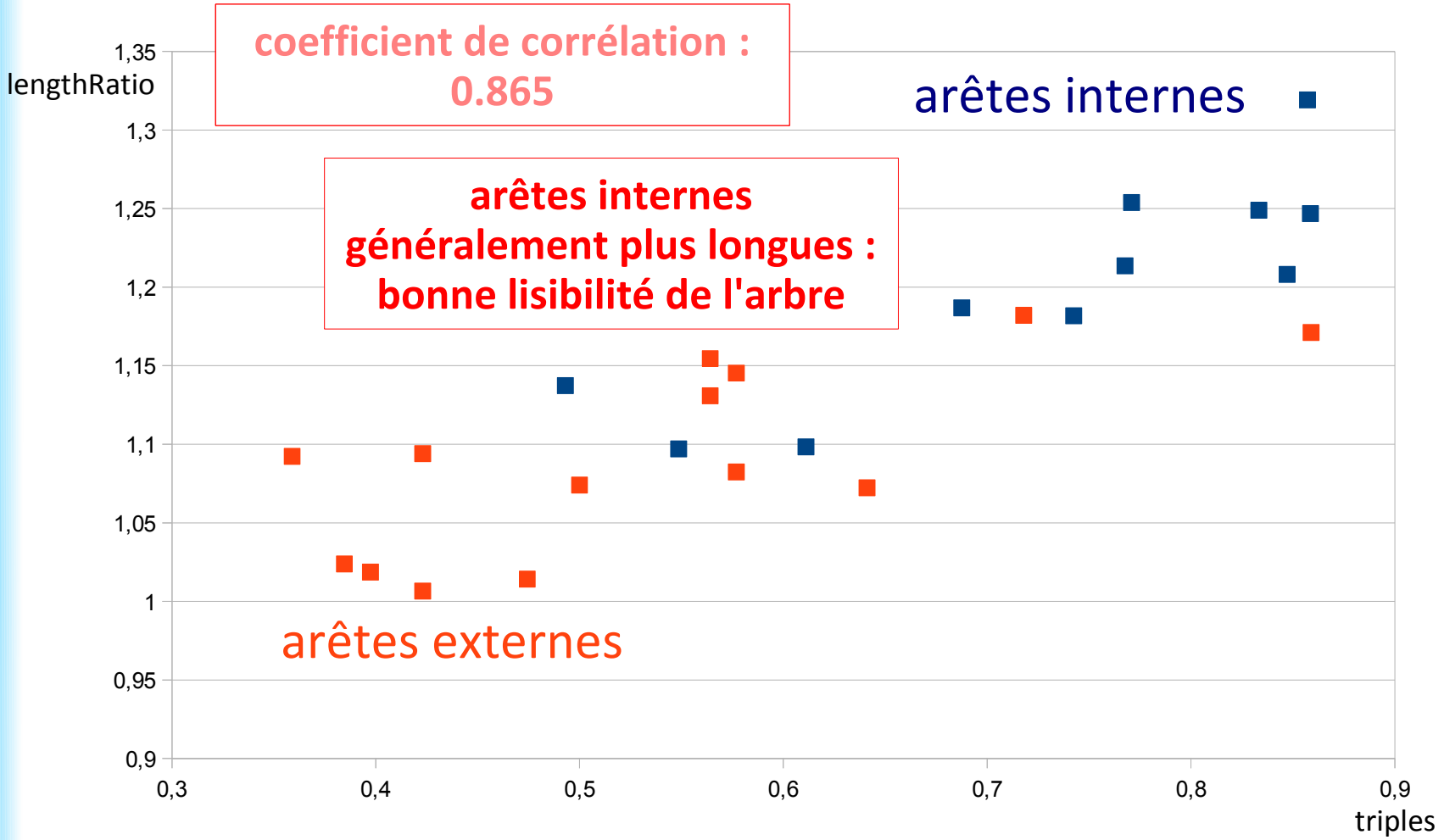
Les formules de longueur d'arête sont-elles **cohérentes** ?



Longueur selon la formule *lengthRatio* en fonction de celle selon la formule *triples* pour l'arbre de la famille de *art*

