Séminaires du pôle R&D d'Adoc Talent Management
10/09/2014 – Paris

Visualisation en nuages arborés pour l'analyse de textes et de données RH

Philippe Gambette

LIGM
Université Paris-Est
Marne-la-Vallée

1. Quelques utilisations des nuages arborés

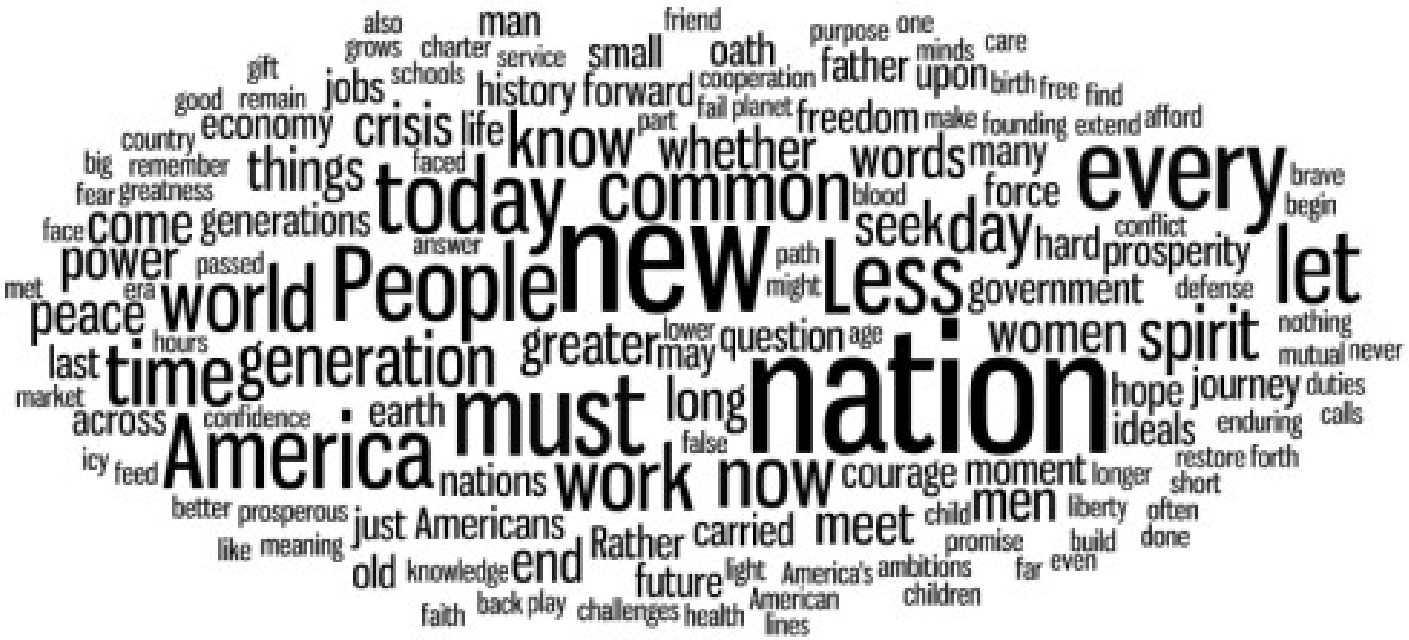
2. Construction des nuages arborés

3. Qualité des nuages arborés

1.
Quelques utilisations des nuages arborés

Le « nuage arboré », une information double

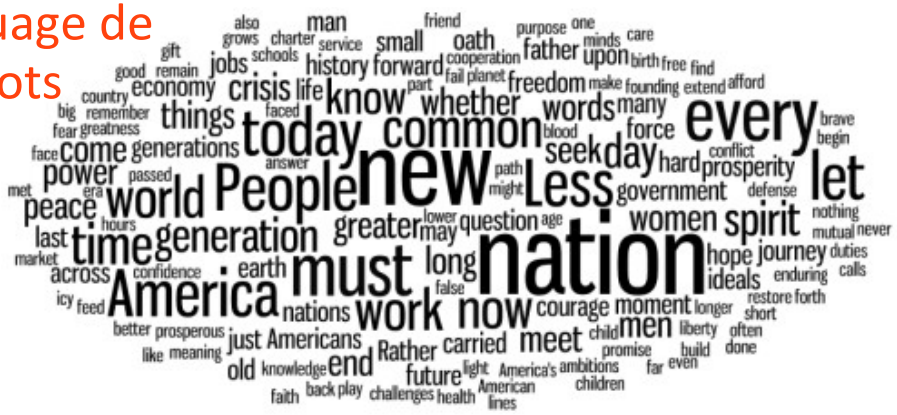
nuage de mots



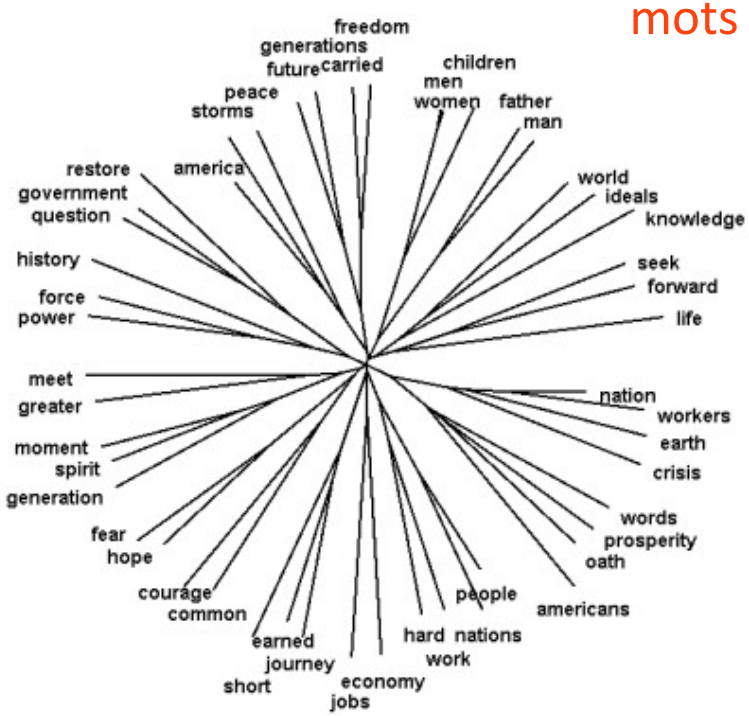
Discours inaugural de Barack Obama en 2008, Wordle

Le « nuage arboré », une information double

nuage de mots



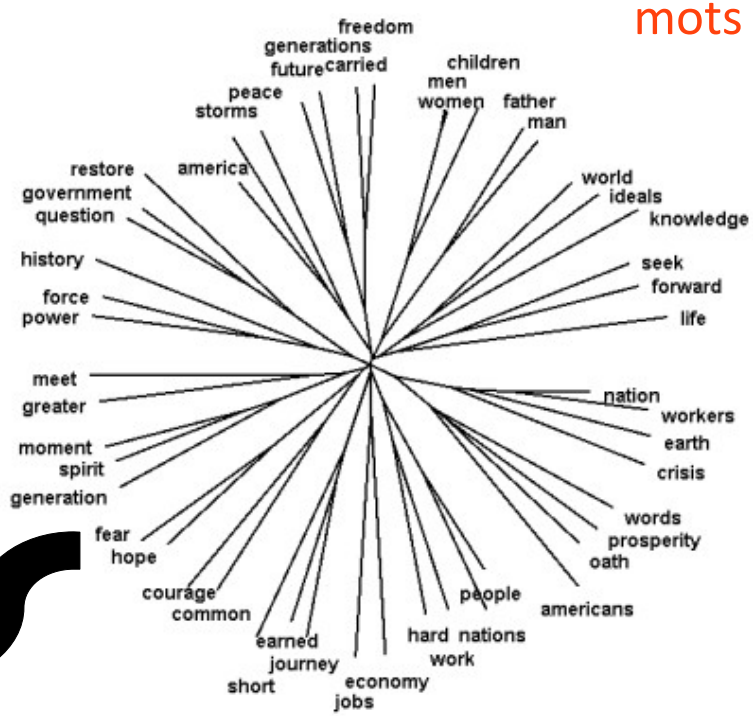
arbre de mots

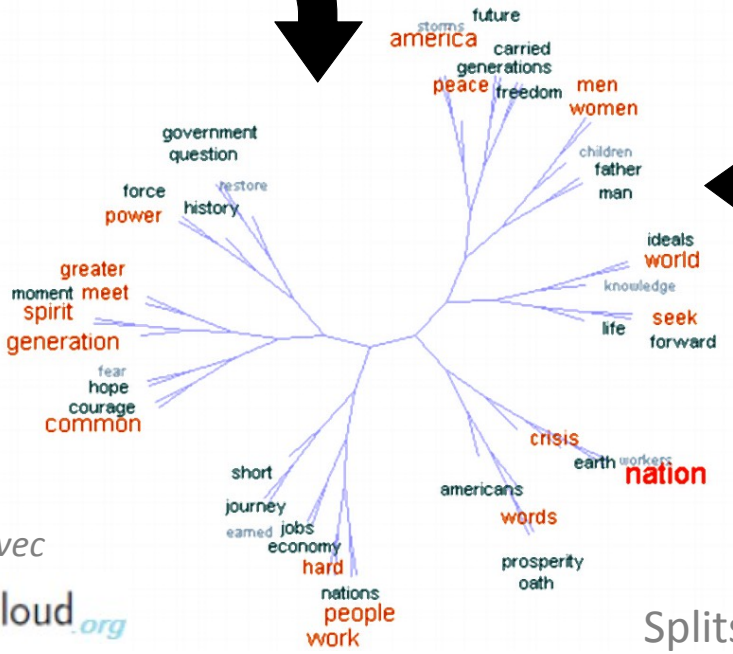


Le « nuage arboré », une information double

nuage de mots

arbre de mots





Discours inaugural de Barack Obama

construit avec

TreeCloud.org

SplitsTree4

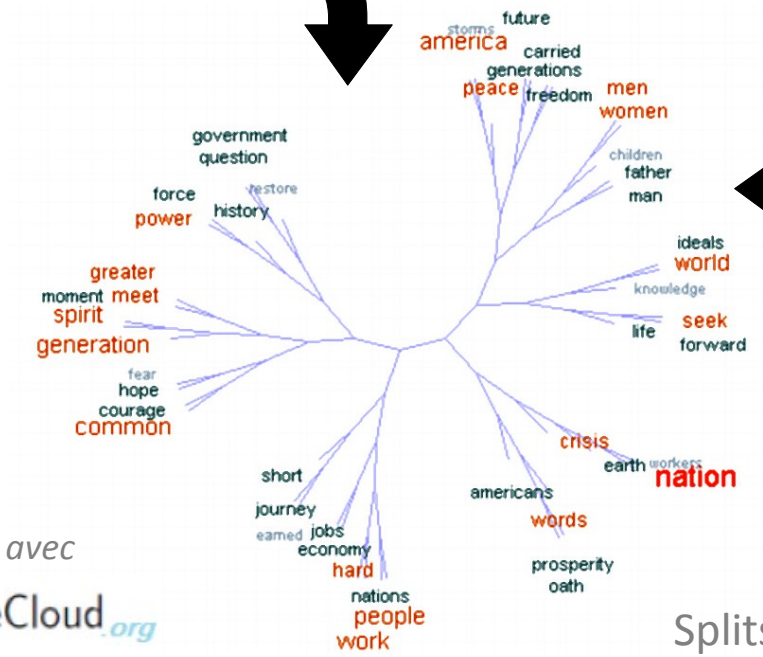
SplitsTree : Huson & Bryant, *Bioinformatics*, 2006

TreeCloud : Gambette & Véronis, *IFCS'09*

Le « nuage arboré », une information double

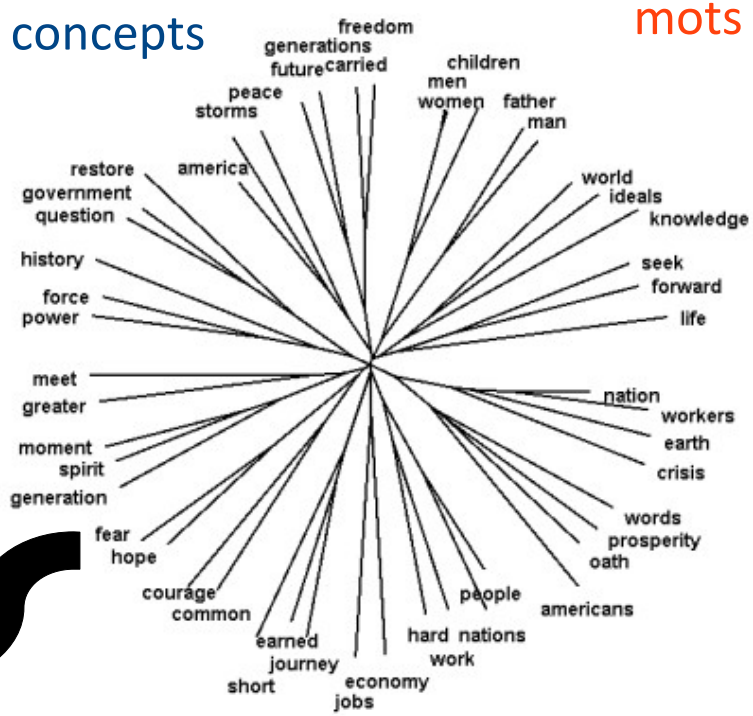
nuage de mots

hiérarchie des mots



hiérarchie des concepts

arbre de mots



Discours inaugural de Barack Obama

construit avec

TreeCloud.org

SplitsTree4

SplitsTree : Huson & Bryant, *Bioinformatics*, 2006

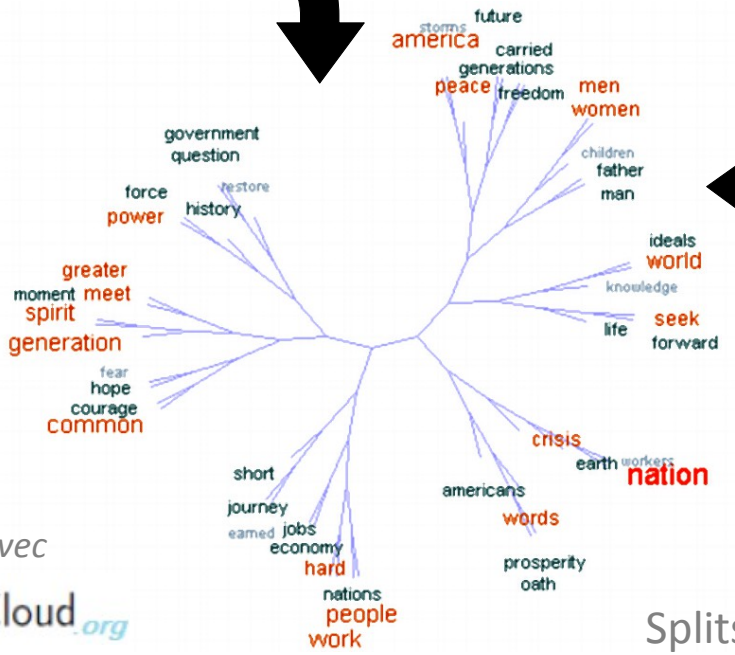
TreeCloud : Gambette & Véronis, *IFCS'09*

Le « nuage arboré », une information double

nuage de mots

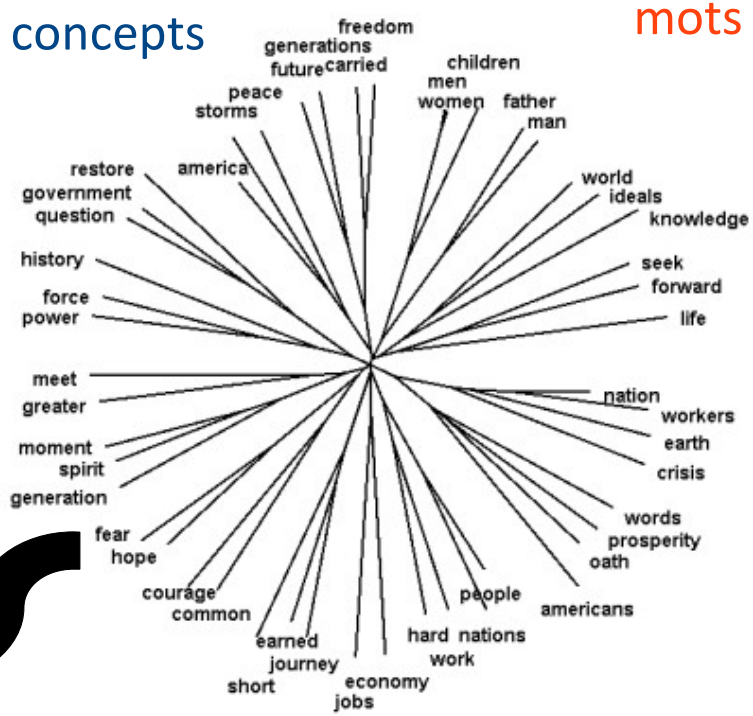
hiérarchie des mots

occurrences



hiérarchie des concepts

arbre de mots



cooccurrences

Discours inaugural de Barack Obama

construit avec

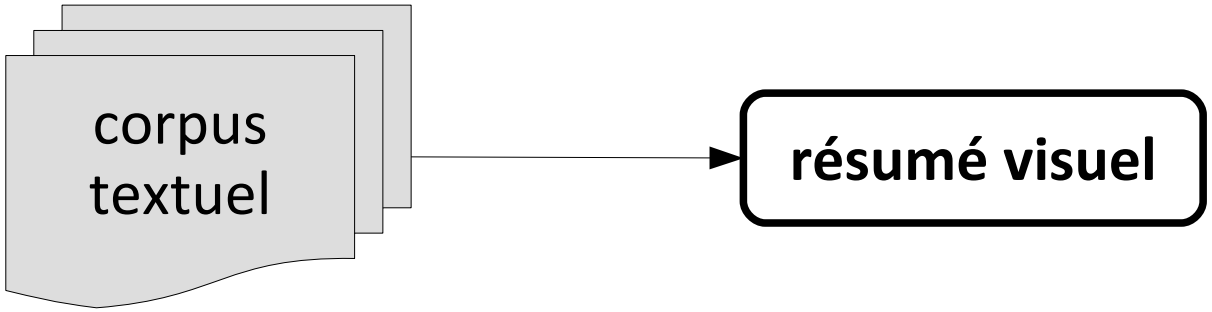
TreeCloud.org

SplitsTree4

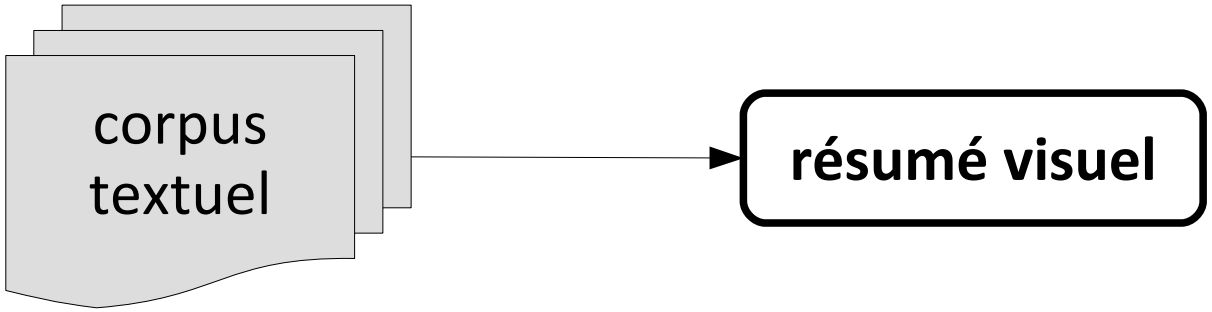
SplitsTree : Huson & Bryant, *Bioinformatics*, 2006

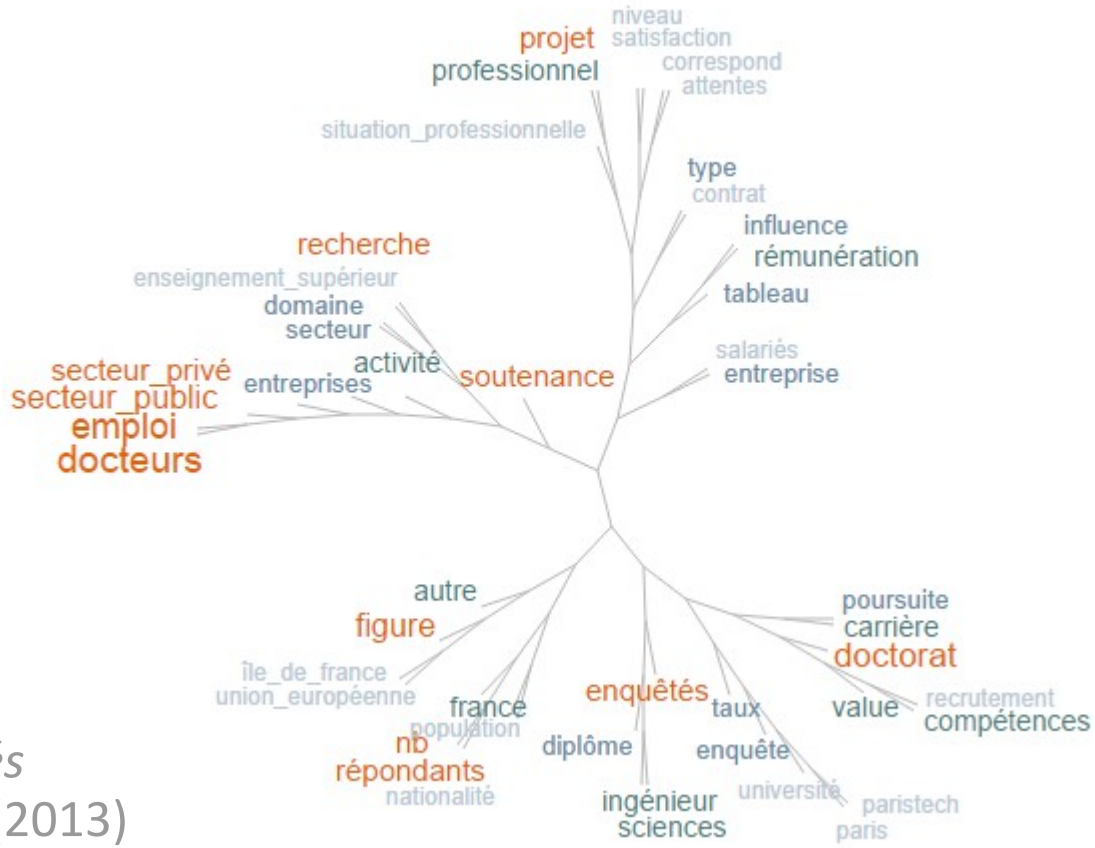
TreeCloud : Gambette & Véronis, *IFCS'09*

Le « nuage arboré », pour quoi faire ?



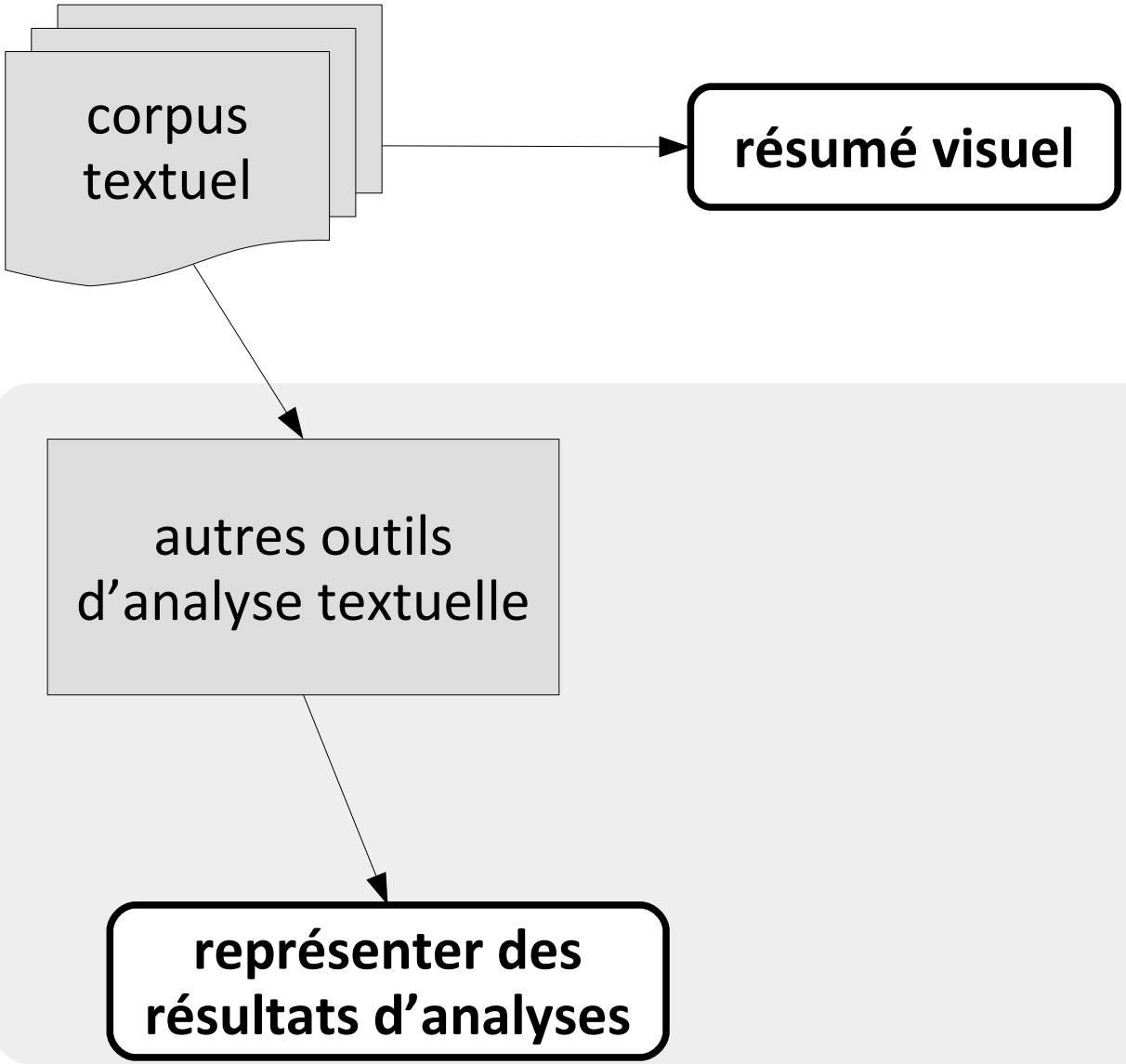
Le « nuage arboré », pour quoi faire ?





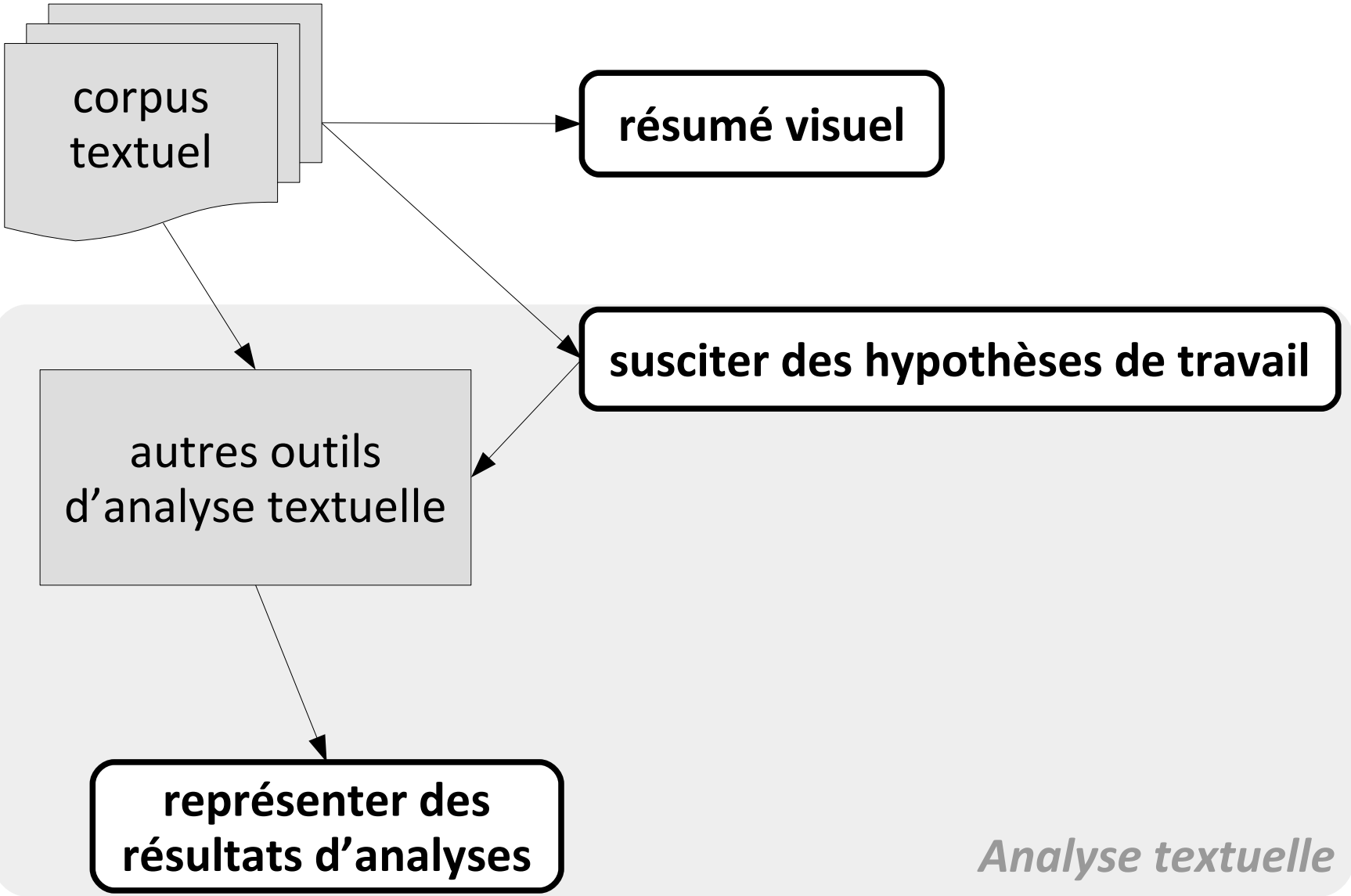
Nuage arboré construit sur Treecloud.org du rapport *La poursuite de carrière des docteurs récemment diplômés* d'Adoc Talent Management (2013)

Le « nuage arboré », pour quoi faire ?



