Module de formation doctorale en informatique textuelle

Séance 4 - De la lexicométrie au traitement automatique des langues (TAL)

06/03/2021 – UPEC, Créteil

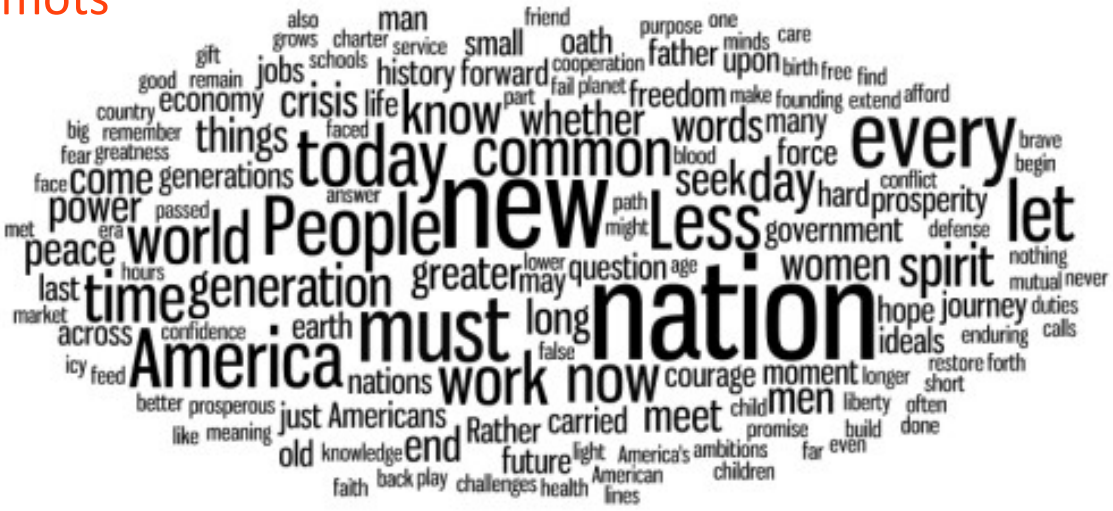
TreeCloud pour la visualisation et l'analyse de données textuelles

Philippe Gambette

LIGM, Université Gustave Eiffel

Le nuage de mots pour résumer un discours

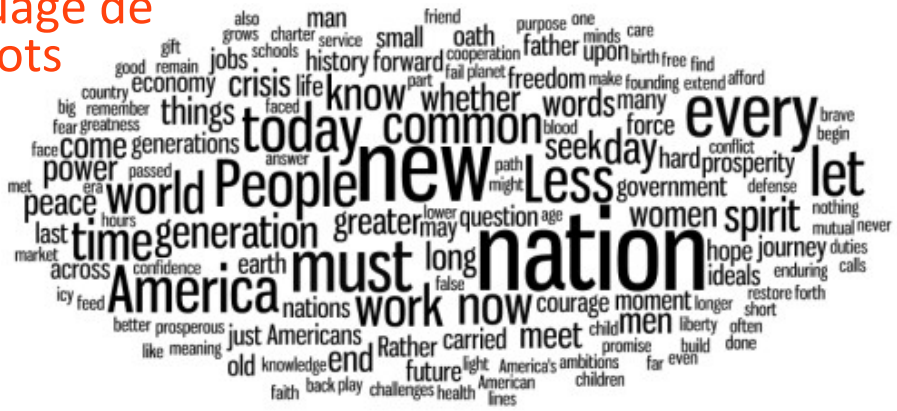
nuage de mots



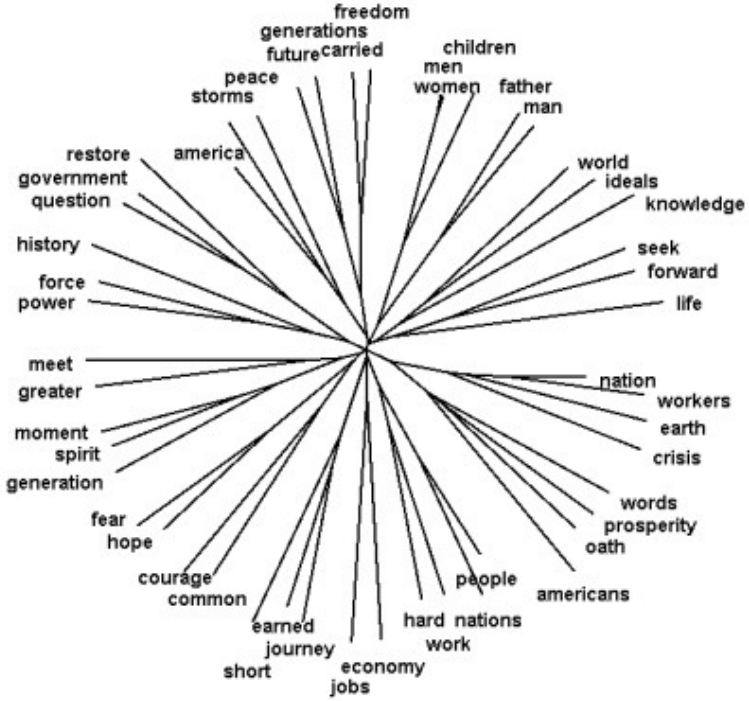
Discours inaugural de Barack Obama en 2008, Wordle

Le « nuage arboré », une information double

nuage de mots

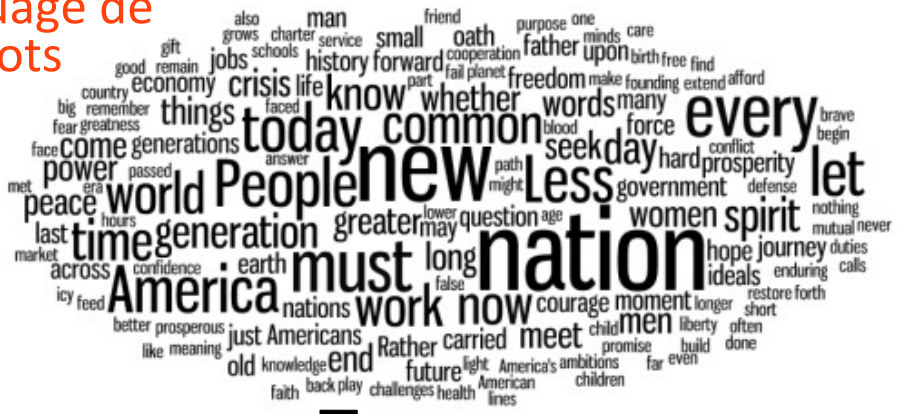


arbre de mots

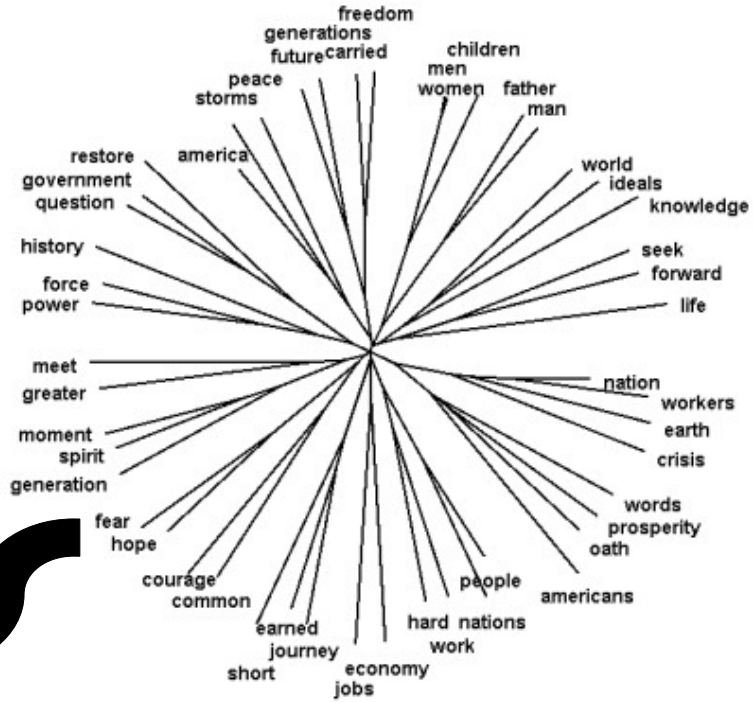


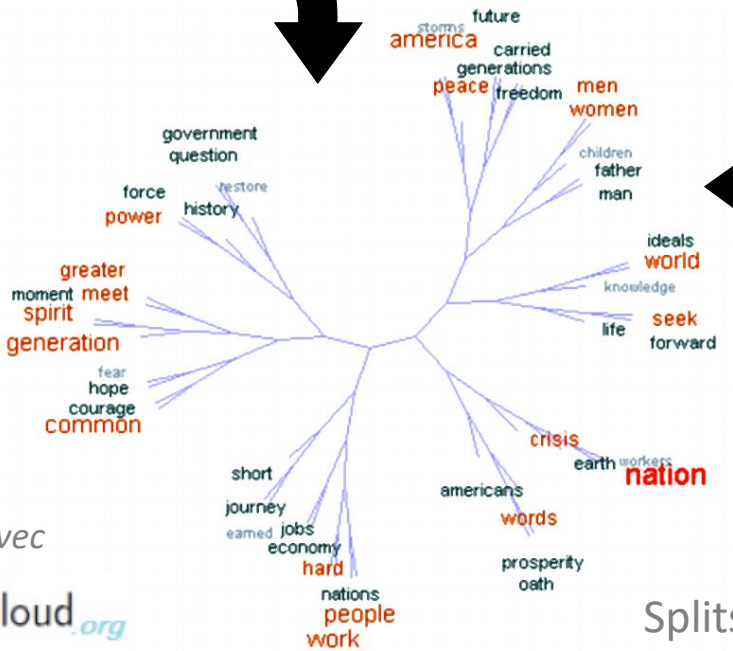
Le « nuage arboré », une information double

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*

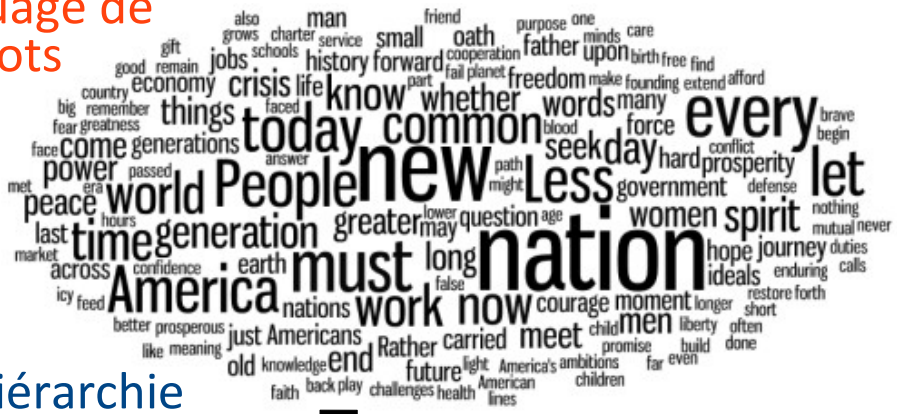
Le « nuage arboré », une information double

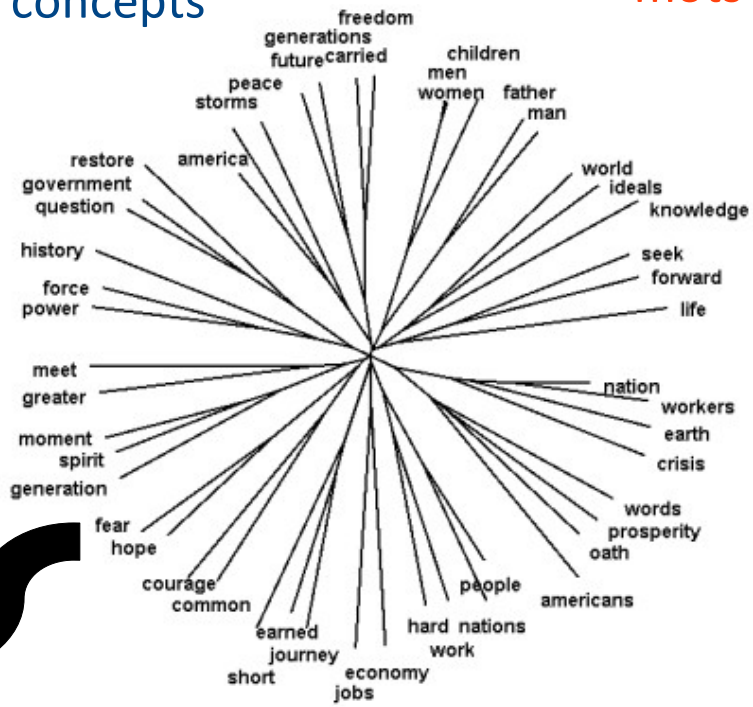
nuage de mots

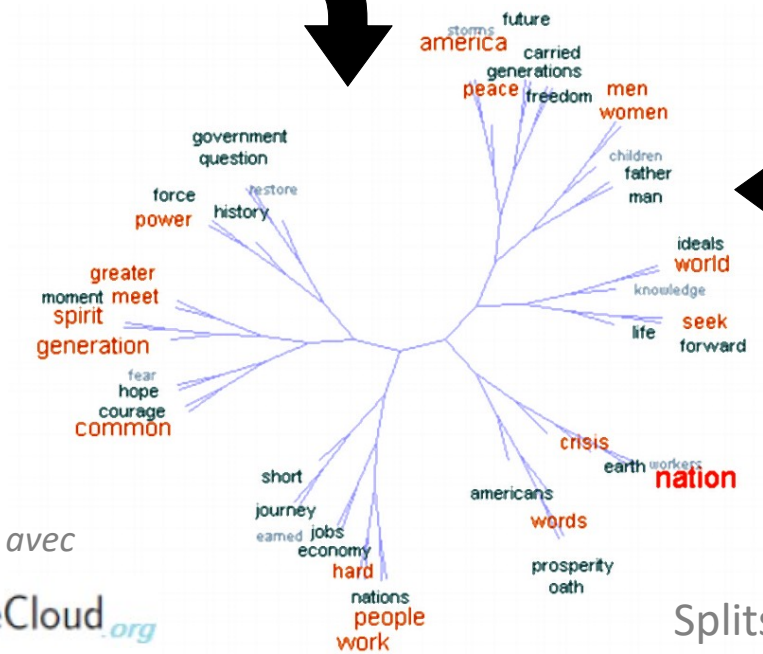
hiérarchie des mots

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*

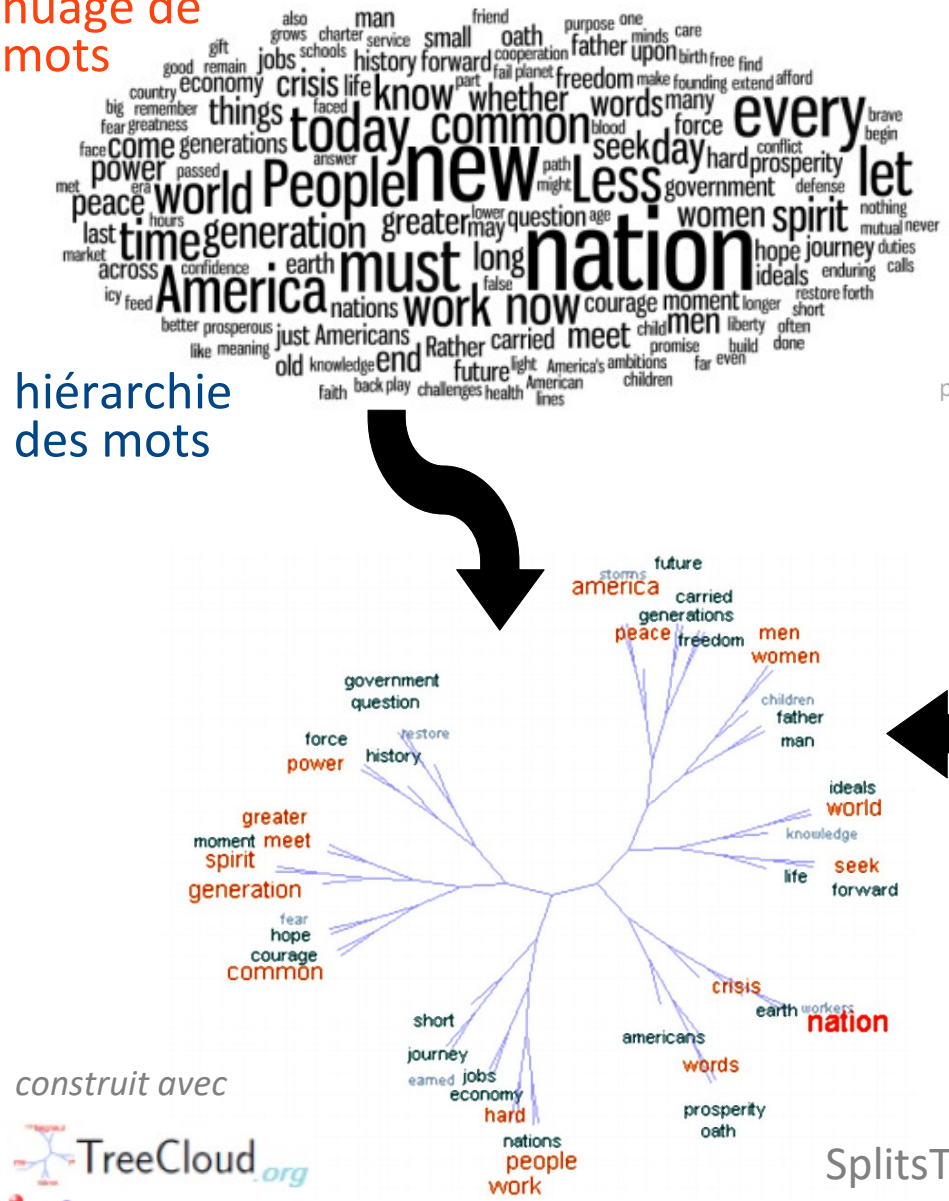
Le « nuage arboré », une information double

nuage de mots

hiérarchie des mots

hiérarchie des concepts

arbre de mots



Discours inaugural de Barack Obama

construit avec

TreeCloud.org

SplitsTree4

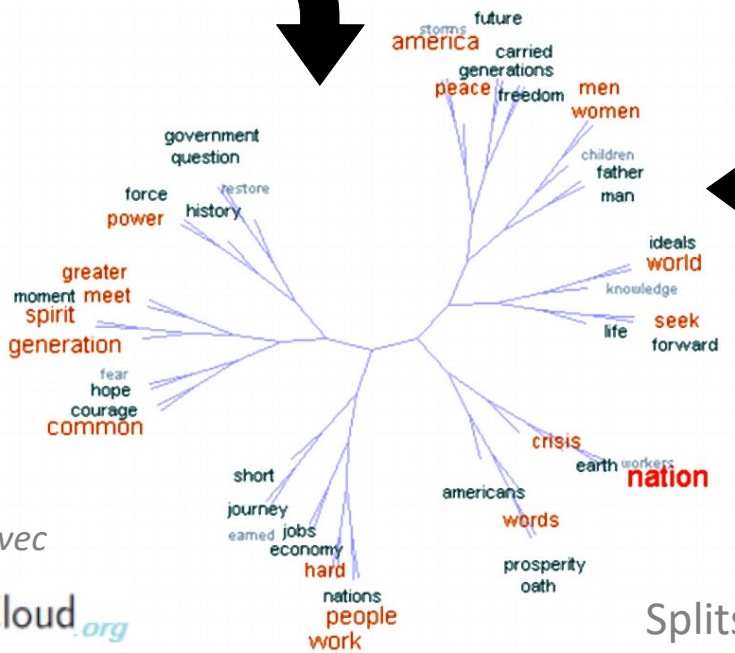
SplitsTree : Huson & Bryant, *Bioinformatics*, 2006
TreeCloud : Gambette & Véronis, *IFCS'09*