Scores de chaque formule

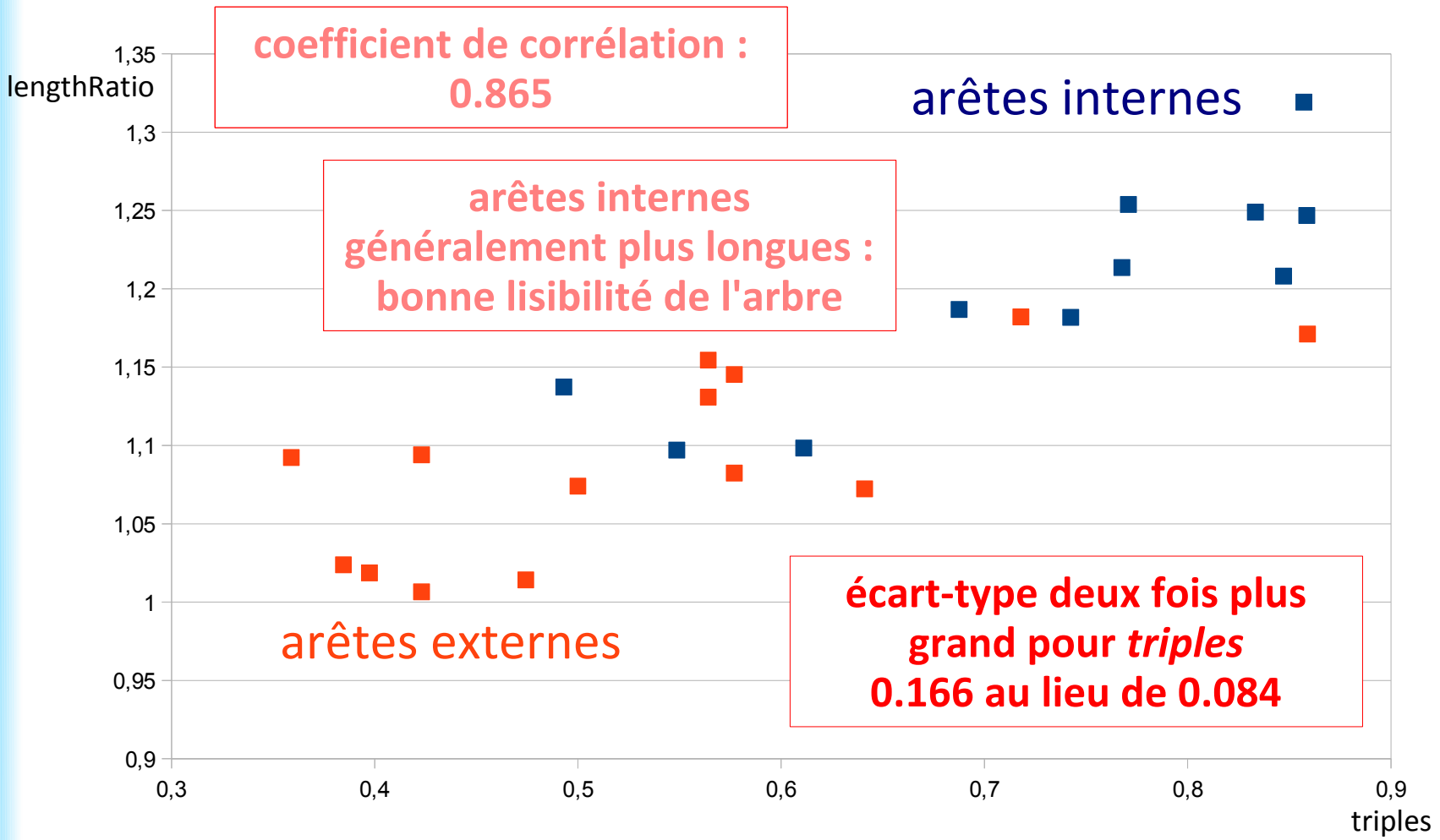
Les formules de longueur d'arête sont-elles **cohérentes** ?



Longueur selon la formule *lengthRatio* en fonction de celle selon la formule *triples* pour l'arbre de la famille de *art*

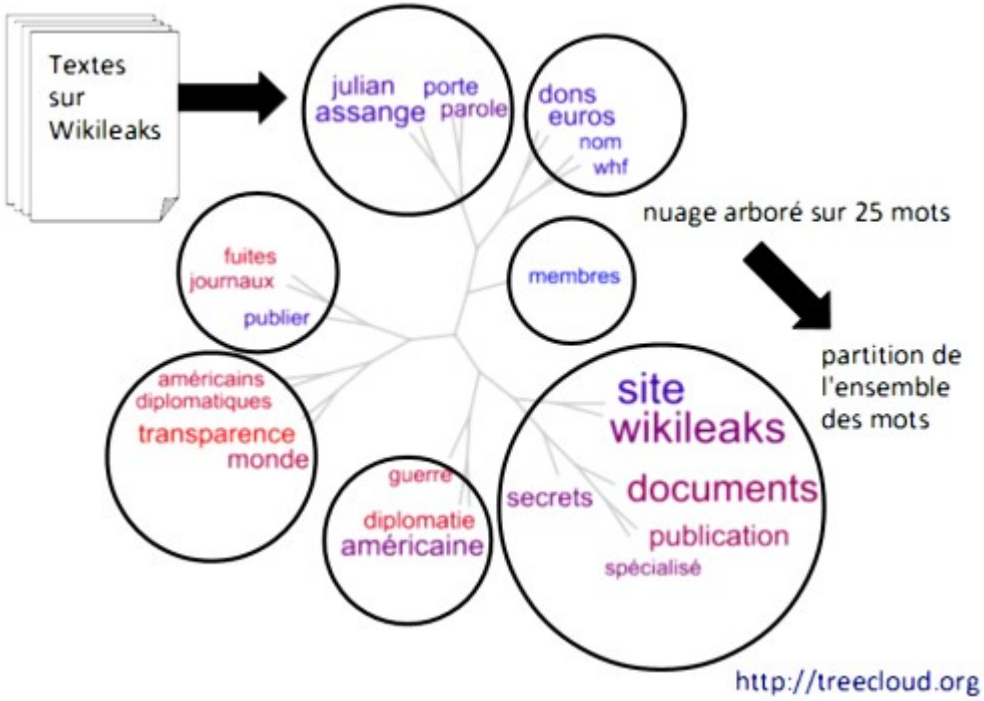
Scores de chaque formule

Les formules de longueur d'arête sont-elles **cohérentes** ?



Longueur selon la formule *lengthRatio* en fonction de celle selon la formule *triples* pour l'arbre de la famille de *art*