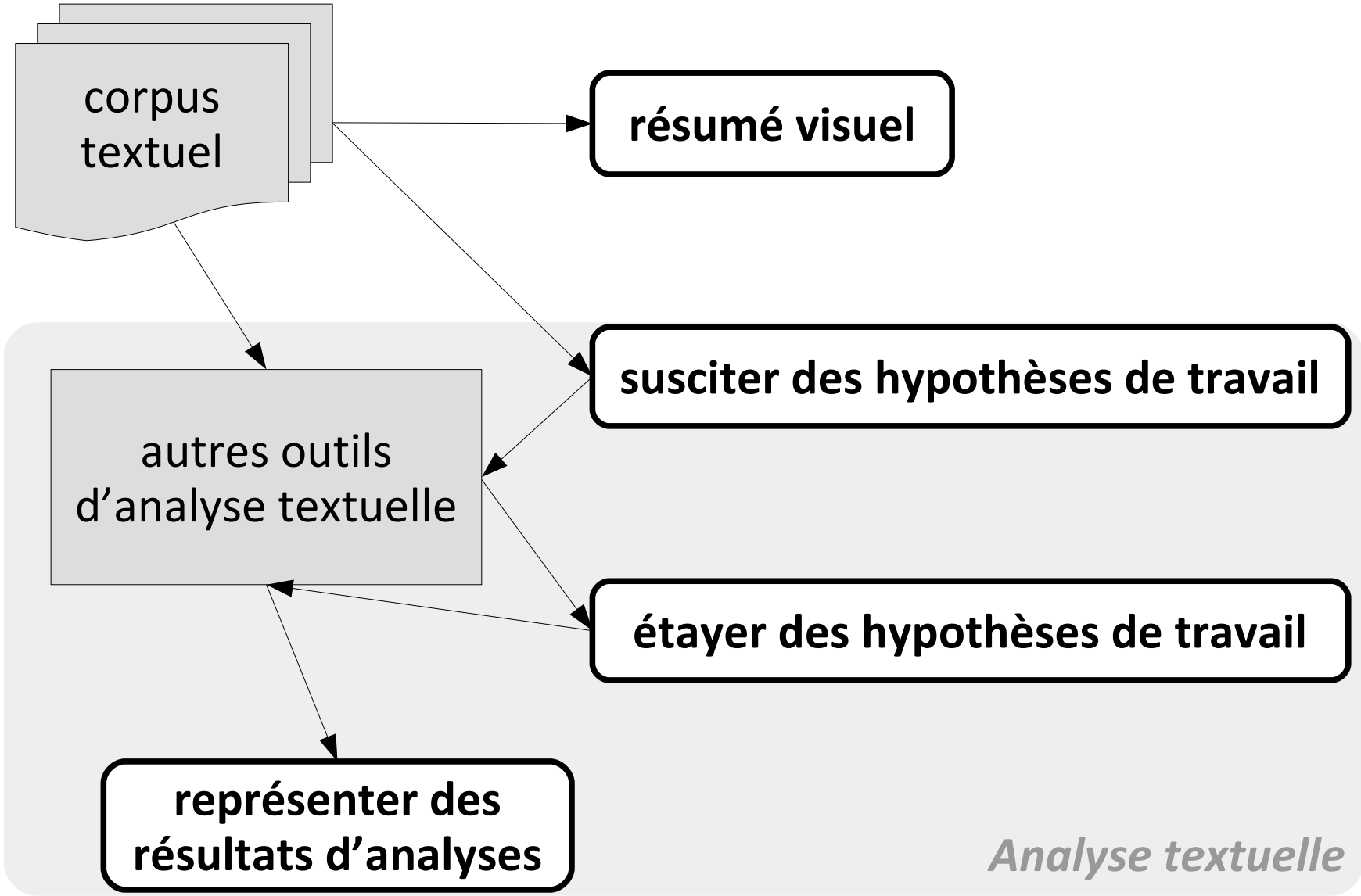
Amstutz & Gambette, *JADT 2010*

Le « nuage arboré », pour quoi faire ?



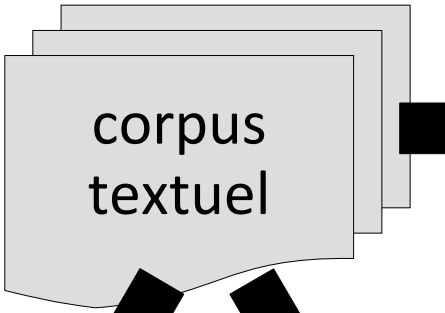
Amstutz & Gambette, *JADT 2010*

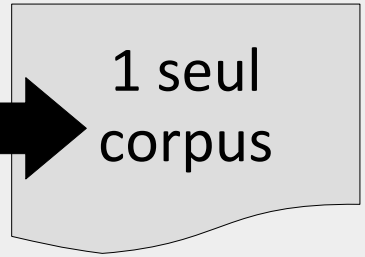
Le « nuage arboré », pour quoi faire ?



Amstutz & Gambette, *JADT 2010*

Exploration de corpus avec TreeCloud

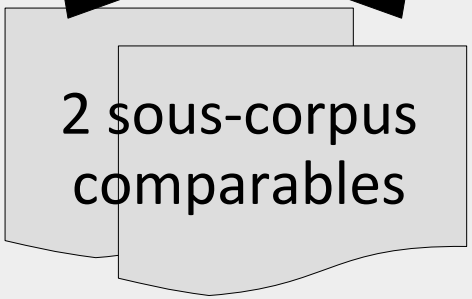




- Interpréter les regroupements
- Examiner les voisinages des mots fréquents dans l'arbre
- Examiner le voisinage de mots ciblés dans l'arbre («acteurs», verbes, etc.)

Si le corpus est assez grand :

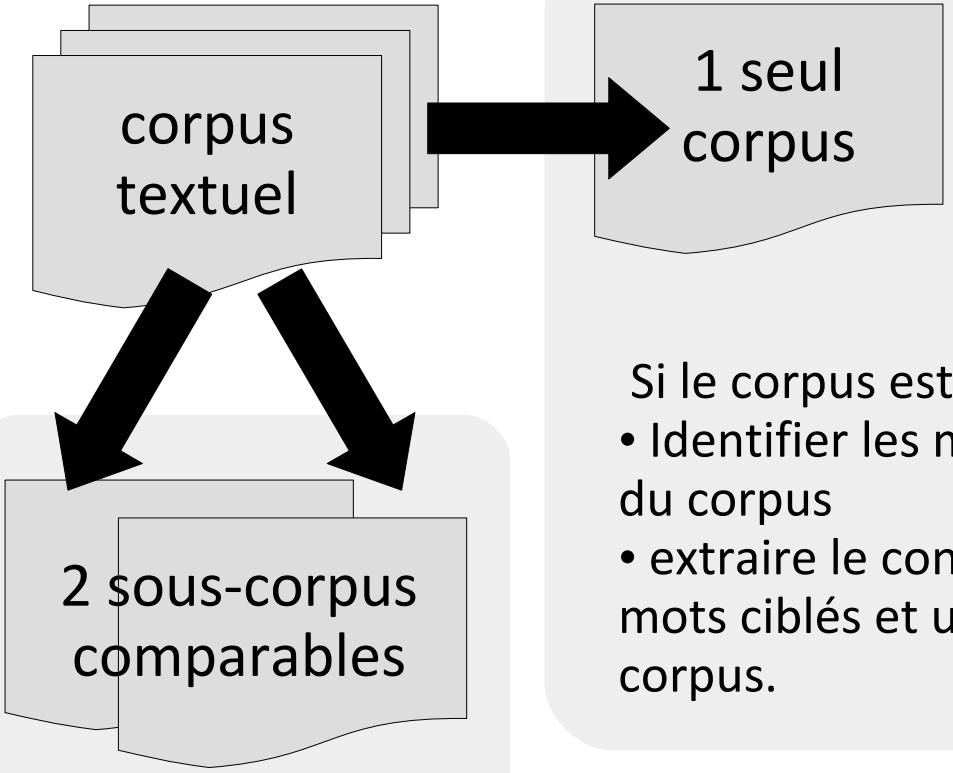
- Identifier les mots plutôt au début ou à la fin du corpus
- extraire le contexte dans le corpus de certains mots ciblés et utiliser ces données comme corpus.



- Identifier des regroupements communs ou différents
- Comparer la fréquence des mots fréquents dans les deux arbres
- Comparer (quanti+quali) les regroupements dans les deux arbres
- Construire le nuage arboré des mots spécifiques de chaque sous-corpus

Travail en cours avec Edna Hernandez

Exploration de corpus avec TreeCloud



- **Interpréter les regroupements**
- Examiner les voisinages des mots fréquents dans l'arbre
- Examiner le voisinage de mots ciblés dans l'arbre («acteurs», verbes, etc.)

Si le corpus est assez grand :

- Identifier les mots plutôt au début ou à la fin du corpus
- extraire le contexte dans le corpus de certains mots ciblés et utiliser ces données comme corpus.

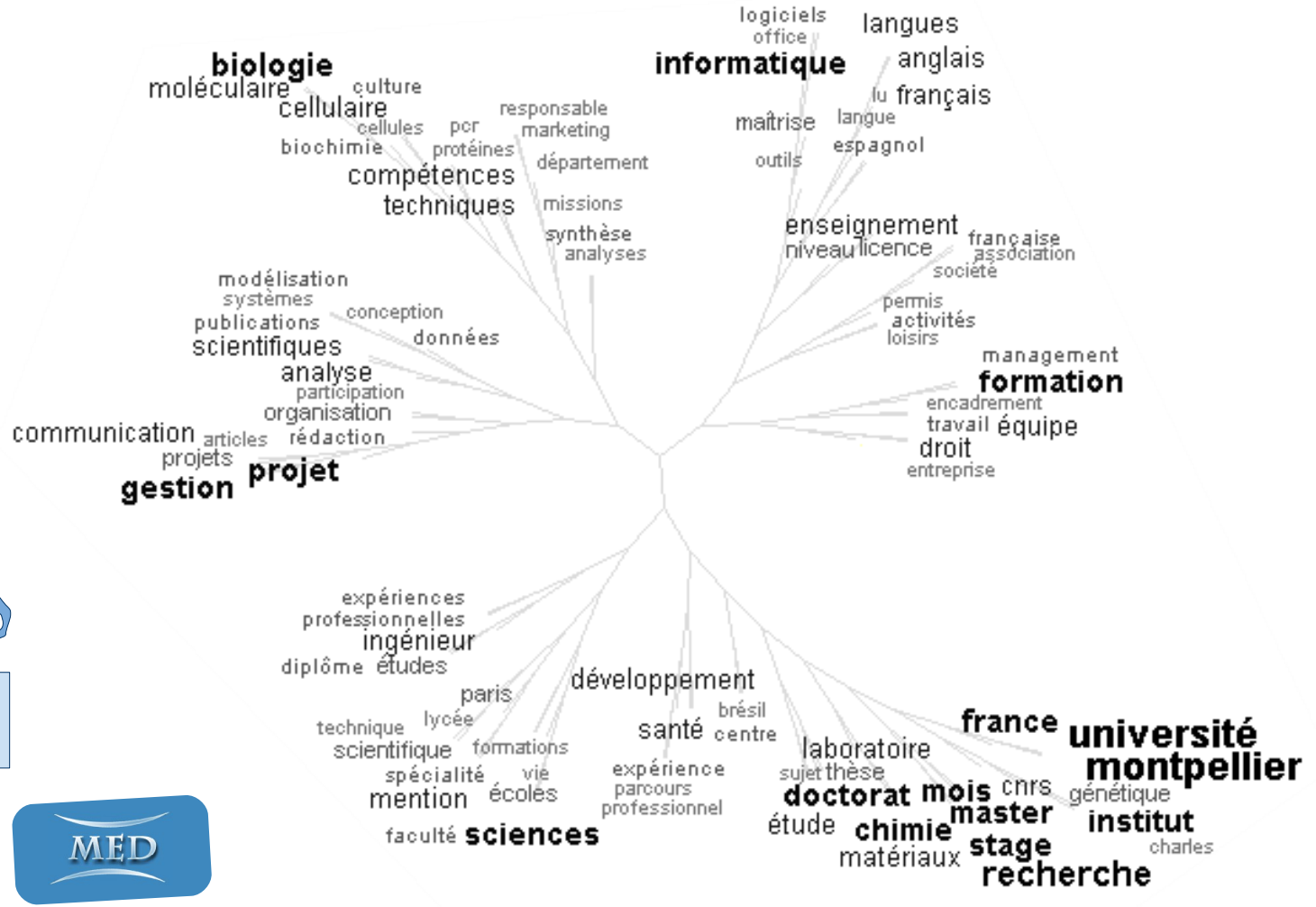
- Identifier des regroupements communs ou différents
- Comparer la fréquence des mots fréquents dans les deux arbres
- Comparer (quanti+quali) les regroupements dans les deux arbres
- Construire le nuage arboré des mots spécifiques de chaque sous-corpus

Travail en cours avec Edna Hernandez

Méthode : interpréter les regroupements

Dessiner des « patates »

Corpus : une centaine de CV soumis à une rencontre docteurs-entreprises



Rencontre Docteurs & Entreprises

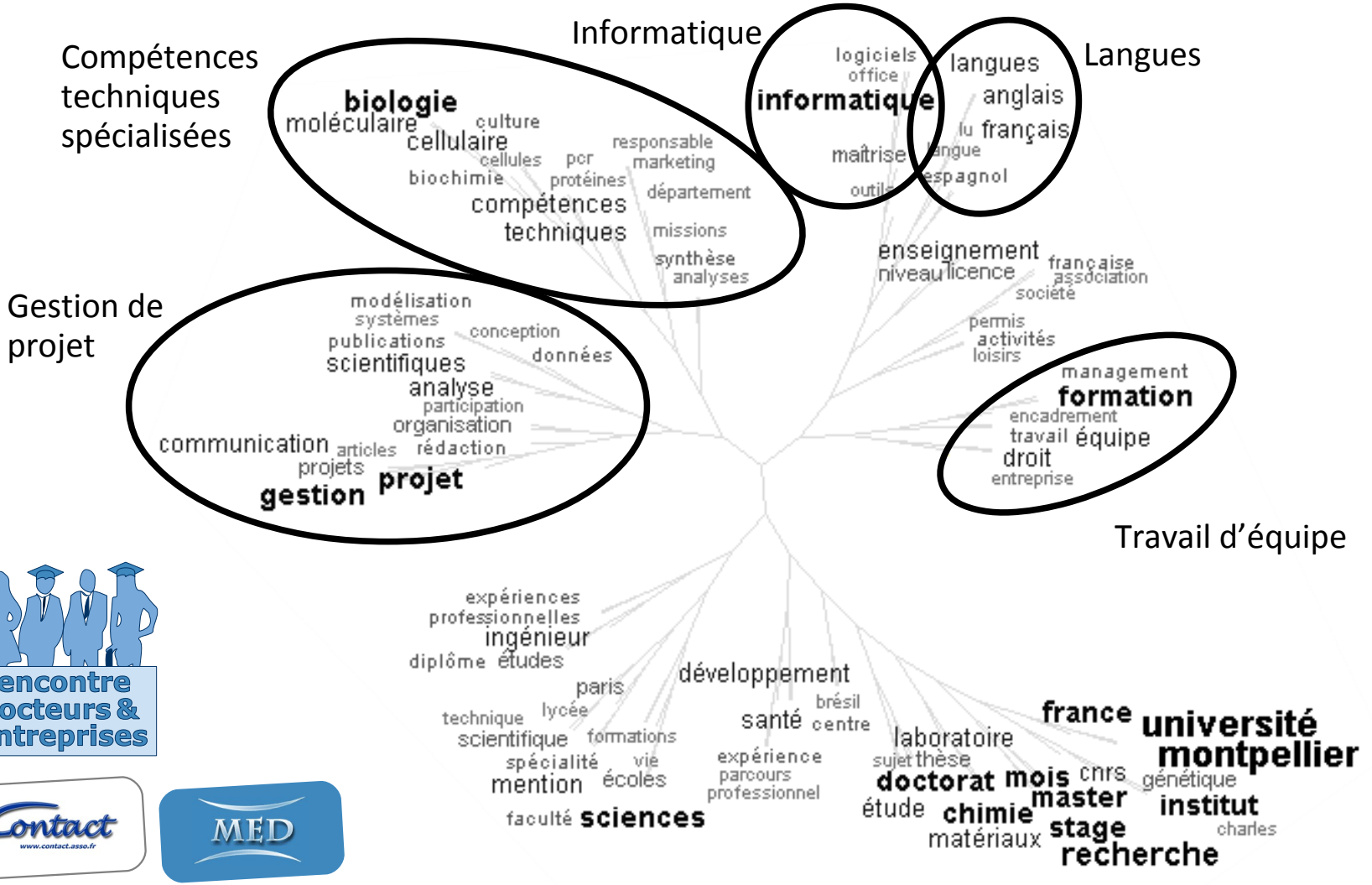
Contact
www.contact.asso.fr

MED

Méthode : interpréter les regroupements

Dessiner des « patates »

Corpus : une centaine de CV soumis à une rencontre docteurs-entreprises

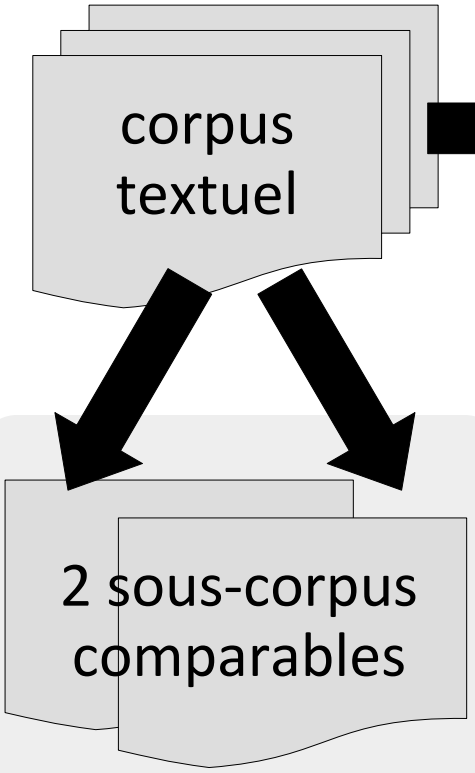


Rencontre Docteurs & Entreprises

Contact
www.contact.asso.fr

MED

Exploration de corpus avec TreeCloud



1 seul corpus

- Interpréter les regroupements
- **Examiner les voisinages des mots fréquents dans l'arbre**
- Examiner le voisinage de mots ciblés dans l'arbre («acteurs», verbes, etc.)

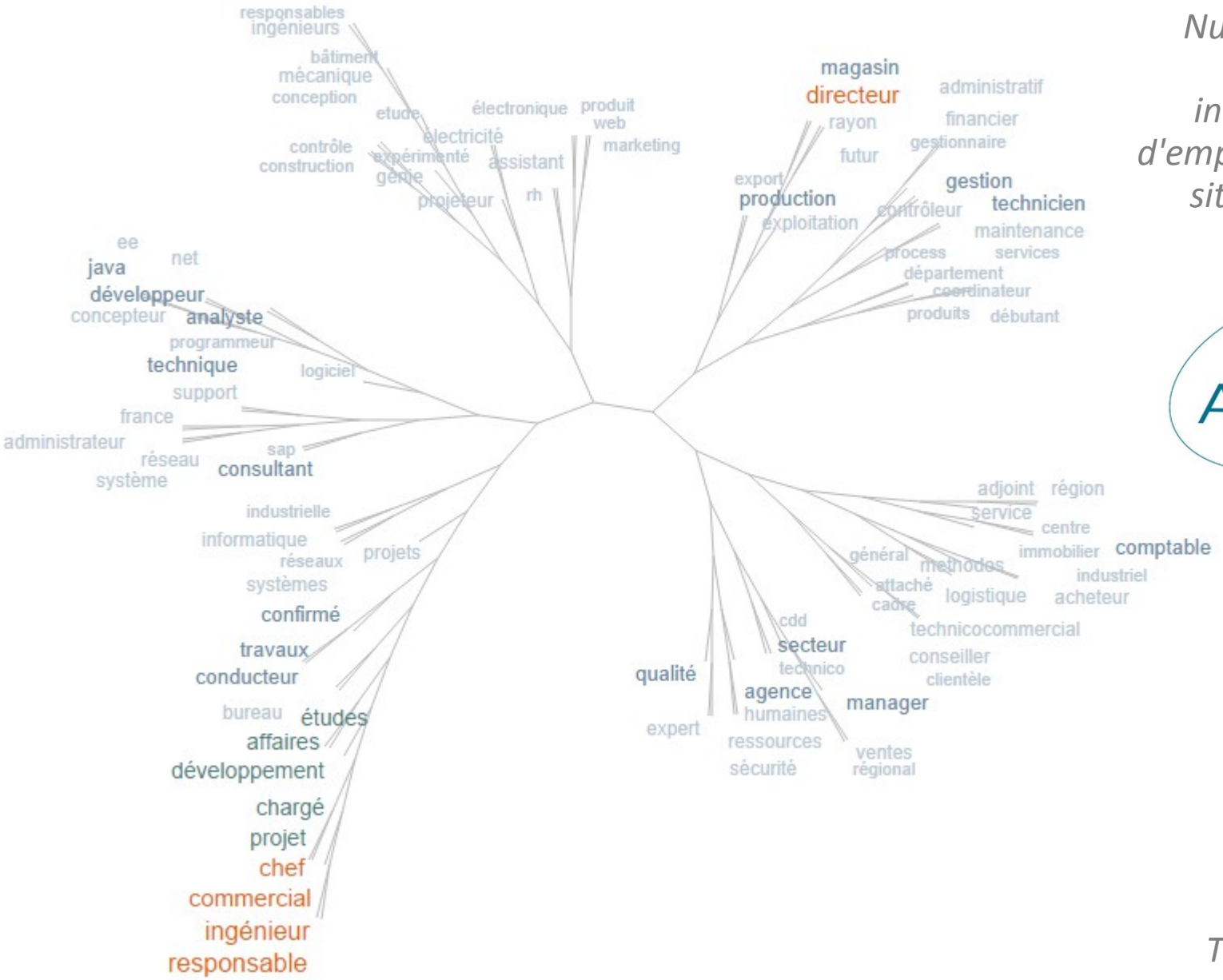
Si le corpus est assez grand :

- Identifier les mots plutôt au début ou à la fin du corpus
- extraire le contexte dans le corpus de certains mots ciblés et utiliser ces données comme corpus.

- Identifier des regroupements communs ou différents
- Comparer la fréquence des mots fréquents dans les deux arbres
- Comparer (quanti+quali) les regroupements dans les deux arbres
- Construire le nuage arboré des mots spécifiques de chaque sous-corpus

Travail en cours avec Edna Hernandez

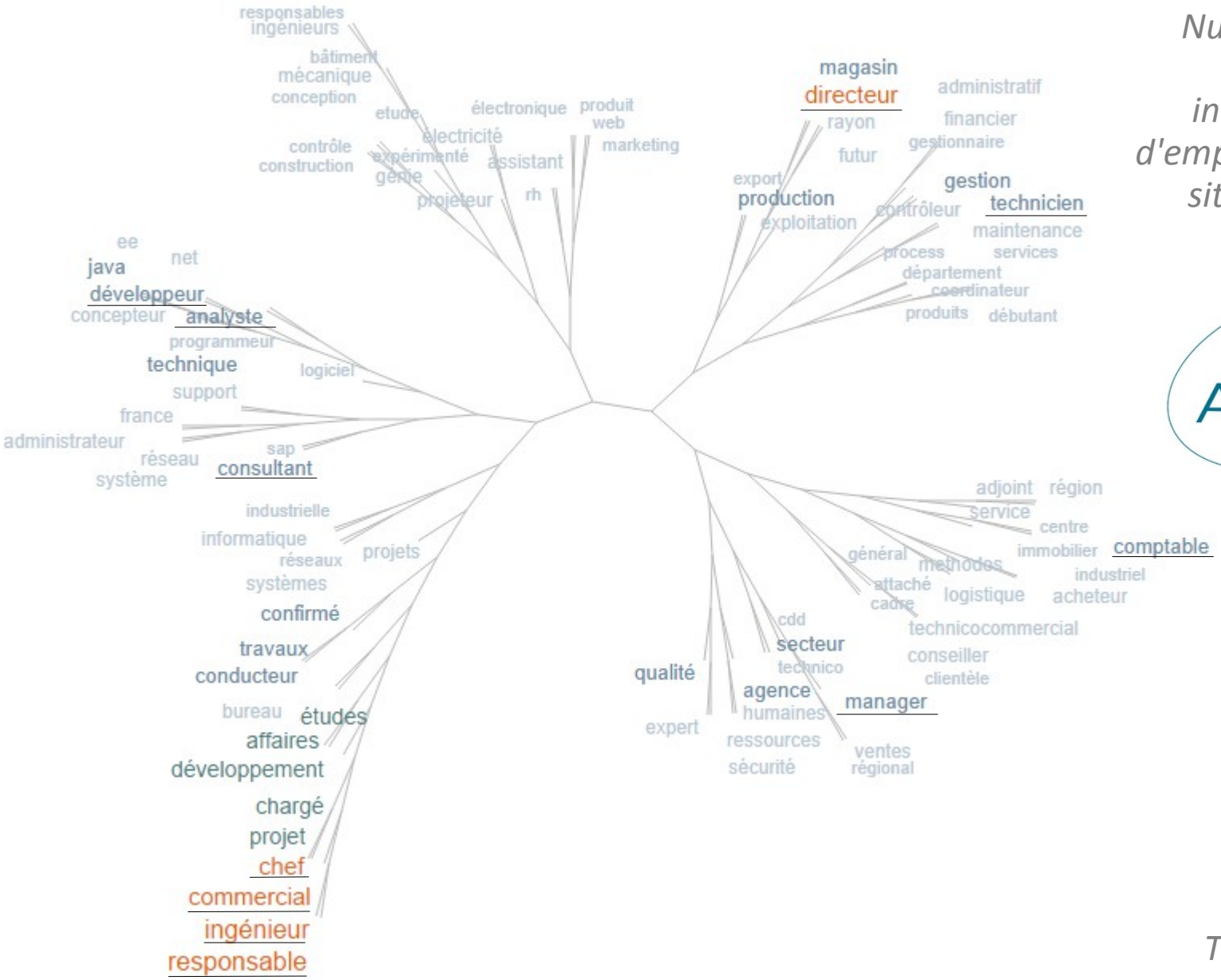
Méthode : voisinage des mots fréquents



Nuage arboré de plus de 4800 intitulés d'offres d'emploi extraits du site de l'APEC en avril 2011.

Travail de 2011 avec Paola Salle

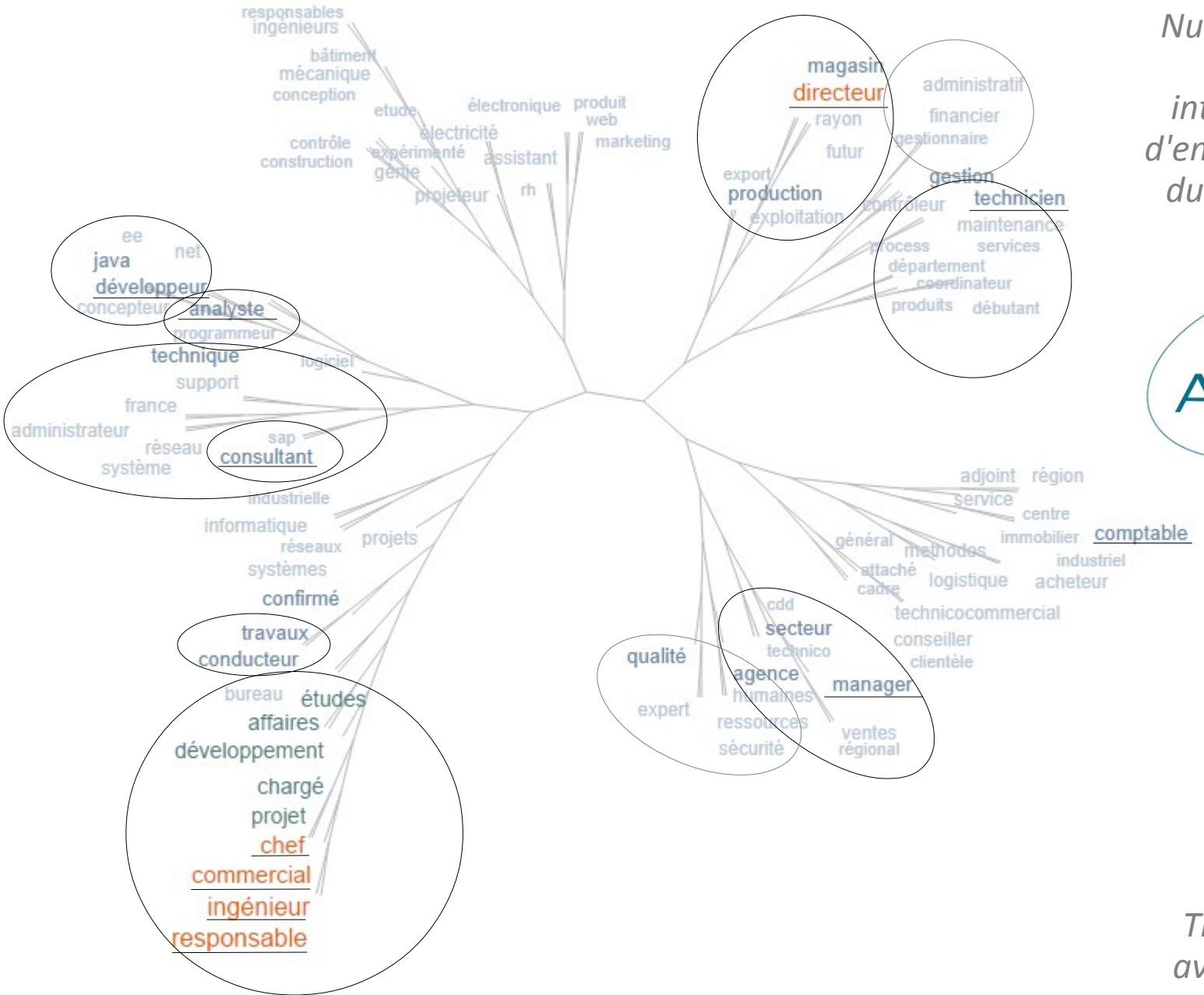
Méthode : voisinage des mots fréquents



Nuage arboré de plus de 4800 intitulés d'offres d'emploi extraits du site de l'APEC en avril 2011.