Le « nuage arboré », une information double

nuage de mots

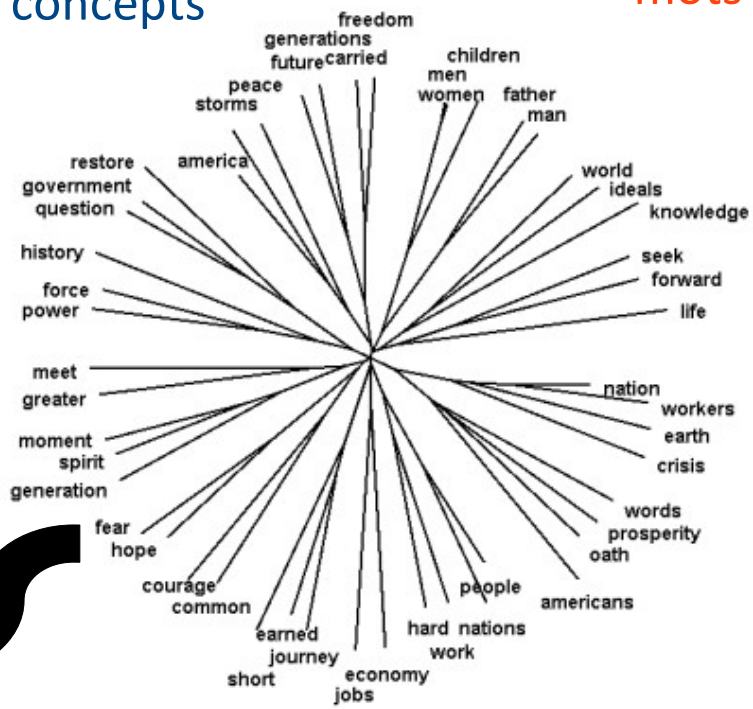
hiérarchie des mots

occurrences



hiérarchie des concepts

arbre de mots



cooccurrences

Discours inaugural de Barack Obama

construit avec

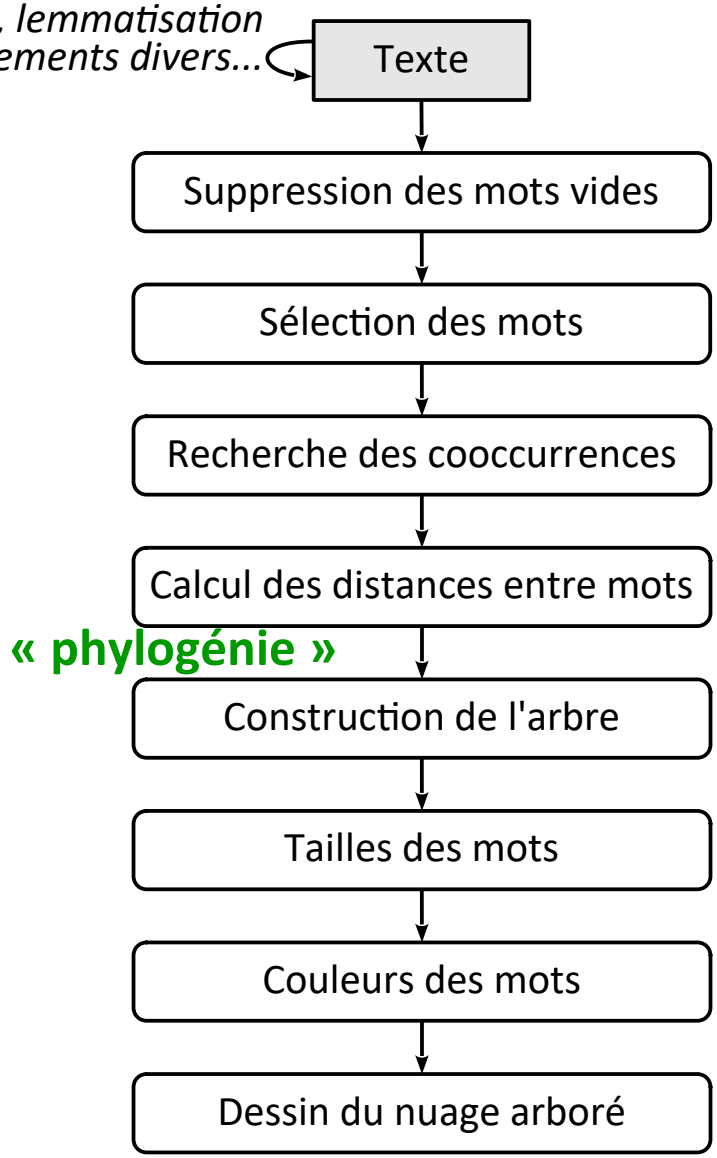
TreeCloud.org

SplitsTree4

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

Processus de construction

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



Proposé dans la version téléchargeable de TreeCloud

antidico anglais, français

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

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

12 formules de distance de cooccurrence

Appel transparent au logiciel SplitsTree

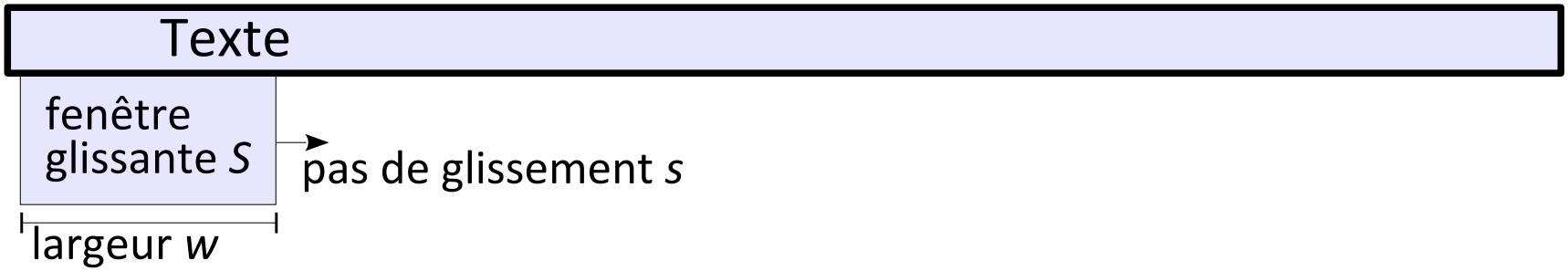
Fréquences ou valeurs personnalisées

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

Appel transparent au logiciel SplitsTree ou Dendroscope

Calcul des proximités entre mots

Déplacement d'une « fenêtre glissante » tout au long du texte pour compter, pour chaque paire de mots u et v, leurs cooccurrences :



Nombre de cooccurrences entre *u* et *v*

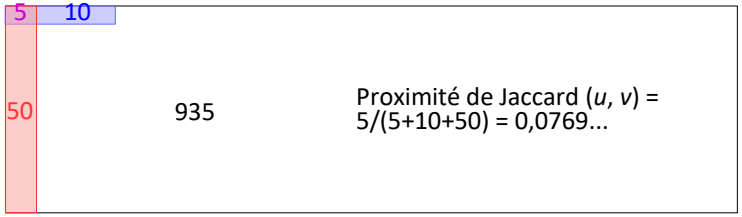
nombre de fenêtres	mot *u* dans *S*	mot *u* **pas** dans *S*
mot *v* dans *S*	5	10
mot *v* **pas** dans *S*	50	935

Exemple : texte de 991 mots avec 55 occurrences de *u* et 15 de *v*, *w*=10, *s*=1

Score de cooccurrence entre *u* et *v*

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



Arbres phylogénétiques et arbres de mots

ESPÈCES

Séquences ADN

Données sur les feuilles

MOTS

Position des mots

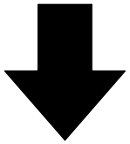
Arbres phylogénétiques et arbres de mots

ESPÈCES

Séquences ADN

Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

Données sur les feuilles



Distances entre les feuilles

	A	B	C	D
A	0	2	5	6
B	2	0	5	6
C	5	5	0	3
D	6	6	3	0

MOTS

Position des mots

Distances fondées sur la cooccurrence entre les deux mots

Arbres phylogénétiques et arbres de mots

ESPÈCES

Séquences ADN

Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

Données sur les feuilles

Distances entre les feuilles

	A	B	C	D
A	0	**2**	5	6
B	**2**	0	5	6
C	5	5	0	3
D	6	6	3	0

classification hiérarchique ascendante algorithme UPGMA

Arbre

MOTS

Position des mots

Distances fondées sur la cooccurrence entre les deux mots

Algorithme UPGMA : https://fr.wikipedia.org/wiki/Unweighted_pair_group_method_with_arithmetic_mean

Arbres phylogénétiques et arbres de mots

ESPÈCES

Séquences ADN

Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

Données sur les feuilles

↓

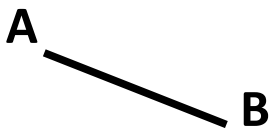
Distances entre les feuilles

	A+B	C	D
A+B	**0**	**5**	**6**
C	**5**	0	3
D	**6**	3	0

↓

classification hiérarchique ascendante
algorithme UPGMA

Arbre



MOTS

Position des mots

Distances fondées sur la cooccurrence entre les deux mots

Algorithme UPGMA : https://fr.wikipedia.org/wiki/Unweighted_pair_group_method_with_arithmetic_mean

Arbres phylogénétiques et arbres de mots

ESPÈCES

Séquences ADN

Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

Données sur les feuilles

↓

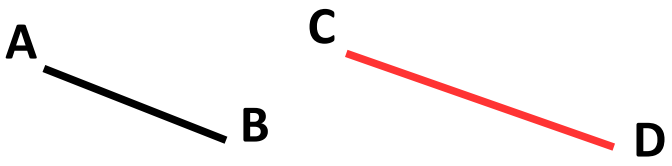
Distances entre les feuilles

	A+B	C	D
A+B	0	5	6
C	5	0	**3**
D	6	**3**	0

↓

classification hiérarchique ascendante
algorithme UPGMA

Arbre



MOTS

Position des mots

Distances fondées sur la cooccurrence entre les deux mots

Algorithme UPGMA : https://fr.wikipedia.org/wiki/Unweighted_pair_group_method_with_arithmetic_mean

Arbres phylogénétiques et arbres de mots

ESPÈCES

Séquences ADN

Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

Données sur les feuilles

↓

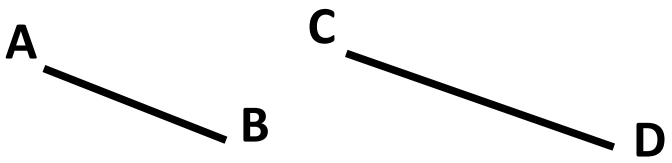
Distances entre les feuilles

	A+B	C+D
A+B	0	5,5
C+D	5,5	0

↓

classification hiérarchique ascendante
algorithme UPGMA

Arbre



MOTS

Position des mots

Distances fondées sur la cooccurrence entre les deux mots

Algorithme UPGMA : https://fr.wikipedia.org/wiki/Unweighted_pair_group_method_with_arithmetic_mean

Arbres phylogénétiques et arbres de mots

ESPÈCES

Séquences ADN

Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

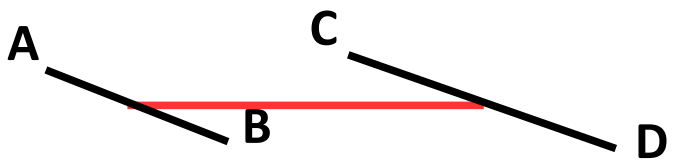
Données sur les feuilles

Distances entre les feuilles

	A+B	C+D
A+B	0	**5,5**
C+D	**5,5**	0

classification hiérarchique ascendante
algorithme UPGMA

Arbre



MOTS

Position des mots

Distances fondées sur la cooccurrence entre les deux mots

Algorithme UPGMA : https://fr.wikipedia.org/wiki/Unweighted_pair_group_method_with_arithmetic_mean

Arbres phylogénétiques et arbres de mots

ESPÈCES

Séquences ADN

Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

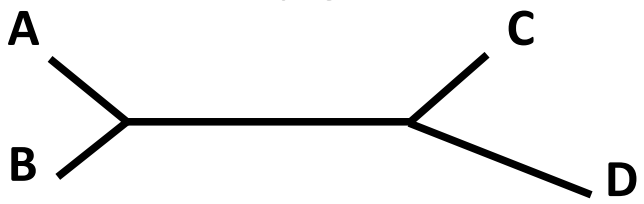
Données sur les feuilles

Distances entre les feuilles

	A	B	C	D
A	0	2	5	6
B	2	0	5	6
C	5	5	0	3
D	6	6	3	0

classification hiérarchique ascendante
algorithme UPGMA

Arbre



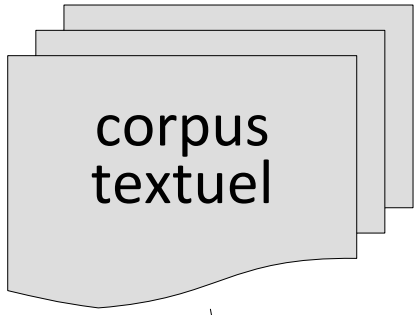
MOTS

Position des mots

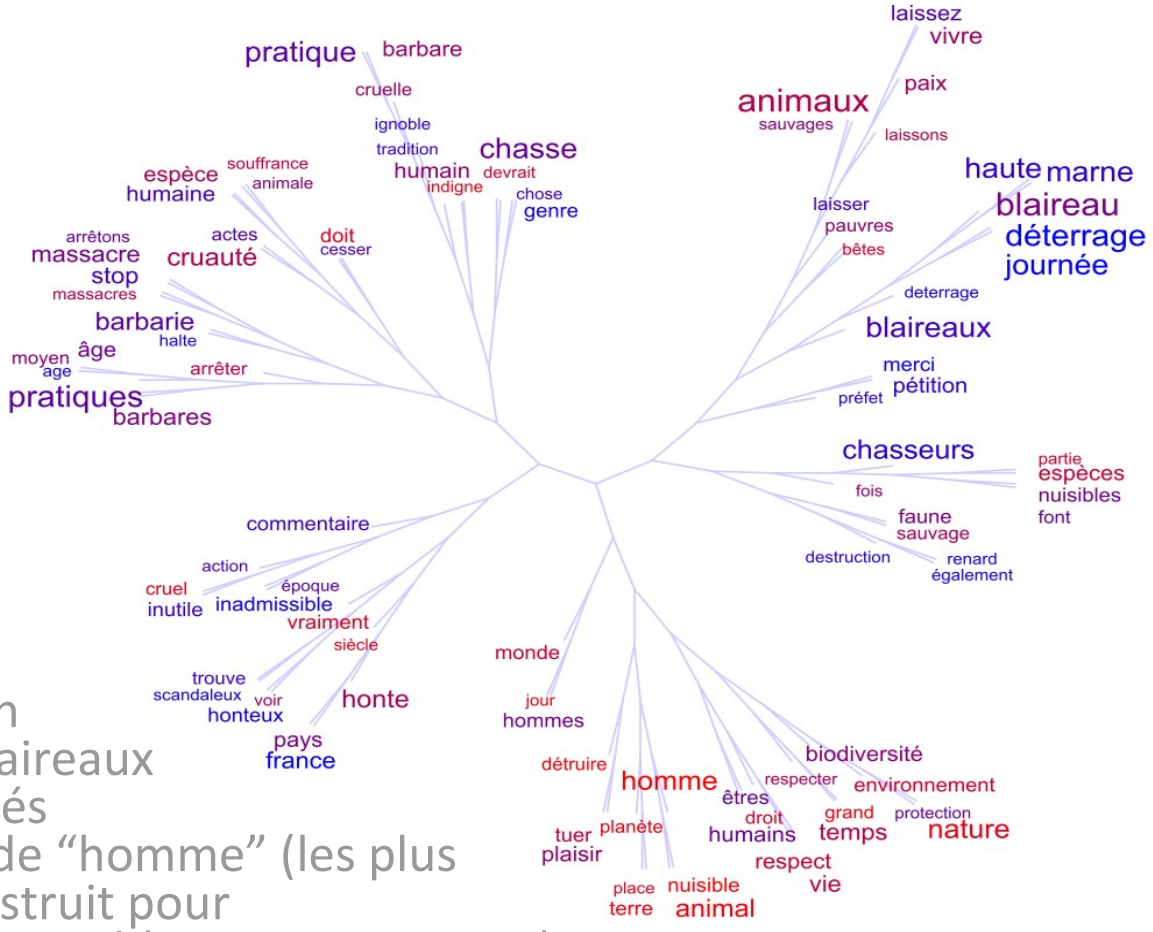
Distances fondées sur la cooccurrence entre les deux mots

Algorithme UPGMA : https://fr.wikipedia.org/wiki/Unweighted_pair_group_method_with_arithmetic_mean

Le « nuage arboré », pour quoi faire ?

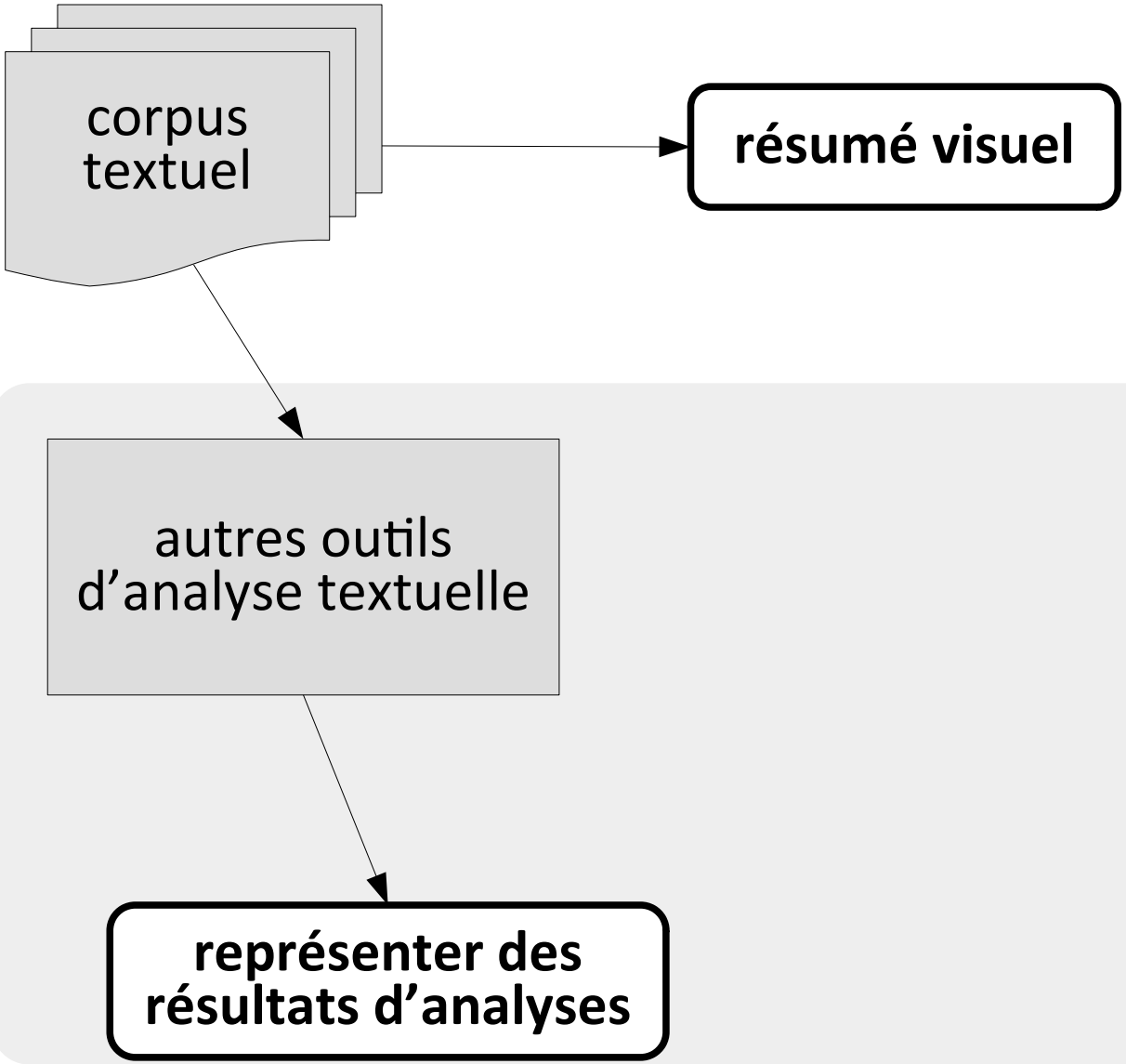


résumé visuel



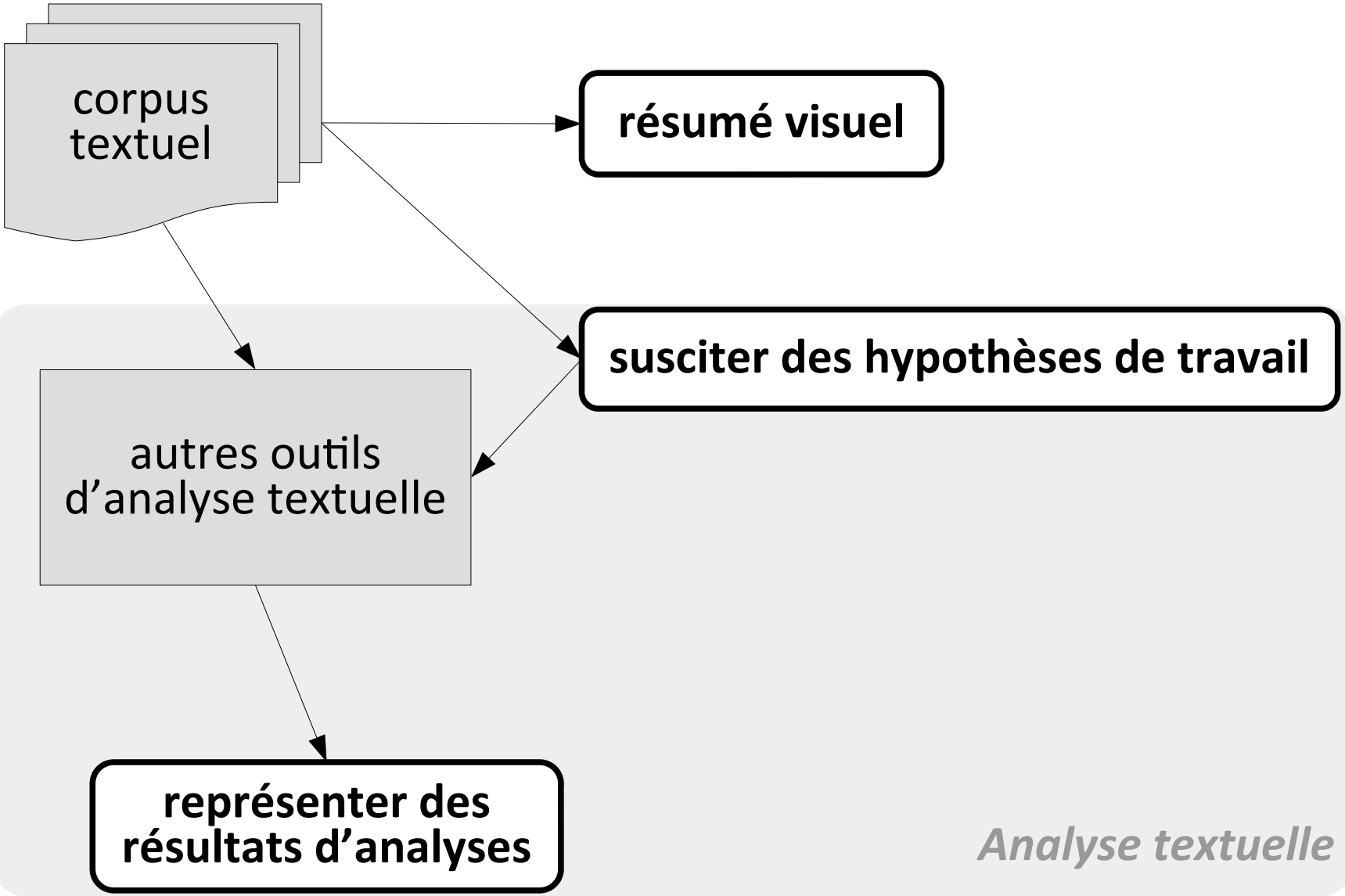
Nuage arboré des 100 mots les plus fréquents des commentaires d'une pétition contre l'extermination de blaireaux sur *lapetition.be*, mots colorés en fonction de la proximité de “homme” (les plus cooccurrents en rouge), construit pour R. Matuszewicz et M. Legris-Revel (projet ANR APPEL)

Le « nuage arboré », pour quoi faire ?