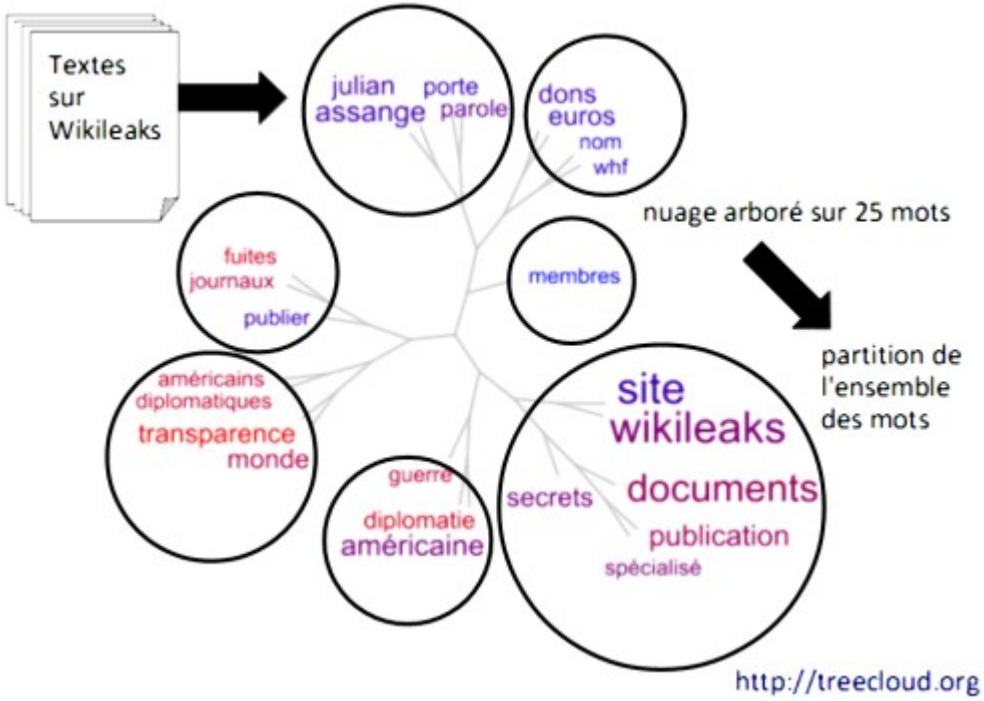
Second protocole d'évaluation



• 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

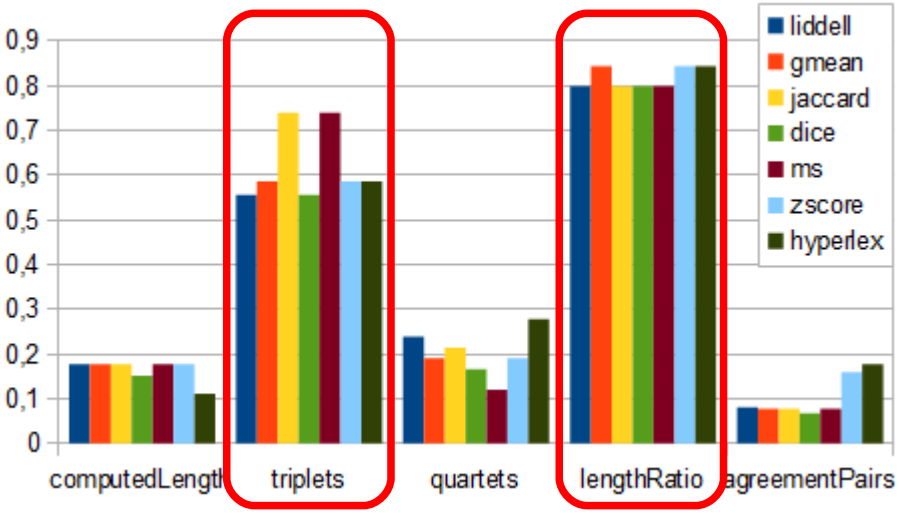
Second protocole d'évaluation



- 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

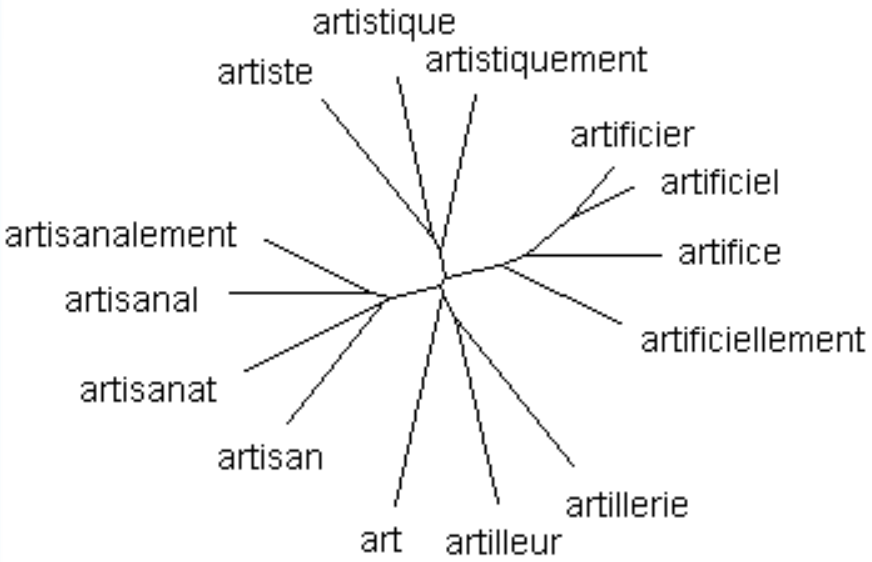


Calcul des longueurs d'arêtes

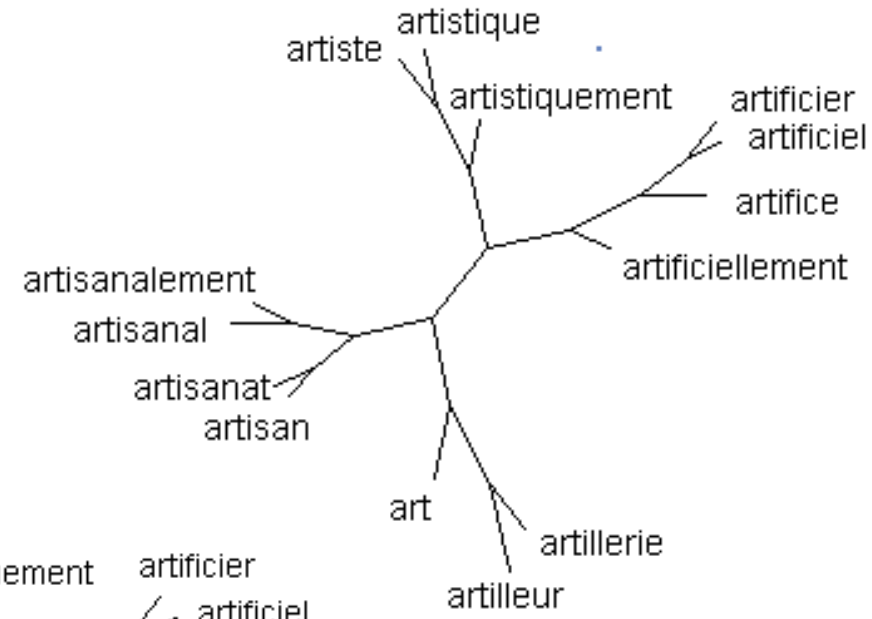
- Interprétation visuelle
- Formules de longueurs d'arêtes
- Protocole d'évaluation
- Résultats
- Visualisations