Travail de 2011 avec Paola Salle

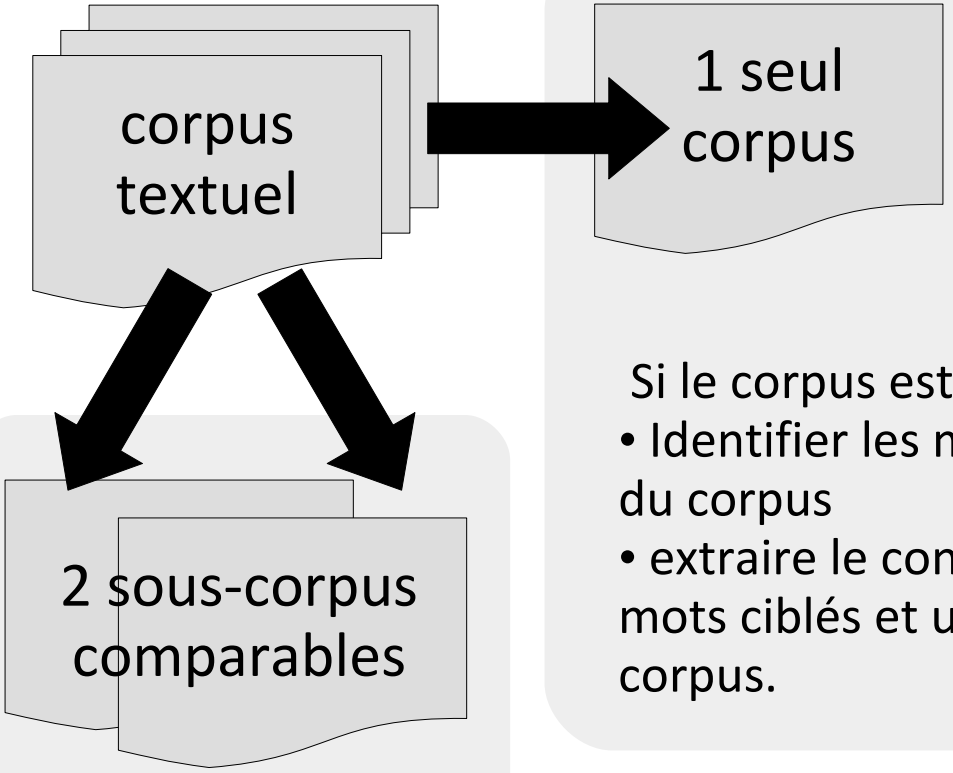
Méthode : voisinage des mots fréquents



Nuage arboré de plus de 4800 intitulés d'offres d'emploi extraites du site de l'APEC en avril 2011.

Travail de 2011 avec Paola Salle

Exploration de corpus avec TreeCloud



- Interpréter les regroupements
- Examiner les voisinages des mots fréquents dans l'arbre
- **Examiner le voisinage de mots ciblés dans l'arbre («acteurs», verbes, etc.)**

Si le corpus est assez grand :

- Identifier les mots plutôt au début ou à la fin du corpus
- extraire le contexte dans le corpus de certains mots ciblés et utiliser ces données comme corpus.

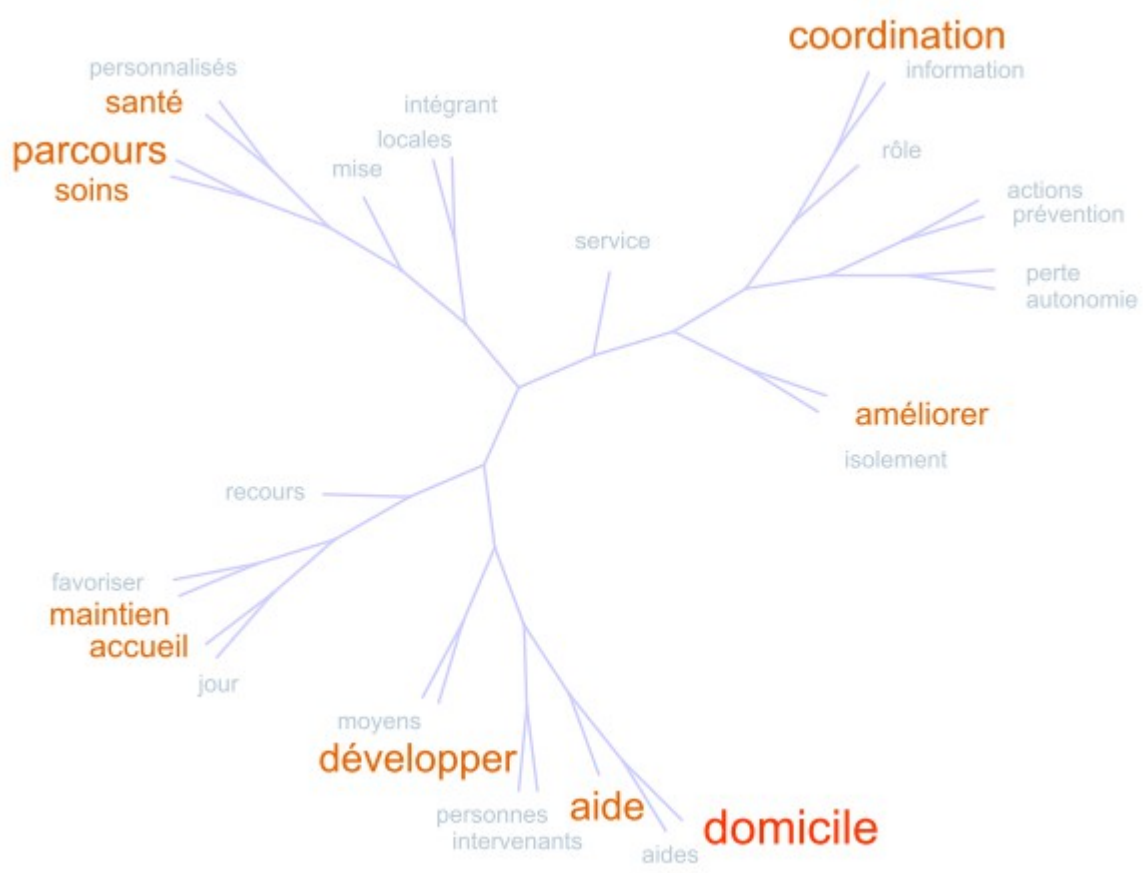
- Identifier des regroupements communs ou différents
- Comparer la fréquence des mots fréquents dans les deux arbres
- Comparer (quanti+quali) les regroupements dans les deux arbres
- Construire le nuage arboré des mots spécifiques de chaque sous-corpus

Travail en cours avec Edna Hernandez

Méthode : voisinage des verbes

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

Suggestions d'améliorations :

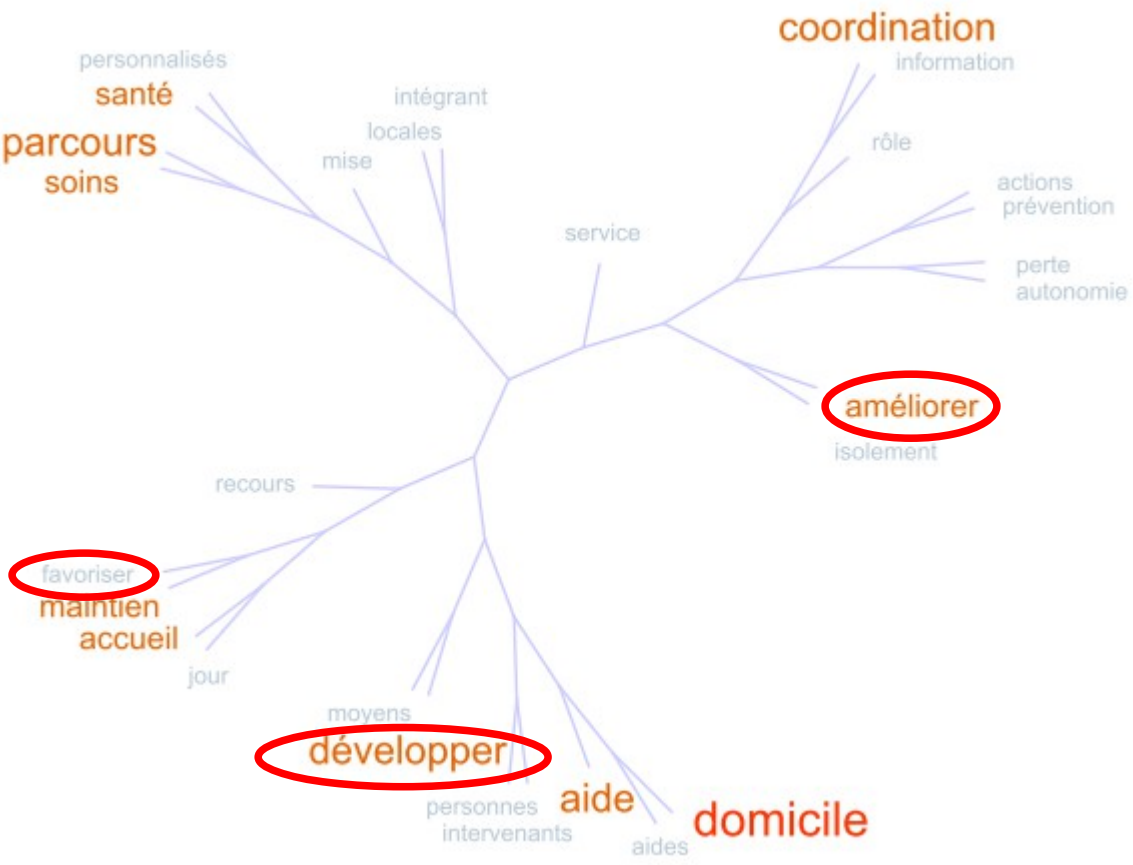


P. Grenier-Tisserand et
l'équipe projet de la DT 04

Méthode : voisinage des verbes

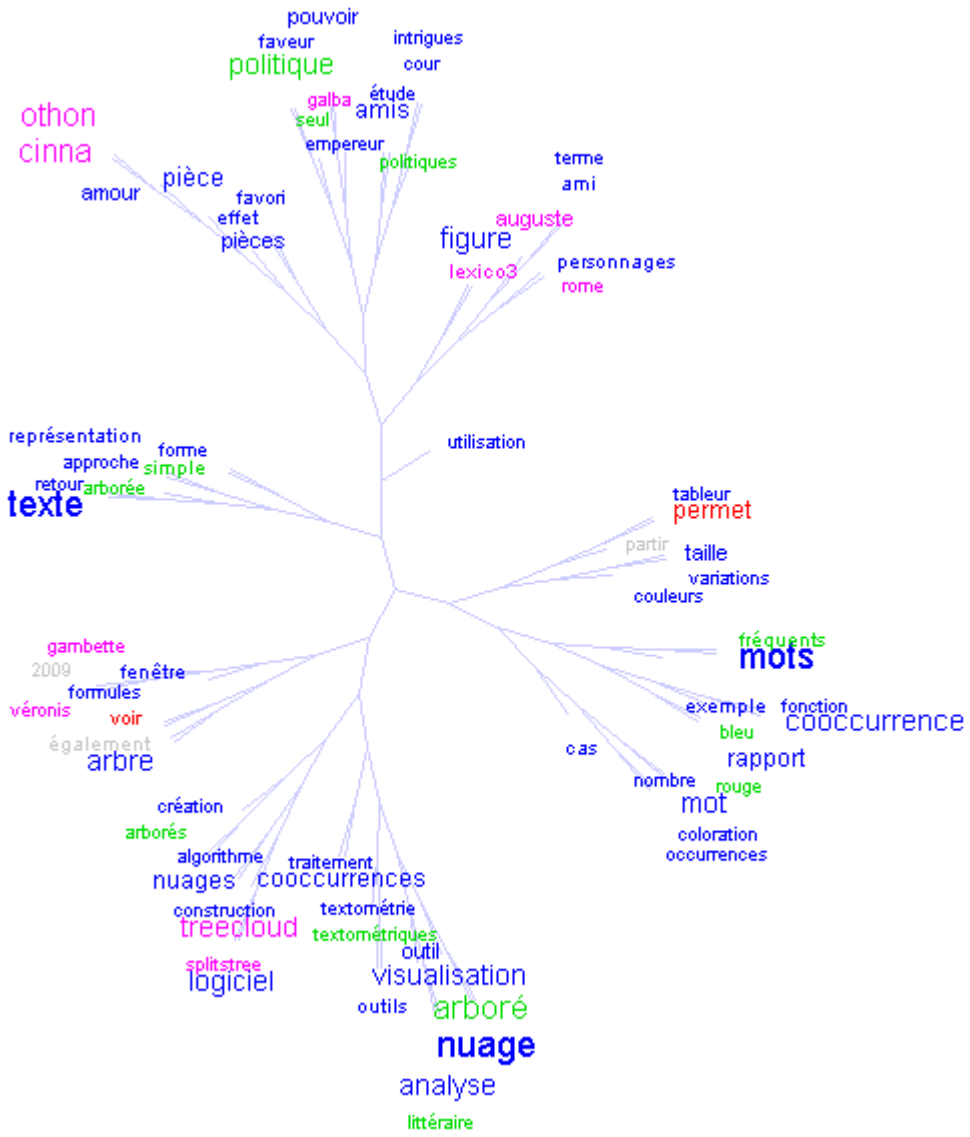
Corpus : réponses à des questions ouvertes à des professionnels de la santé sur le parcours de santé des personnes âgées dans les Alpes de Haute-Provence

Suggestions d'améliorations :



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

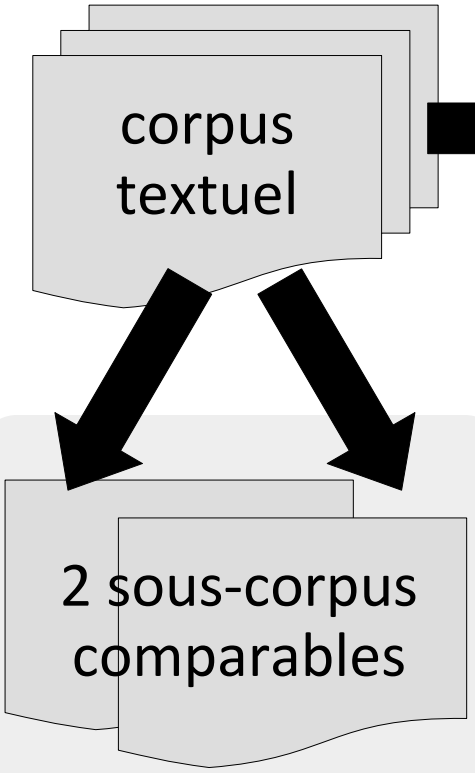
Perspective : coloration grammaticale



noms
adjectifs
verbes
noms propres

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

Exploration de corpus avec TreeCloud



1 seul corpus

- Interpréter les regroupements
- Examiner les voisinages des mots fréquents dans l'arbre
- Examiner le voisinage de mots ciblés dans l'arbre («acteurs», verbes, etc.)

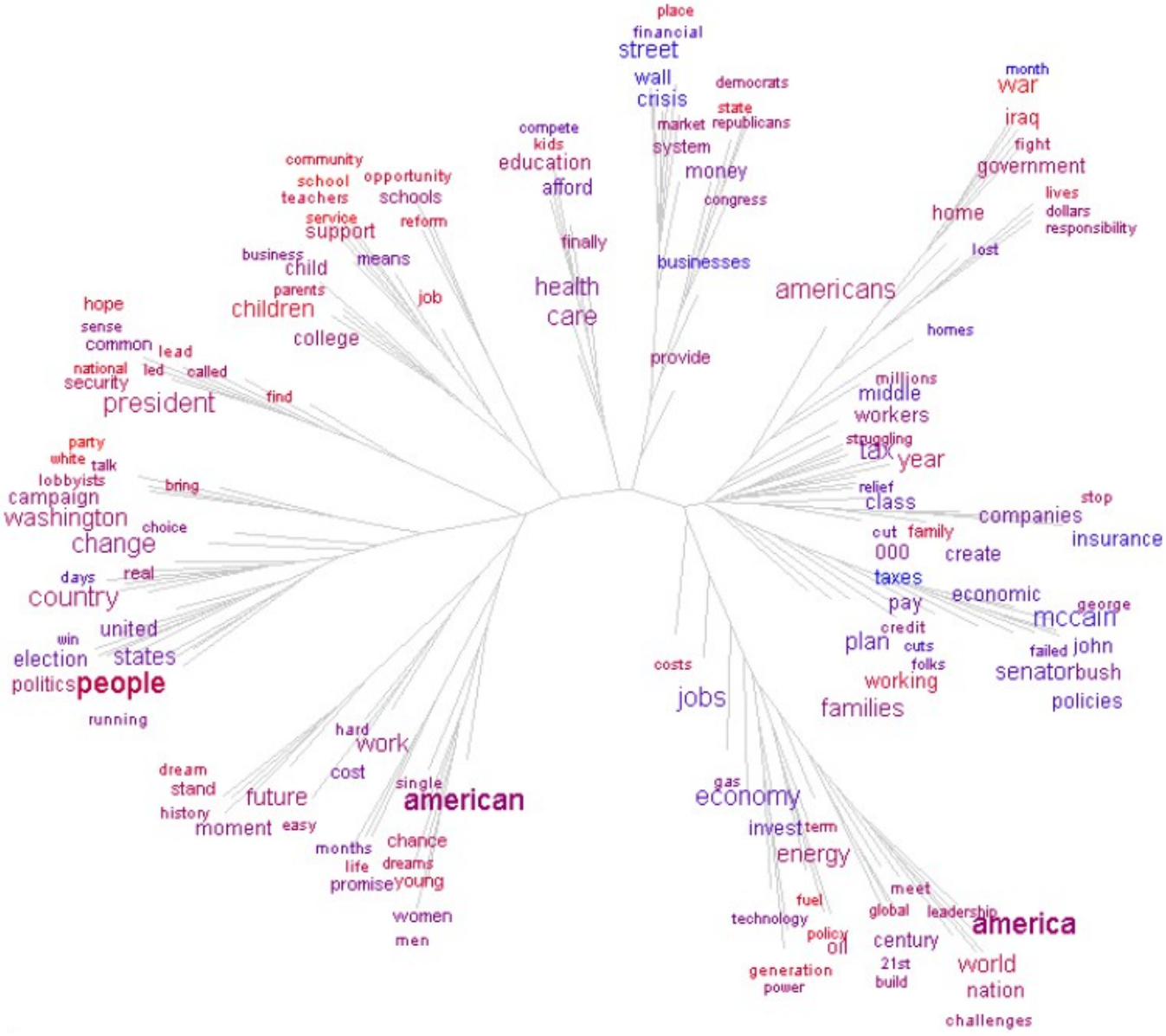
Si le corpus est assez grand :

- **Identifier les mots plutôt au début ou à la fin du corpus**
- extraire le contexte dans le corpus de certains mots ciblés et utiliser ces données comme corpus.

- Identifier des regroupements communs ou différents
- Comparer la fréquence des mots fréquents dans les deux arbres
- Comparer (quanti+quali) les regroupements dans les deux arbres
- Construire le nuage arboré des mots spécifiques de chaque sous-corpus

Travail en cours avec Edna Hernandez

Méthode : mots au début ou à la fin

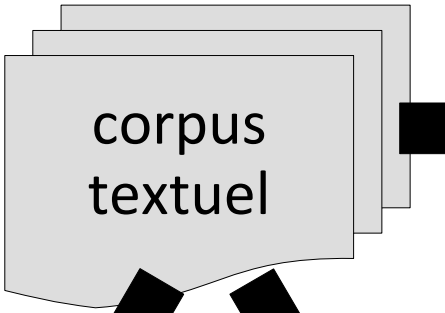


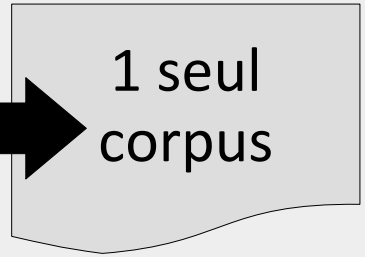
Nuage arboré de l'ensemble des discours de campagne de 2008 de Barack Obama, coloration chronologique

début de la campagne
fin de la campagne

Gambette & Véronis, IFCS 2009

Exploration de corpus avec TreeCloud

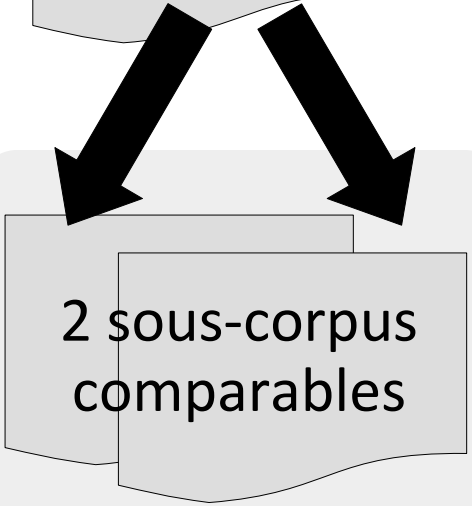




- Interpréter les regroupements
- Examiner les voisinages des mots fréquents dans l'arbre
- Examiner le voisinage de mots ciblés dans l'arbre («acteurs», verbes, etc.)

Si le corpus est assez grand :

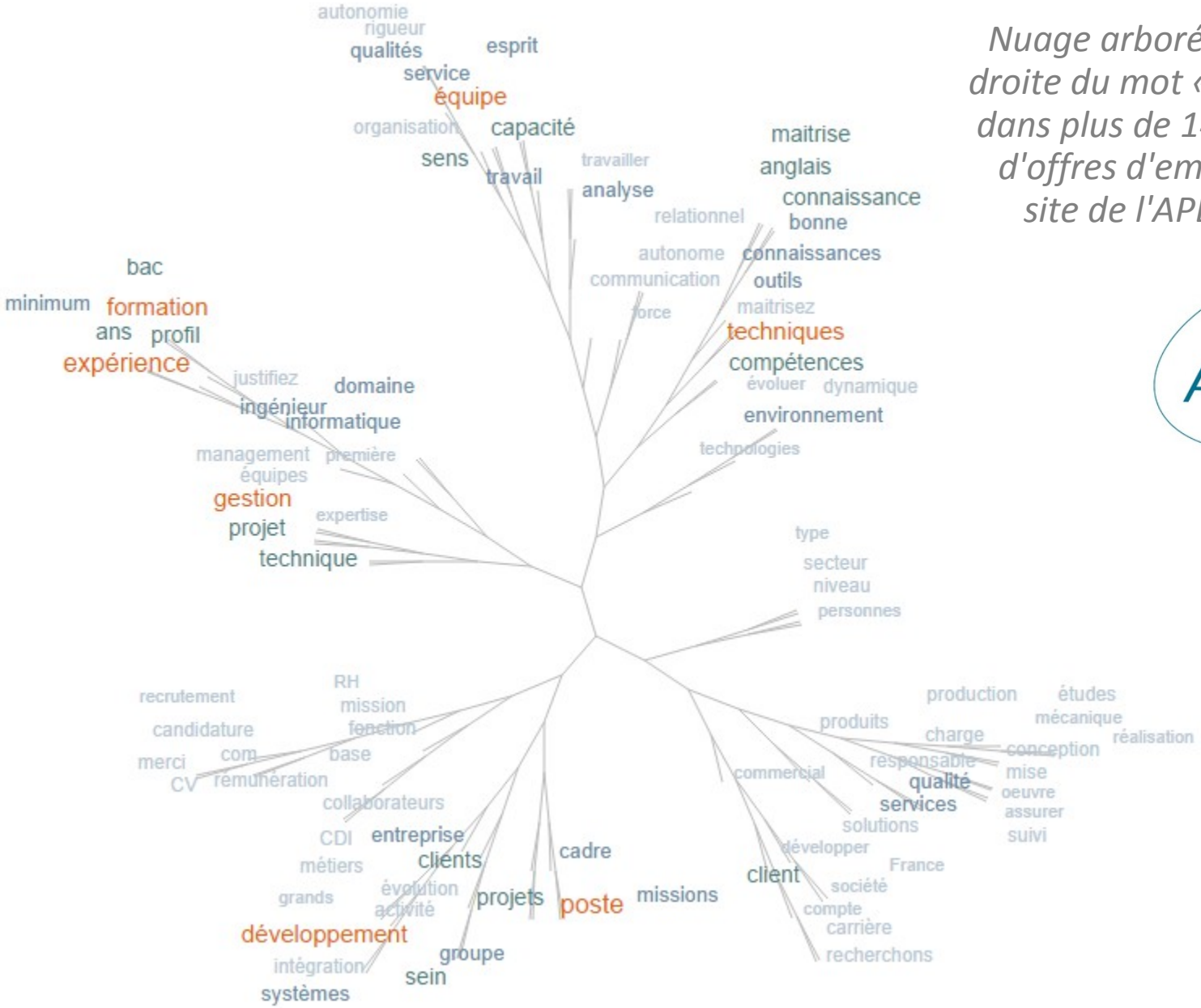
- Identifier les mots plutôt au début ou à la fin du corpus
- **extraire le contexte dans le corpus de certains mots ciblés et utiliser ces données comme corpus.**



- Identifier des regroupements communs ou différents
- Comparer la fréquence des mots fréquents dans les deux arbres
- Comparer (quanti+quali) les regroupements dans les deux arbres
- Construire le nuage arboré des mots spécifiques de chaque sous-corpus

Travail en cours avec Edna Hernandez

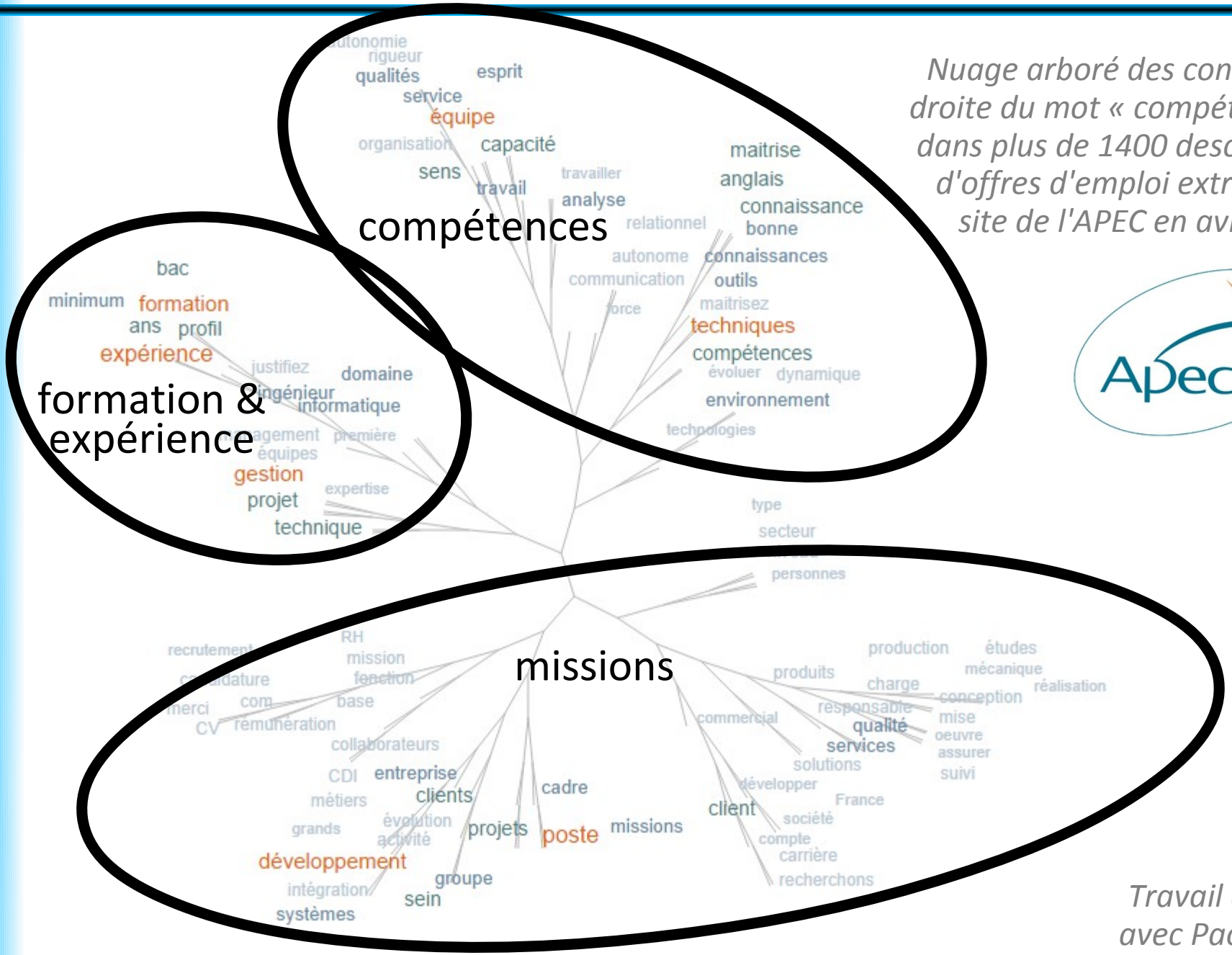
Contextes à droite de « compétences »



Nuage arboré des contextes à droite du mot « compétences » dans plus de 1400 descriptions d'offres d'emploi extraites du site de l'APEC en avril 2011.