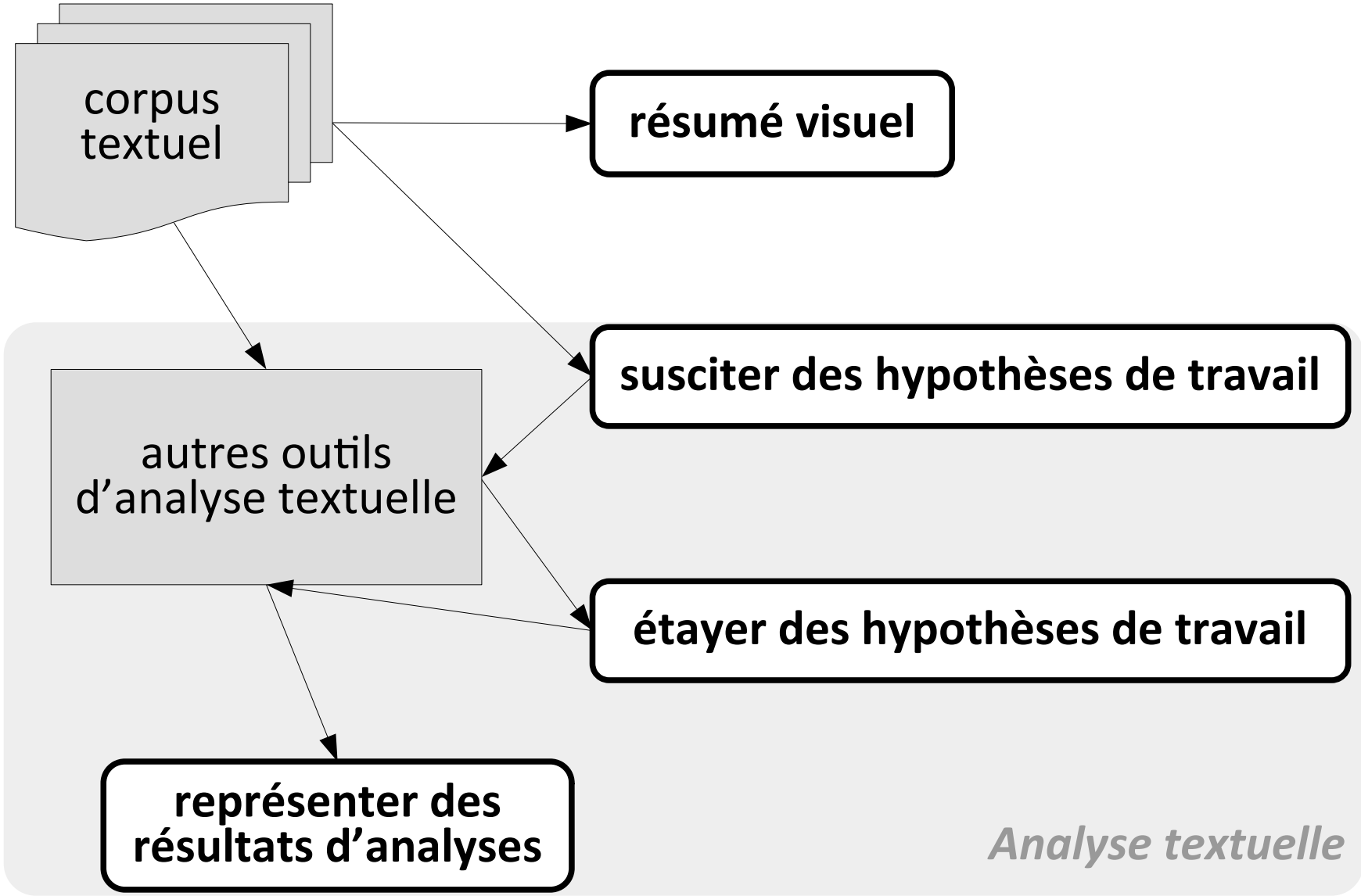
Amstutz & Gambette, *JADT 2010*

Le « nuage arboré », pour quoi faire ?



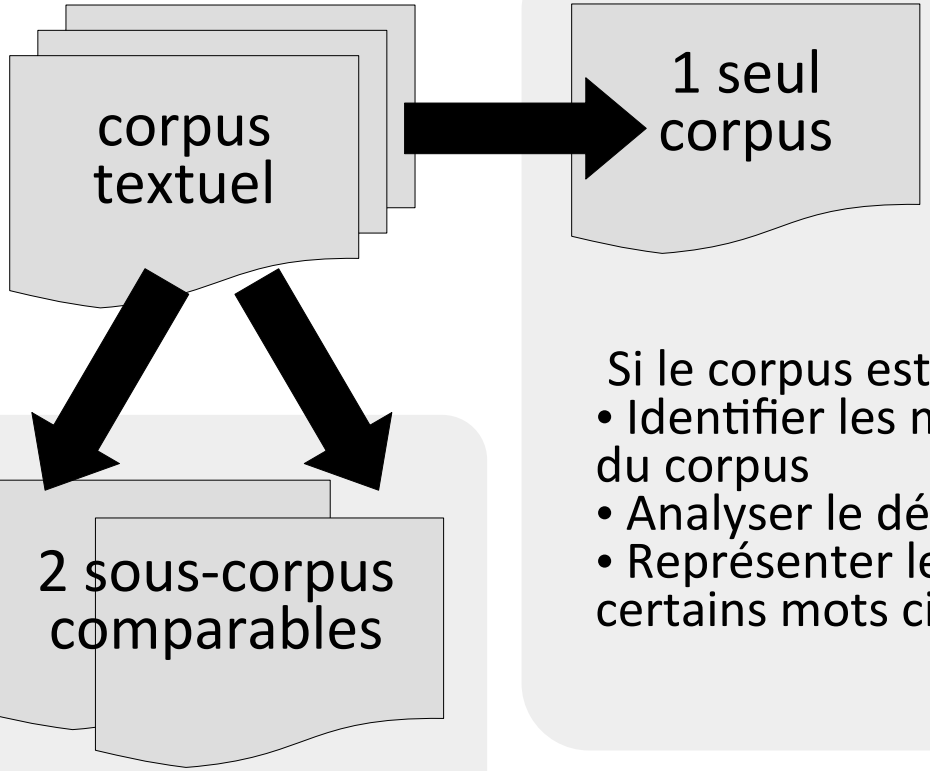
Amstutz & Gambette, *JADT 2010*

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.)

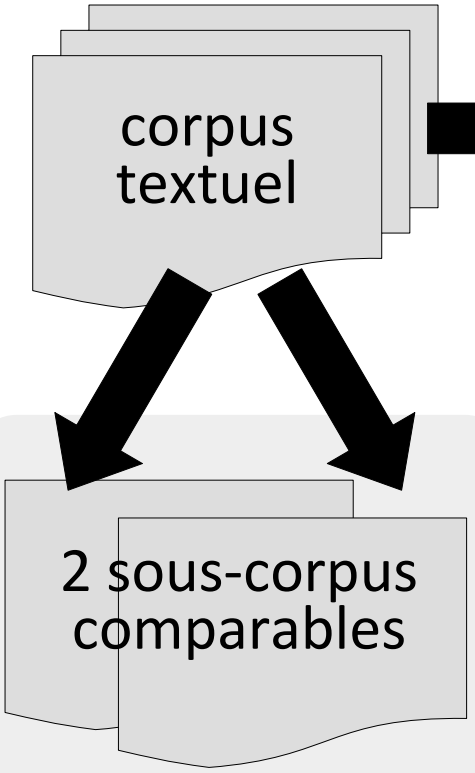
Si le corpus est assez grand :

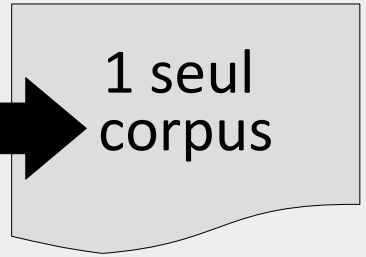
- Identifier les mots plutôt au début ou à la fin du corpus
- Analyser le début de chaque phrase du corpus
- Représenter le contexte dans le corpus de certains mots ciblés

- Identifier des regroupements communs ou différents
- Comparer les voisinages d'un mot cible dans les deux corpus
- Comparer (quanti+quali) les regroupements dans les deux arbres
- Construire le nuage arboré des mots spécifiques de chaque sous-corpus

Travail mené 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
- Analyser le début de chaque phrase du corpus
- Représenter le contexte dans le corpus de certains mots ciblés

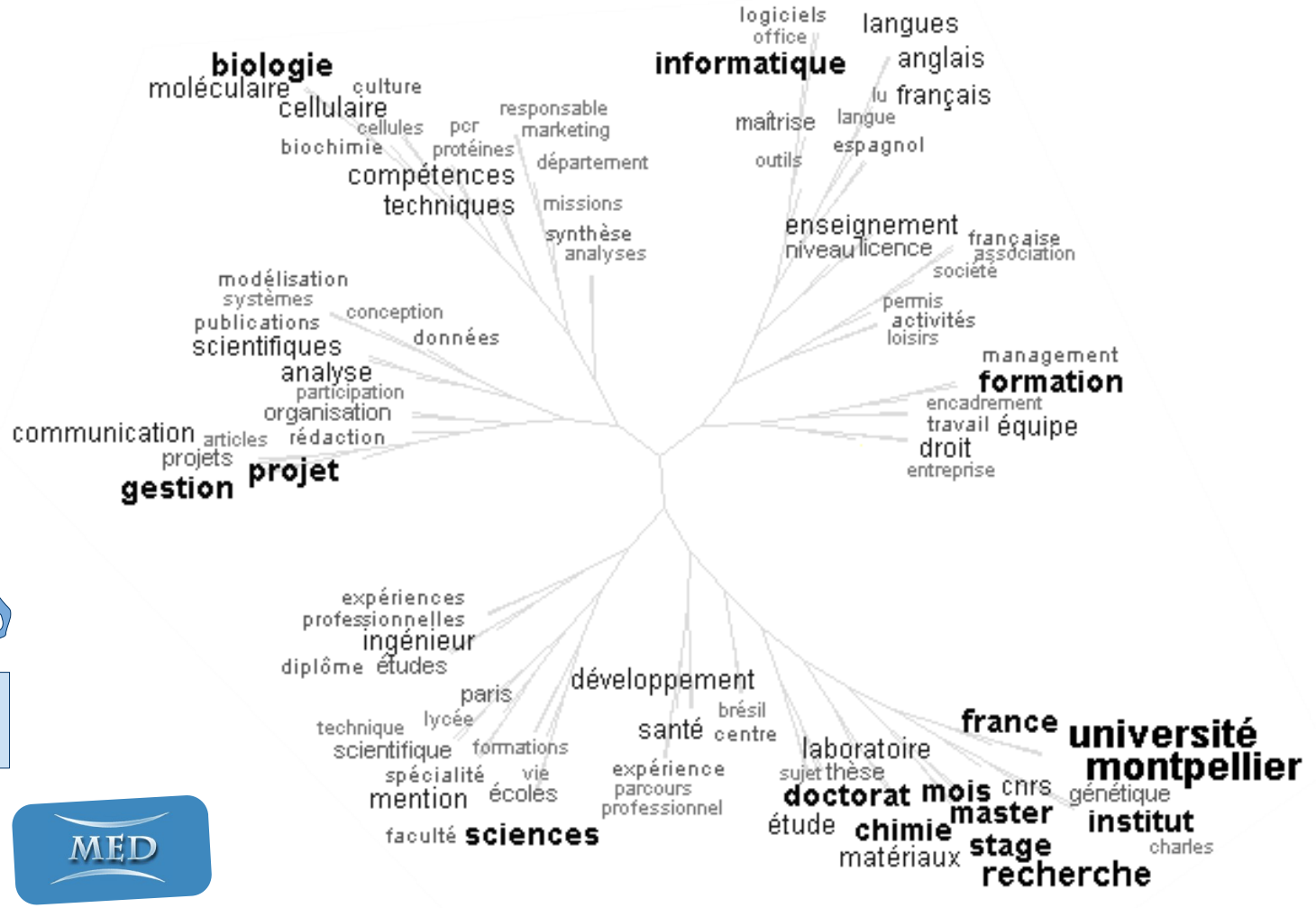
- Identifier des regroupements communs ou différents
- Comparer les voisinages d'un mot cible dans les deux corpus
- Comparer (quanti+quali) les regroupements dans les deux arbres
- Construire le nuage arboré des mots spécifiques de chaque sous-corpus

Travail mené avec Edna Hernandez

Méthode : interpréter les regroupements

Dessiner des « patates »

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



Rencontre Docteurs & Entreprises

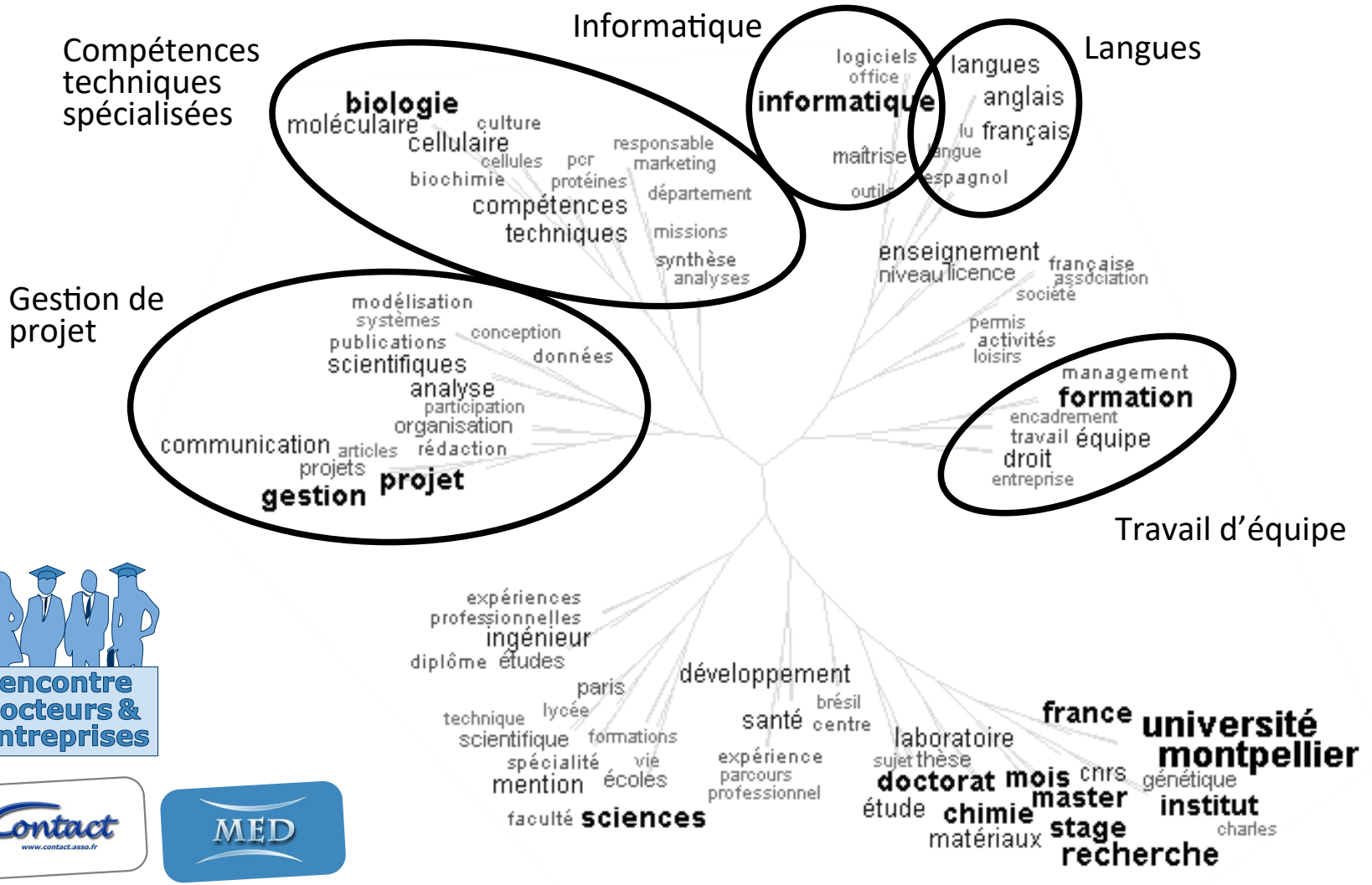
Contact
www.contact.asso.fr

MED

Méthode : interpréter les regroupements

Dessiner des « patates »