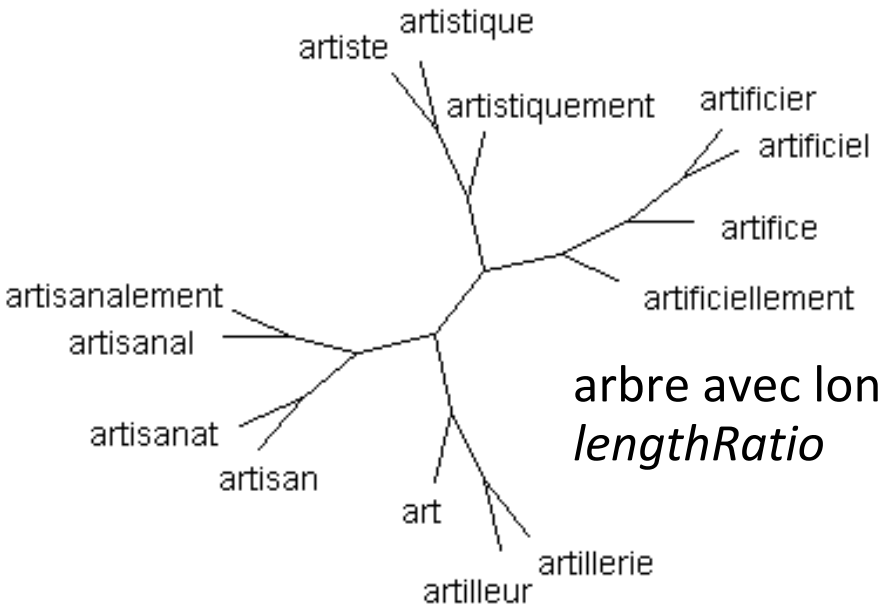
Visualisations

arbre original *computedLength*



arbre avec longueurs d'arêtes *triples*

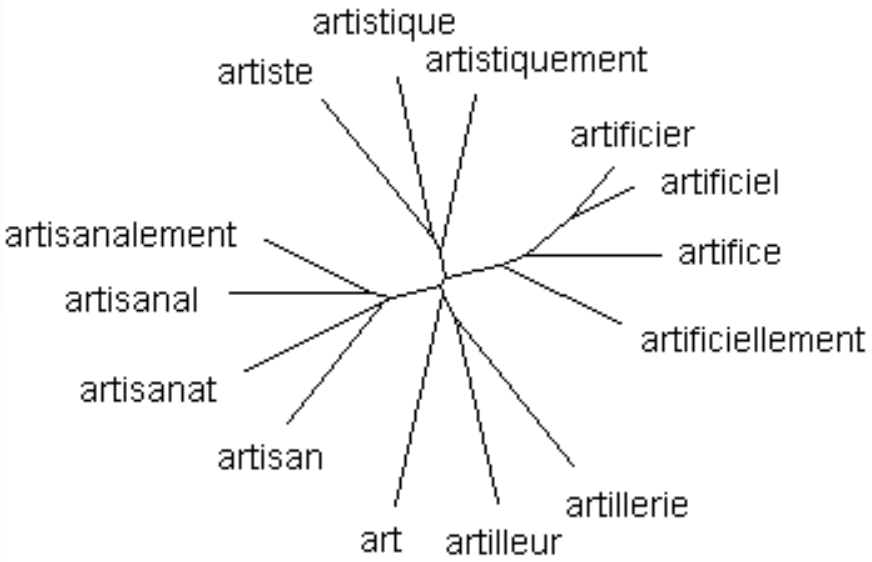




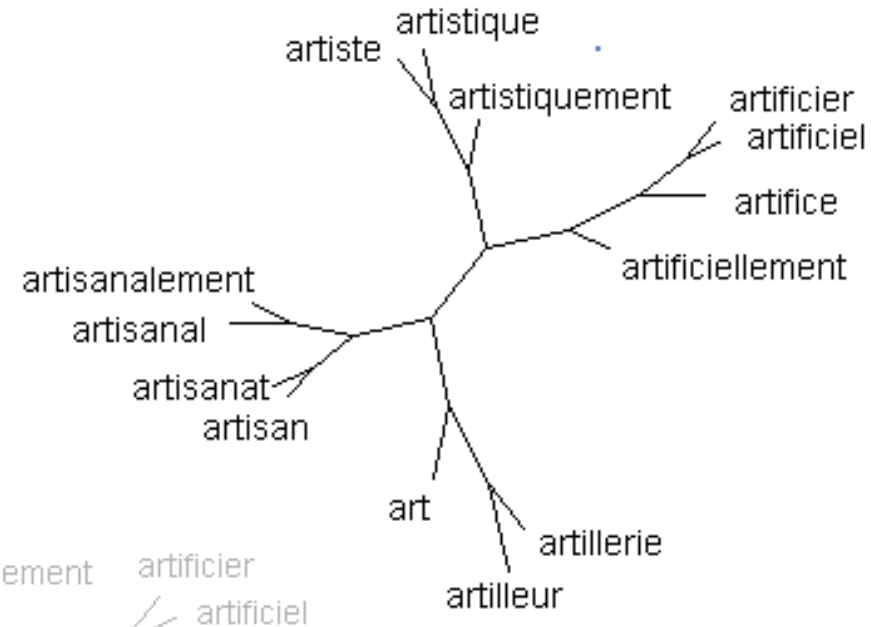
arbre avec longueurs d'arêtes *lengthRatio*