Travail de 2011 avec Paola Salle

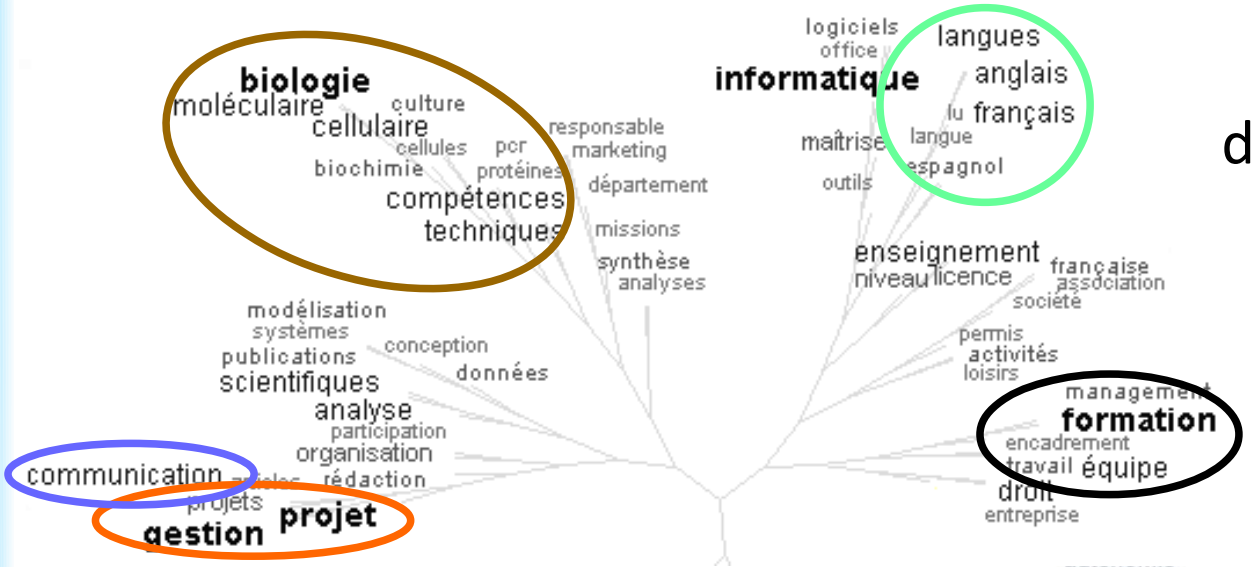
Contextes à droite de « compétences »



Nuage arboré des contextes à droite du mot « compétences » dans plus de 1400 descriptions d'offres d'emploi extraites du site de l'APEC en avril 2011.

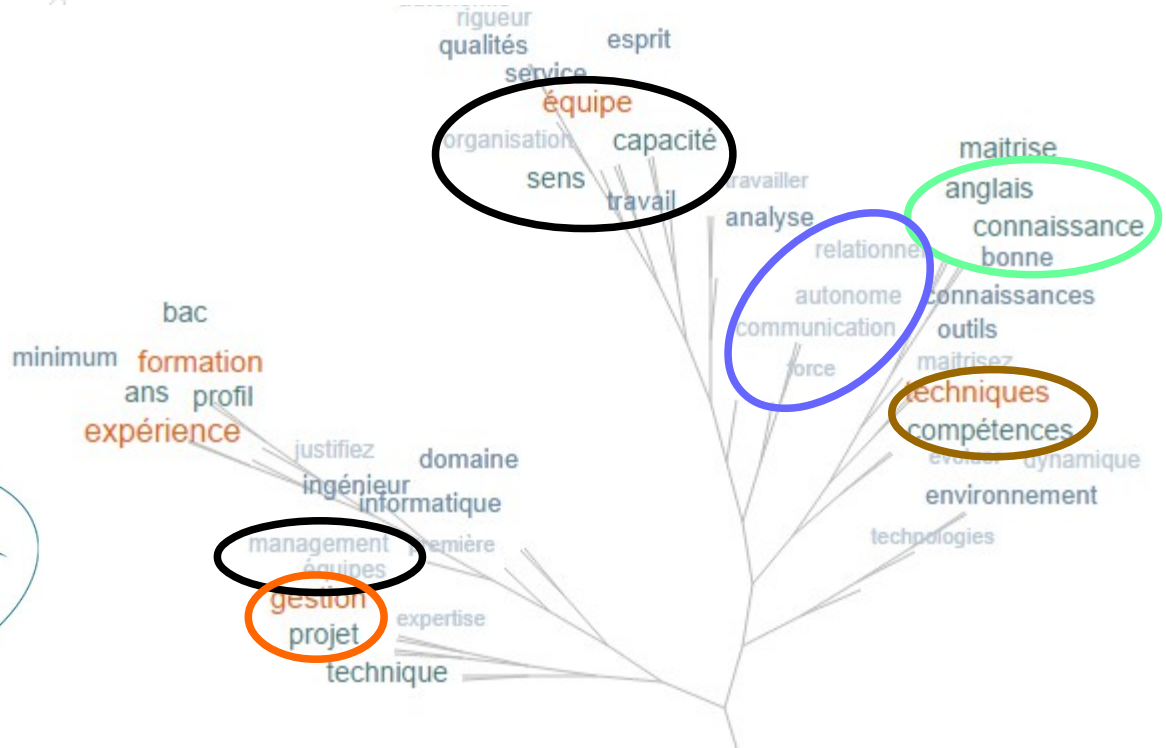
Travail de 2011 avec Paola Salle

Nuages arborés des compétences

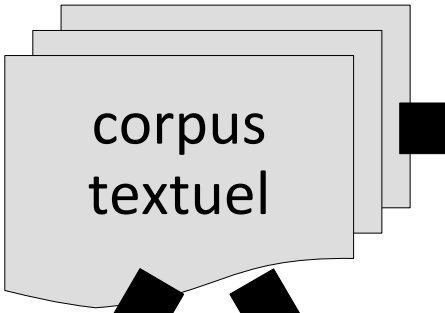


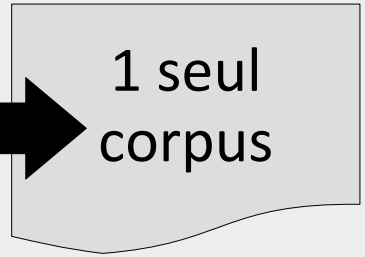
Déclarées par les docteurs dans leur CV

Demandées dans les offres d'emploi de l'APEC



Exploration de corpus avec TreeCloud

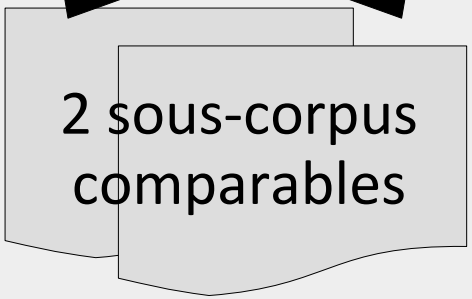




- Interpréter les regroupements
- Examiner les voisinages des mots fréquents dans l'arbre
- Examiner le voisinage de mots ciblés dans l'arbre («acteurs», verbes, etc.)

Si le corpus est assez grand :

- Identifier les mots plutôt au début ou à la fin du corpus
- extraire le voisinage dans le corpus de certains mots ciblés et utiliser ces données comme corpus.



- **Identifier des regroupements communs ou différents**
- Comparer la fréquence des mots fréquents dans les deux arbres
- Comparer (quanti+quali) les regroupements dans les deux arbres
- Construire le nuage arboré des mots spécifiques de chaque sous-corpus

Travail en cours avec Edna Hernandez

Illustration sur le corpus Mediator

Comparer les articles d'agences et articles de journalistes

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

Ensemble des articles

Gambette & Martinez, *Texto!*, 2013

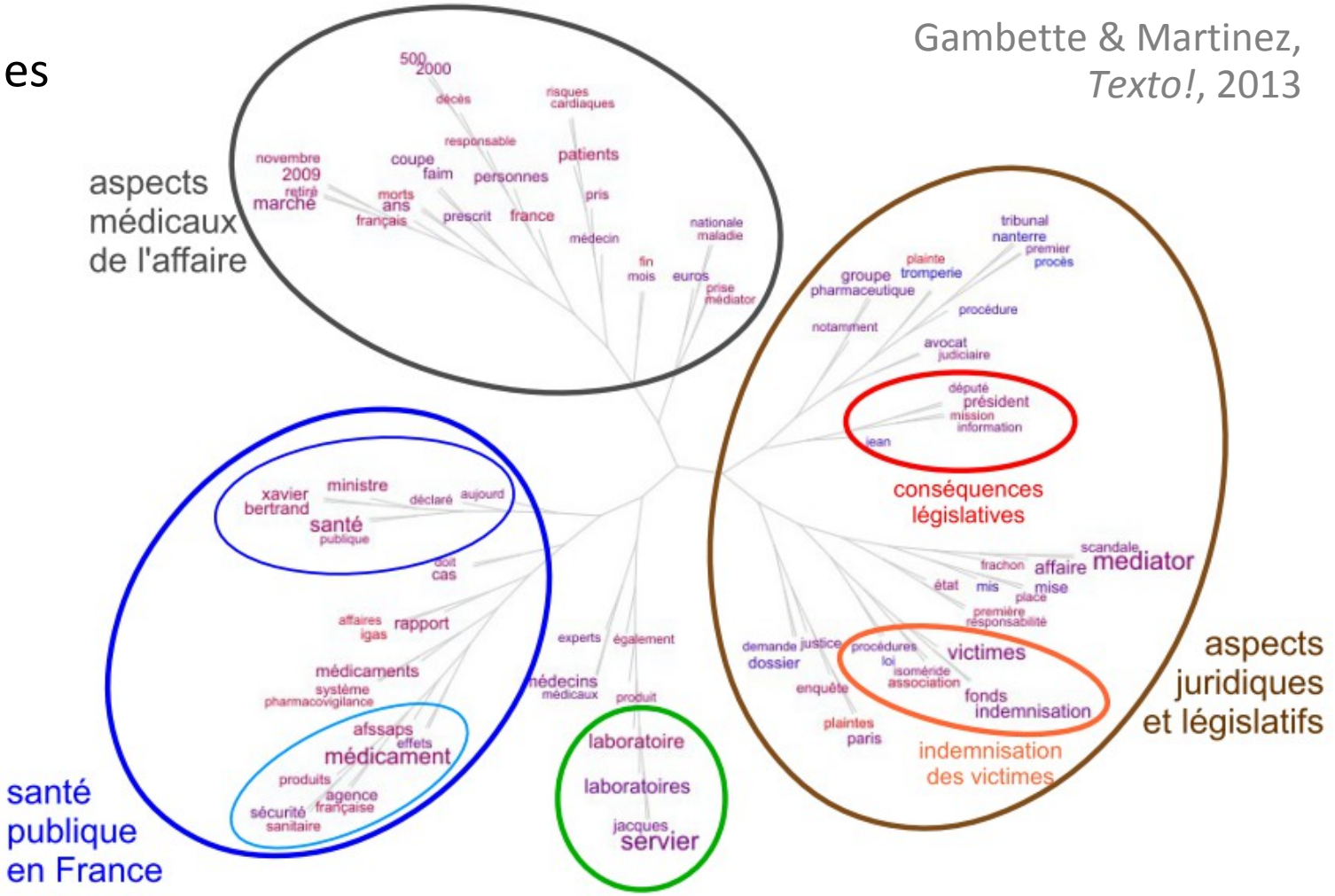
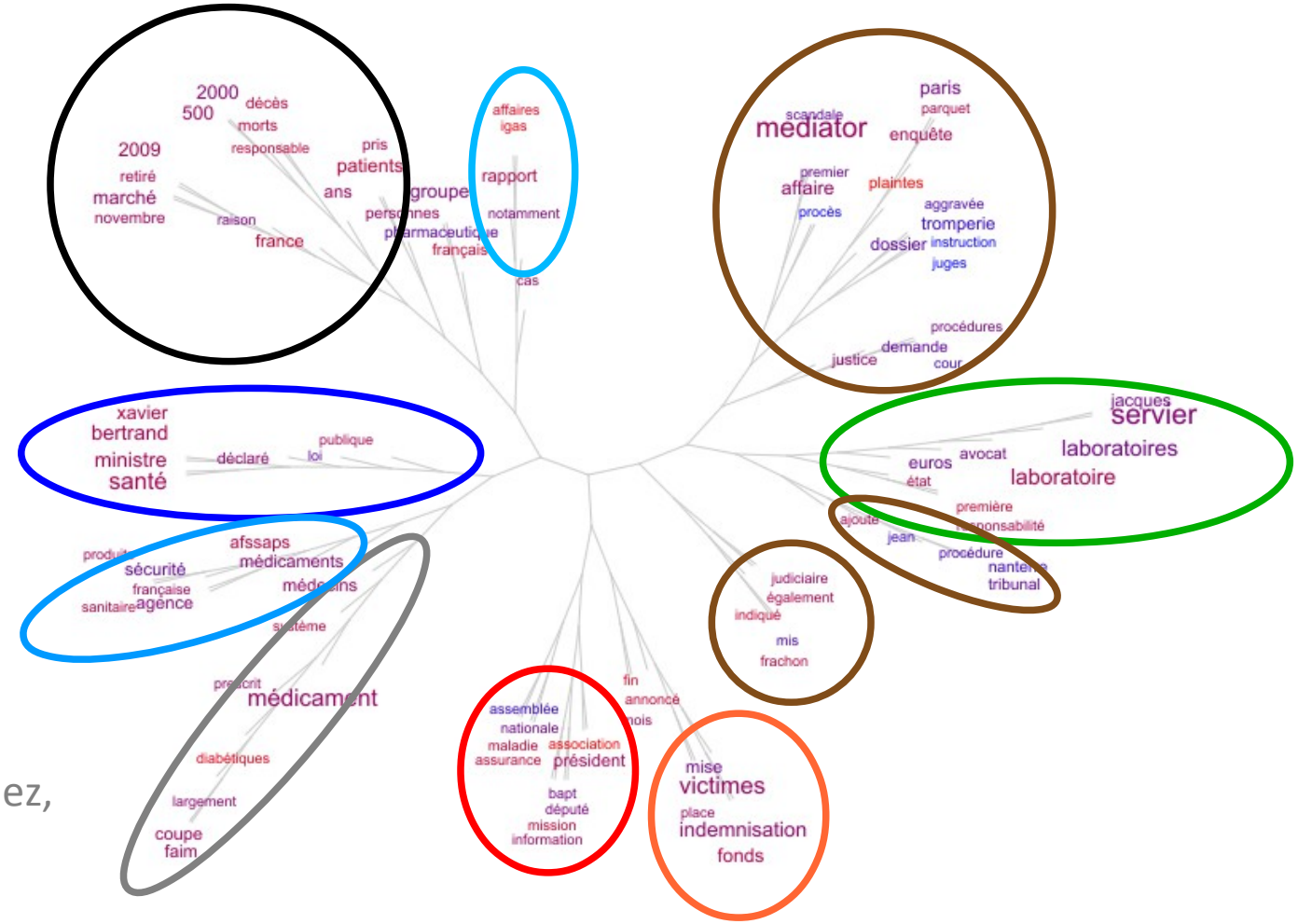


Illustration sur le corpus Mediator

Comparer les articles d'agences et articles de journalistes

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

Articles d'agences



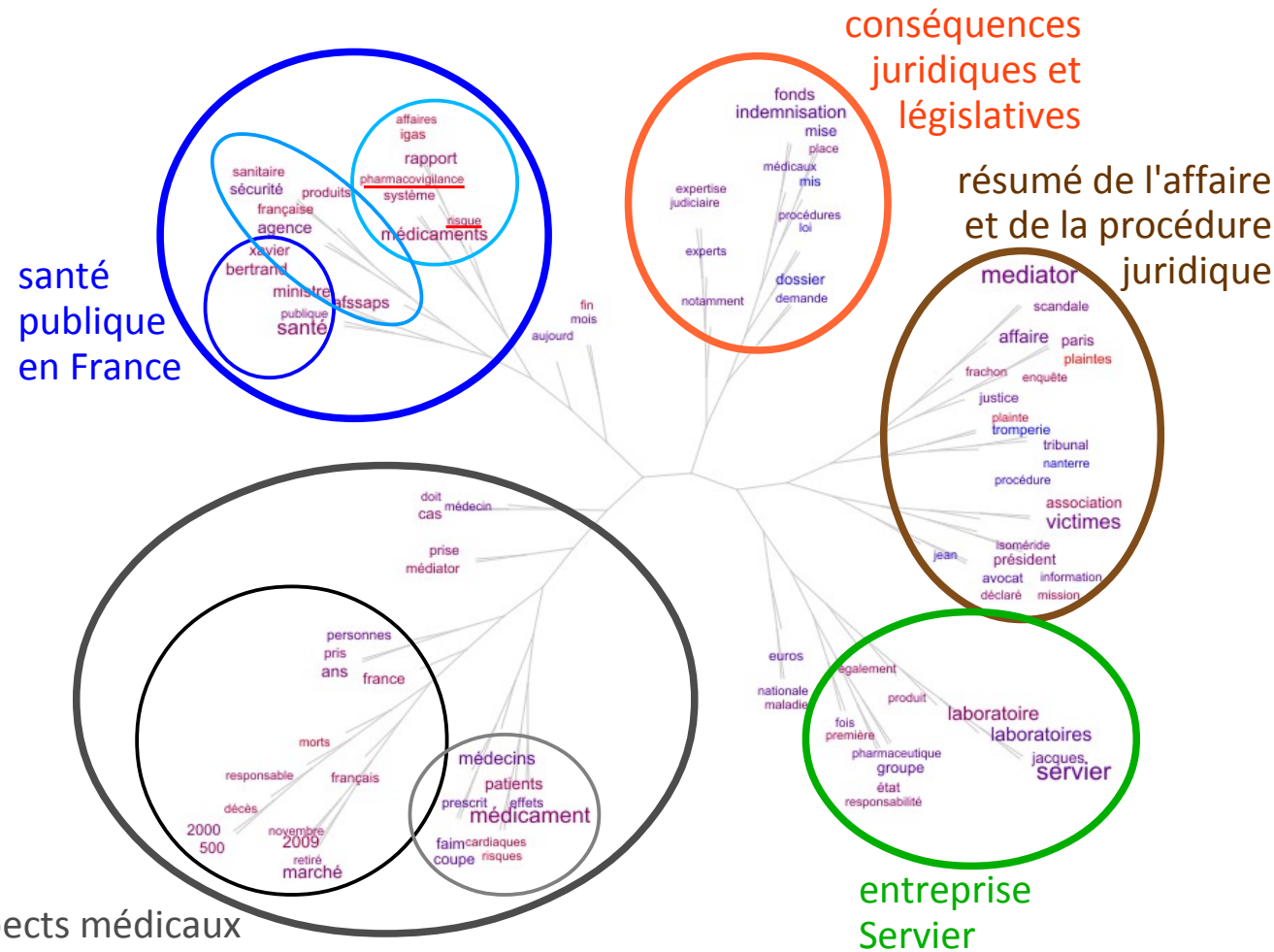
Gambette & Martinez, *Texto!*, 2013

Illustration sur le corpus Mediator

Comparer les articles d'agences et articles de journalistes

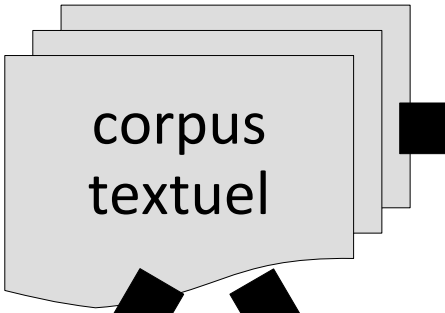
Corpus : 595 articles d'agences contre 1496 articles de journalistes de 2011 évoquant l'affaire du Mediator dans la presse française.

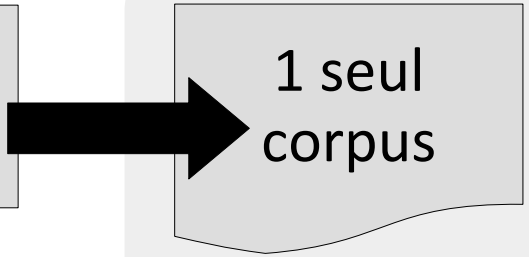
Articles
de journalistes



Gambette & Martinez,
Texto!, 2013

Exploration de corpus avec TreeCloud

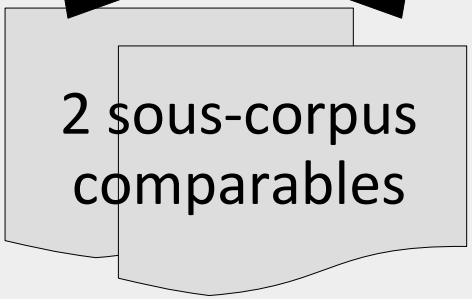




- Interpréter les regroupements
- Examiner les voisinages des mots fréquents dans l'arbre
- Examiner le voisinage de mots ciblés dans l'arbre («acteurs», verbes, etc.)

Si le corpus est assez grand :

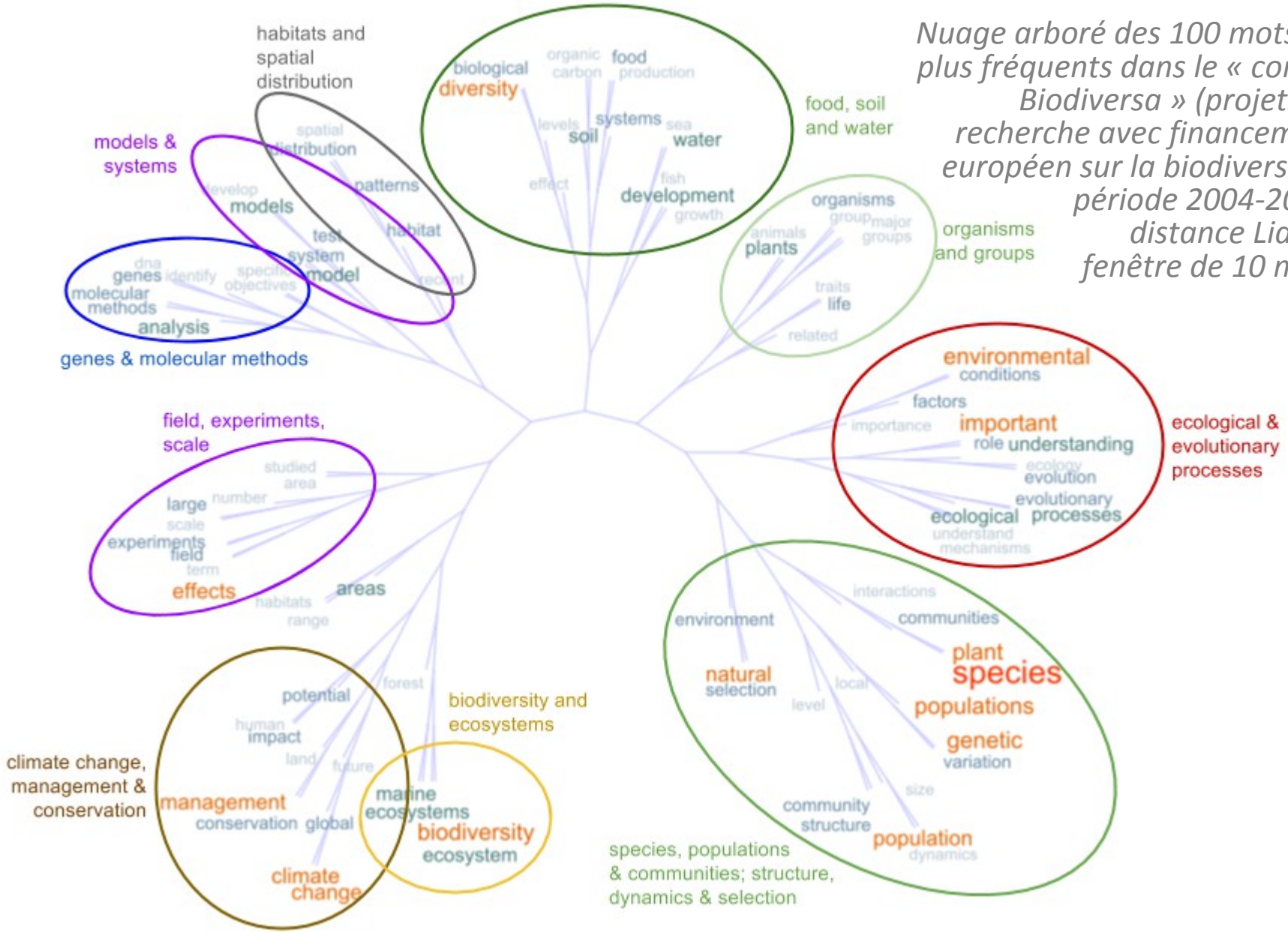
- Identifier les mots plutôt au début ou à la fin du corpus
- extraire le voisinage dans le corpus de certains mots ciblés et utiliser ces données comme corpus.



- Identifier des regroupements communs ou différents
- Comparer la fréquence des mots fréquents dans les deux arbres
- **Comparer (quanti+quali) les regroupements dans les deux arbres**
- Construire le nuage arboré des mots spécifiques de chaque sous-corpus

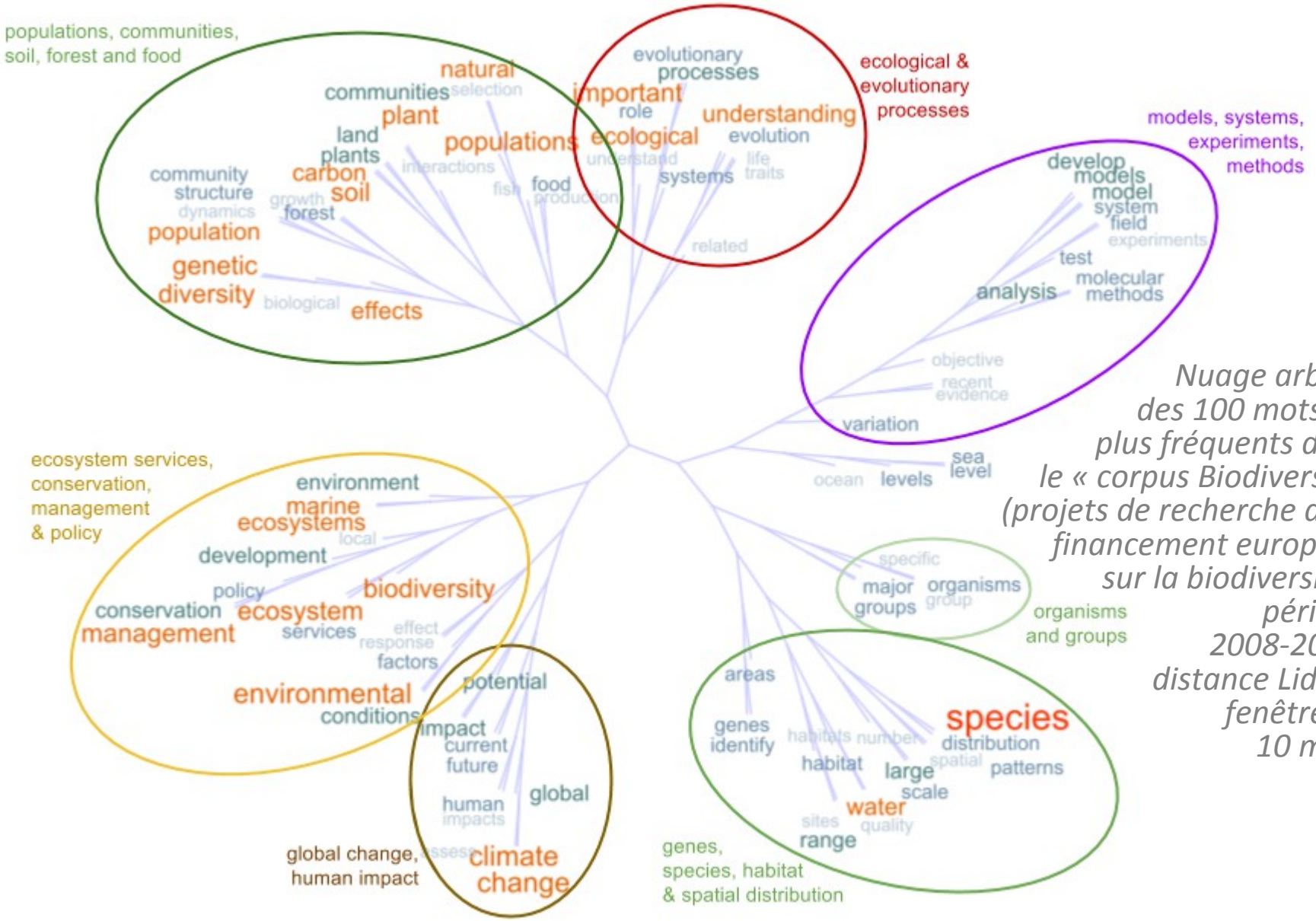
Travail en cours avec Edna Hernandez

Méthode : comparaison de voisinages dans l'arbre



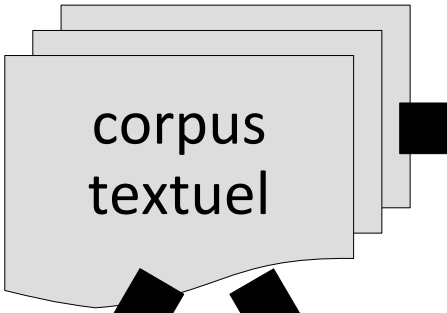
Nuage arboré des 100 mots les plus fréquents dans le « corpus Biodiversa » (projets de recherche avec financement européen sur la biodiversité), période 2004-2007, distance Liddel, fenêtre de 10 mots

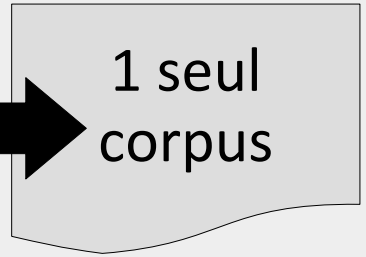
Méthode : comparaison de voisinages dans l'arbre



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

Exploration de corpus avec TreeCloud

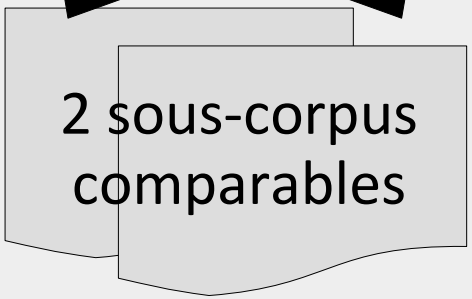




- Interpréter les regroupements
- Examiner les voisinages des mots fréquents dans l'arbre
- Examiner le voisinage de mots ciblés dans l'arbre («acteurs», verbes, etc.)