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

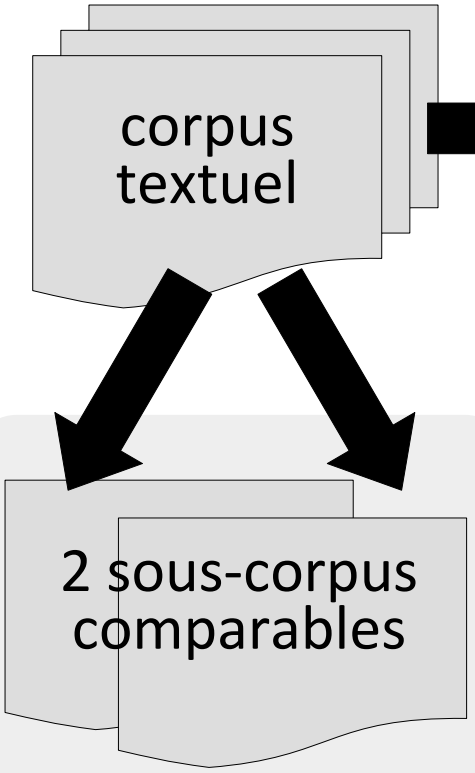


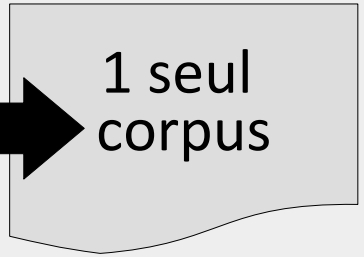
Rencontre Docteurs & Entreprises

Contact www.contact.asso.fr

MED

Exploration de corpus avec TreeCloud





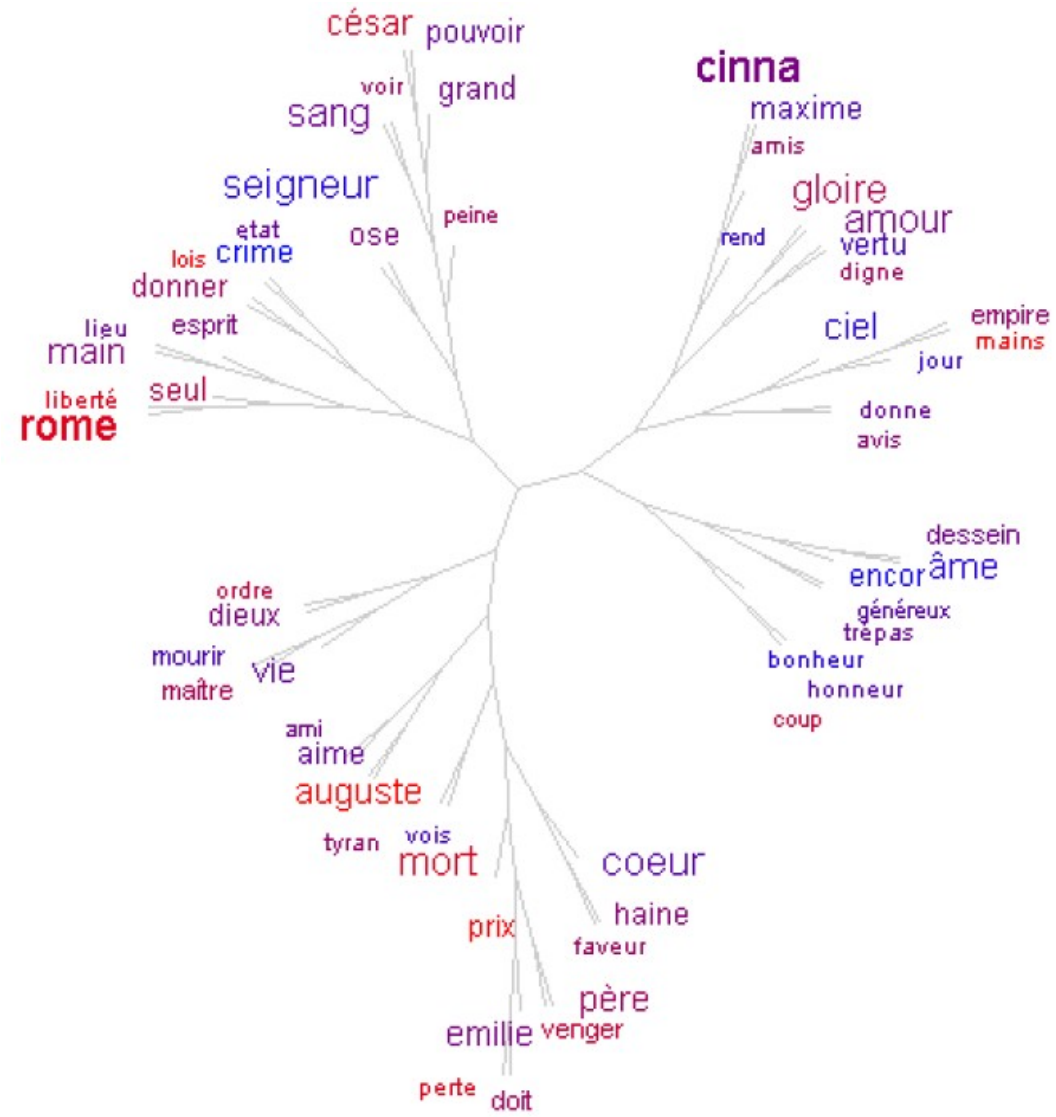
- Interpréter les regroupements
- **Examiner les voisinages des mots fréquents dans l'arbre**
- Examiner le voisinage de mots ciblés dans l'arbre («acteurs», verbes, etc.)

Si le corpus est assez grand :
- Identifier les mots plutôt au début ou à la fin du corpus
- Analyser le début de chaque phrase du corpus
- Représenter le contexte dans le corpus de certains mots ciblés

- Identifier des regroupements communs ou différents
- Comparer les voisinages d'un mot cible dans les deux corpus
- Comparer (quanti+quali) les regroupements dans les deux arbres
- Construire le nuage arboré des mots spécifiques de chaque sous-corpus

Travail mené avec Edna Hernandez

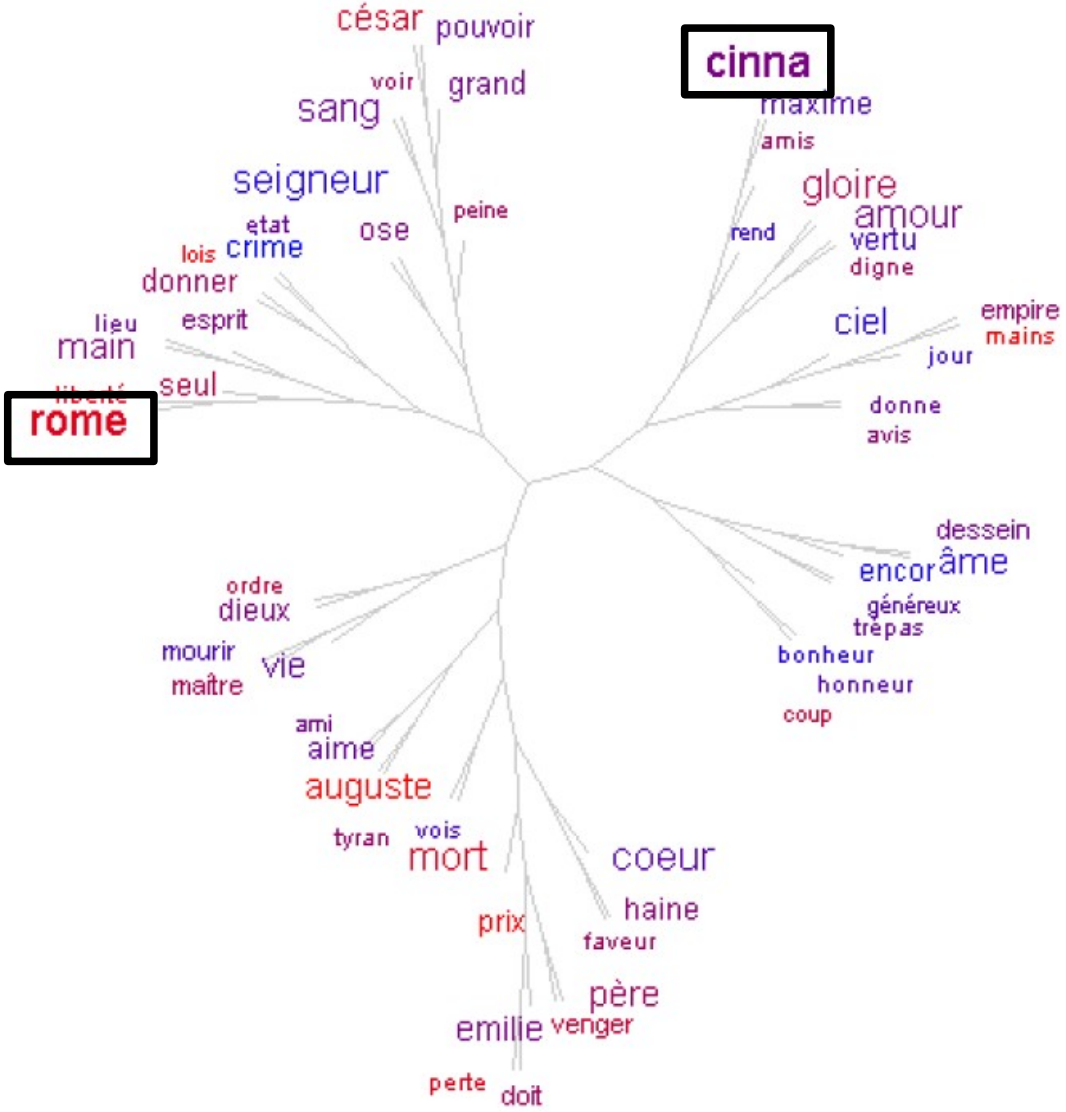
Méthode : voisinage des mots fréquents



Amstutz & Gambette, JADT 2010

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

Méthode : voisinage des mots fréquents

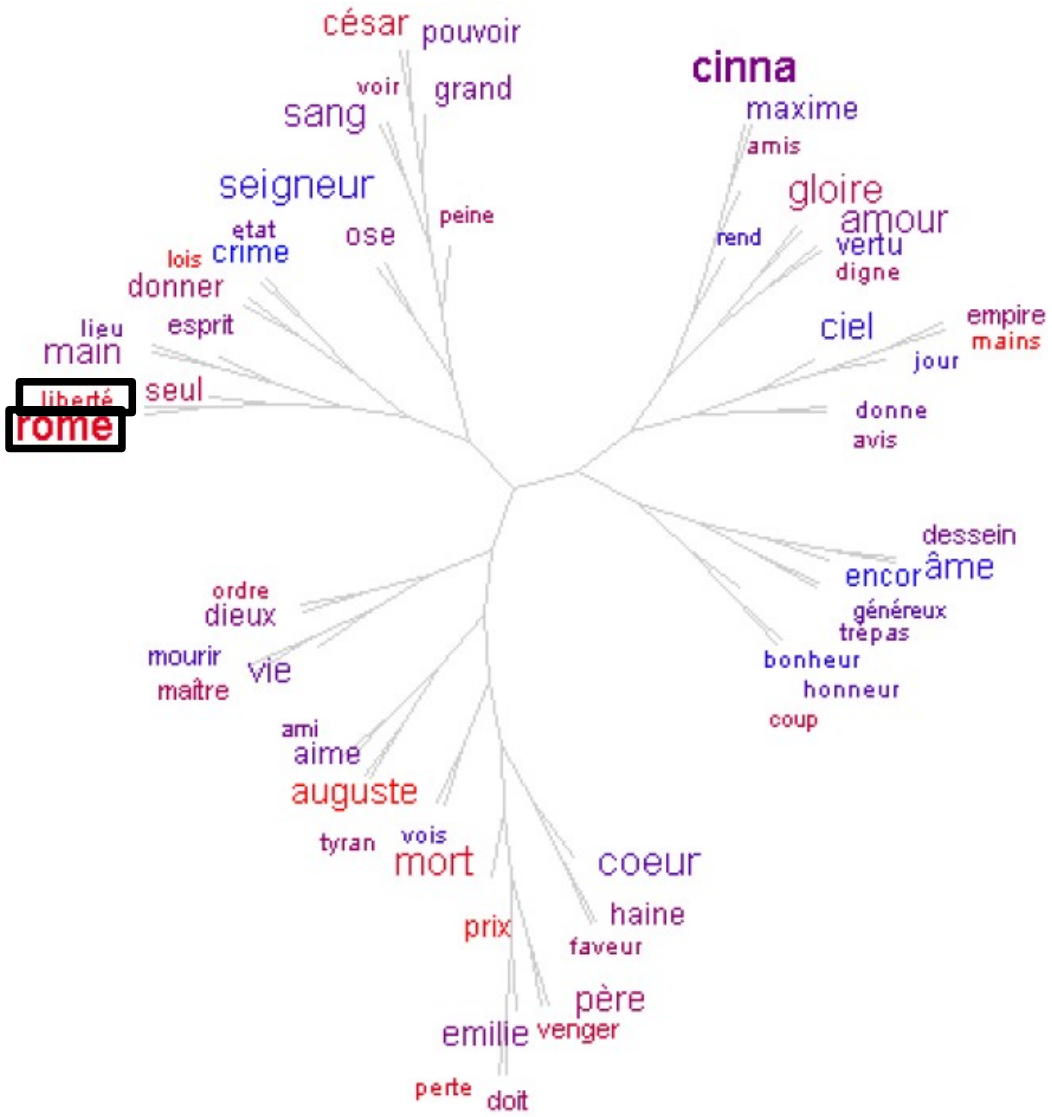


Amstutz & Gambette, JADT 2010

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

Méthode : voisinage des mots fréquents

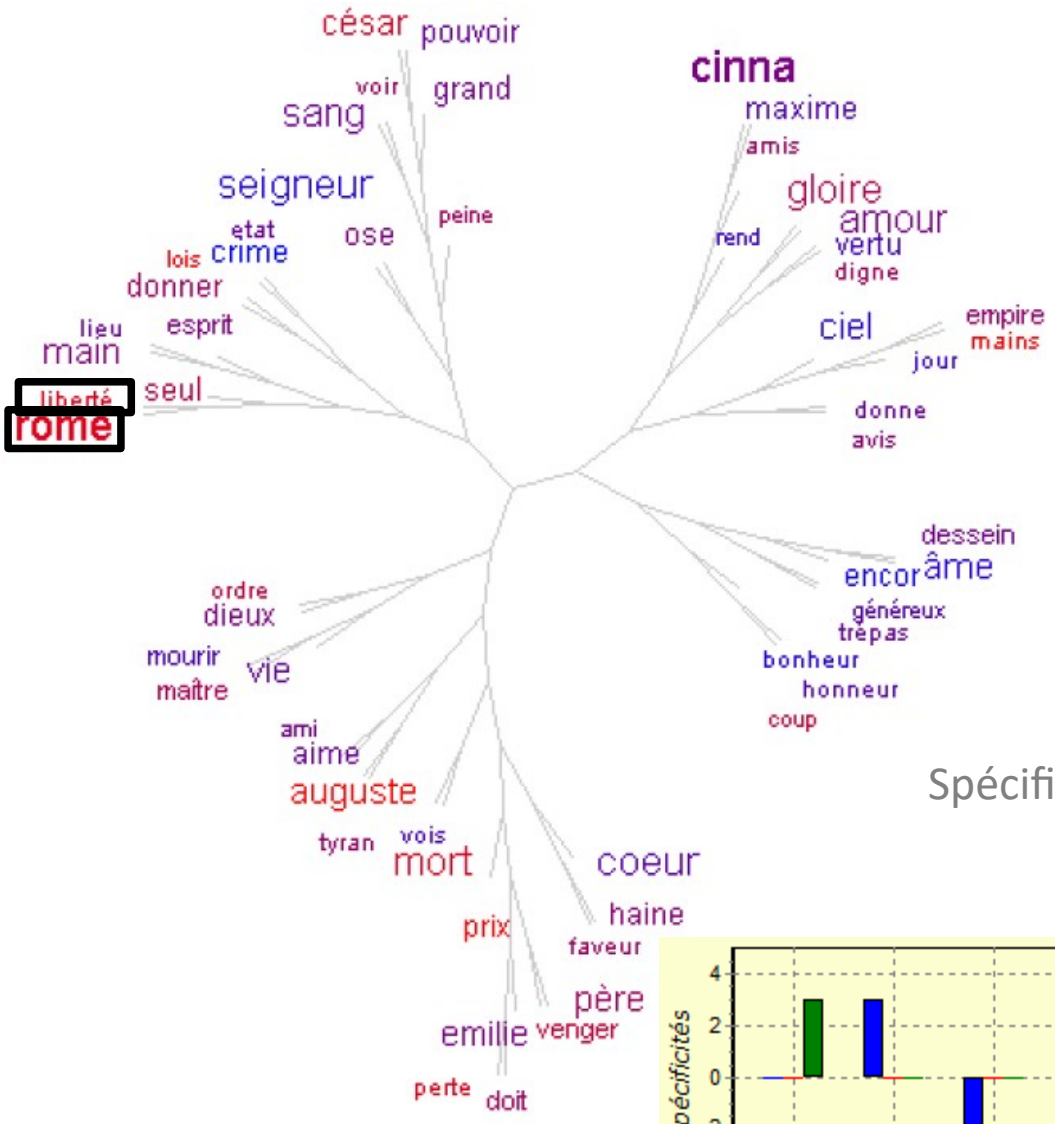
Amstutz & Gambette, JADT 2010



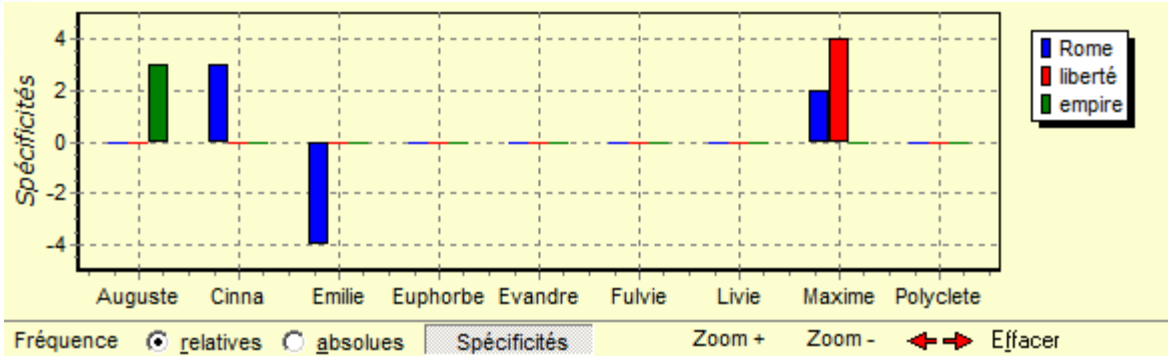
Nuage arboré globaux des 60 mots les plus fréquents dans *Cinna* de Corneille (distance Liddell, fenêtre de largeur 20), colorés chronologiquement (rouge au début, bleu à la fin)