Visualisations

arbre original *computedLength*



arbre avec longueurs d'arêtes *triples*

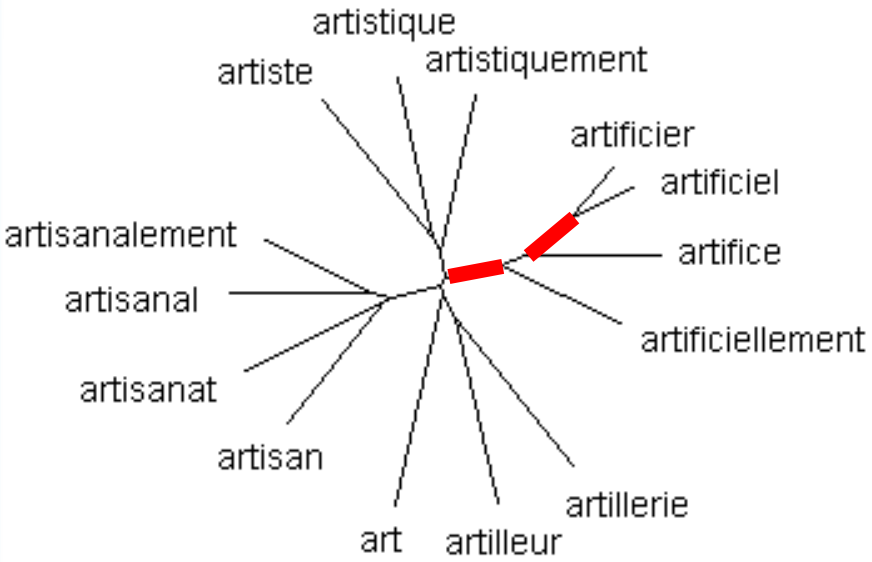


variance trop faible des longueurs d'arêtes !

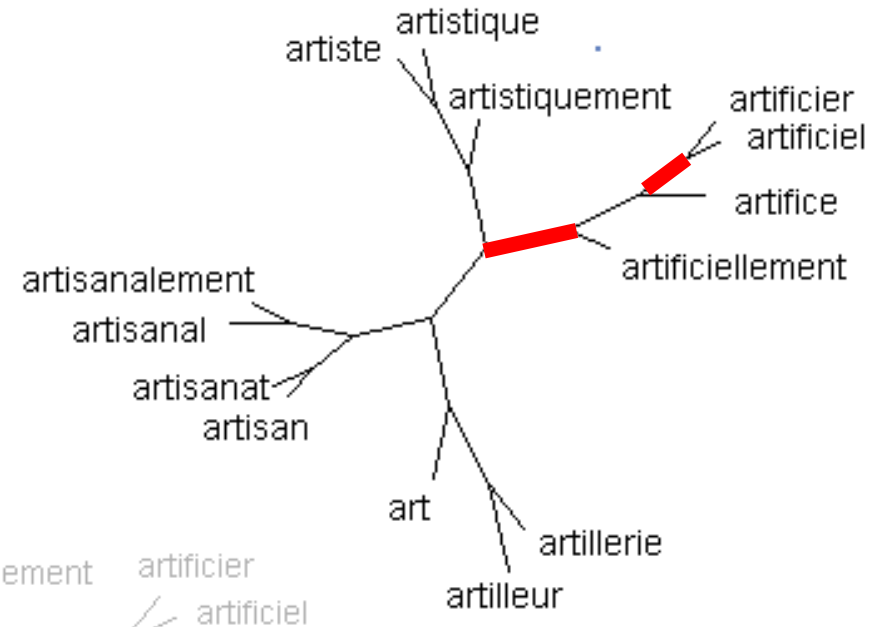
arbre avec longueurs d'arêtes *lengthRatio*

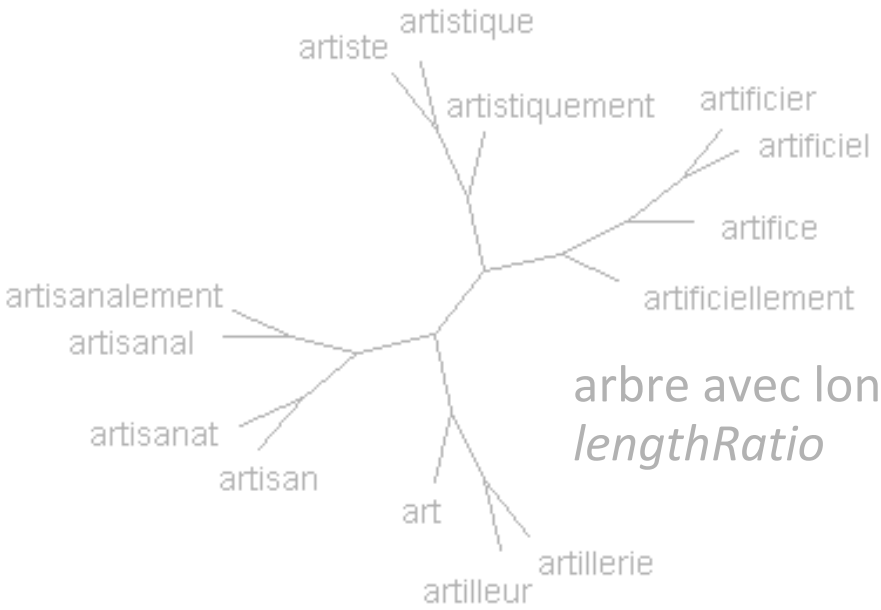
Visualisations

arbre original *computedLength*



arbre avec longueurs d'arêtes *triples*





arbre avec longueurs d'arêtes *lengthRatio*

Plan

- Nuages de mots et nuages arborés
- Caractéristiques du nuage arboré
- Utilisations du nuage arboré
- Construction d'un nuage arboré
- Évaluation de la robustesse de l'arbre
- Calcul des longueurs d'arêtes de l'arbre
- Perspectives