Si le corpus est assez grand :

- Identifier les mots plutôt au début ou à la fin du corpus
- extraire le voisinage dans le corpus de certains mots ciblés et utiliser ces données comme corpus.

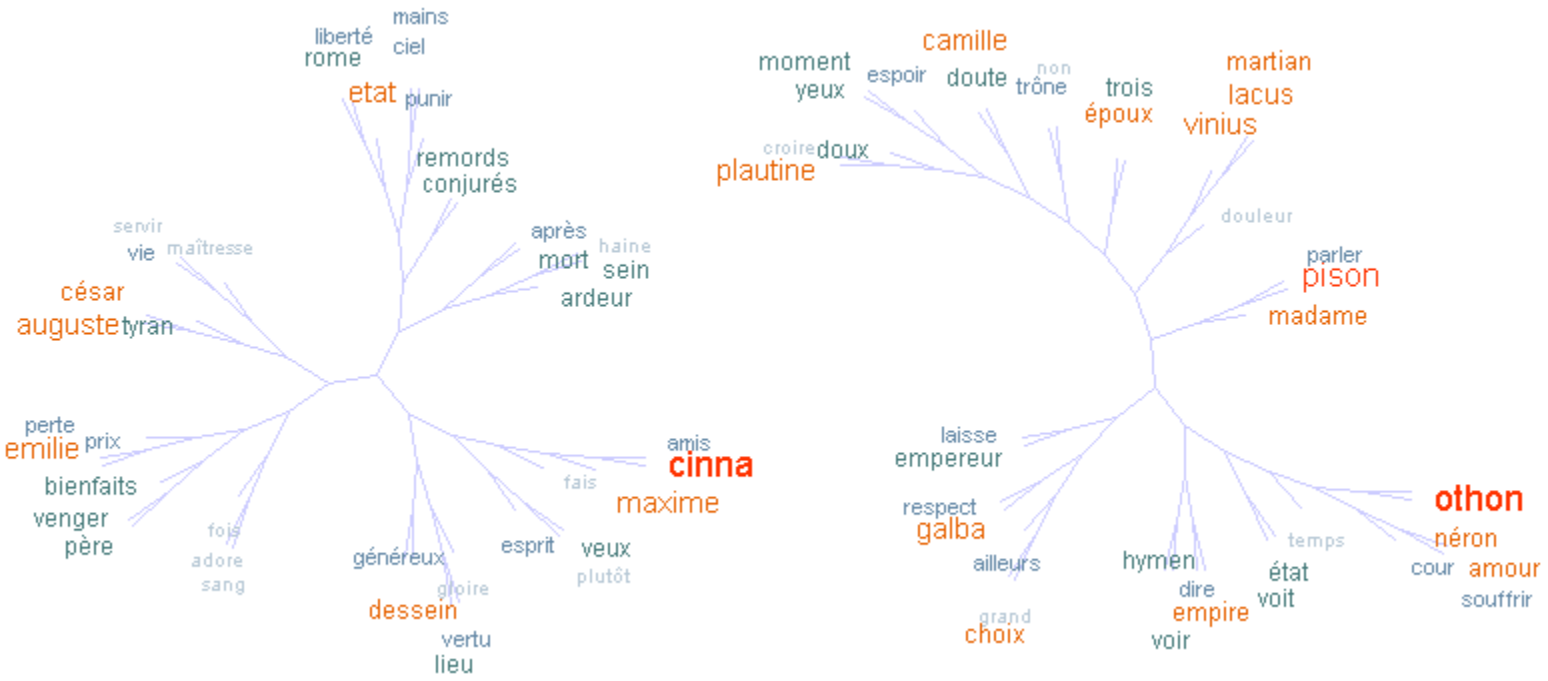


- Identifier des regroupements communs ou différents
- Comparer la fréquence des mots fréquents dans les deux arbres
- Comparer (quanti+quali) les regroupements dans les deux arbres
- **Construire le nuage arboré des mots spécifiques de chaque sous-corpus**

Travail en cours avec Edna Hernandez

Méthode : comparaison des spécifiques

Amstutz & Gambette,
JADT 2010

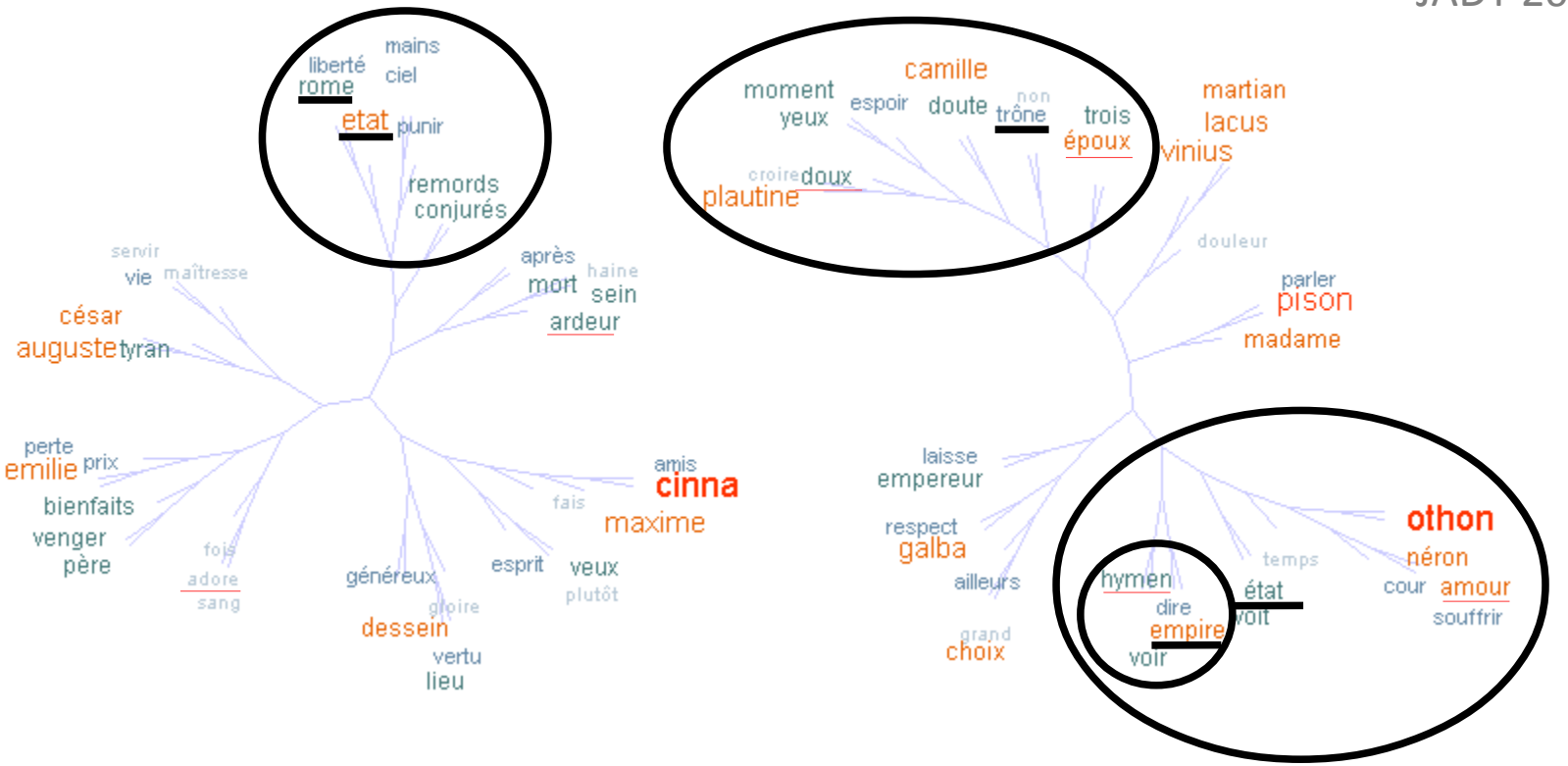


*Nuages arborés des **mots spécifiques** de Cinna et Othon, dimensionnés et colorés d'après leur spécificité calculée dans Lexico3.*

Quels moyens au service de la cause politique ?

Méthode : comparaison des spécifiques

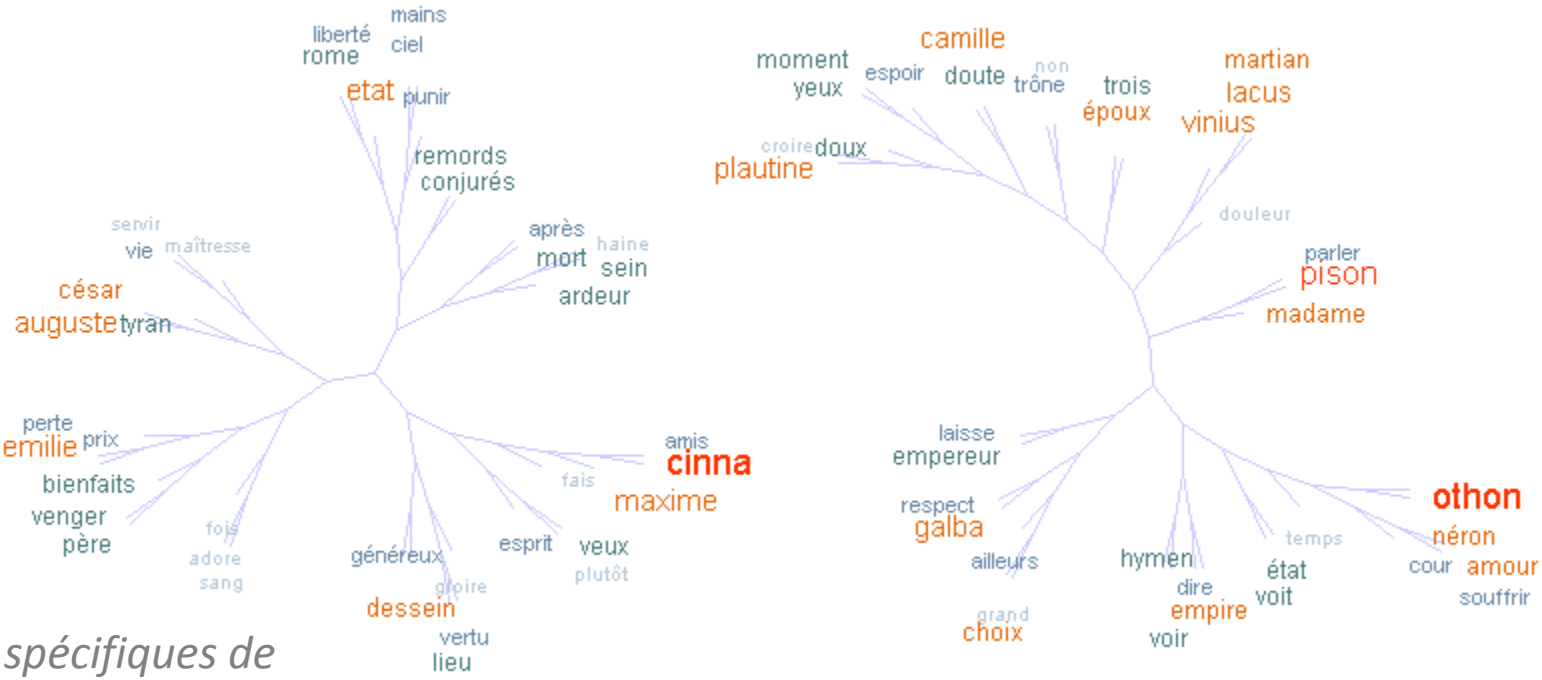
Amstutz & Gambette,
JADT 2010



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

Quels moyens au service de la cause politique ?

Méthode : comparaison des spécifiques



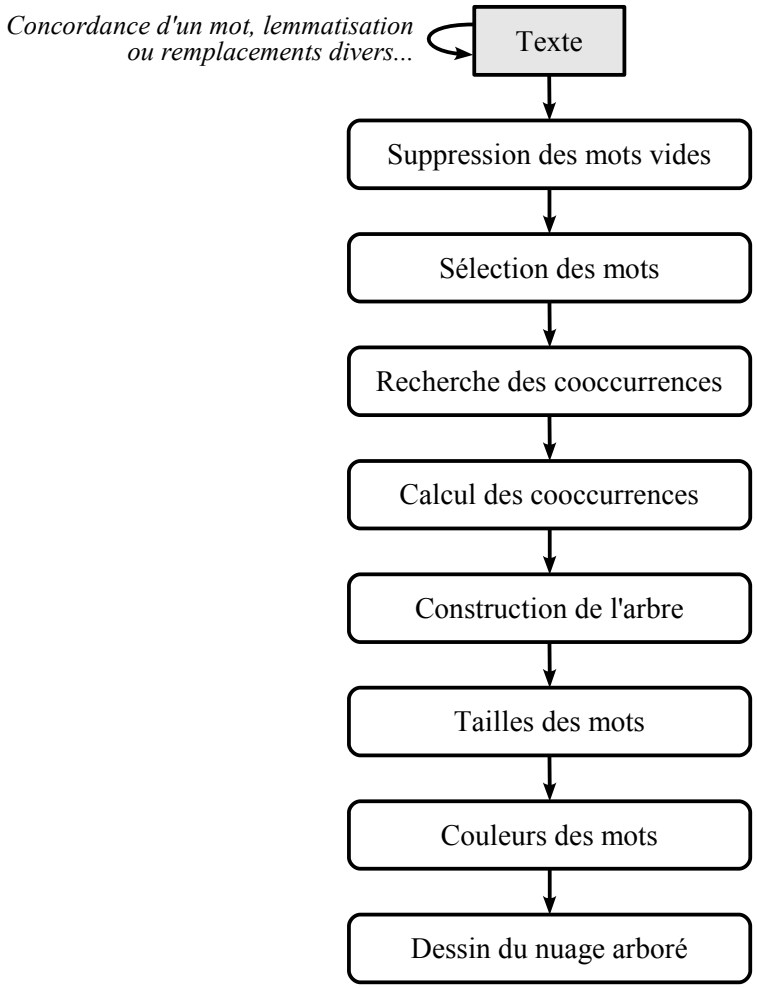
mots spécifiques de Cinna et Othon d'après Lexico3

	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

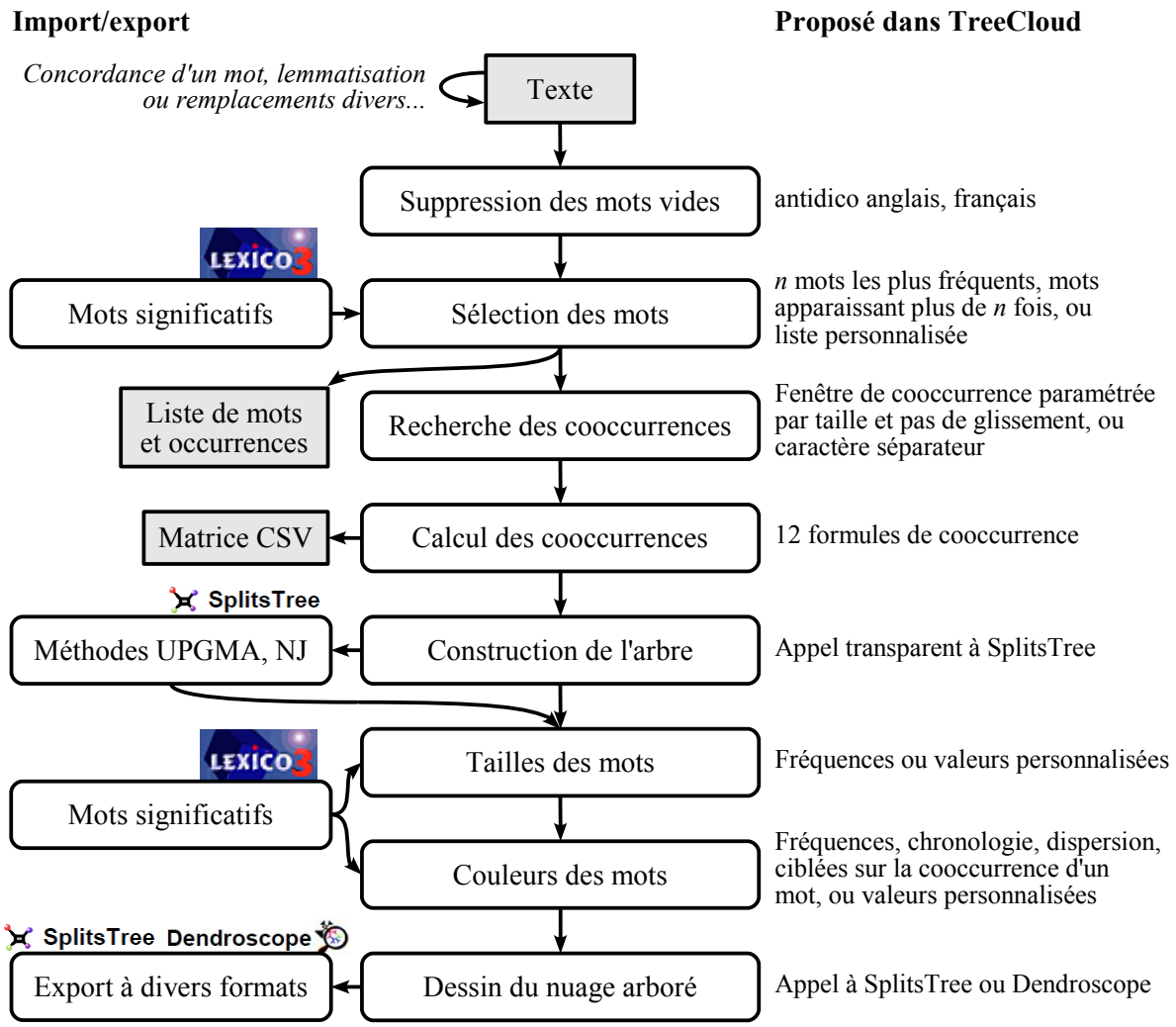
2.
Construction des nuages arborés

Processus de construction

Import/export

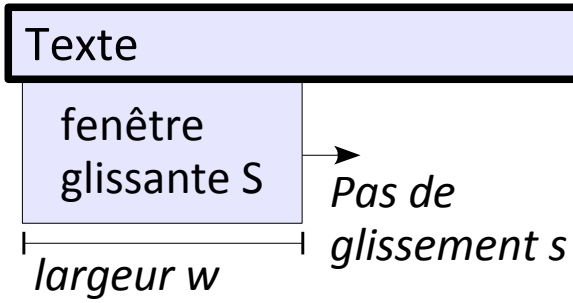


Processus de construction



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}

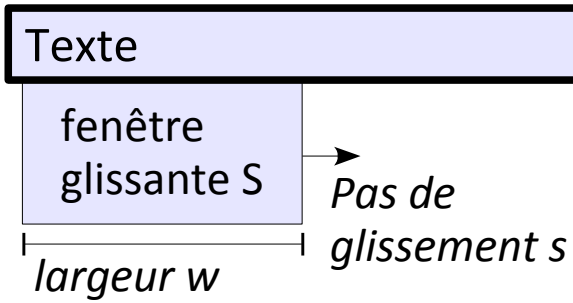
matrice de dissimilarité sémantique
chi squared, mutual information, liddel, dice, jaccard, gmean, hyperlex, minimum sensitivity, odds ratio, zscore, log likelihood, poisson-stirling...

Evert, *Statistics of words cooccurrences*, thèse, 2005
Gambette, *User manual for TreeCloud,* 2009

Calcul des scores de cooccurrence

La diapo des formules !

Calcul de la matrice de distance entre m



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}

- jaccard: $1 - O_{11}/(O_{11} + O_{12} + O_{21})$
- liddell: $1 - (O_{11}O_{22} - O_{12}O_{21})/(C_1C_2)$
- dice: $1 - 2O_{11}/(R_1 + C_1)$
- hyperlex: $1 - \max(O_{11}/R_1, O_{11}/C_1)$
- poissonstirling: $O_{11}(\log O_{11} - \log E_{11} - 1)$
- chisquared: $1000 - N(O_{11} - E_{11})^2/(E_{11}E_{22})$
- zscore: $1 - (O_{11} - E_{11})/\sqrt{E_{11}}$
- ms: $1 - \min(O_{11}/R_1, O_{11}/C_1)$
- oddsratio: $1 - \log((O_{11}O_{22})/(O_{12}O_{21}))$
- loglikelihood: $1 - 2(O_{11}\log(O_{11}/E_{11}) + O_{12}\log(O_{12}/E_{12}) + O_{21}\log(O_{21}/E_{21}) + O_{22}\log(O_{22}/E_{22}))$
- gmean: $1 - O_{11}/\sqrt{R_1C_1} = 1 - O_{11}/\sqrt{NE_{11}}$
- mi (mutual information): $1 - \log(O_{11}/E_{11})$
- ngd (normalized Google distance): $(max(\log R_1, \log C_1) - \log O_{11})/(N - min(\log R_1, \log C_1))$

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]

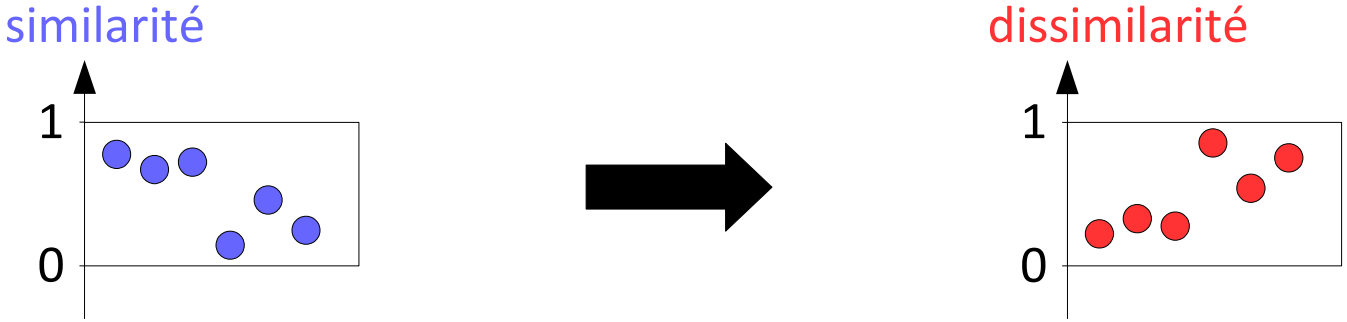
Normalisation des scores de similarité sur [0,1] :

- 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)

Calcul des distances de cooccurrence

Les formules statistiques fournissent un score de **similarité**.

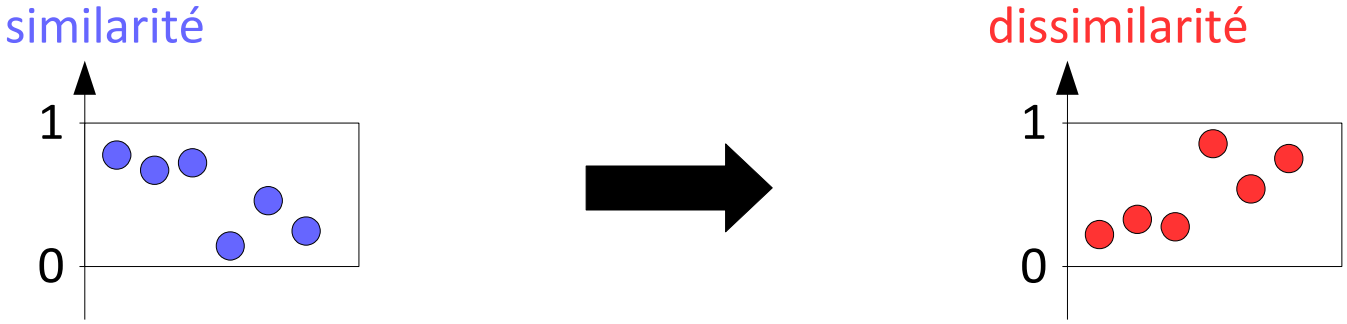
Comment obtenir des dissimilarités, dans l'intervalle [0,1] ?



Calcul des distances de cooccurrence

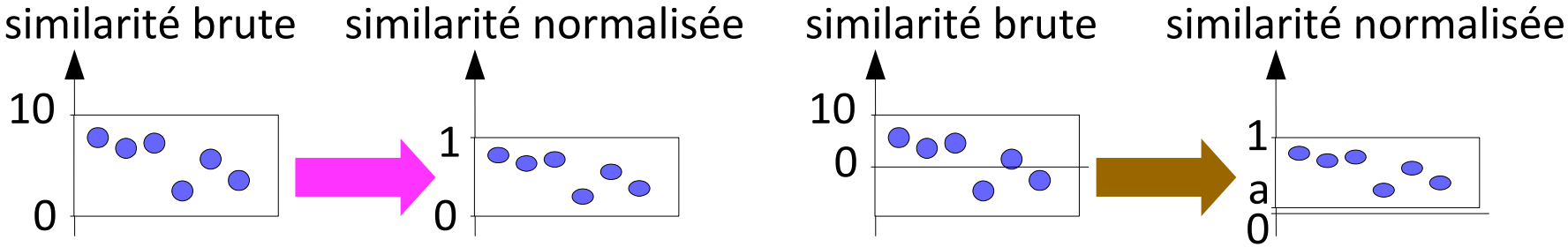
Les formules statistiques fournissent un score de **similarité**.