Méthode : voisinage des mots fréquents

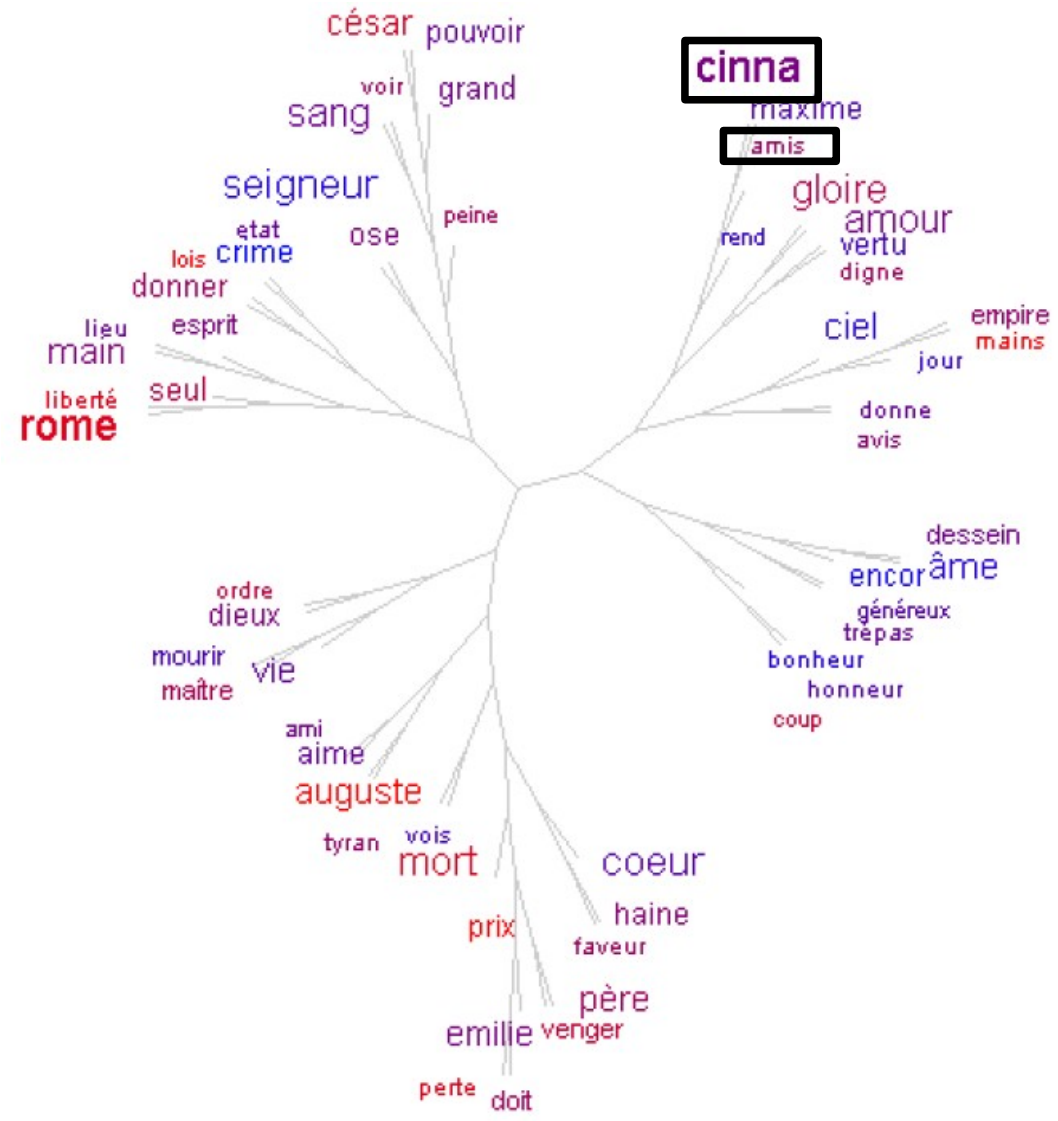
Amstutz & Gambette, JADT 2010



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



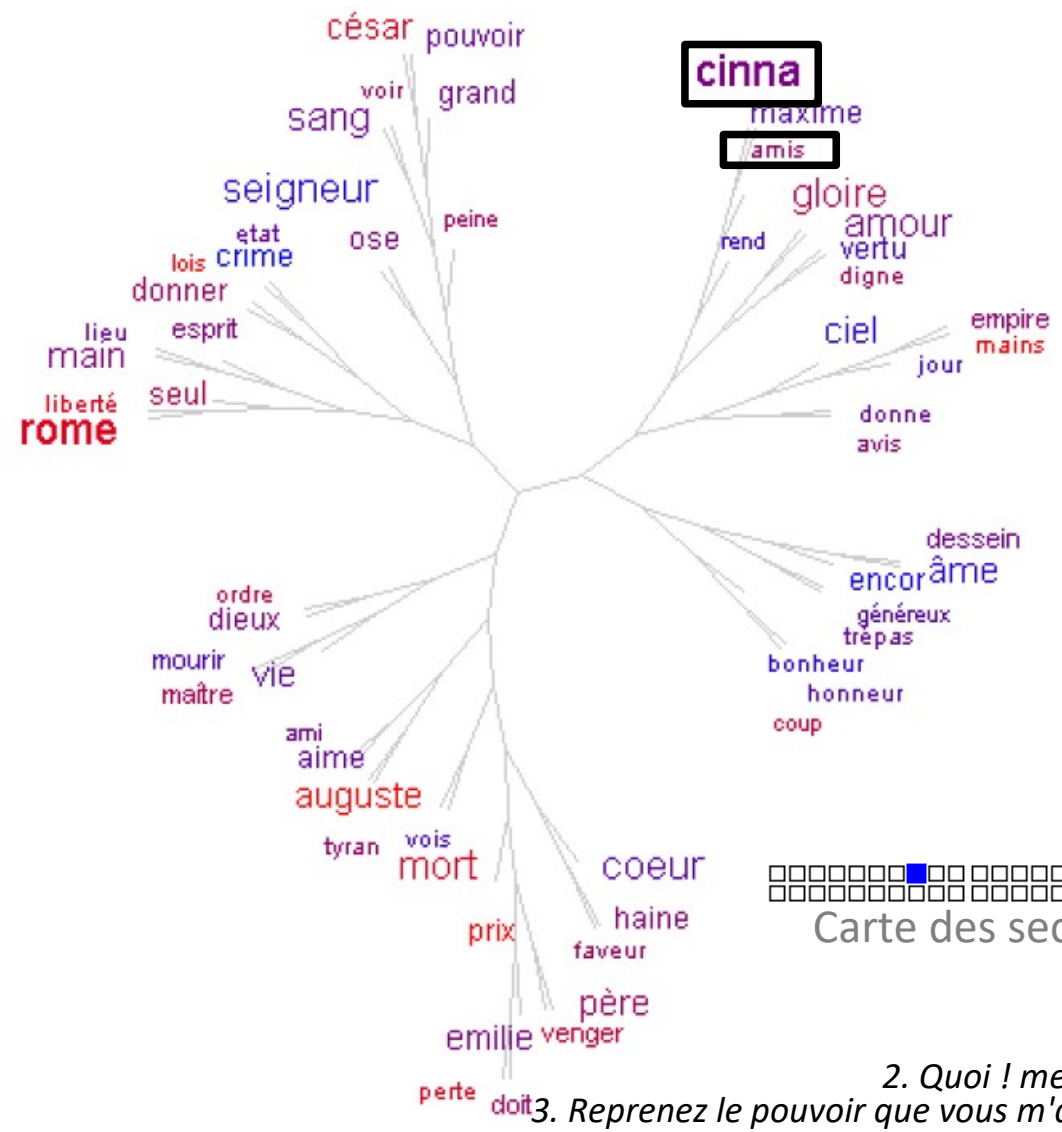
Méthode : voisinage des mots fréquents

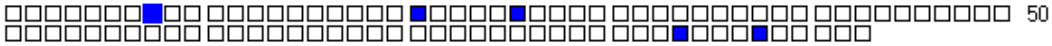


Amstutz & Gambette, JADT 2010

Méthode : voisinage des mots fréquents

Amstutz & Gambette, JADT 2010

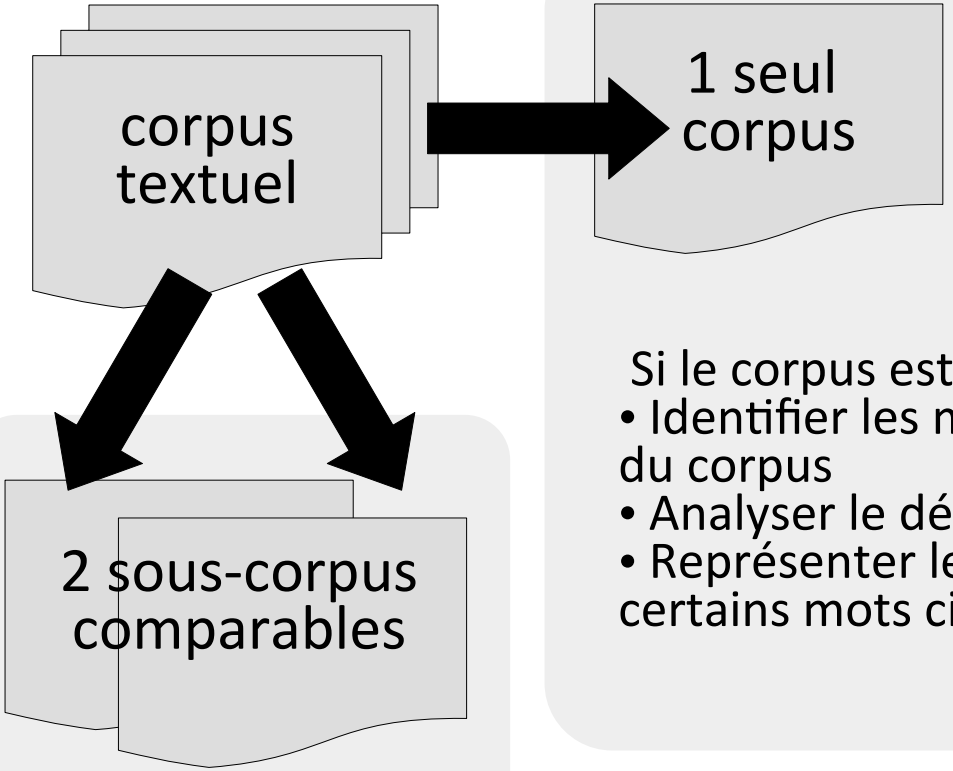




Carte des sections Lexico3 et contextes de « amis » dans les paroles d'Auguste dans *Cinna*

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

Exploration de corpus avec TreeCloud



- 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
- Analyser le début de chaque phrase du corpus
- Représenter le contexte dans le corpus de certains mots ciblés

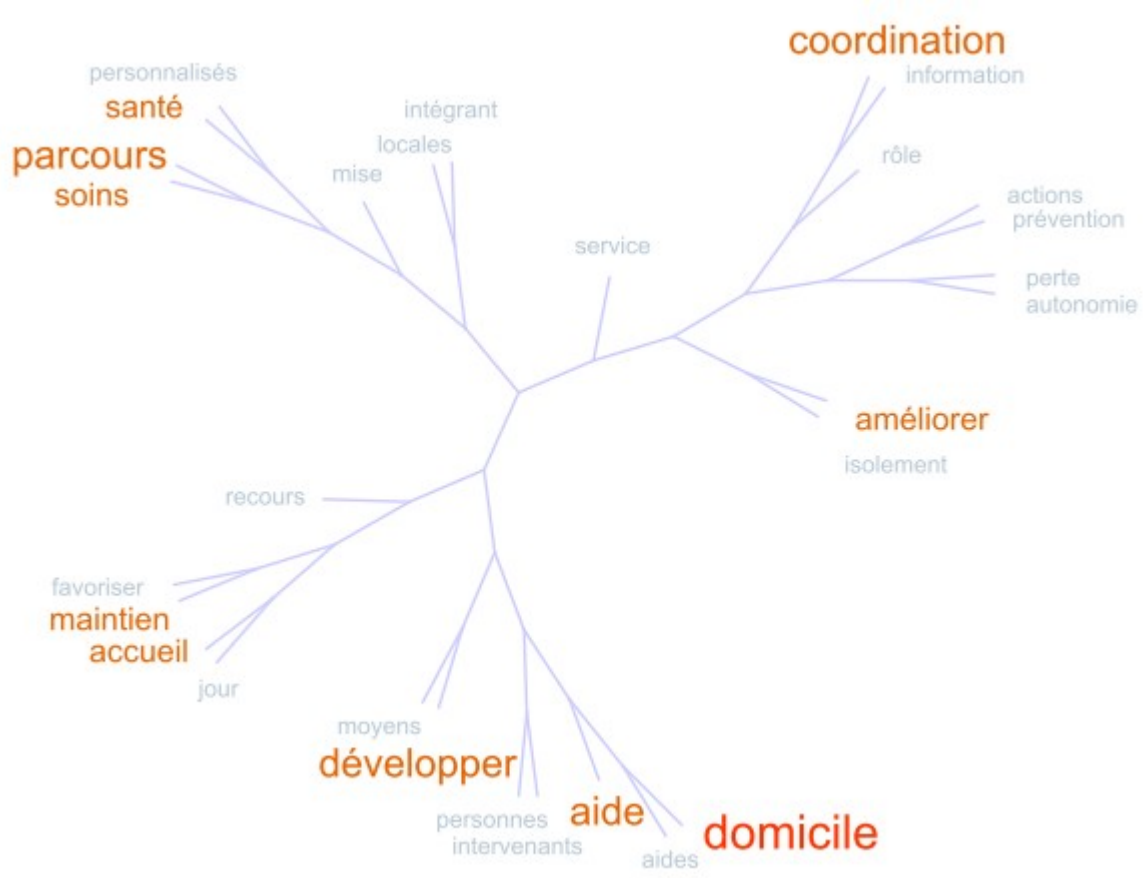
- Identifier des regroupements communs ou différents
- Comparer les voisinages d'un mot cible dans les deux corpus
- Comparer (quanti+quali) les regroupements dans les deux arbres
- Construire le nuage arboré des mots spécifiques de chaque sous-corpus

Travail mené 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 :

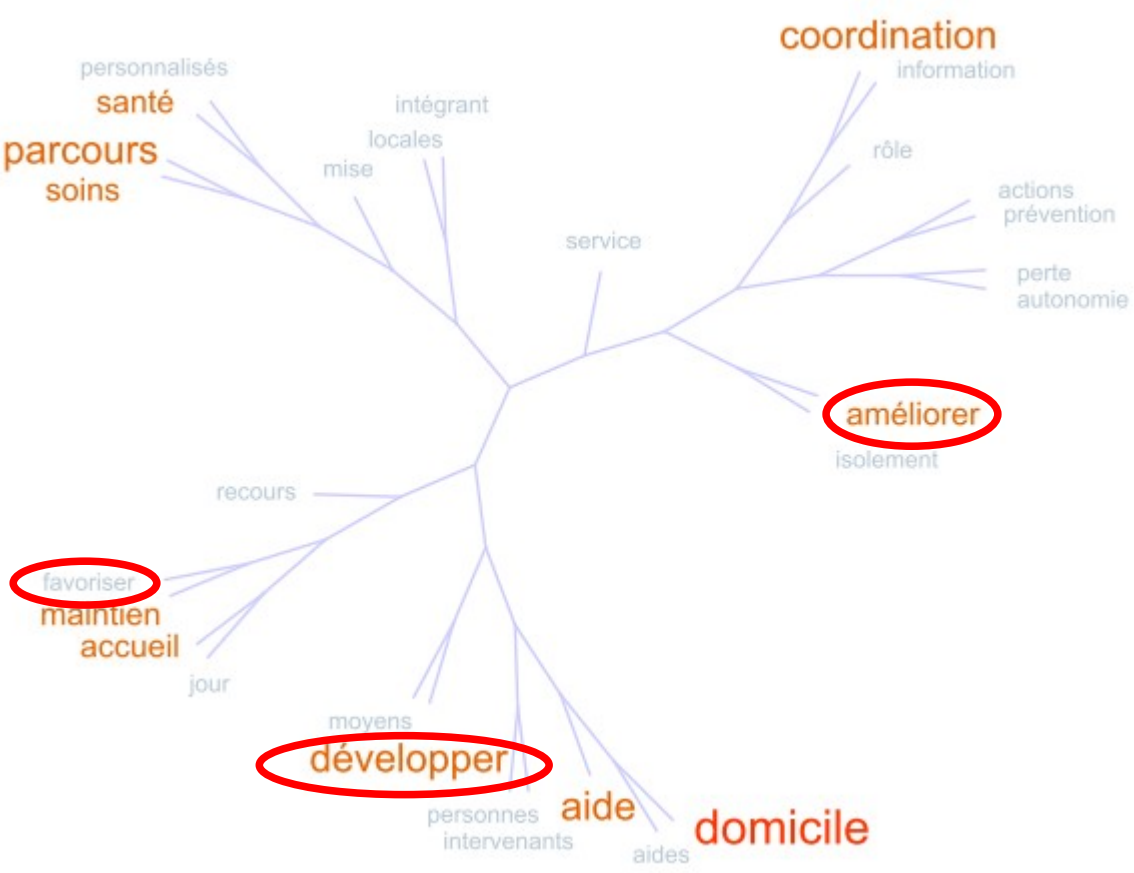


P. Grenier-Tisserand et
l'équipe projet de la DT 04

Méthode : voisinage des verbes

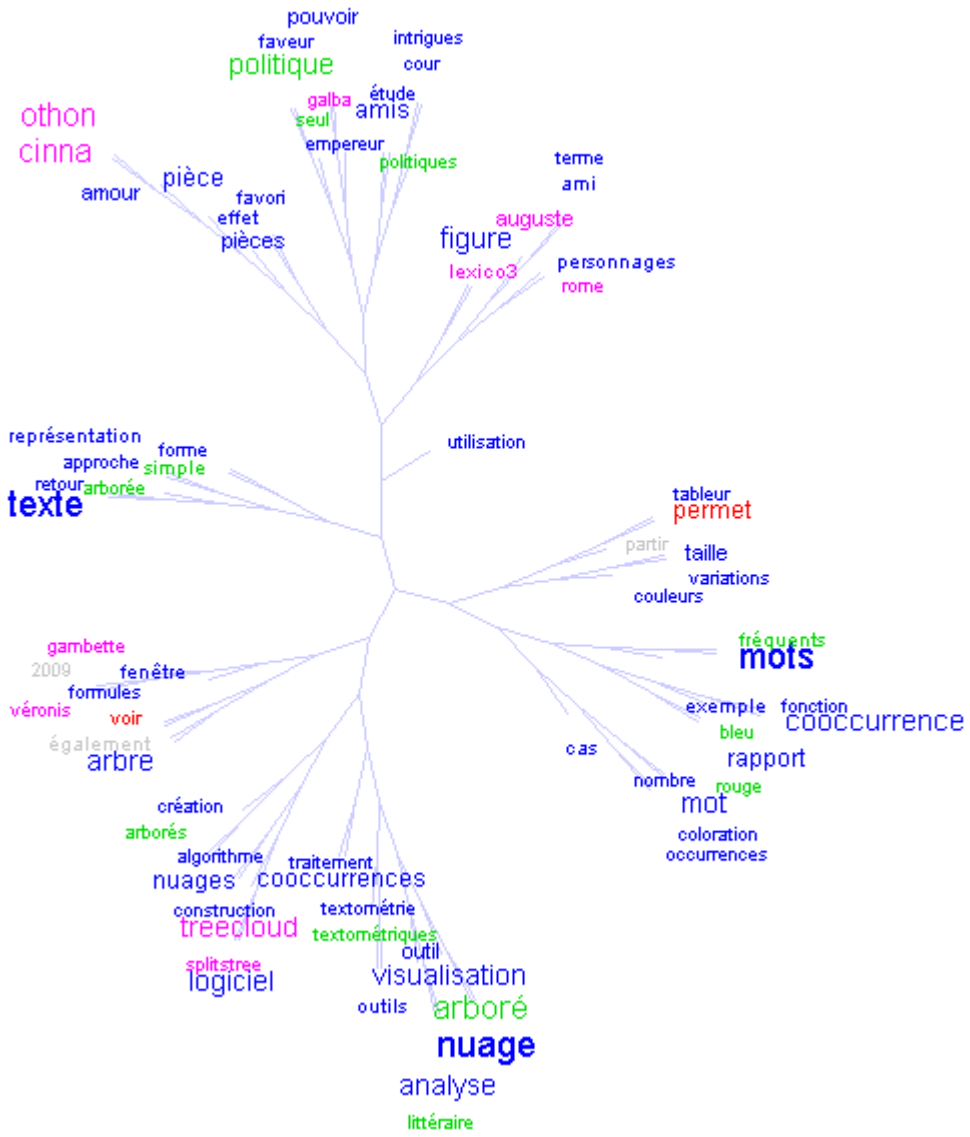
Corpus : réponses à des questions ouvertes à des professionnels de la santé sur le parcours de santé des personnes âgées dans les Alpes de Haute-Provence

Suggestions d'améliorations :



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

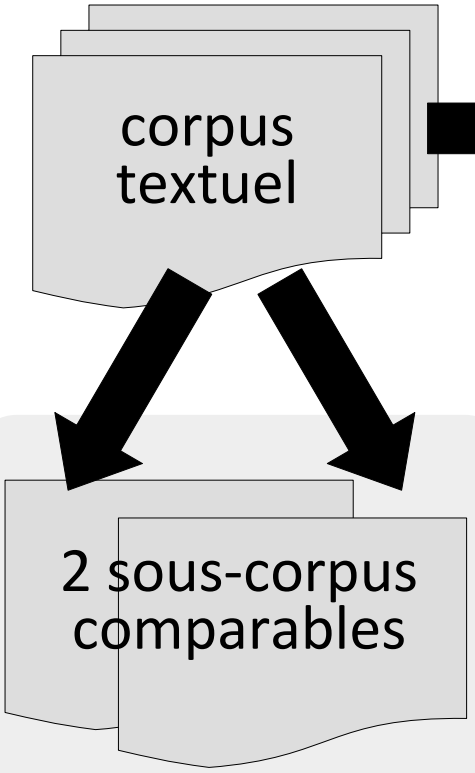
Perspective : coloration grammaticale

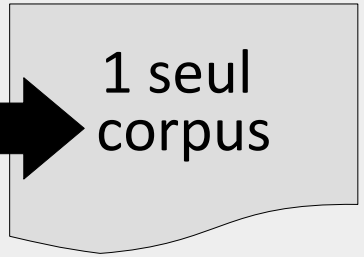


noms
adjectifs
verbes
noms propres

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





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

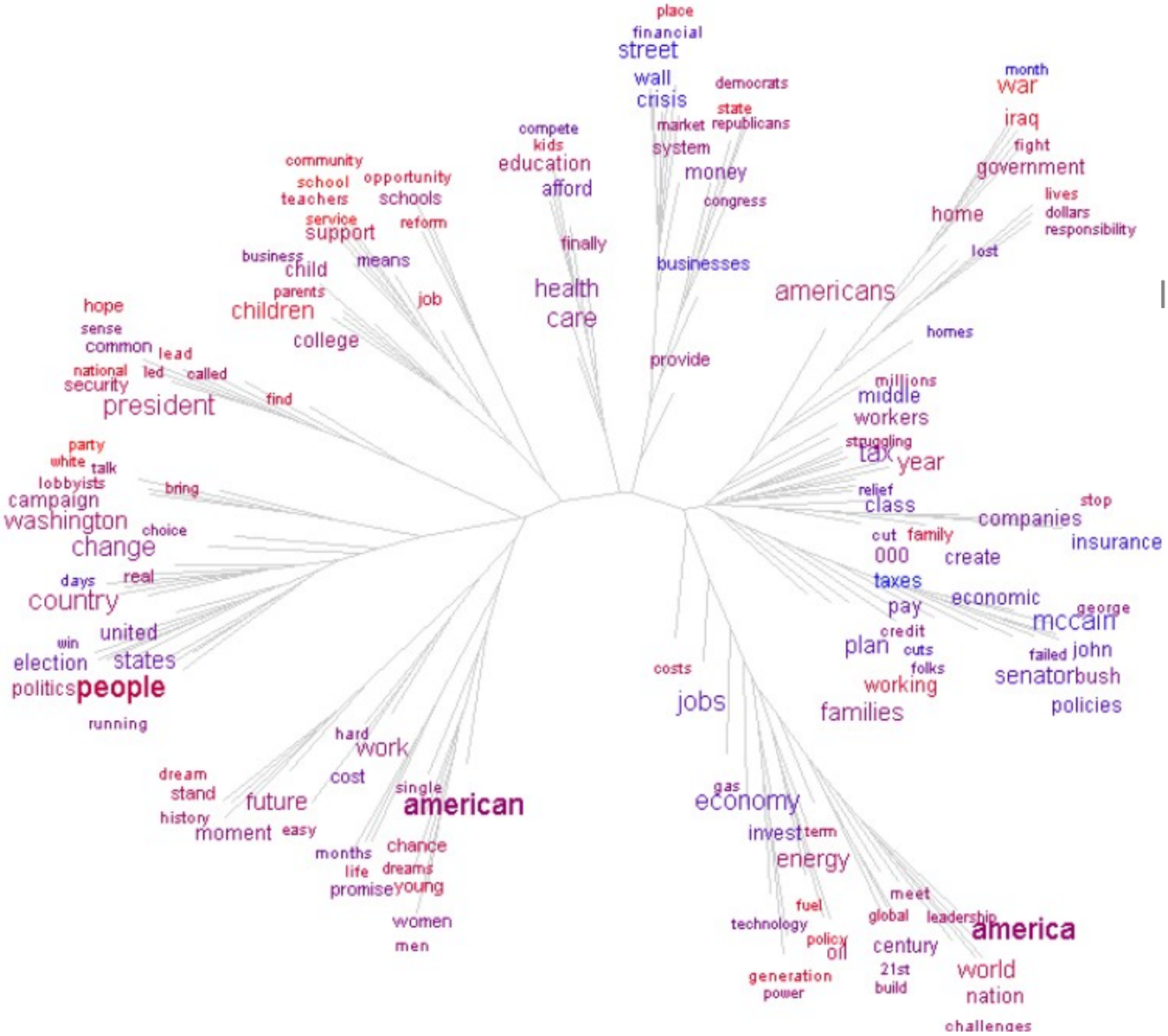
Si le corpus est assez grand :