Perspectives

• intégration de la **visualisation en nuages arborés** avec longueurs de branches post-calculées :
    • 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

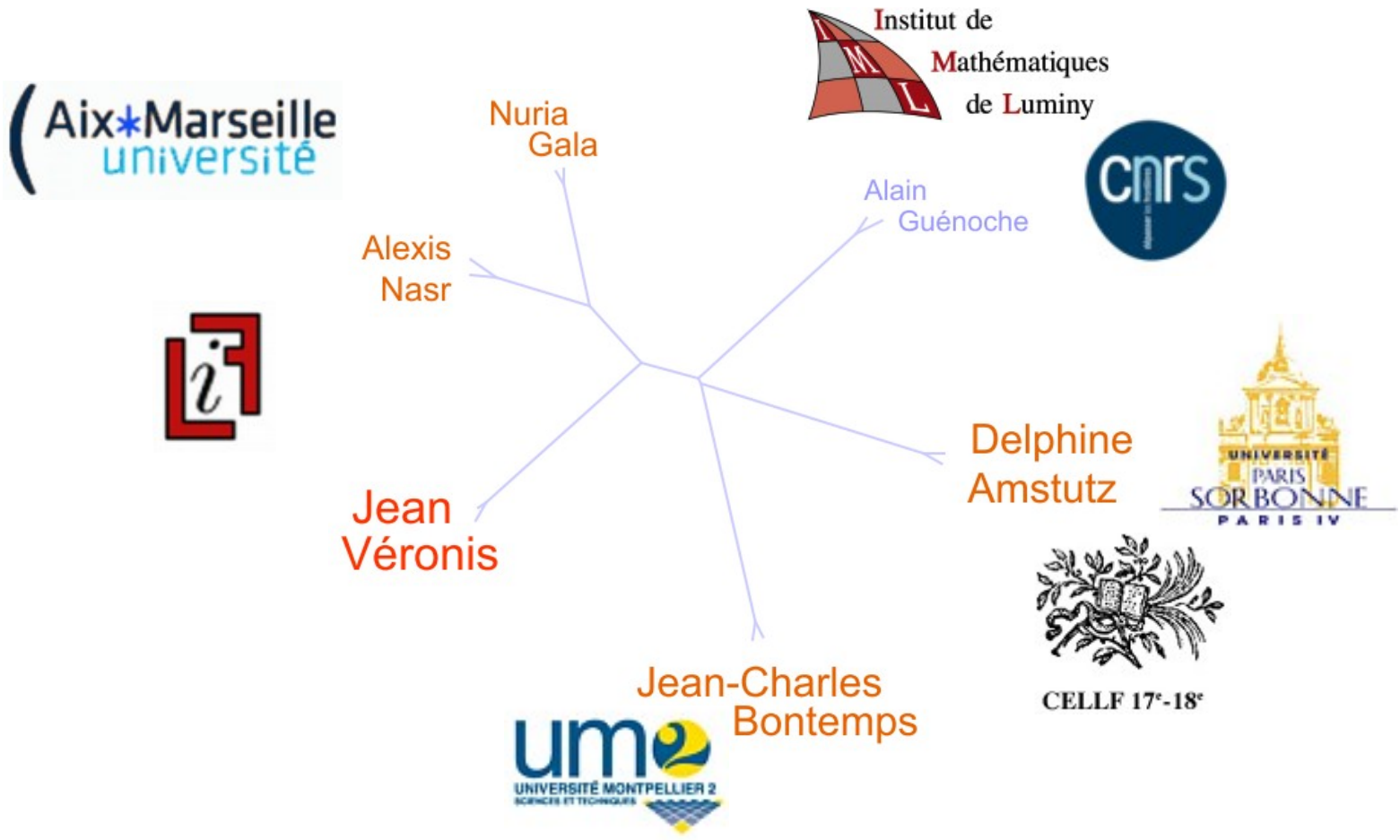
• étude du lien entre les distances de cooccurrences extraits d'un corpus et les données de Jeux de mots
    • réseau de plus de 200 000 mots et 1 200 000 liens pondérés
    • cooccurrences dans la production spontanée de mots par rapport à un mot cible
    • cohérence avec les distances de cooccurrence calculées à partir d'un texte ?

Lafourcade, *JADT'08*

Questions ?

Merci pour votre attention !

Coauteurs de ces travaux :