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



Normalisation des scores de similarité sur [0,1] :

- **normalisation linéaire** pour les matrices positives
- **normalisation affine** pour les matrices contenant des valeurs négatives, afin d'obtenir des distances dans l'intervalle [a,1] (a=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

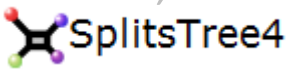
Plusieurs méthodes pour construire un arbre à partir d'une matrice de distances (classification hiérarchique) :

- UPGMA

Sokal & Michener, 1958

- Neighbor-Joining

Saitou & Nei, 1987



- Variantes d'Addtree

Barthelemy & Luong, 1987

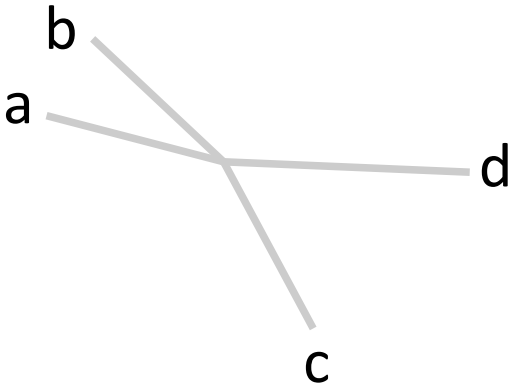
- Heuristique des quadruplets

Cilibrasi & Vitanyi, 2007

Construction arbre : UPGMA vs Neighbor-Joining

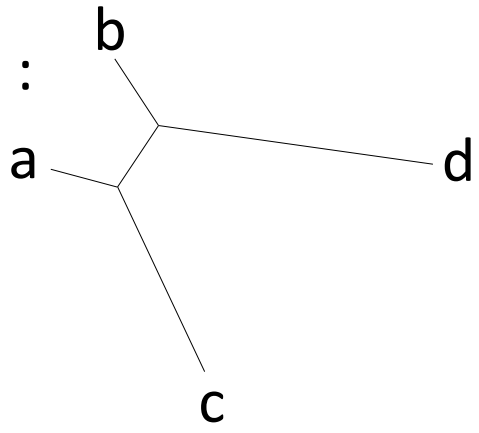
Partir d'un arbre en étoile, puis répéter l'étape suivante :

- UPGMA : fusionner les deux feuilles les plus proches
- Neighbor-Joining : fusionner les deux feuilles qui minimisent la longueur totale des branches de l'arbre



Arbre UPGMA
(étape 1)

Vrai arbre :

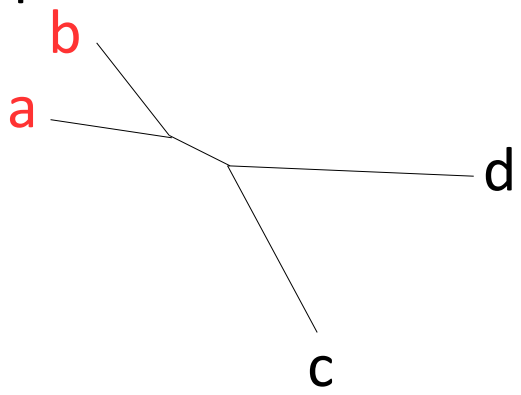


Construction arbre : UPGMA vs Neighbor-Joining

Partir d'un arbre en étoile, puis répéter l'étape suivante :

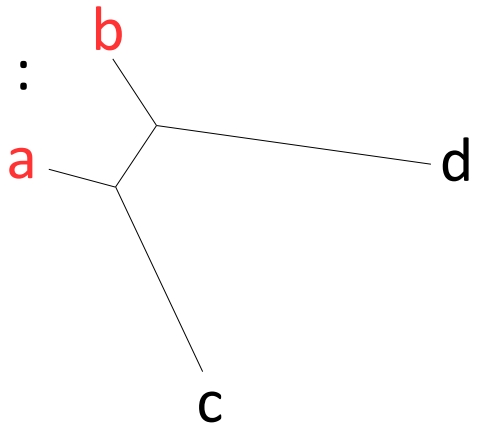
- UPGMA : fusionner les deux feuilles les plus proches

- Neighbor-Joining : fusionner les deux feuilles qui minimisent la longueur totale des branches de l'arbre



Arbre UPGMA
(étape 1)

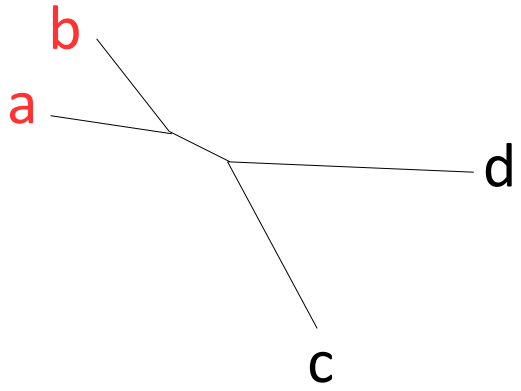
Vrai arbre :



Construction arbre : UPGMA vs Neighbor-Joining

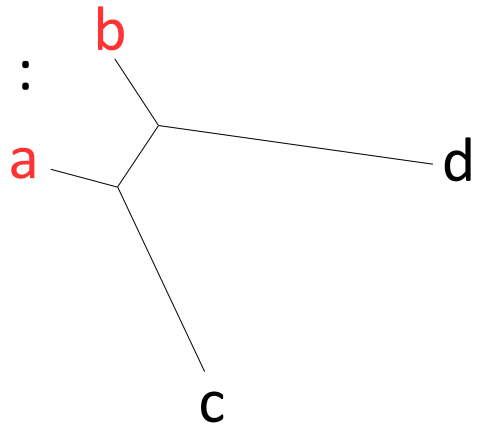
Partir d'un arbre en étoile, puis répéter l'étape suivante :

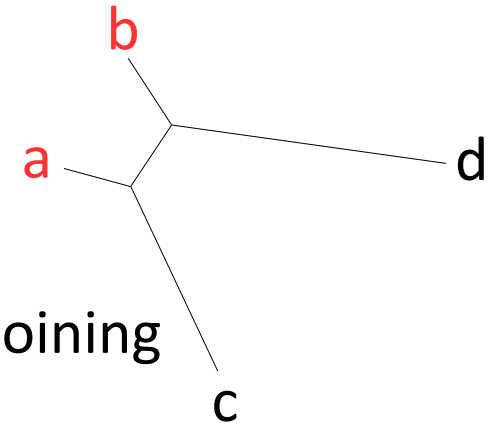
- UPGMA : fusionner les deux feuilles les plus proches
- Neighbor-Joining : fusionner les deux feuilles qui minimisent la longueur totale des branches de l'arbre



Arbre UPGMA
(étape 1)

Vrai arbre :





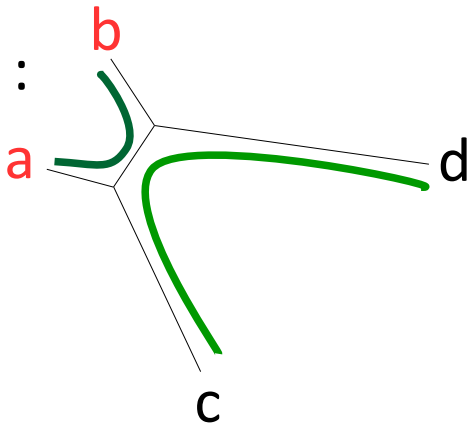
Arbre Neighbor-Joining
(étape 1)

Construction arbre : UPGMA vs Neighbor-Joining

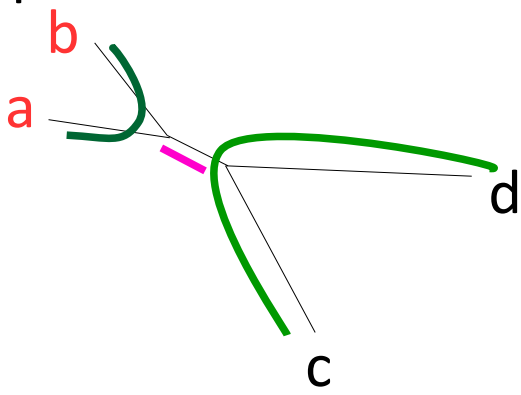
Partir d'un arbre en étoile, puis répéter l'étape suivante :

- UPGMA : fusionner les deux feuilles les plus proches
- Neighbor-Joining : fusionner les deux feuilles qui minimisent la **longueur totale des branches de l'arbre**

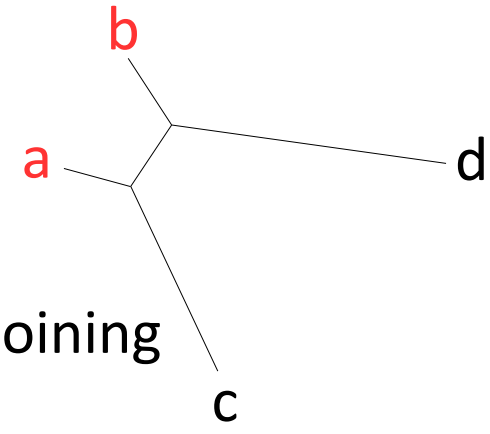
Vrai arbre :



Arbre UPGMA (étape 1)



Arbre Neighbor-Joining (étape 1)



Décoration de l'arbre

Tailles des mots :

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

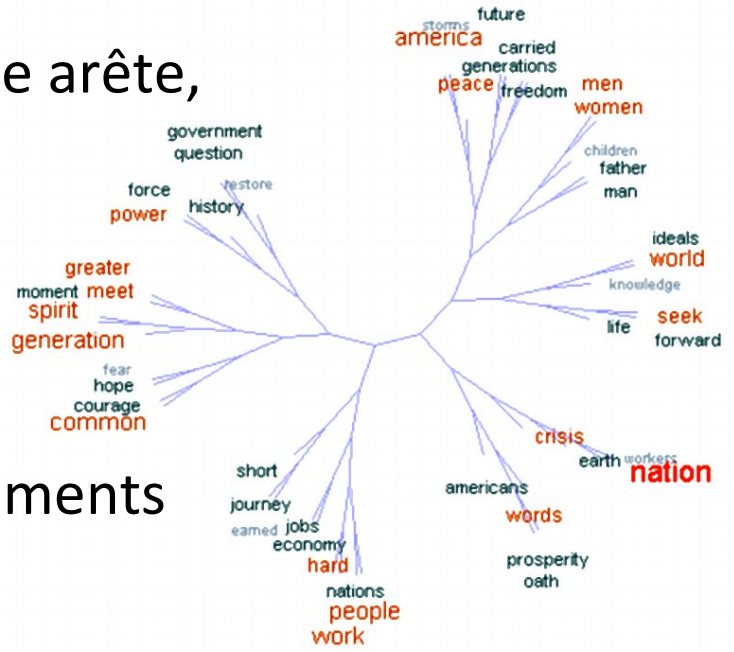
- calculées à partir des **rangs des fréquences**
  (distribution exponentielle)

- **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

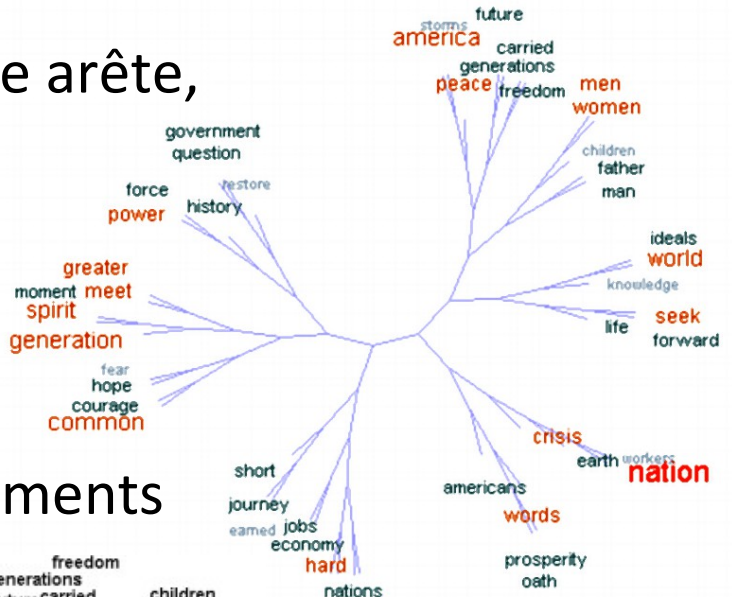
Algorithme “equal angle” :

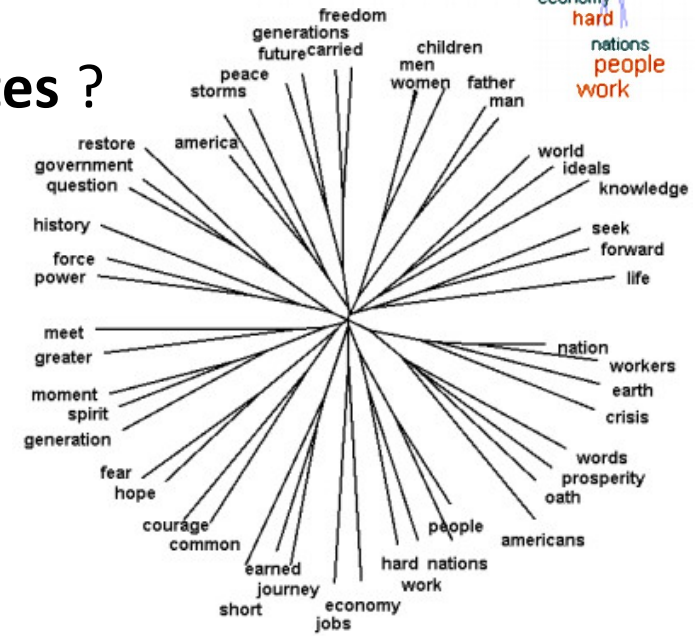
- montant pour calculer l'angle de chaque arête,
  en partant des feuilles
- descendant pour placer chaque arête
  en partant d'un sommet interne

Placement automatique des étiquettes :

→ heuristique pour éviter les chevauchements

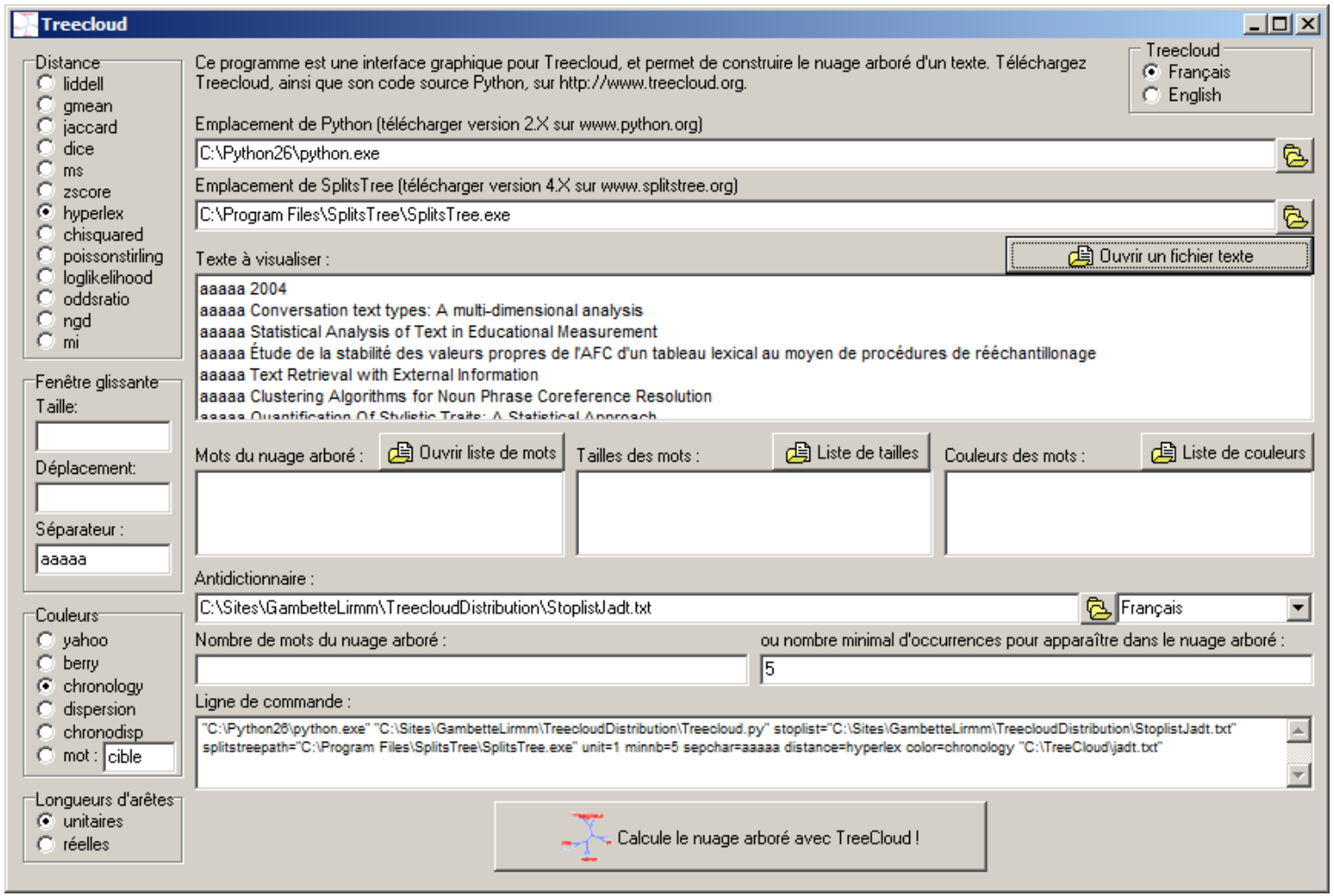
Question des longueurs d'arêtes ?





Implémentations

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

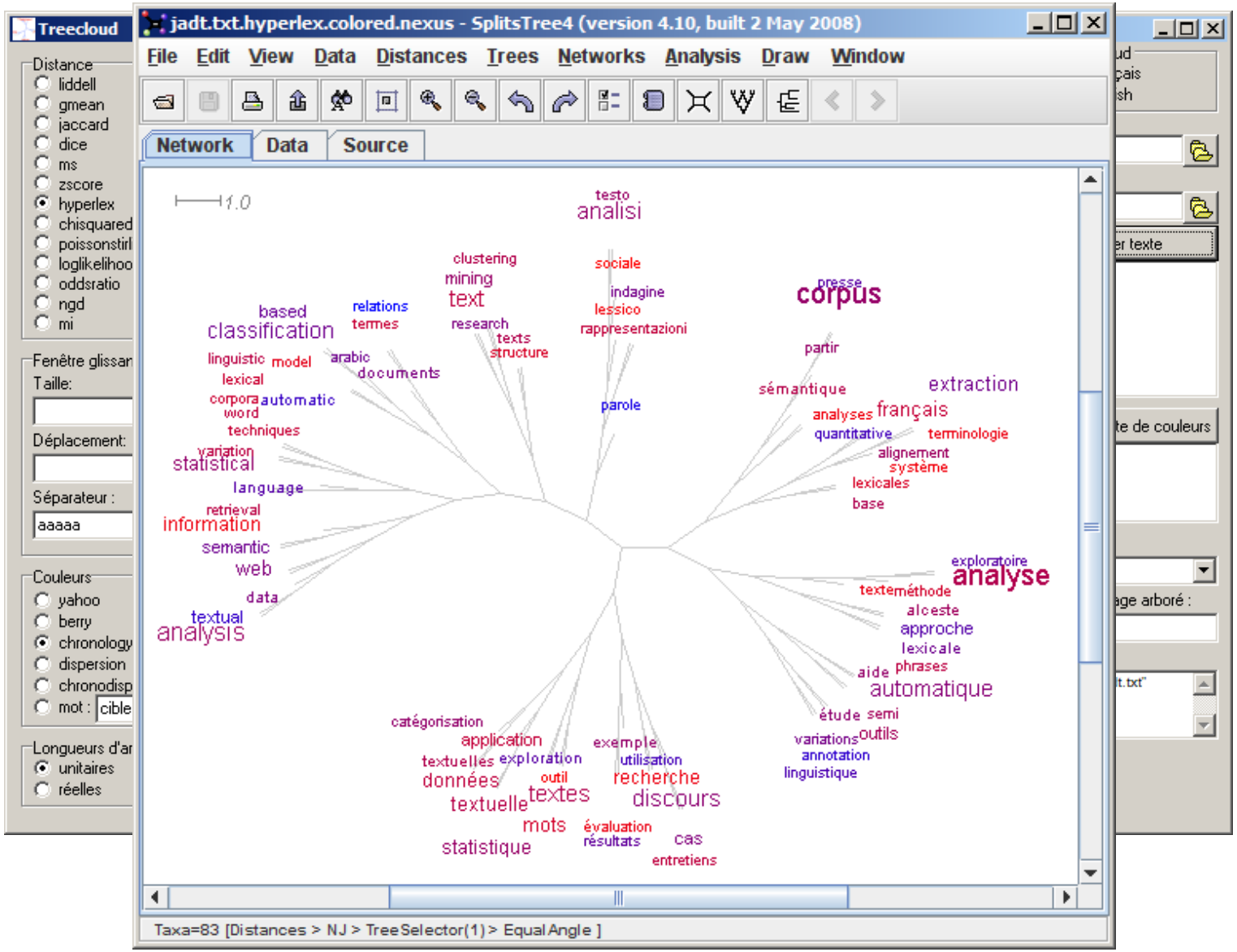


www.treecloud.org

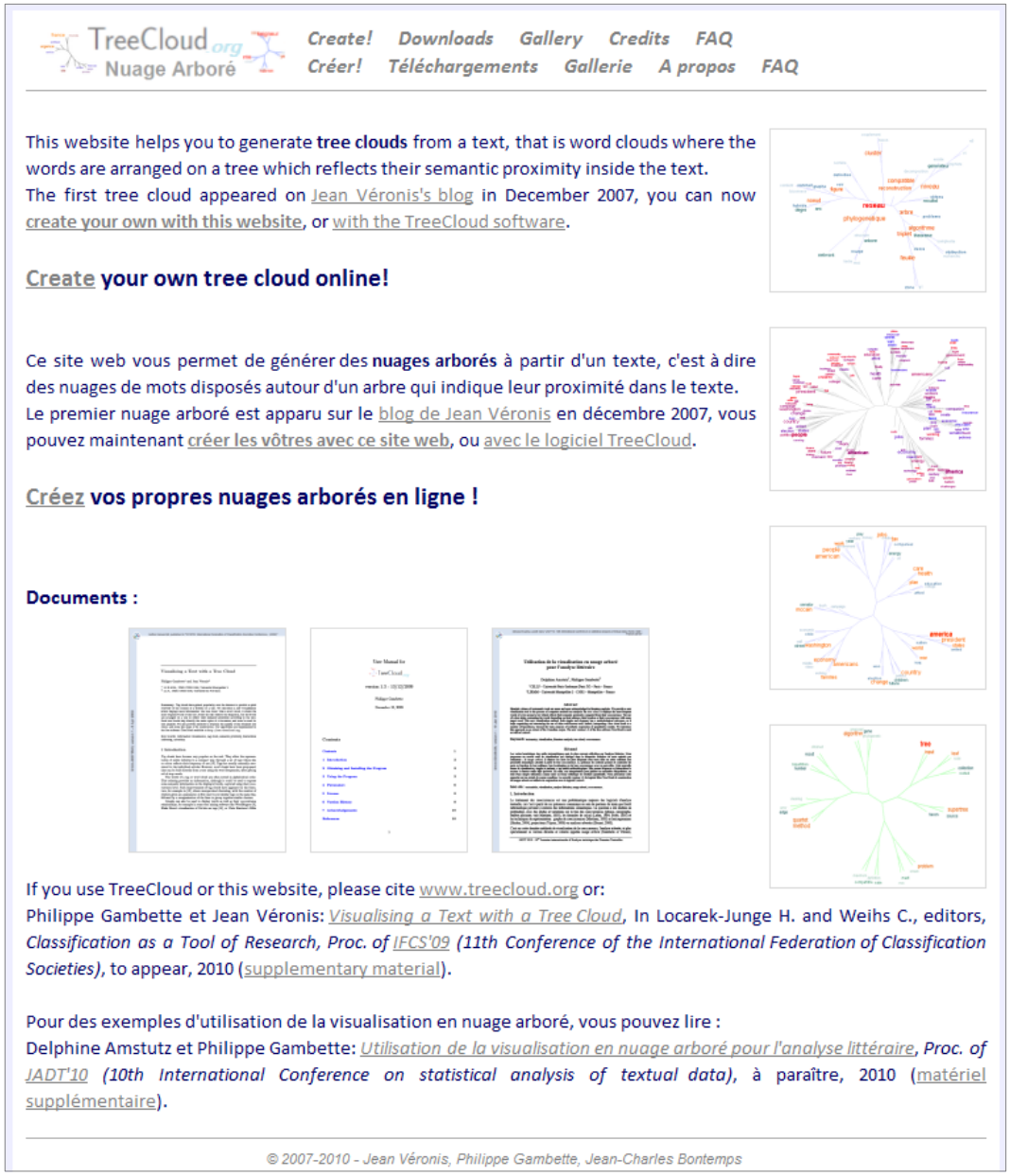
www.splitstree.org

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/

Développements supplémentaires avec d3.js par Deepak Srinivas

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 clo**
words are arranged on a tree which reflects
The first tree cloud appeared on Jean Vé
create your own with this website, or with t

Create **your own tree cloud online**

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

Créez **vos propres nuages arborés**

Documents :

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

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

TreeCloud org
Nuage Arboré

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

Créez vos propres nuages arborés !

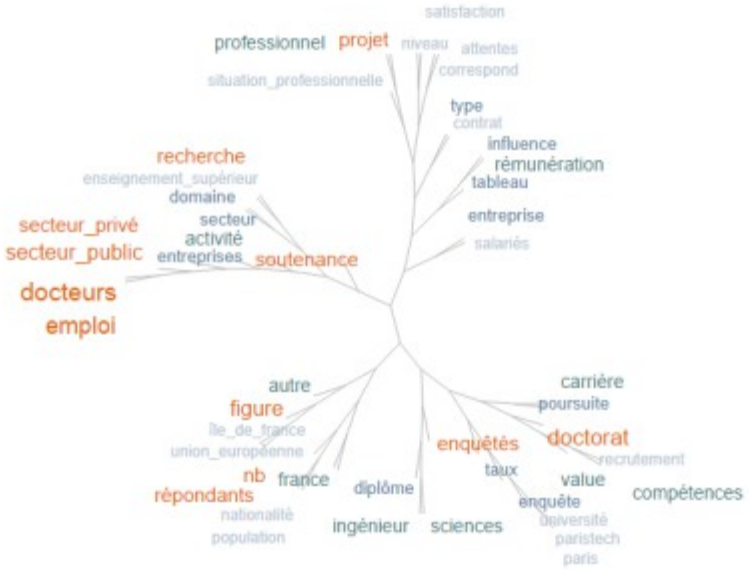
Collez votre texte dans le cadre ci-dessous, puis cliquez sur *Envoyer* ! Attention, l'utilisateur suivant verra votre texte quand il se connectera au site, si vous ne voulez pas faire apparaître vos textes, installez plutôt TreeCloud sur votre machine.

Texte :

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

Envoyer

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



Interface web

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 clo
words are arranged on a tree which reflects
The first tree cloud appeared on Jean Vé
create your own with this website, or with t

Create your own tree cloud online

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

Créez vos propres nuages arborés

Documents :

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

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

TreeCloud org
Nuage Arboré

Créer! Téléchargements Gallerie A propos FAQ

Créez vos propres nuages arborés !

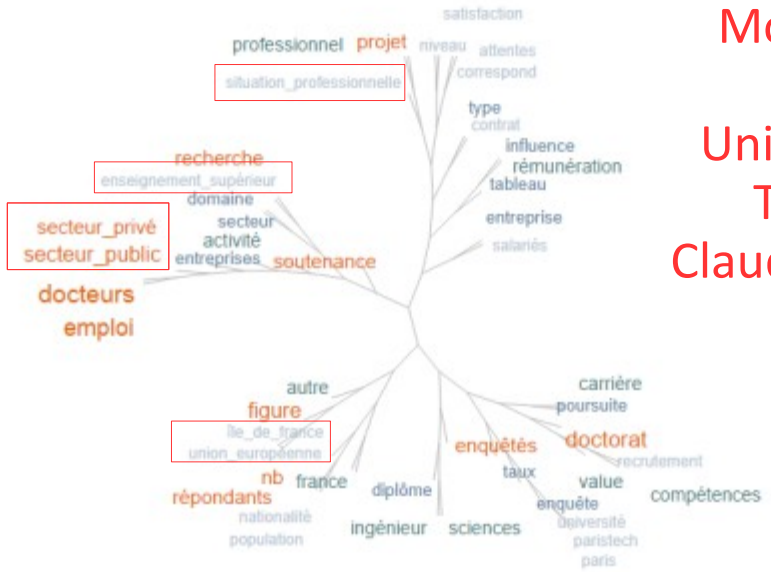
Collez votre texte dans le cadre ci-dessous, puis cliquez sur *Envoyer* ! Attention, l'utilisateur suivant verra votre texte quand il se connectera au site, si vous ne voulez pas faire apparaître vos textes, installez plutôt TreeCloud sur votre machine.

Texte :

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

Envoyer

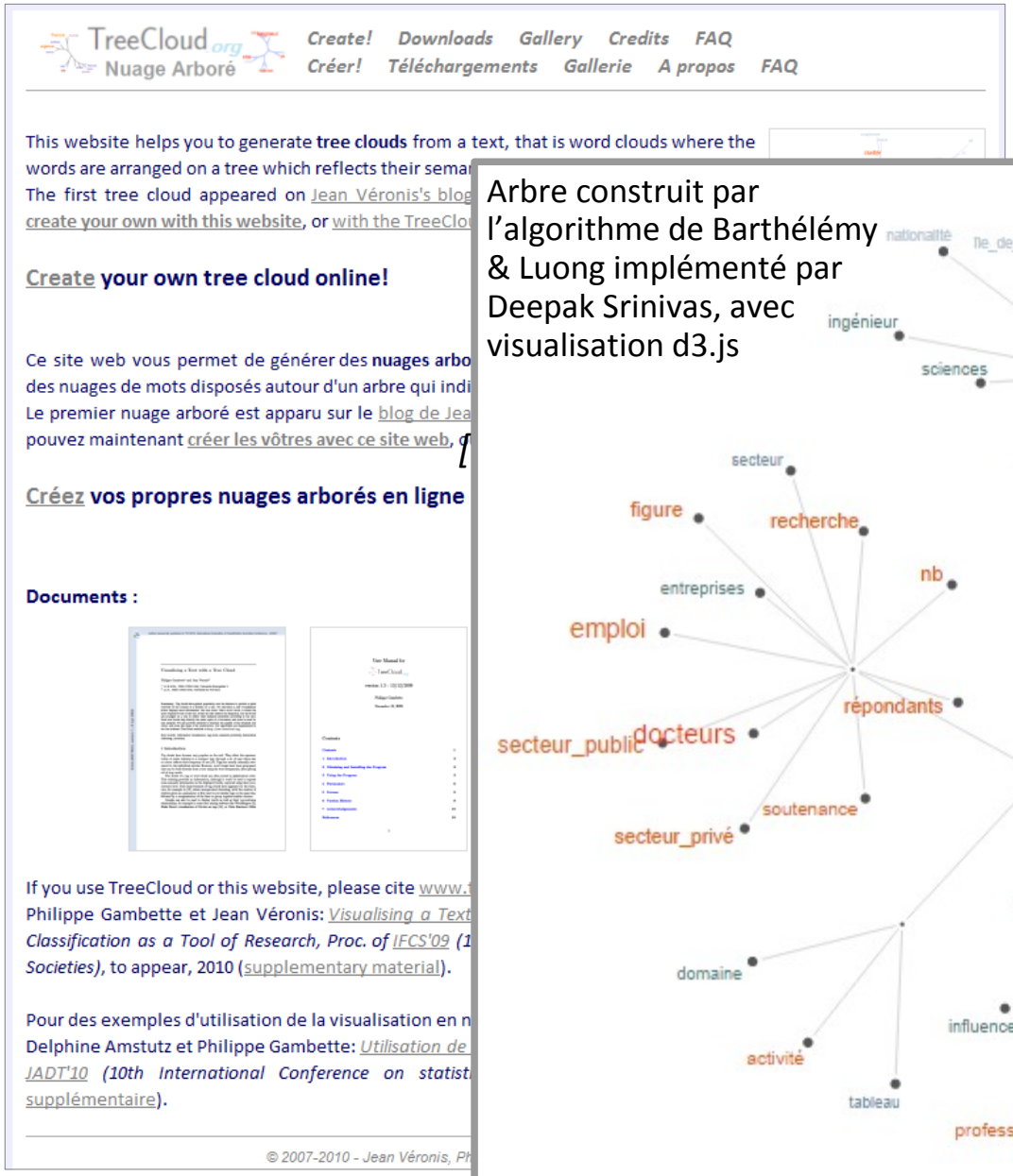
Vous pouvez déplacer les étiquettes par cliquer-glisser, l'étiquette reprend sa place d'origine lors d'un nouveau clic. L'infobulle indique le nombre d'occurrences du mot.