- **Identifier les mots plutôt au début ou à la fin du corpus**
- Analyser le début de chaque phrase du corpus
- Représenter le contexte dans le corpus de certains mots ciblés

- Identifier des regroupements communs ou différents
- Comparer les voisinages d'un mot cible dans les deux corpus
- Comparer (quanti+quali) les regroupements dans les deux arbres
- Construire le nuage arboré des mots spécifiques de chaque sous-corpus

Travail mené avec Edna Hernandez

Méthode : mots au début ou à la fin

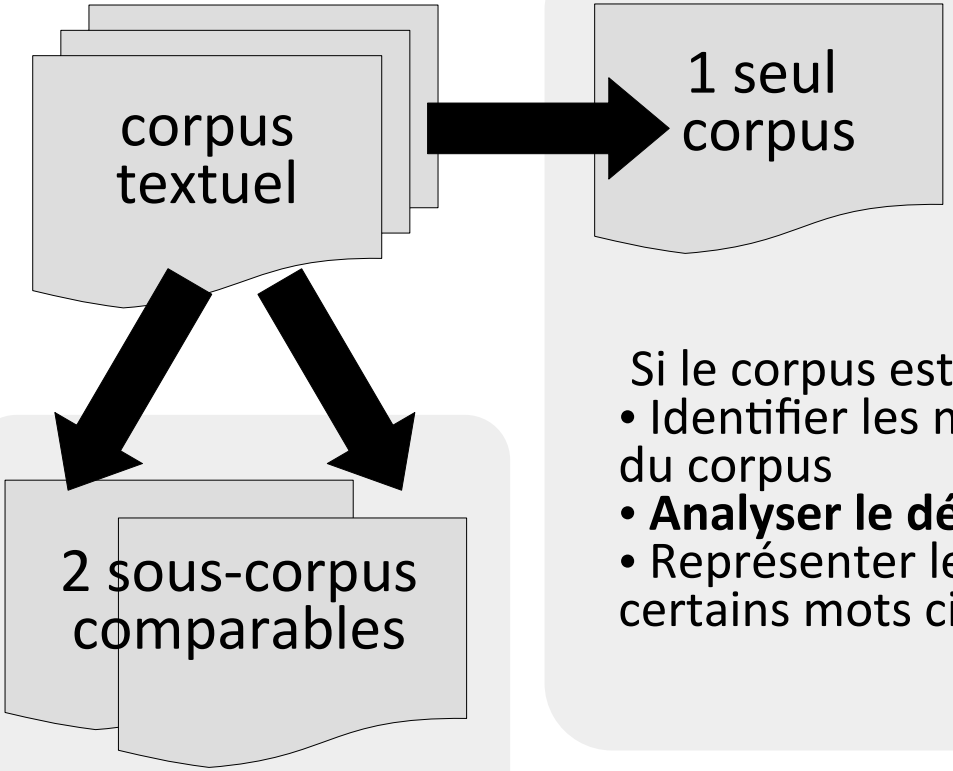


Nuage arboré de l'ensemble des discours de campagne de 2008 de Barack Obama, coloration chronologique

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.)

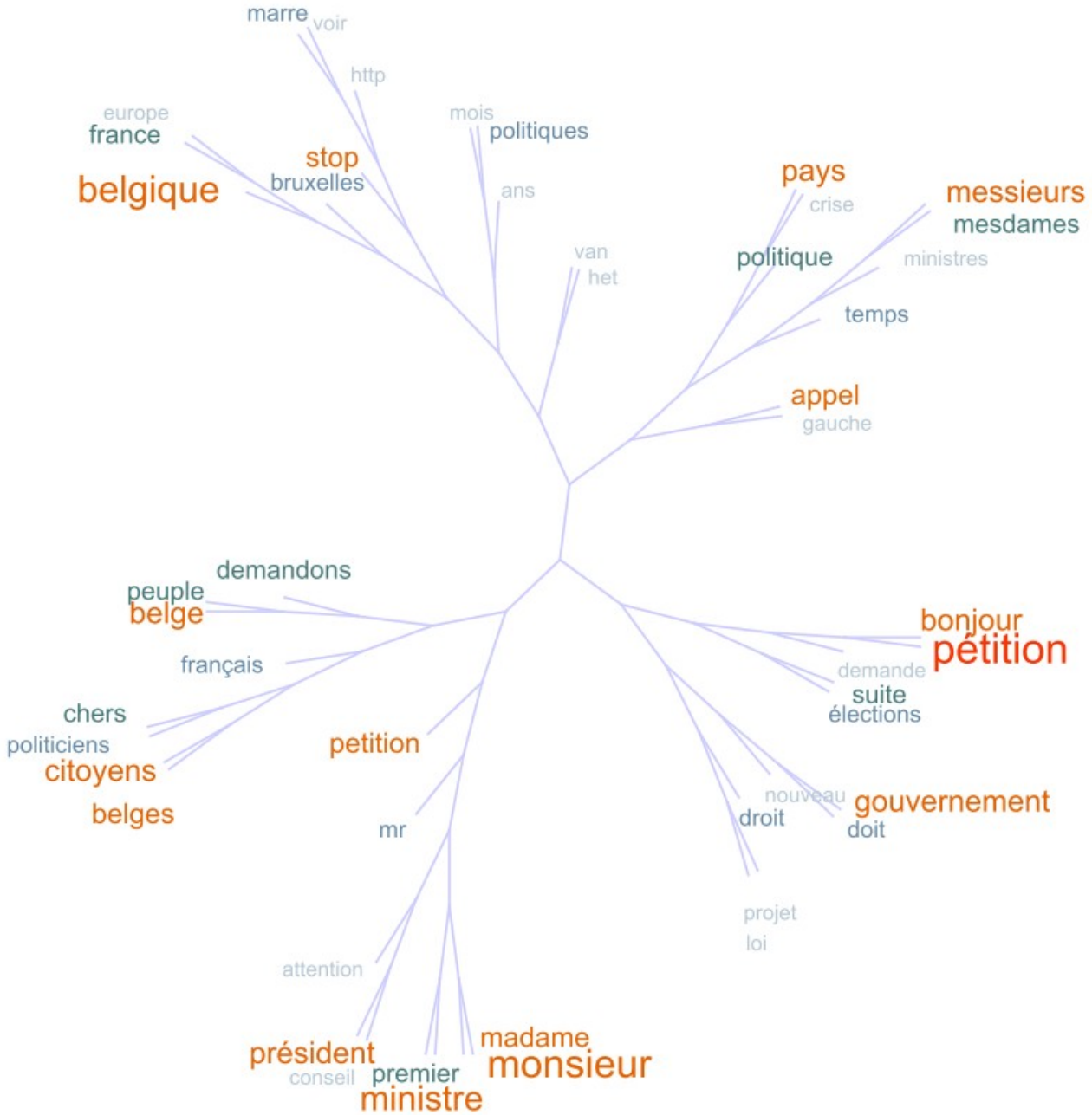
Si le corpus est assez grand :

- Identifier les mots plutôt au début ou à la fin du corpus
- **Analyser le début de chaque phrase du corpus**
- Représenter le contexte dans le corpus de certains mots ciblés

- Identifier des regroupements communs ou différents
- Comparer les voisinages d'un mot cible dans les deux corpus
- Comparer (quanti+quali) les regroupements dans les deux arbres
- Construire le nuage arboré des mots spécifiques de chaque sous-corpus

Travail mené avec Edna Hernandez

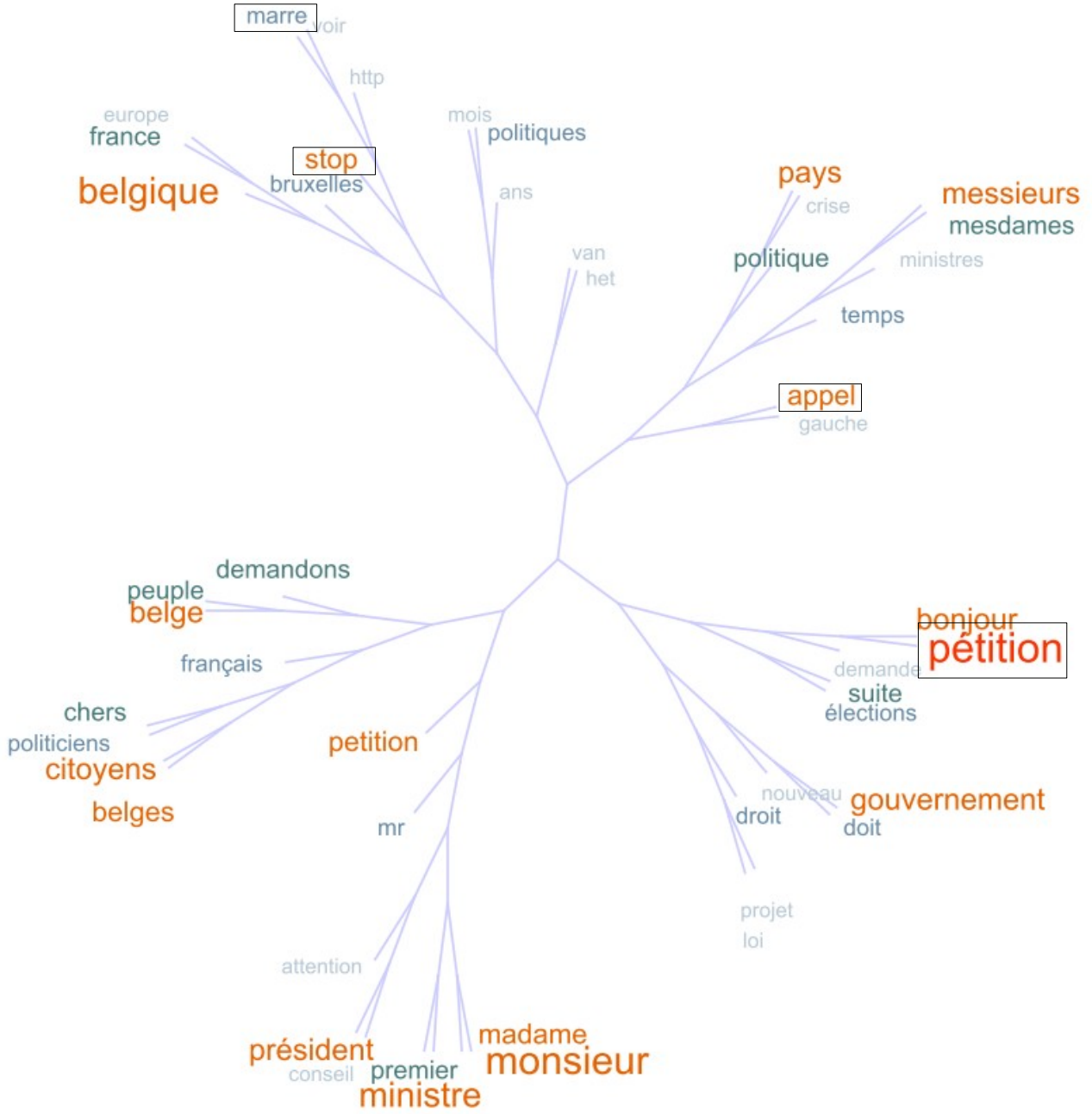
Méthode : mots au début de chaque phrase



50 mots les plus fréquents parmi les débuts (10 premiers mots) des textes de pétitions de la catégorie « politique » du site lapetition.be.

Travaux menés avec Christine Barats, Anne Dister, Jean-Marc Leblanc et Marie Peres-Leblanc dans le cadre de l'ANR APPEL

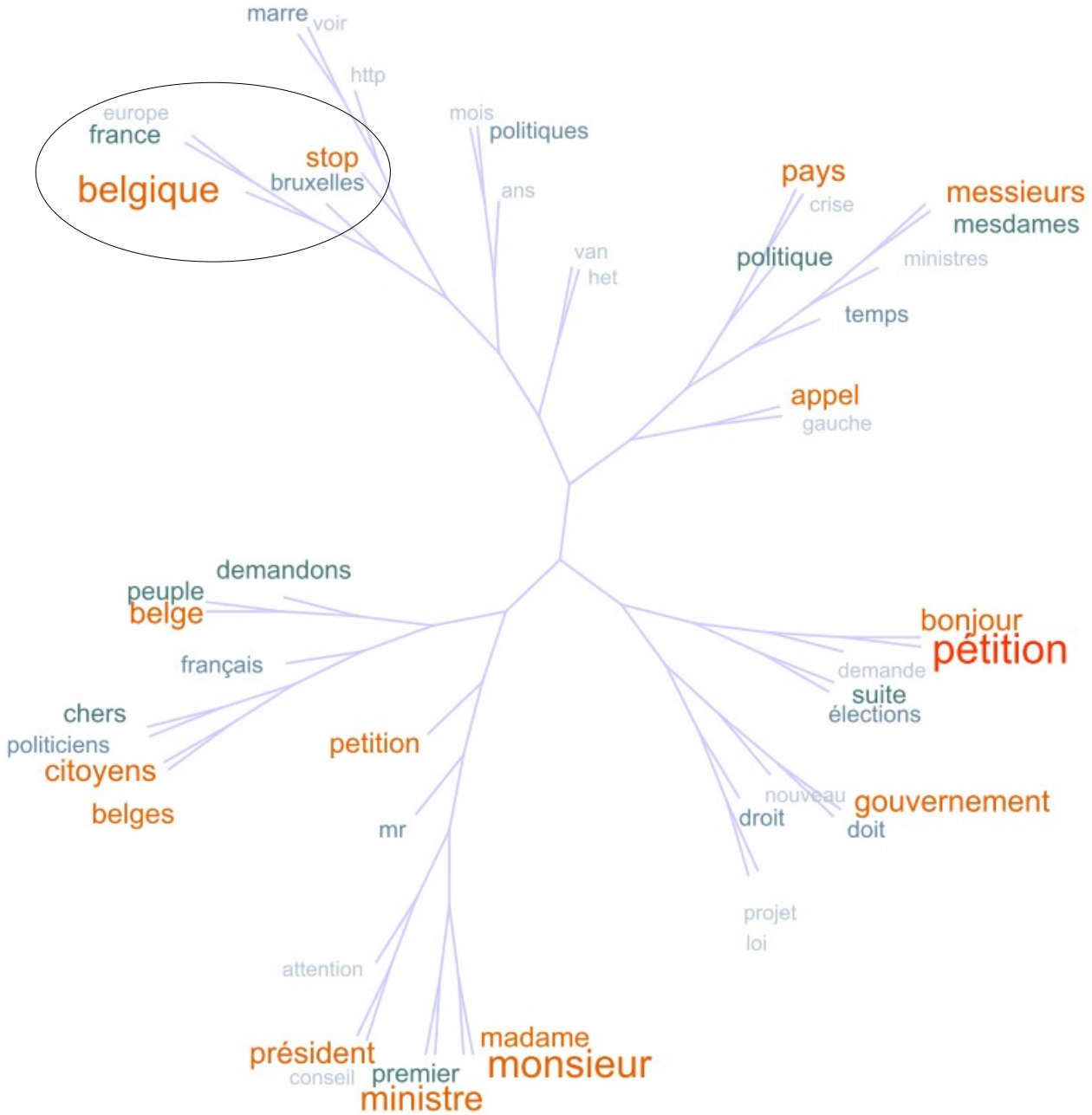
Méthode : mots au début de chaque phrase



50 mots les plus fréquents parmi les débuts (10 premiers mots) des textes de pétitions de la catégorie « politique » du site lapetition.be.

Travaux menés avec Christine Barats, Anne Dister, Jean-Marc Leblanc et Marie Peres-Leblanc dans le cadre de l'ANR APPEL

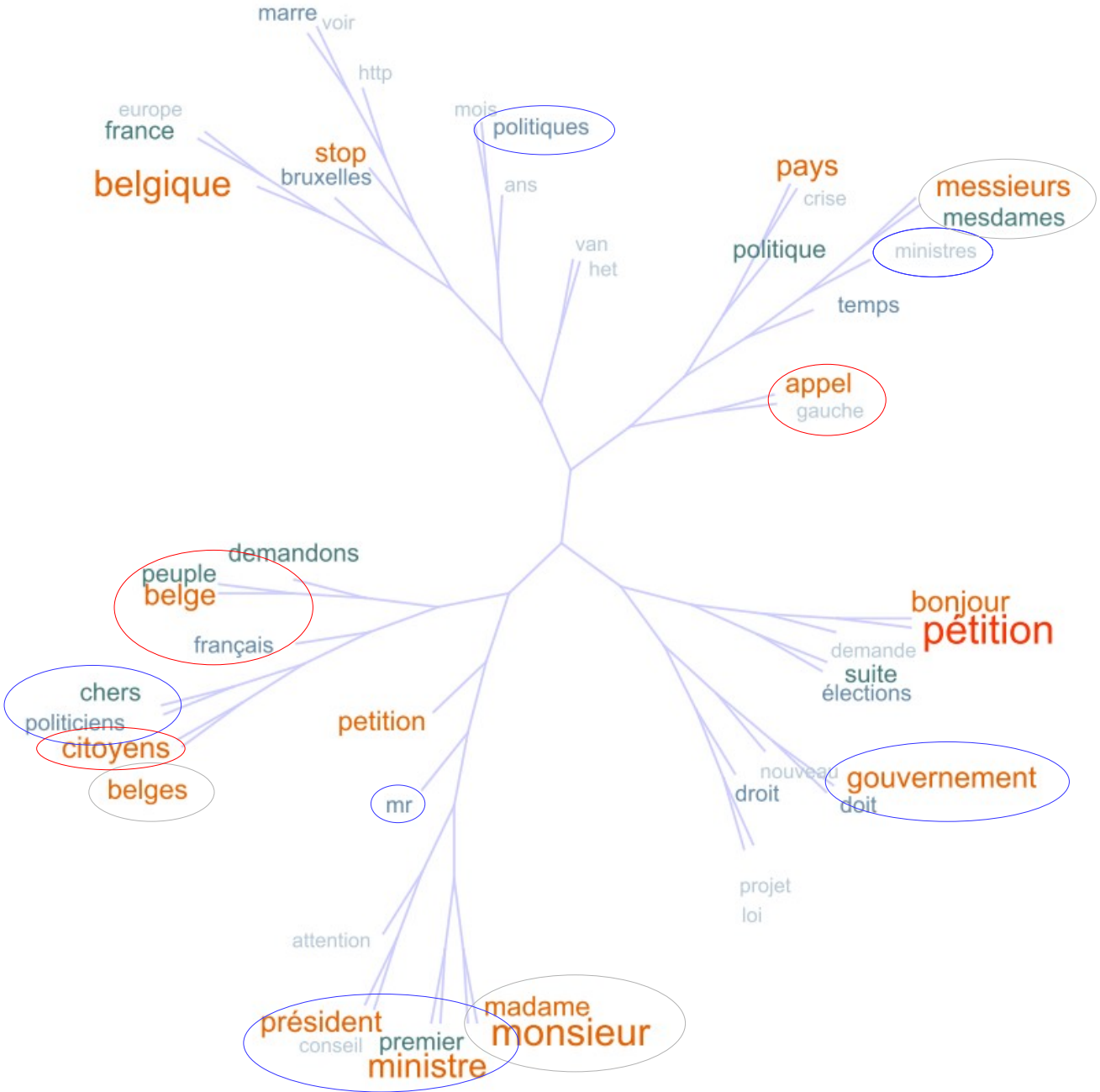
Méthode : mots au début de chaque phrase



50 mots les plus fréquents parmi les débuts (10 premiers mots) des textes de pétitions de la catégorie « politique » du site lapetition.be.