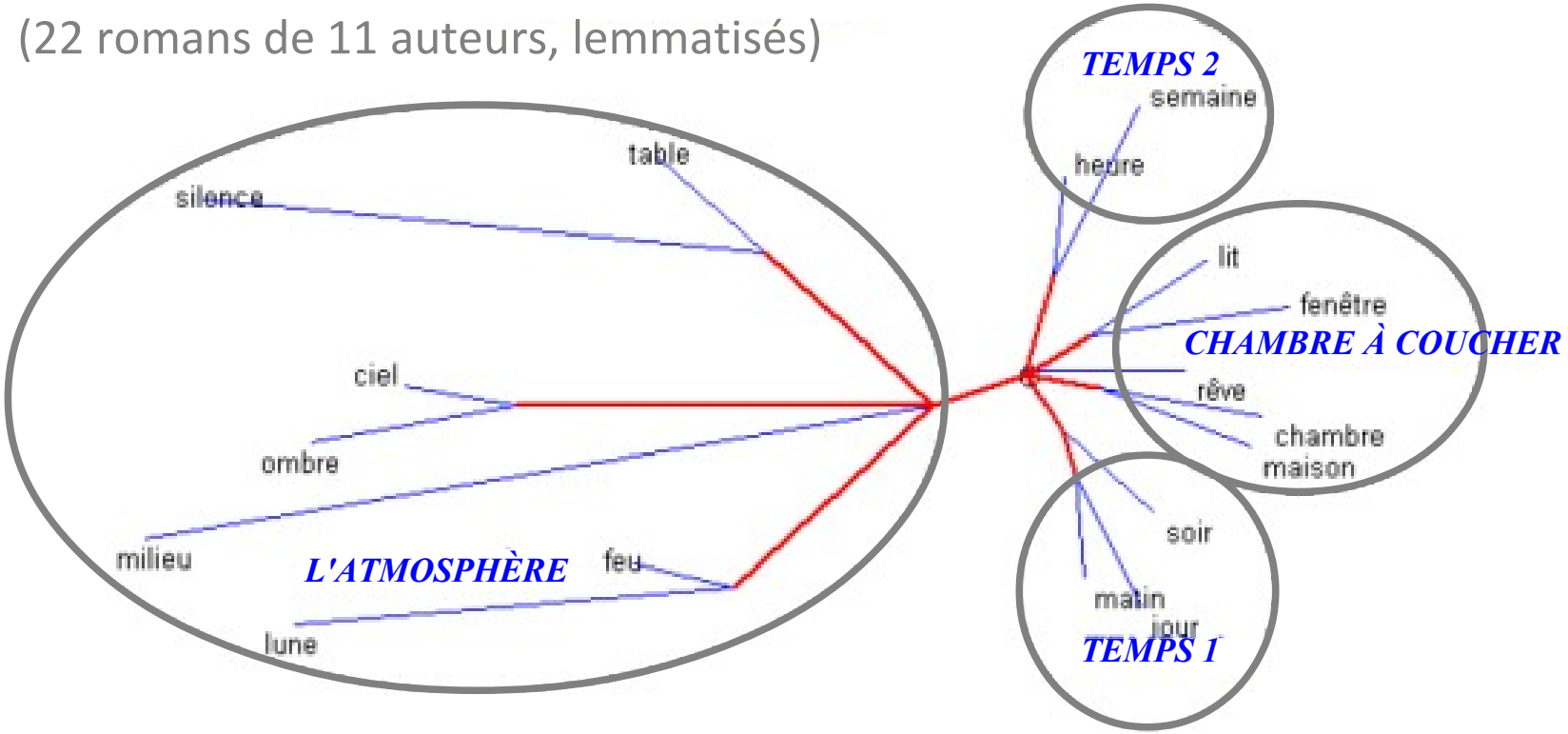


http://www.treecloud.org

Analyses arborées

Rapprochement des mots d'un texte selon leur degré de cooccurrence dans le texte

Constellation du mot “nuit” dans la base EXEMPLEM
(22 romans de 11 auteurs, lemmatisés)