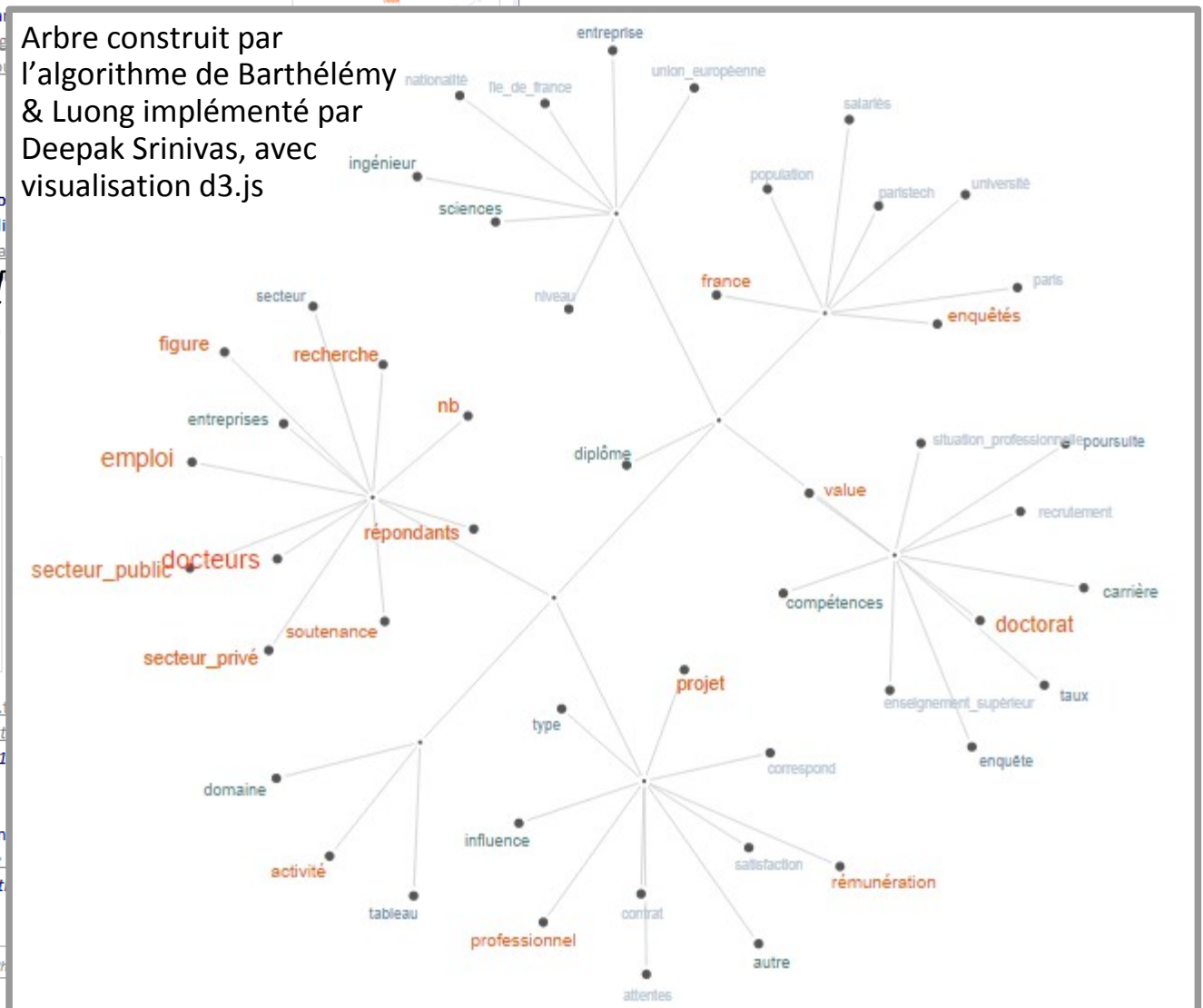
Mots composés identifiés par Unitex, intégré à TreeCloud par Claude Martineau

Interface web

www.treecloud.org

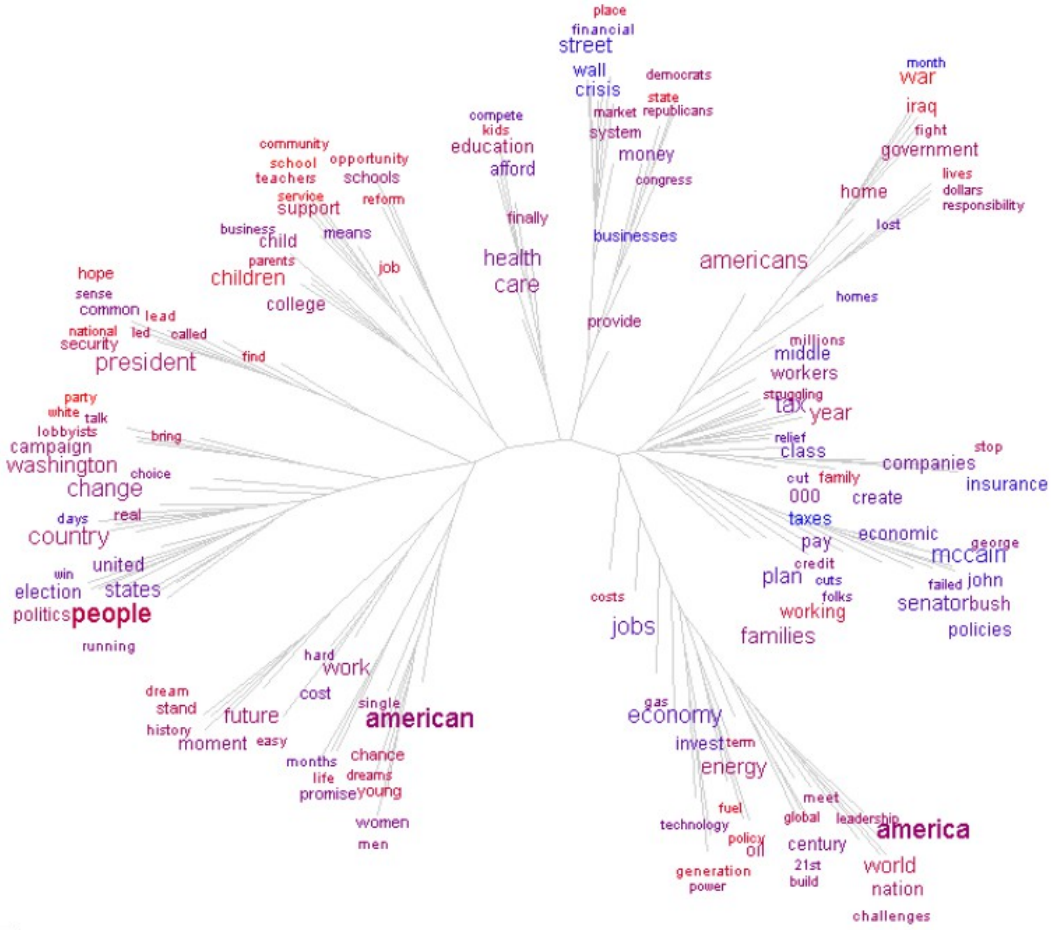


Arbre construit par l'algorithme de Barthélémy & Luong implémenté par Deepak Srinivas, avec visualisation d3.js



Temps d'exécution

Limites sur la taille du corpus pour utiliser TreeCloud ?



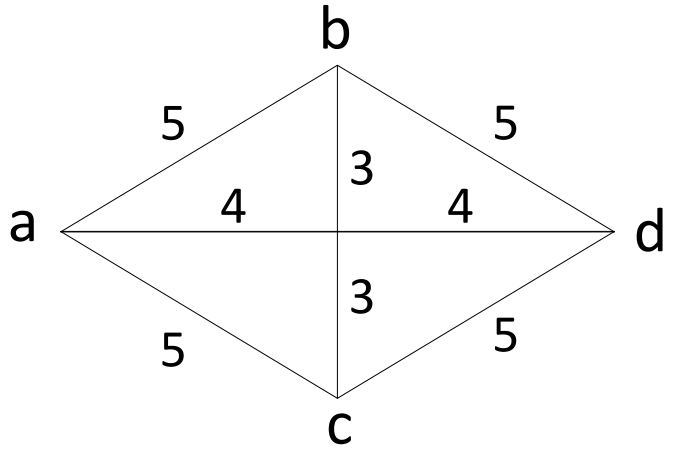
30 secondes pour la construction du nuage arboré de l'ensemble des discours de campagne de Barack Obama (>300 000 mots)

3.
Qualité des nuages arborés

Qualité des arbres

Arbre = **approximation** des distances

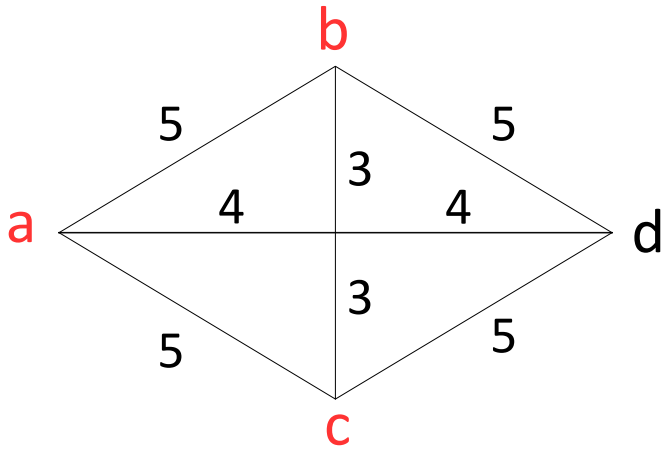
Construire un arbre qui représente ces distances :

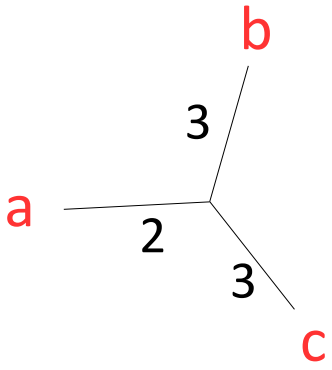


Qualité des arbres

Arbre = **approximation** des distances

Construire un arbre qui représente ces distances :

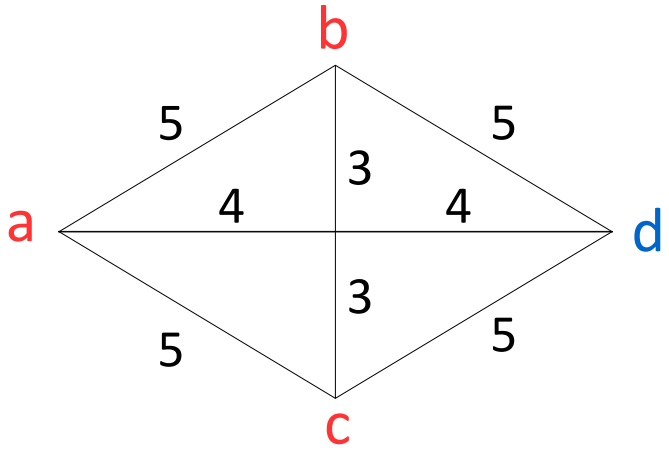


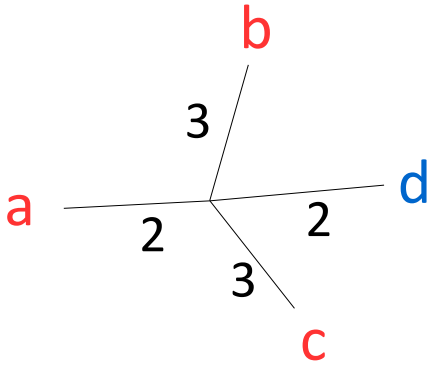


Qualité des arbres

Arbre = **approximation** des distances

Construire un arbre qui représente ces distances :



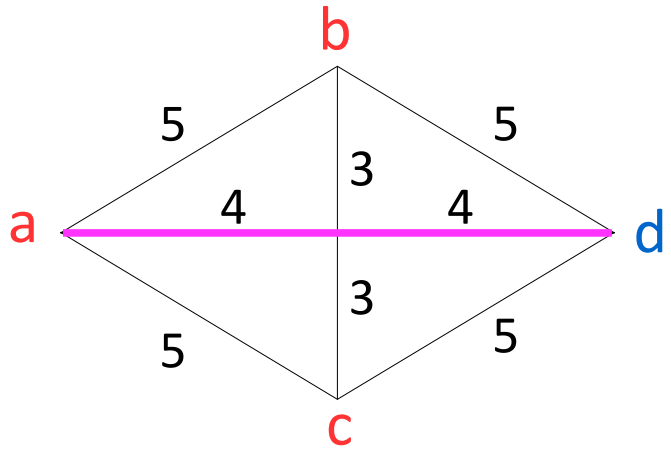


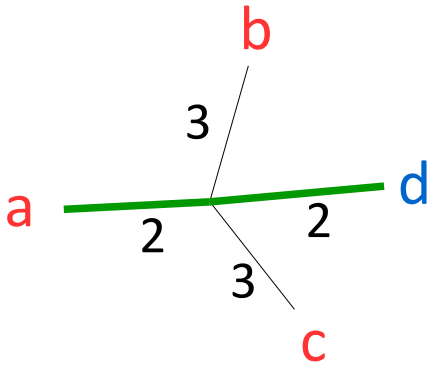
Impossible !

Qualité des arbres

Arbre = **approximation** des distances

Construire un arbre qui représente ces distances :





Impossible !

distance(a,d)=4 alors qu'on devrait avoir distance(a,d)=8

Évaluer la qualité des arbres

Arboricité de la distance de cooccurrence
= proximité de cette distance avec une « distance d'arbre »

Stabilité de l'arbre (si on modifie légèrement le texte ?)

Gain de temps pour une tâche de classification en utilisant l'arbre

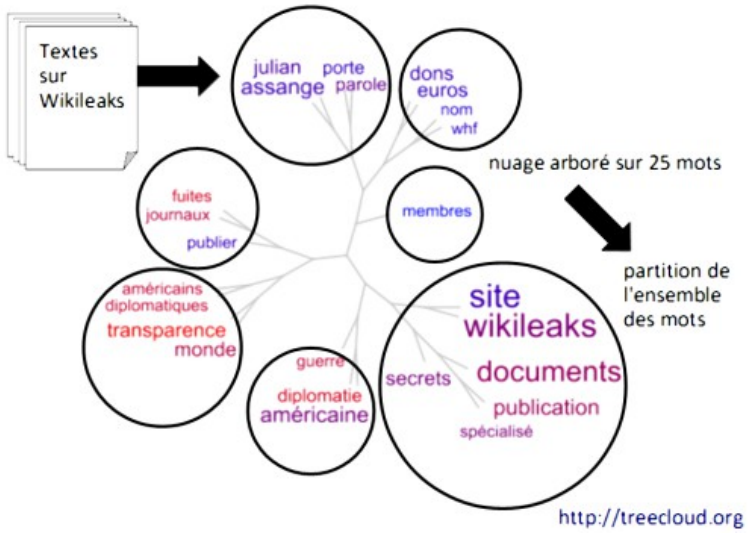
Makki, Brooks & Milios, IVAPP 2014

Protocole de comparaison avec une partition construite manuellement

Gambette, Gala & Nasr, *Corpus*, 2012

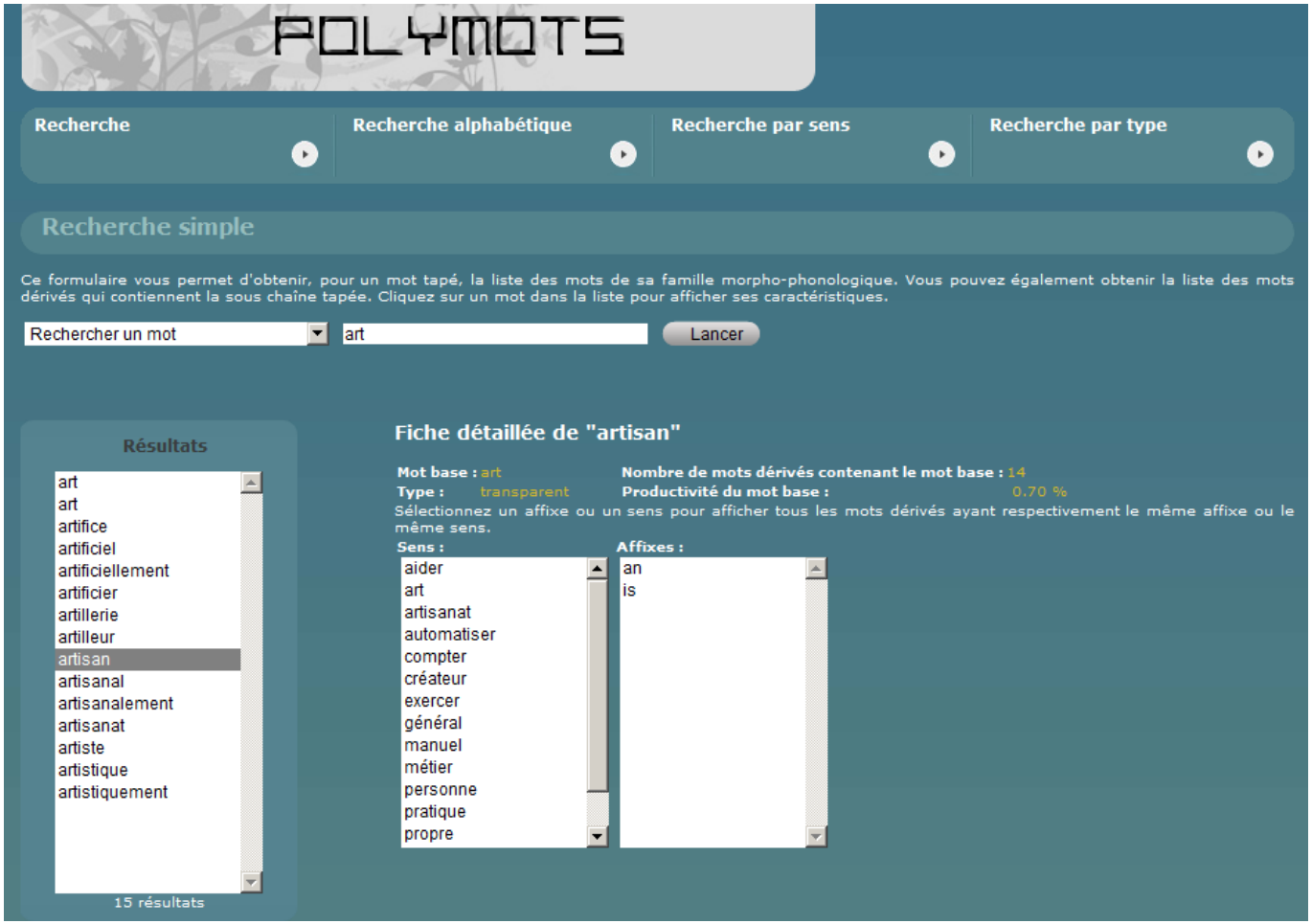
Comparaison avec l'arbre qui a inspiré la création du corpus visualisé

Gambette, Gala & Nasr, *Corpus*, 2012



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” ?

Gambette, Gala & Nasr, *Corpus*, 2012

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 :

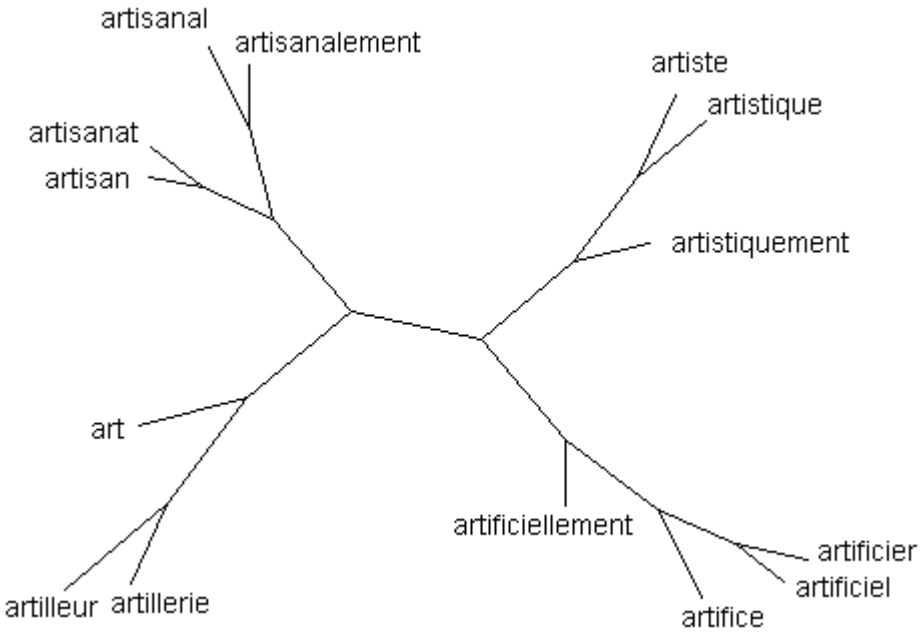
- nombre d'affixes communs
- degré de cooccurrence dans Le Trésor de la Langue Française informatisé

Gambette, Gala & Nasr, *Corpus*, 2012

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}**}**

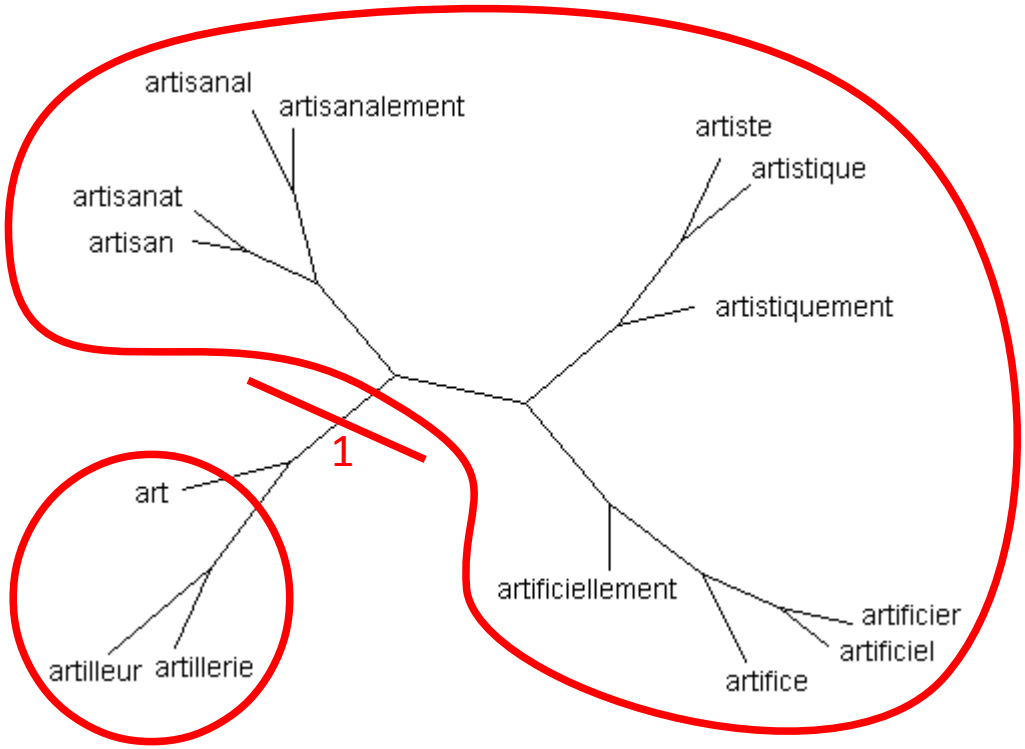
Partition manuelle : **Pm = {**{artificier, artifice, artificiel, artificiellement}, {artillerie, artilleur}, {artisan, artisanal, artisanalement, artisanat}, {artiste, artistique, artistiquement, art}**}**

Gambette, Gala & Nasr, *Corpus*, 2012

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}**}**

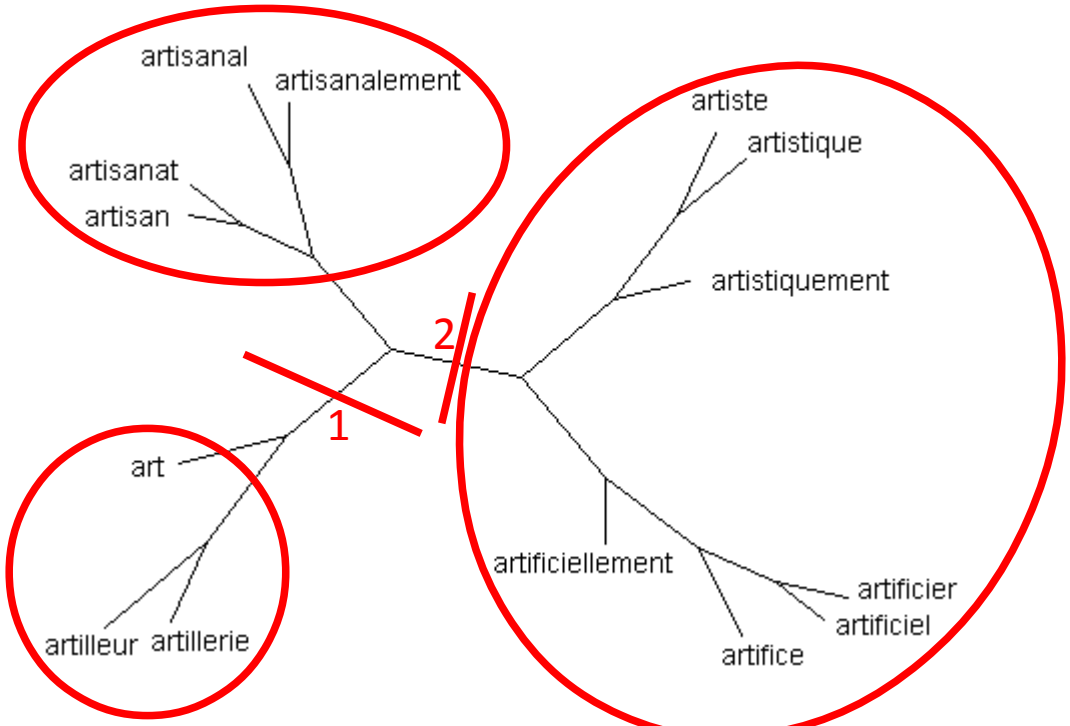
Partition manuelle : **Pm = {**{artificier, artifice, artificiel, artificiellement}, {artillerie, artilleur}, {artisan, artisanal, artisanalement, artisanat}, {artiste, artistique, artistiquement, art}**}**

Gambette, Gala & Nasr, *Corpus*, 2012

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}**}**

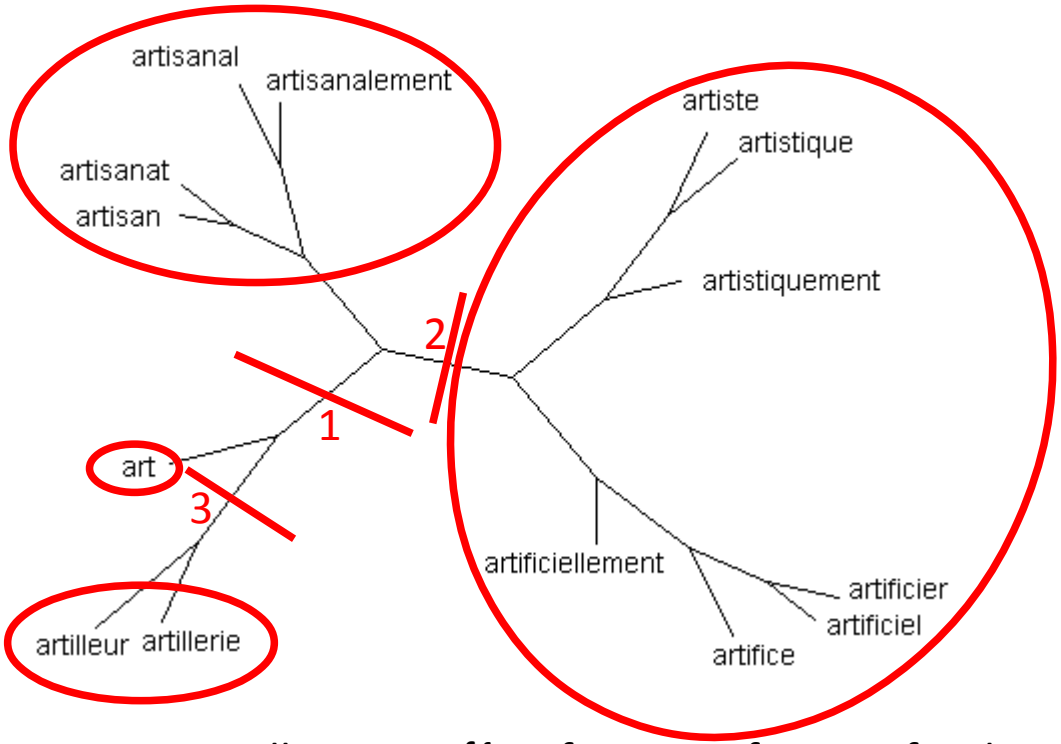
Partition manuelle : **Pm = {**{artificier, artifice, artificiel, artificiellement}, {artillerie, artilleur}, {artisan, artisanal, artisanalement, artisanat}, {artiste, artistique, artistiquement, art}**}**