Travaux menés avec Christine Barats, Anne Dister, Jean-Marc Leblanc et Marie Peres-Leblanc dans le cadre de l'ANR APPEL

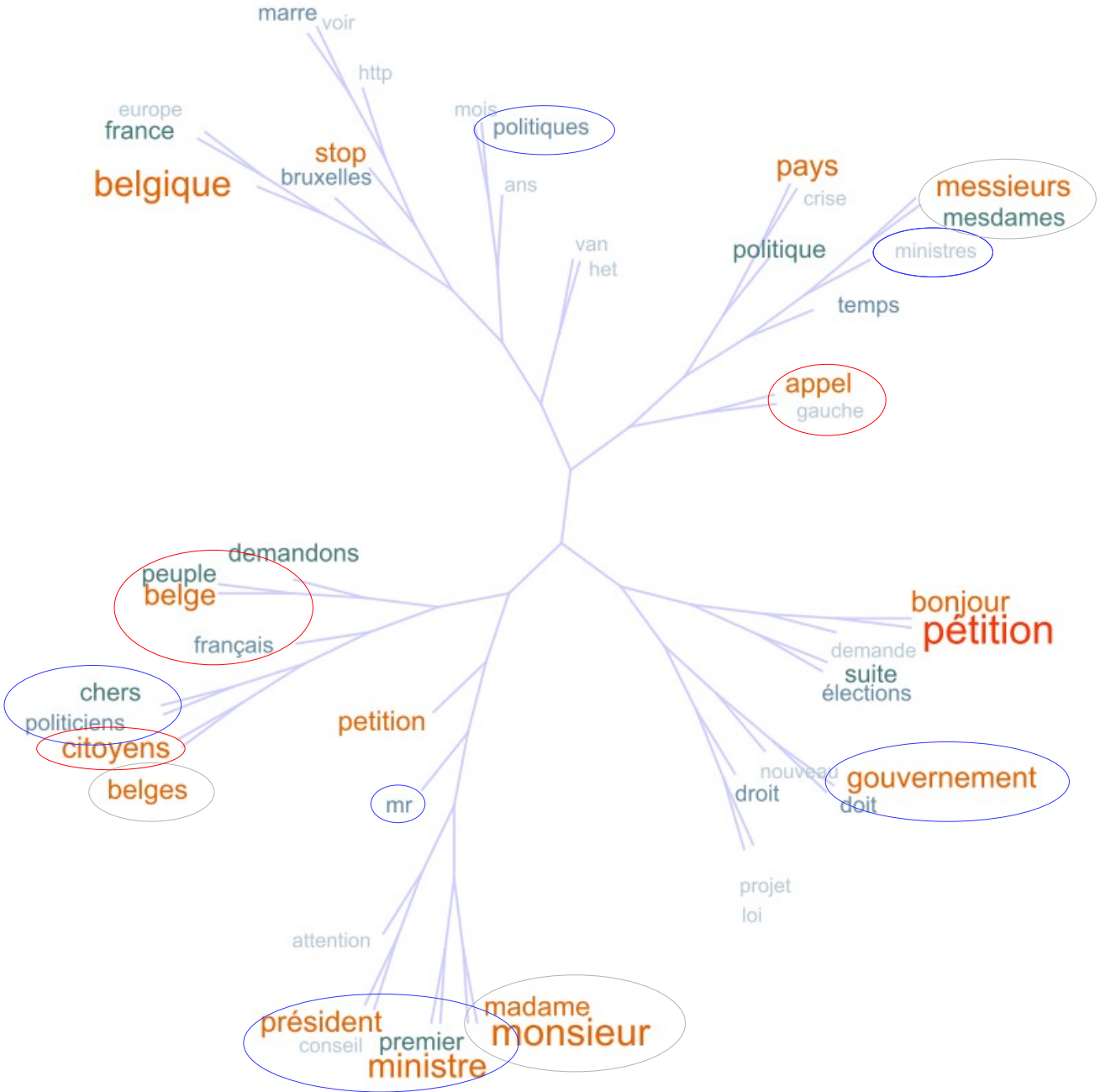
Méthode : mots au début de chaque phrase



50 mots les plus fréquents parmi les débuts (10 premiers mots) des textes de pétitions de la catégorie « politique » du site lapetition.be.

Travaux menés avec Christine Barats, Anne Dister, Jean-Marc Leblanc et Marie Peres-Leblanc dans le cadre de l'ANR APPEL

Méthode : mots au début de chaque phrase

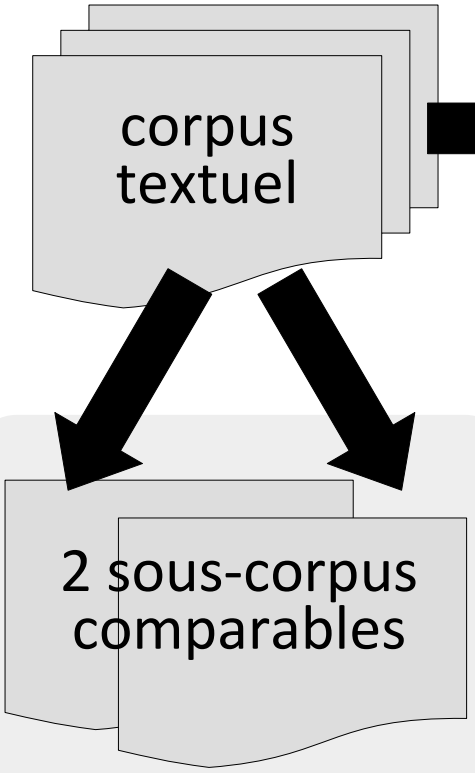


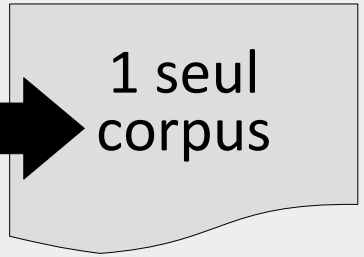
Relevé manuel :
6,5% d'adresse aux décideurs des pétitions
3,8% d'adresse aux signataires des pétitions

50 mots les plus fréquents parmi les débuts (10 premiers mots) des textes de pétitions de la catégorie « politique » du site lapetition.be.

Travaux menés avec Christine Barats, Anne Dister, Jean-Marc Leblanc et Marie Peres-Leblanc dans le cadre de l'ANR APPEL

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.)

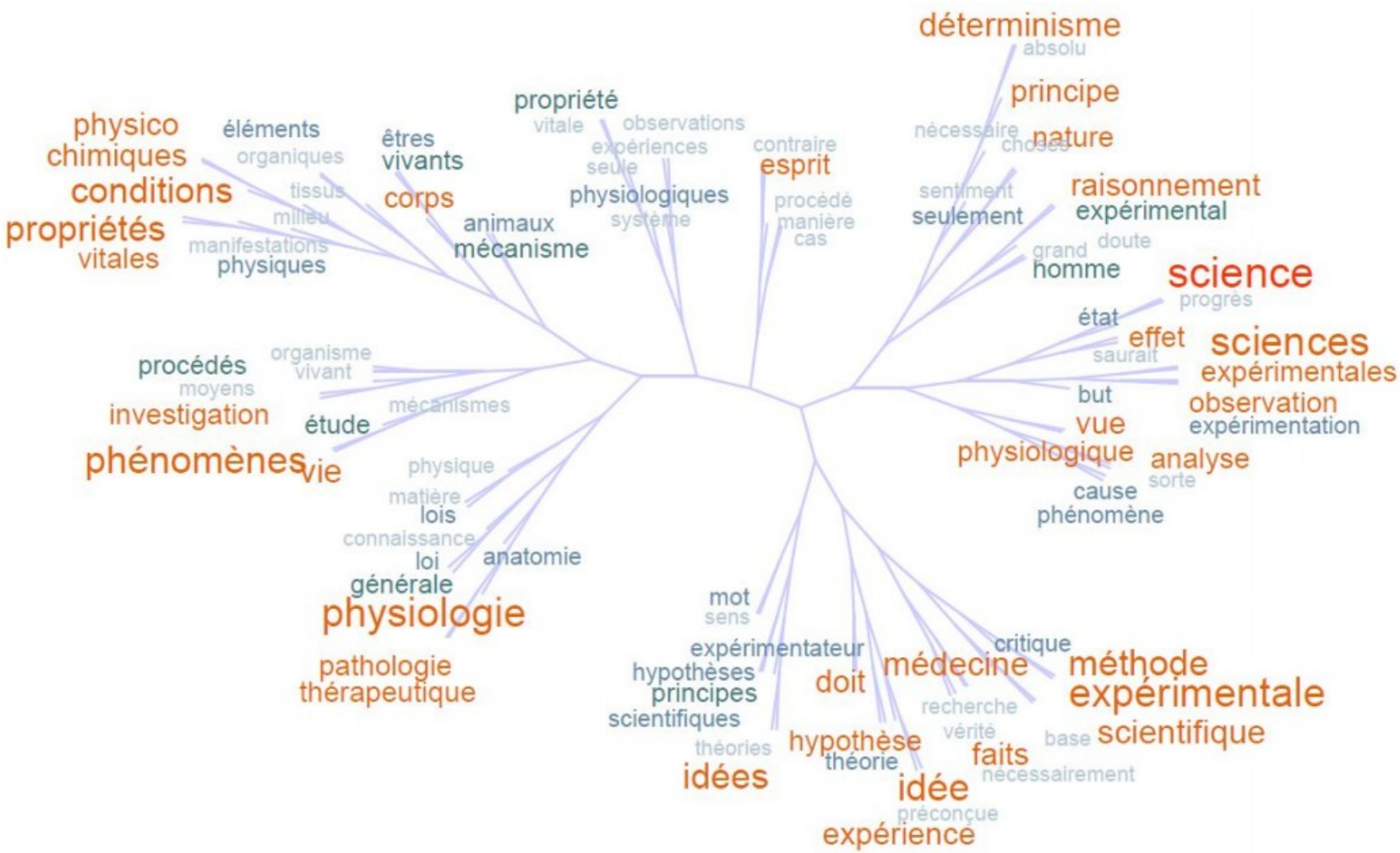
Si le corpus est assez grand :

- Identifier les mots plutôt au début ou à la fin du corpus
- Analyser le début de chaque phrase du corpus
- **Représenter le contexte dans le corpus de certains mots ciblés**

- Identifier des regroupements communs ou différents
- Comparer les voisinages d'un mot cible dans les deux corpus
- Comparer (quanti+quali) les regroupements dans les deux arbres
- Construire le nuage arboré des mots spécifiques de chaque sous-corpus

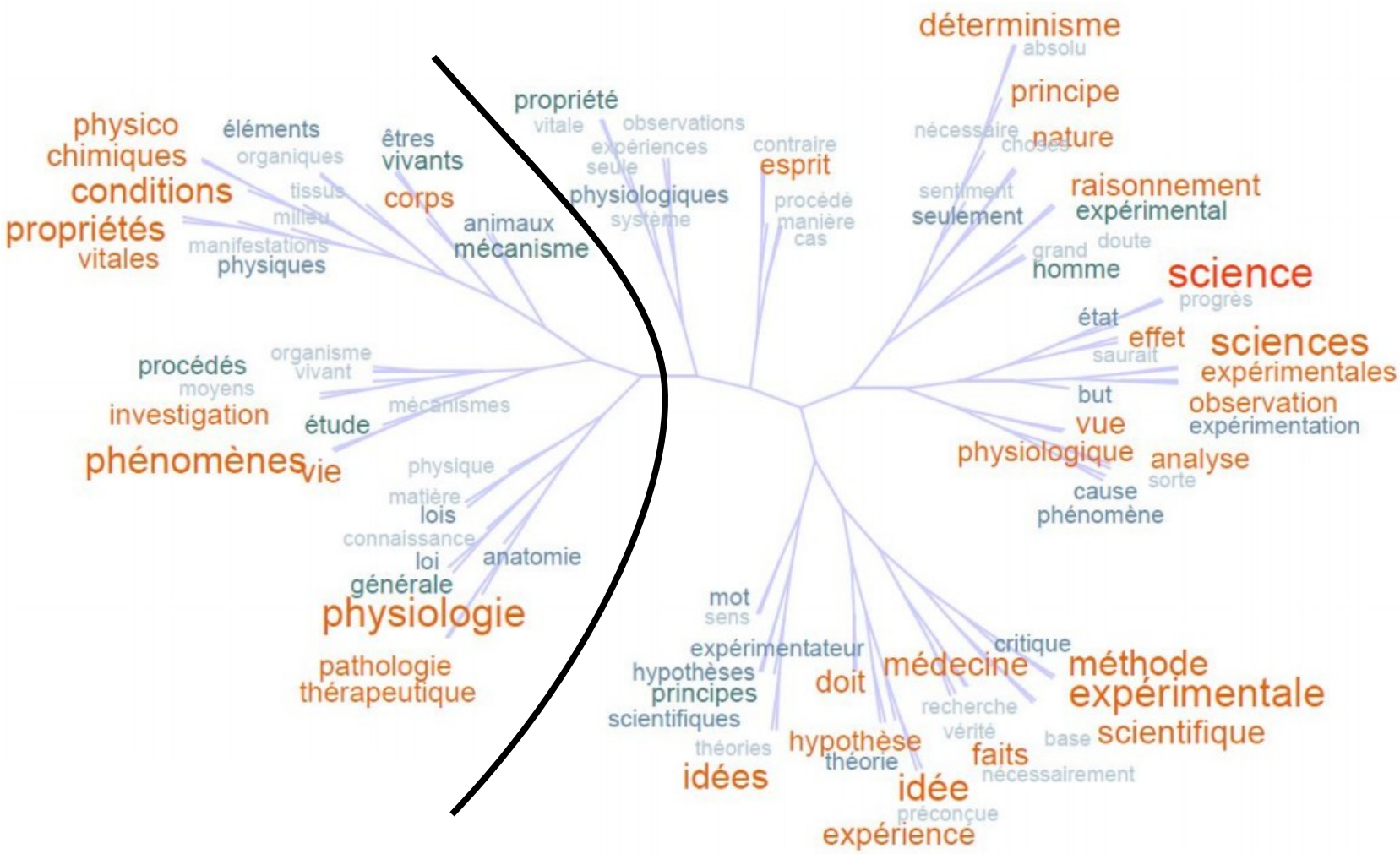
Travail mené avec Edna Hernandez

Nuages arborés des contextes de « étude »



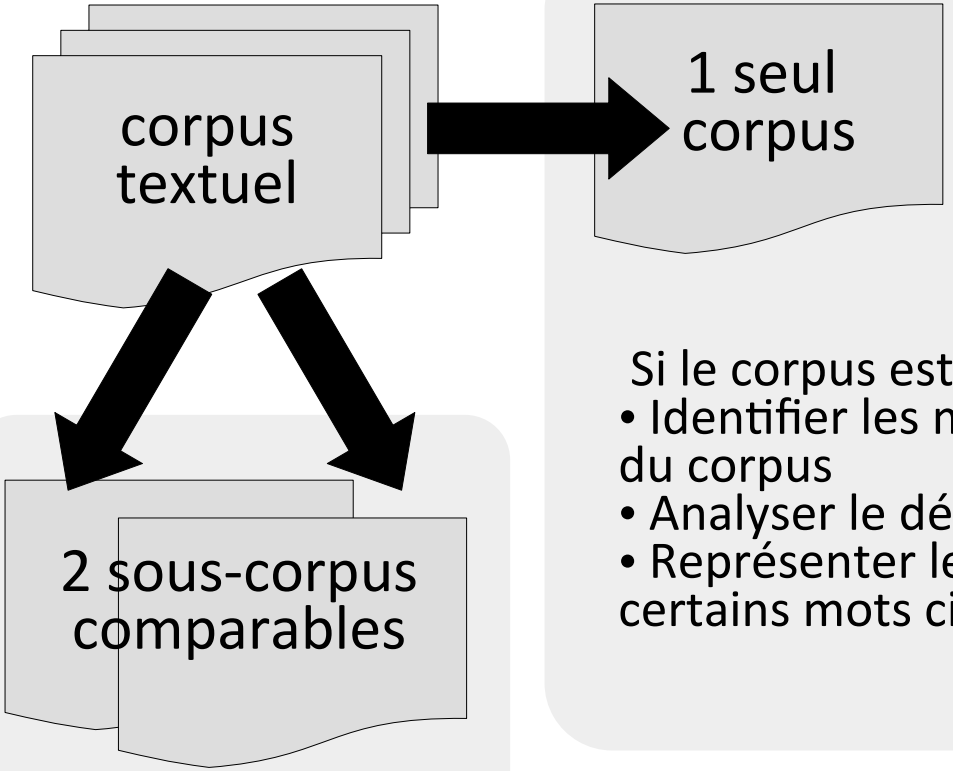
Nuage arboré des 100 mots les plus fréquents dans les contextes (10 mots avant, 10 mots après) des mots de la catégorie "étude" dans un corpus de textes scientifiques et littéraires sur la science (projet AnimalHumanité)

Nuages arborés des contextes de « étude »



Nuage arboré des 100 mots les plus fréquents dans les contextes (10 mots avant, 10 mots après) des mots de la catégorie "étude" dans un corpus de textes scientifiques et littéraires sur la science (projet AnimalHumanité)

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
- Analyser le début de chaque phrase du corpus
- Représenter le contexte dans le corpus de certains mots ciblés

- **Identifier des regroupements communs ou différents**
- Comparer les voisinages d'un mot cible dans les deux corpus
- Comparer (quanti+quali) les regroupements dans les deux arbres
- Construire le nuage arboré des mots spécifiques de chaque sous-corpus