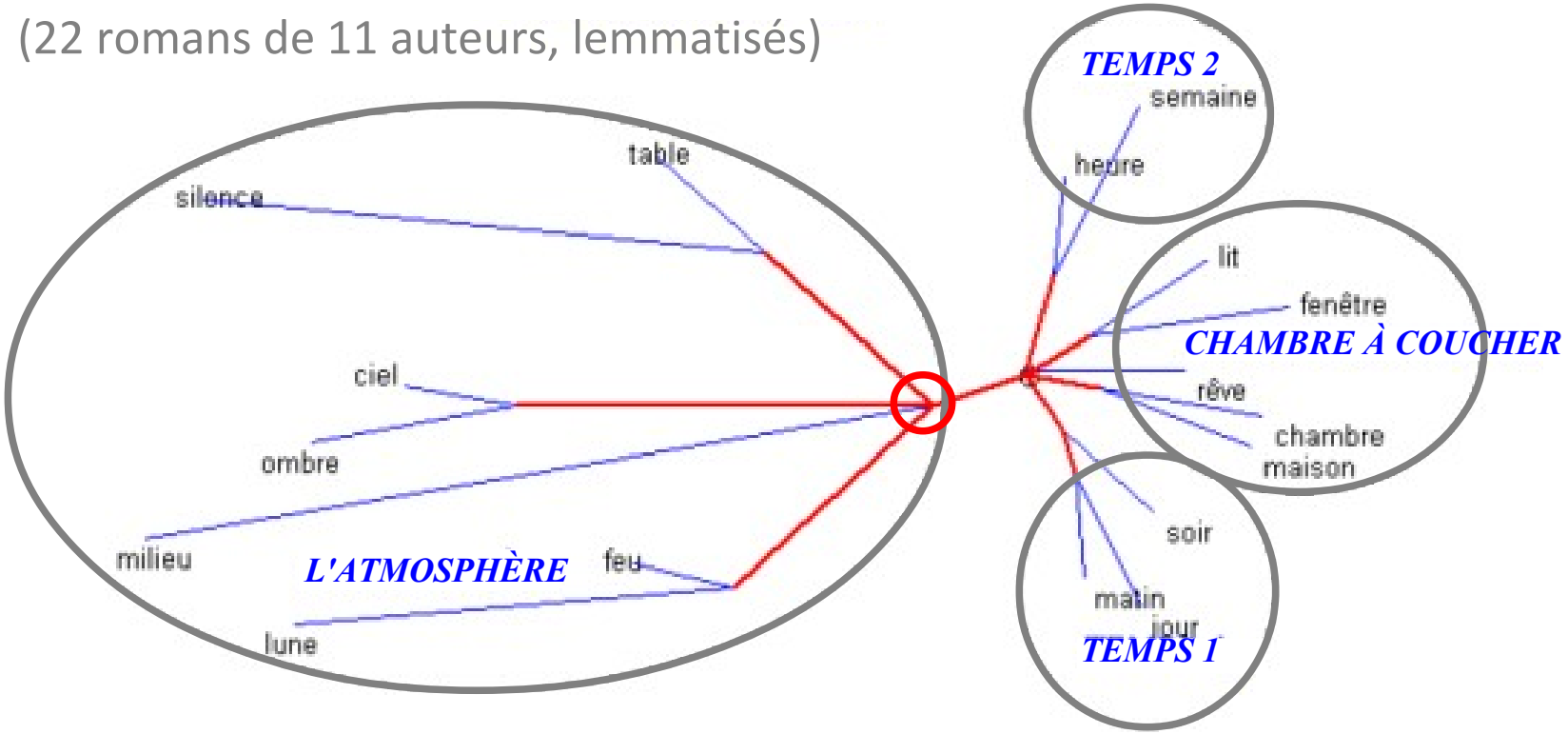


Barthelémy & Luong, *Statistique et Analyse des Données,* 1986
Brunet, *JADT'08*

Analyses arborées

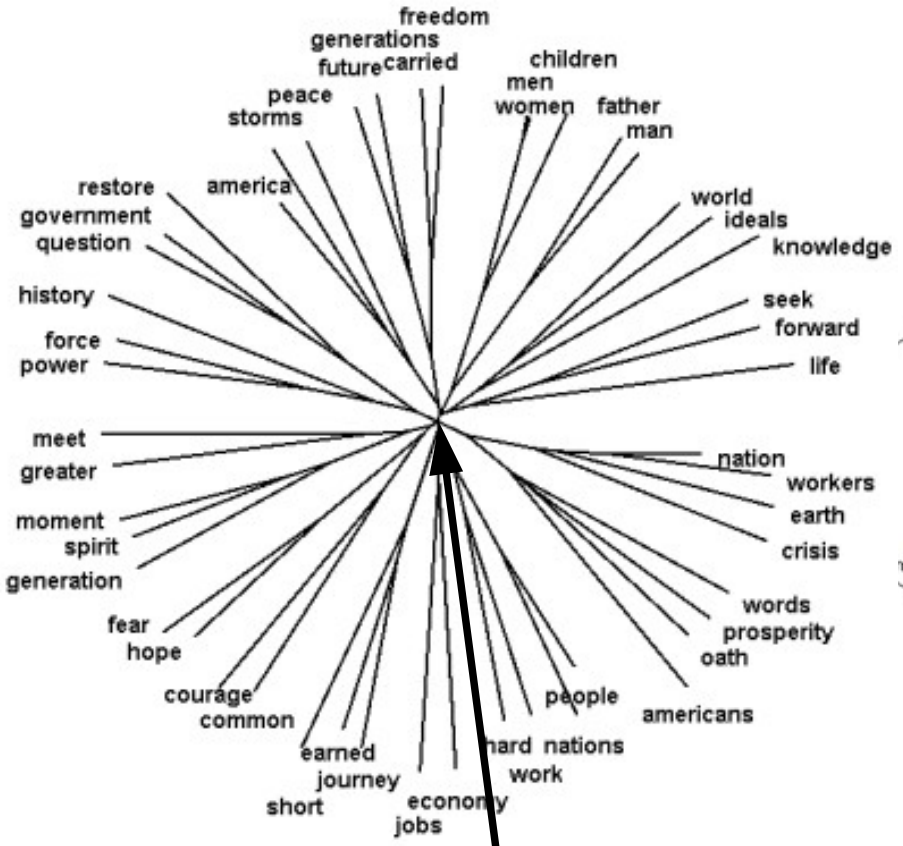
Rapprochement des mots d'un texte selon leur degré de cooccurrence dans le texte

Constellation du mot “nuit” dans la base EXEMPLEM
(22 romans de 11 auteurs, lemmatisés)

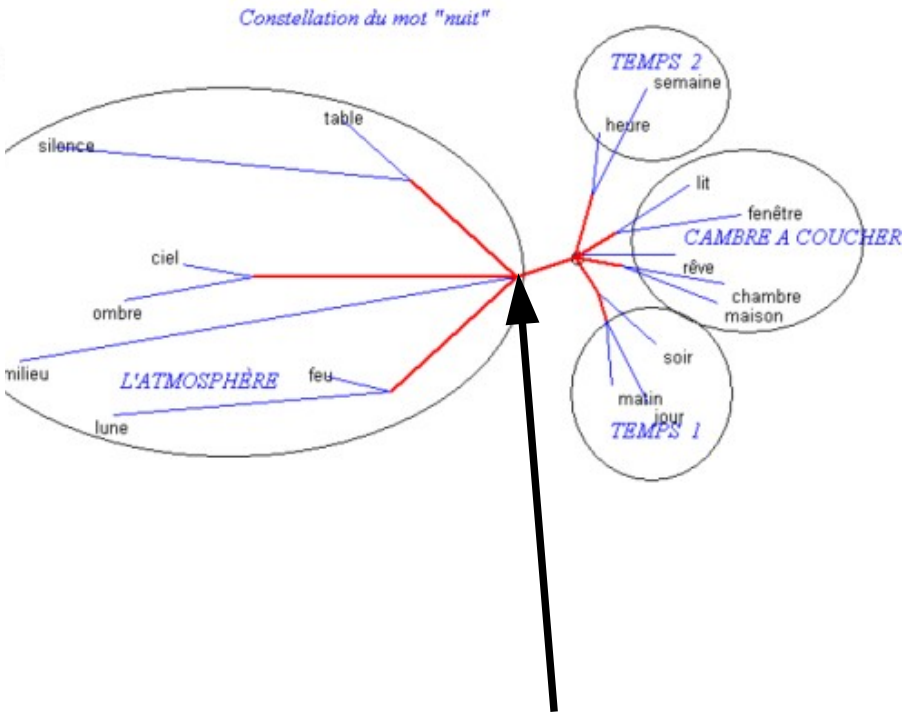


Barthelémy & Luong, *Statistique et Analyse des Données,* 1986
Brunet, *JADT'08*

Ultramétriques, centre d'un arbre



“centre” de l'arbre



arbre “sans centre” (feuilles à gauche plus éloignées de ce point que celles à droit)