Gambette, Gala & Nasr, *Corpus*, 2012

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}**}**

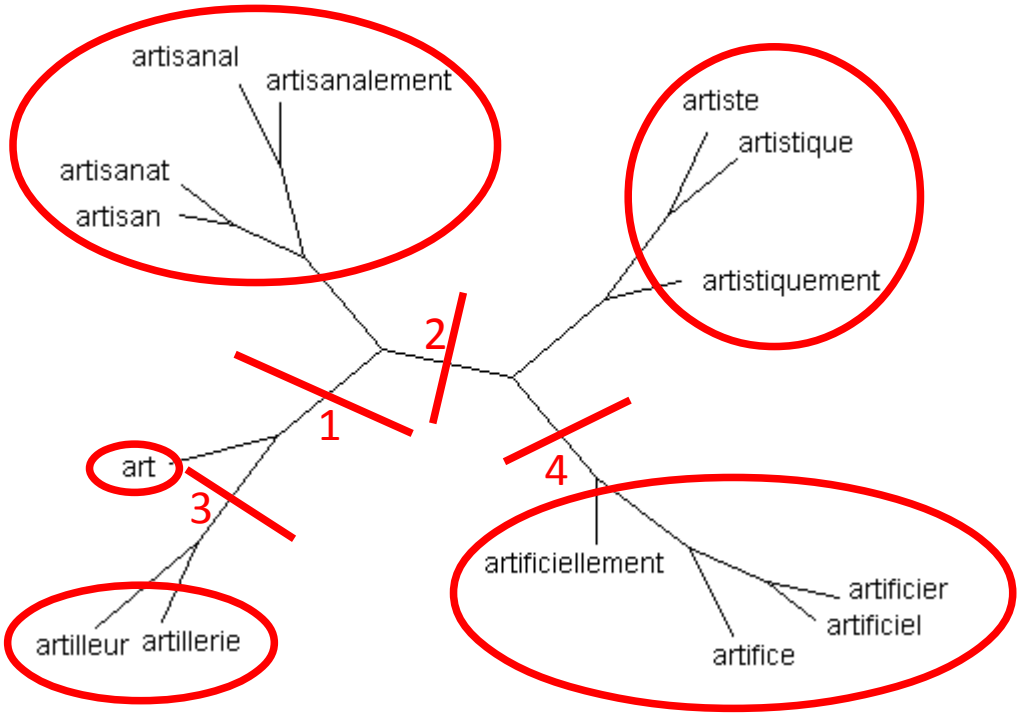
Partition manuelle : **Pm = {**{artificier, artifice, artificiel, artificiellement}, {artillerie, artilleur}, {artisan, artisanal, artisanalement, artisanat}, {artiste, artistique, artistiquement, art}**}**

Gambette, Gala & Nasr, *Corpus*, 2012

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}**}**

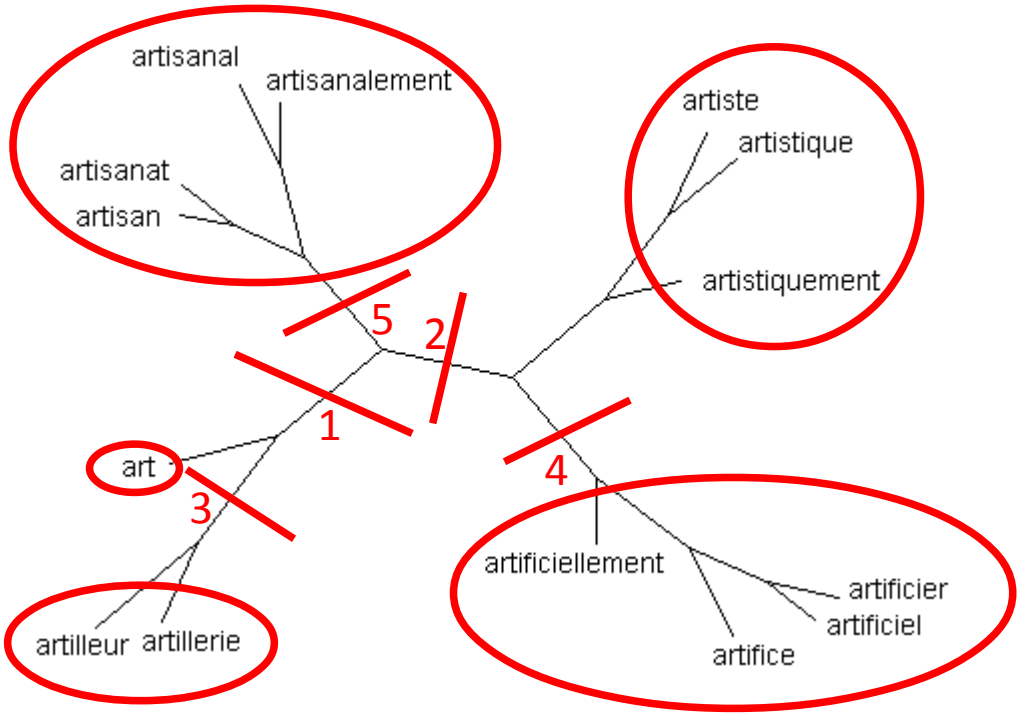
Partition manuelle : **Pm = {**{artificier, artifice, artificiel, artificiellement}, {artillerie, artilleur}, {artisan, artisanal, artisanalement, artisanat}, {artiste, artistique, artistiquement, art}**}**

Gambette, Gala & Nasr, *Corpus*, 2012

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}**}**

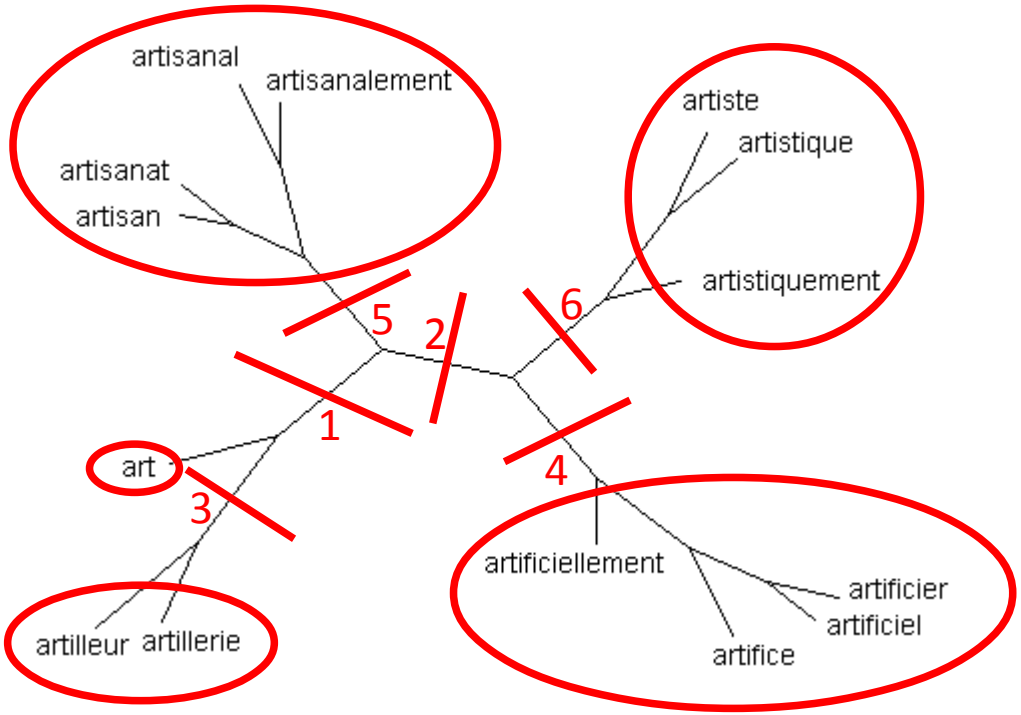
Partition manuelle : **Pm = {**{artificier, artifice, artificiel, artificiellement}, {artillerie, artilleur}, {artisan, artisanal, artisanalement, artisanat}, {artiste, artistique, artistiquement, art}**}**

Gambette, Gala & Nasr, *Corpus*, 2012

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}**}**

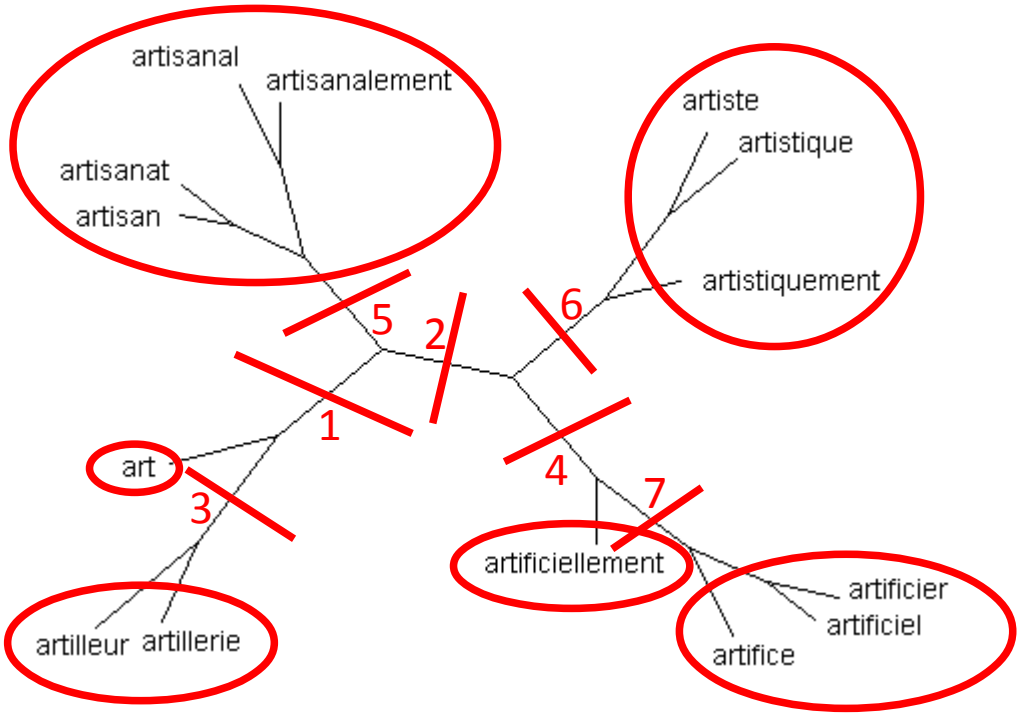
Partition manuelle : **Pm = {**{artificier, artifice, artificiel, artificiellement}, {artillerie, artilleur}, {artisan, artisanal, artisanalement, artisanat}, {artiste, artistique, artistiquement, art}**}**

Gambette, Gala & Nasr, *Corpus*, 2012

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}**}**

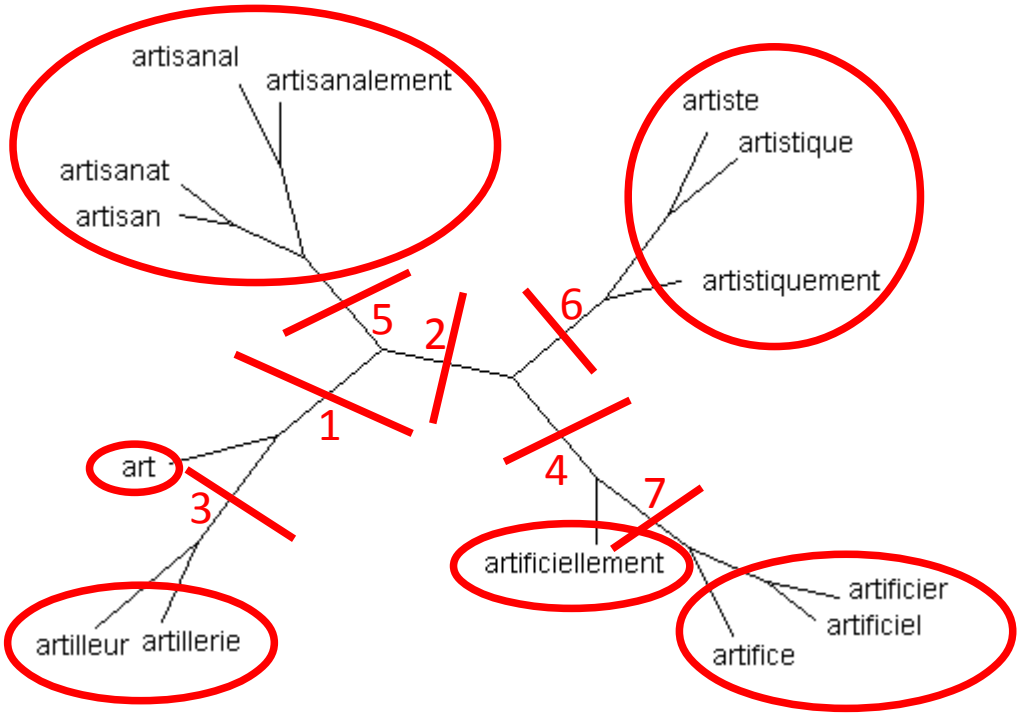
Partition manuelle : **Pm = {**{artificier, artifice, artificiel, artificiellement}, {artillerie, artilleur}, {artisan, artisanal, artisanalement, artisanat}, {artiste, artistique, artistiquement, art}**}**

Gambette, Gala & Nasr, *Corpus*, 2012

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}**}**

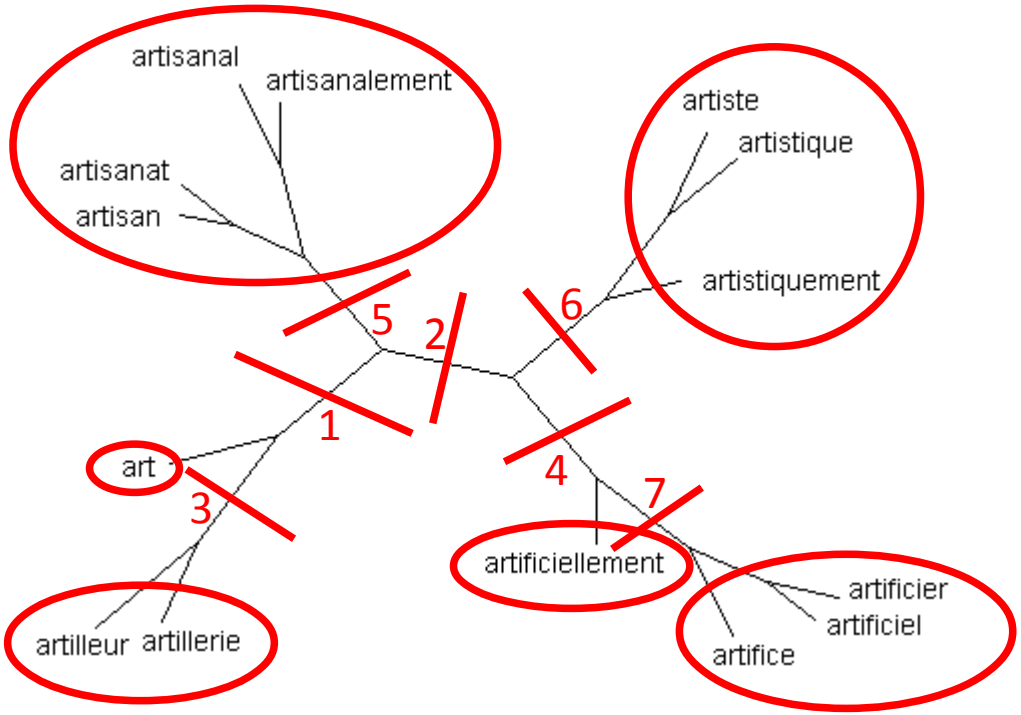
Partition manuelle : **Pm = {**{artificier, artifice, artificiel, artificiellement}, {artillerie, artilleur}, {artisan, artisanal, artisanalement, artisanat}, {artiste, artistique, artistiquement, art}**}**

Gambette, Gala & Nasr, *Corpus*, 2012

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)

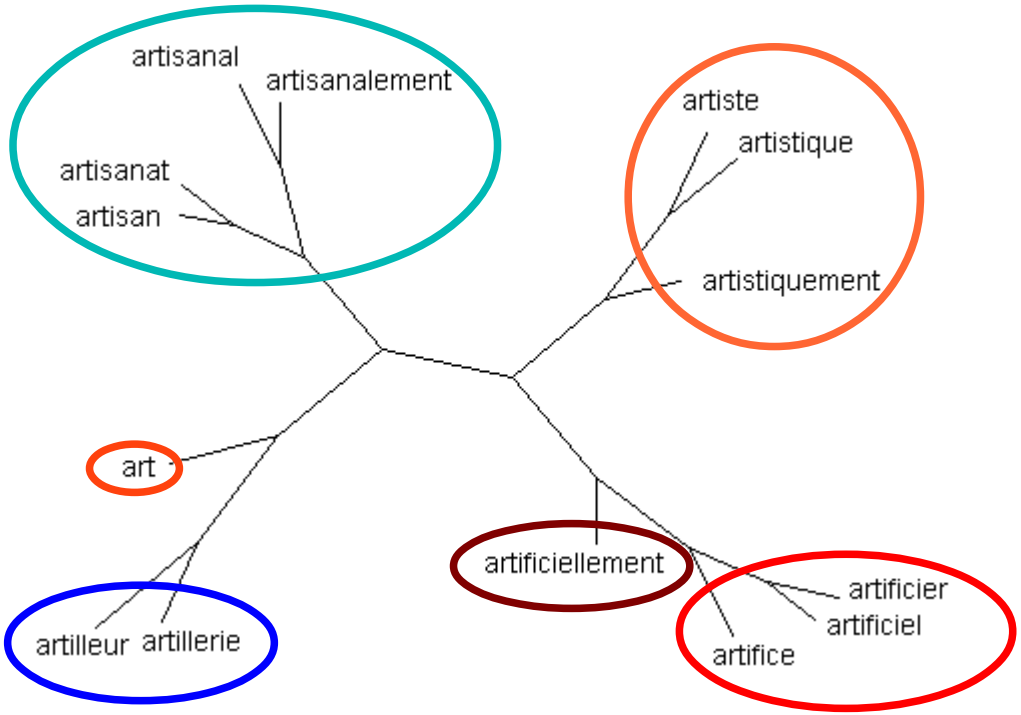
Partition manuelle : **Pm = {**{artificier, artifice, artificiel, artificiellement}, {artillerie, artilleur}, {artisan, artisanal, artisanalement, artisanat}, {artiste, artistique, artistiquement, art}**}**

Gambette, Gala & Nasr, *Corpus*, 2012

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)

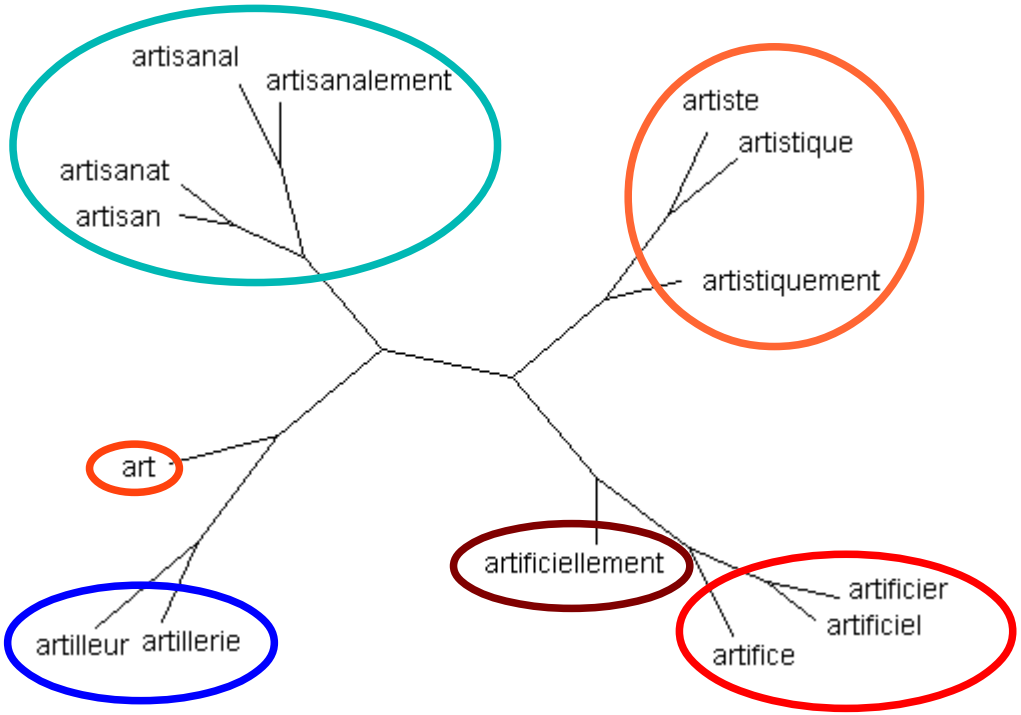
Partition manuelle : **Pm = {**{artificier, artifice, artificiel, artificiellement}, {artillerie, artilleur}, {artisan, artisanal, artisanalement, artisanat}, {artiste, artistique, artistiquement, art}**}**

Gambette, Gala & Nasr, *Corpus*, 2012

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

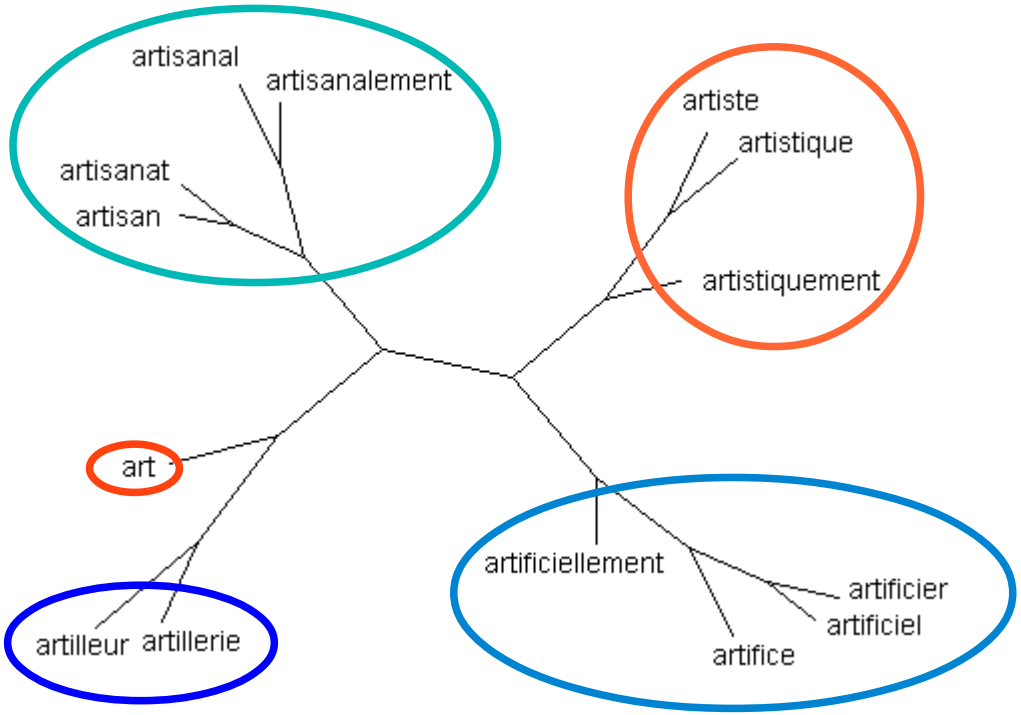
Partition manuelle : **Pm = {**{artificier, artifice, artificiel, artificiellement}, {artillerie, artilleur}, {artisan, artisanal, artisanalement, artisanat}, {artiste, artistique, artistiquement, art}**}**

Gambette, Gala & Nasr, *Corpus*, 2012

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” ?



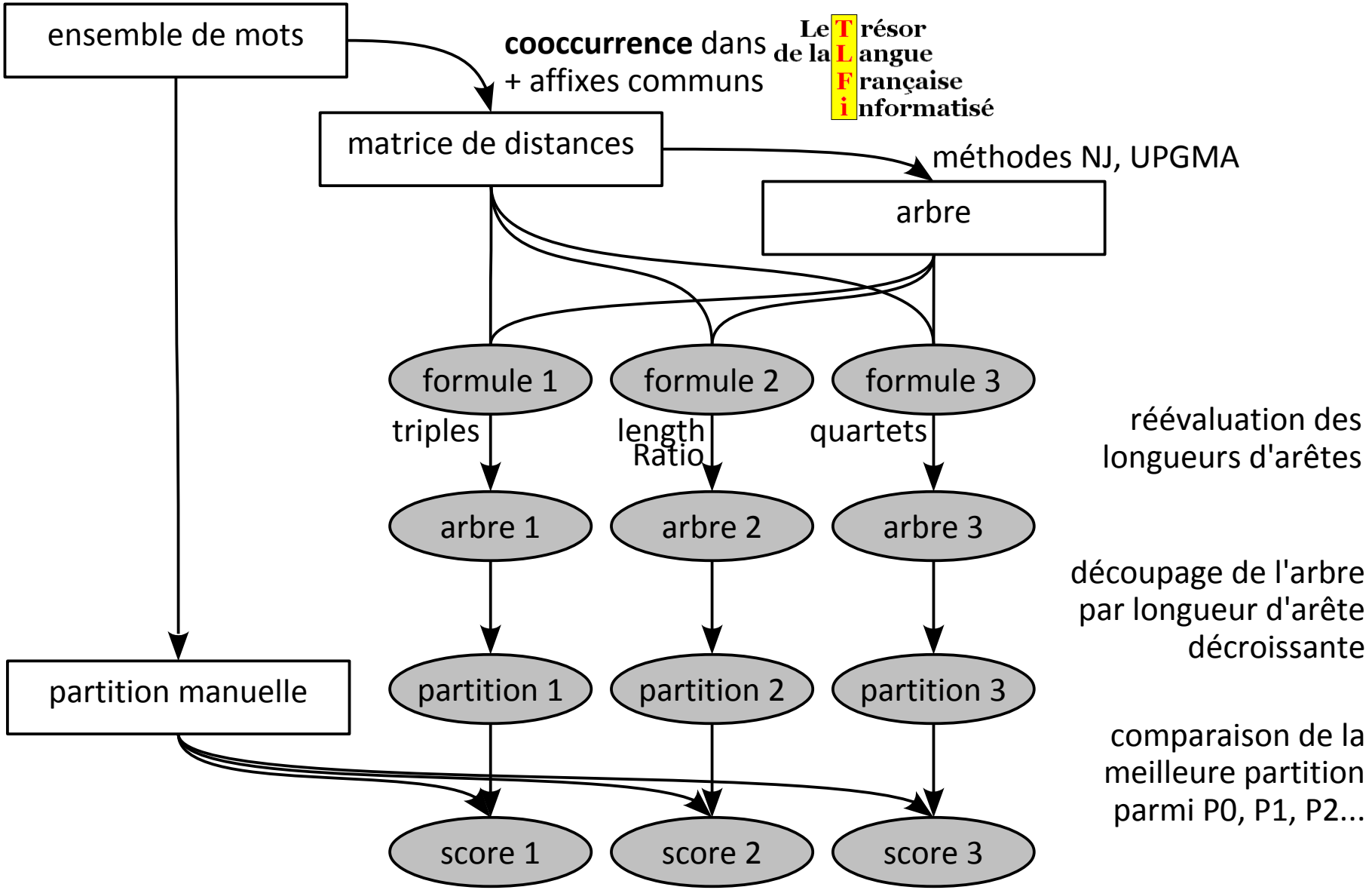
Partition automatique :
P4 = {{artisan, artisanat, artisanal, artisanalement}, {artiste, artistique, artistiquement}, {artificier, artificiel, artifice, artificiellement}, {artillerie, artilleur}, {art}**}**

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}**}**

Gambette, Gala & Nasr, *Corpus*, 2012

Protocole d'évaluation



Gambette, Gala & Nasr, *Corpus*, 2012

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 (2013)
L'affaire du Médiator au prisme de la textométrie,
Texto! XVIII(4)
http://www.revue-texto.net/index.php?id=3318

Co-auteurs des travaux en cours :

- Edna Hernandez : méthodologie d'utilisation de TreeCloud pour les analyses exploratoires
- Xavier Le Roux, Hilde Eggermont : analyse du corpus de projets sur la biodiversité
- Claude Martineau : intégration de prétraitements Unitex dans TreeCloud
- Paola Salle : extraction et analyse du corpus d'offres d'emploi de l'APEC
- Deepak Srinivas : visualisation avec bibliothèque d3.js