Travail mené 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

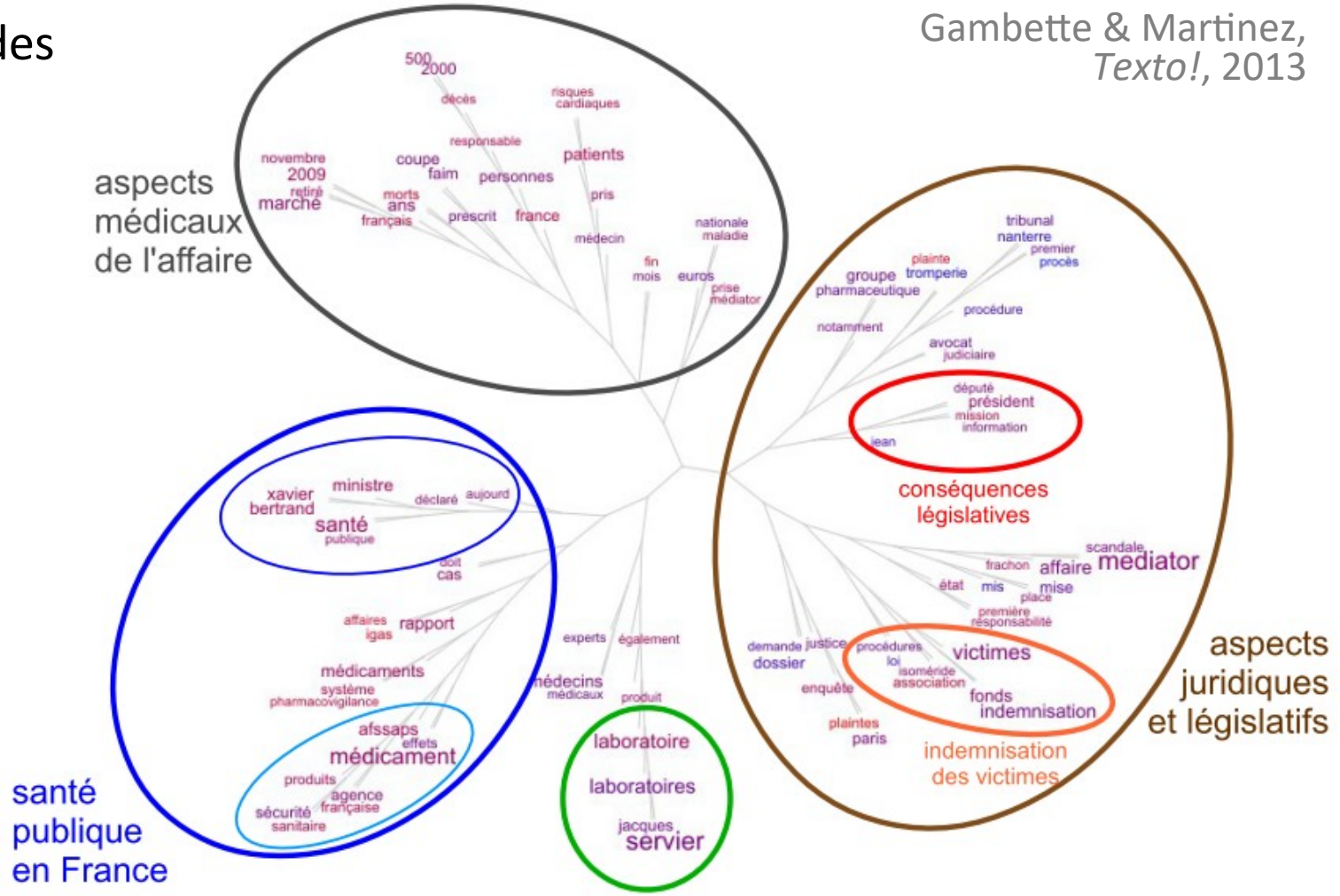
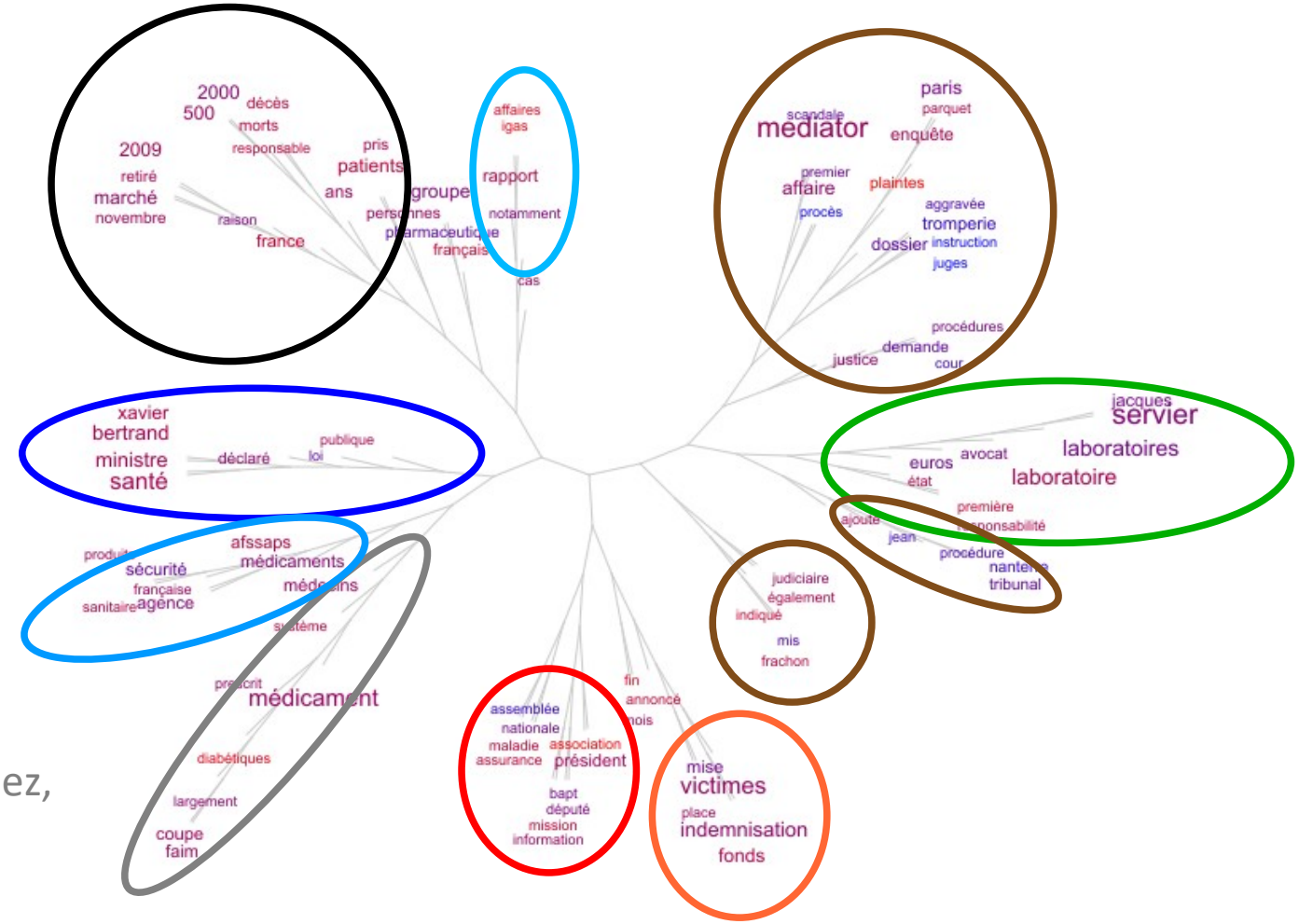


Illustration sur le corpus Mediator

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



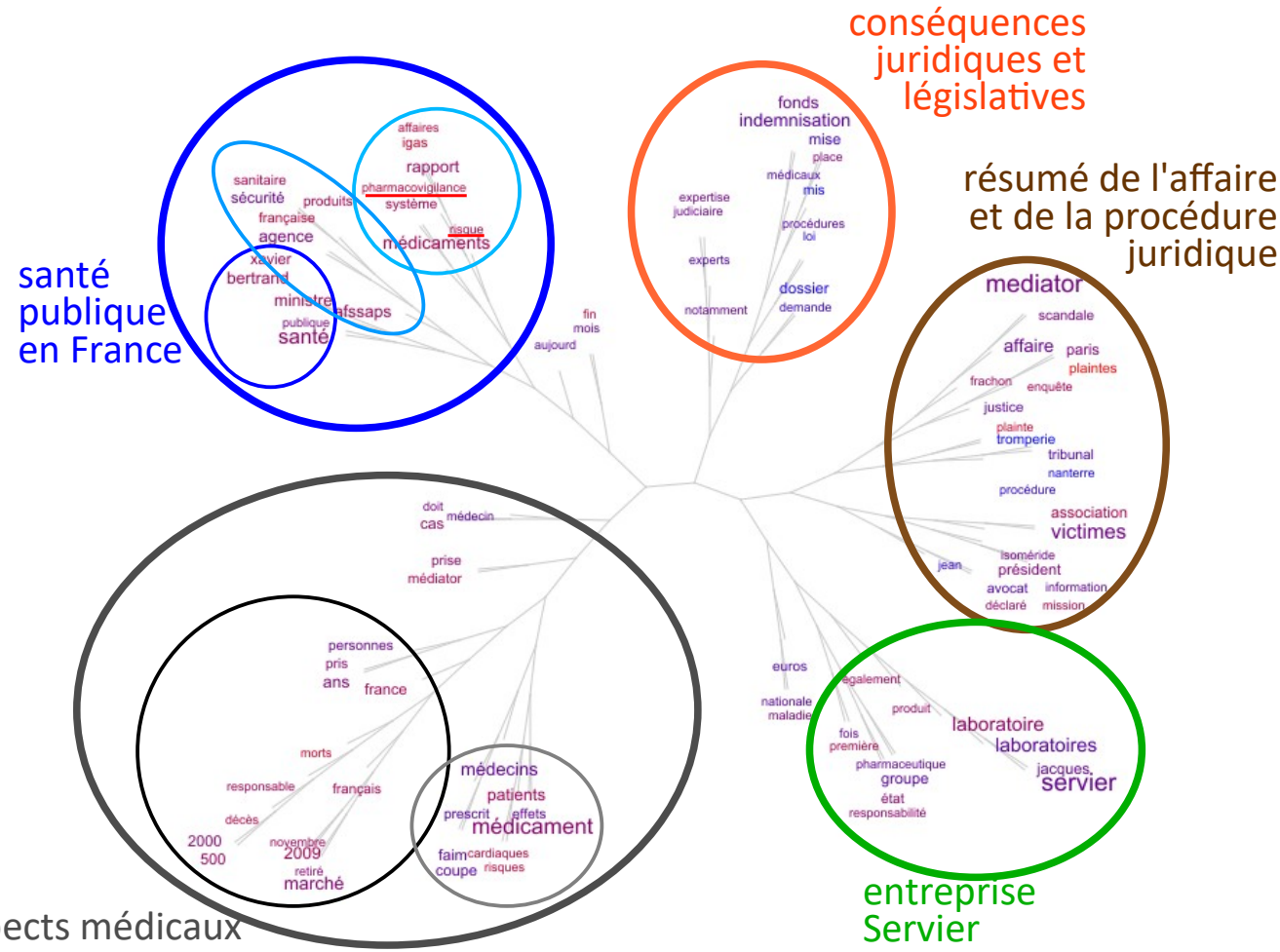
Gambette & Martinez, *Texto!*, 2013

Illustration sur le corpus Mediator

Comparer les articles d'agences et articles de journalistes

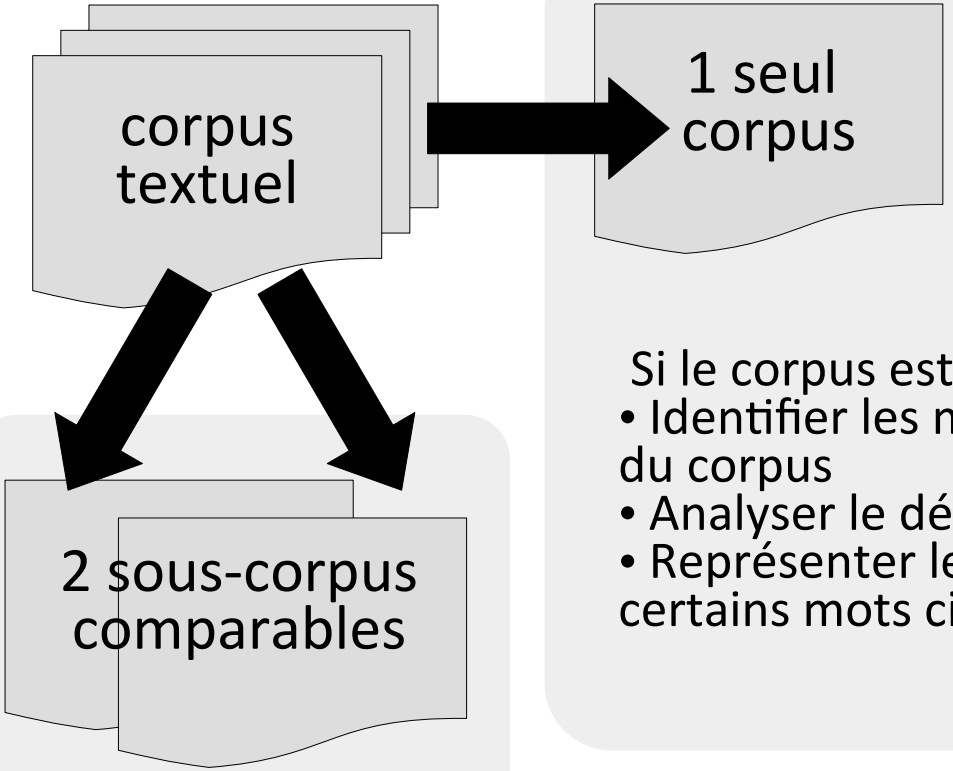
Corpus : 595 articles d'agences contre 1496 articles de journalistes de 2011 évoquant l'affaire du Mediator dans la presse française.

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.)

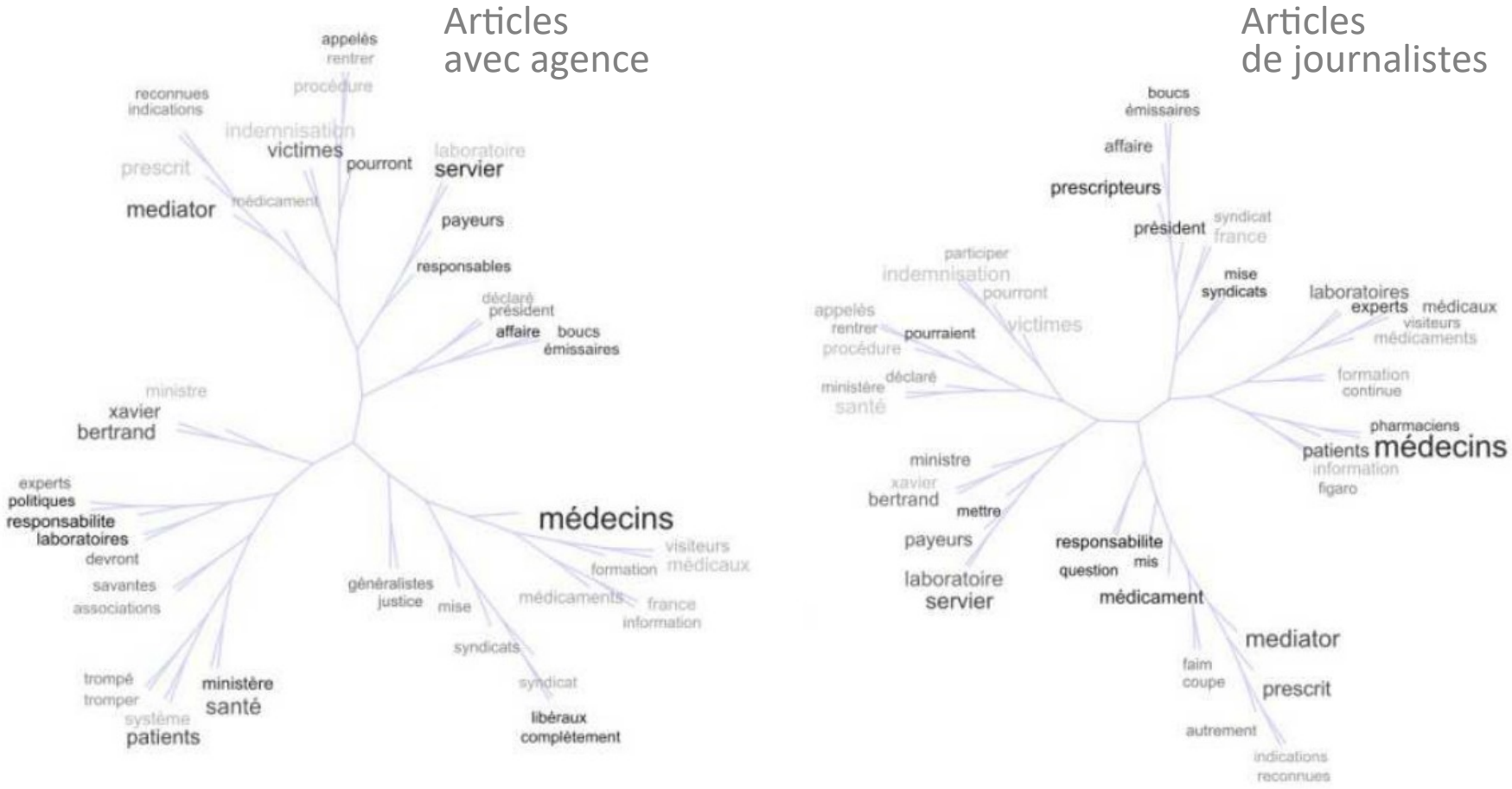
Si le corpus est assez grand :

- Identifier les mots plutôt au début ou à la fin du corpus
- Analyser le début de chaque phrase du corpus
- Représenter le contexte dans le corpus de certains mots ciblés

- Identifier des regroupements communs ou différents
- **Comparer les voisinages d'un mot cible dans les deux corpus**
- Comparer (quanti+quali) les regroupements dans les deux arbres
- Construire le nuage arboré des mots spécifiques de chaque sous-corpus

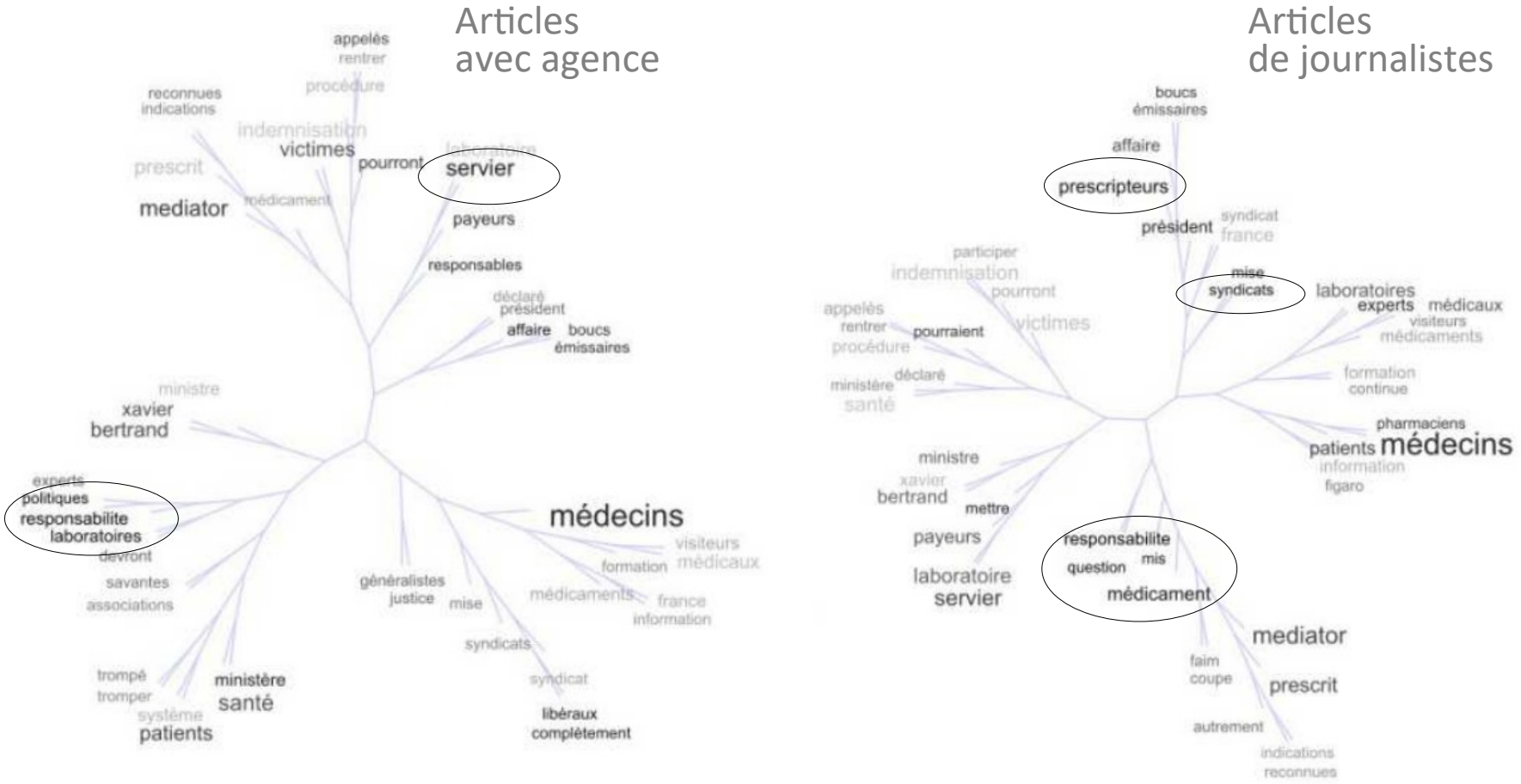
Travail mené avec Edna Hernandez

Nuages arborés des contextes de « médecins »



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

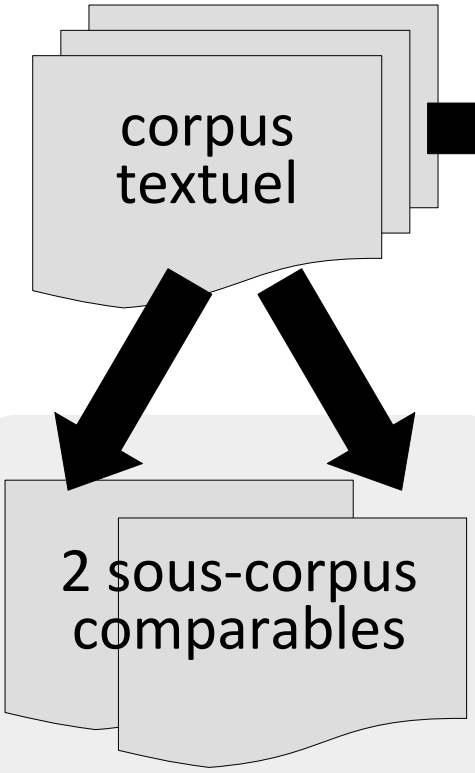
Nuages arborés des contextes de « médecins »

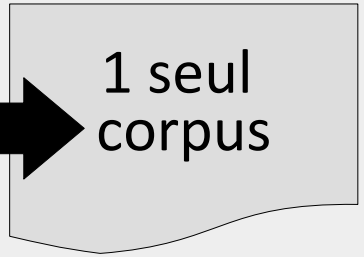


Nuage arboré des 50 mots les plus fréquents des contextes (10 mots avant et 10 mots après) du mot médecins dans le sous-corpus des articles sur le Mediator, colorés par le degré de cooccurrence avec le mot responsabilités (en noir pour les mots les plus cooccurrents), construit par TreeCloud avec la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

Martinez & Gambette, *Corpus,* 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.)

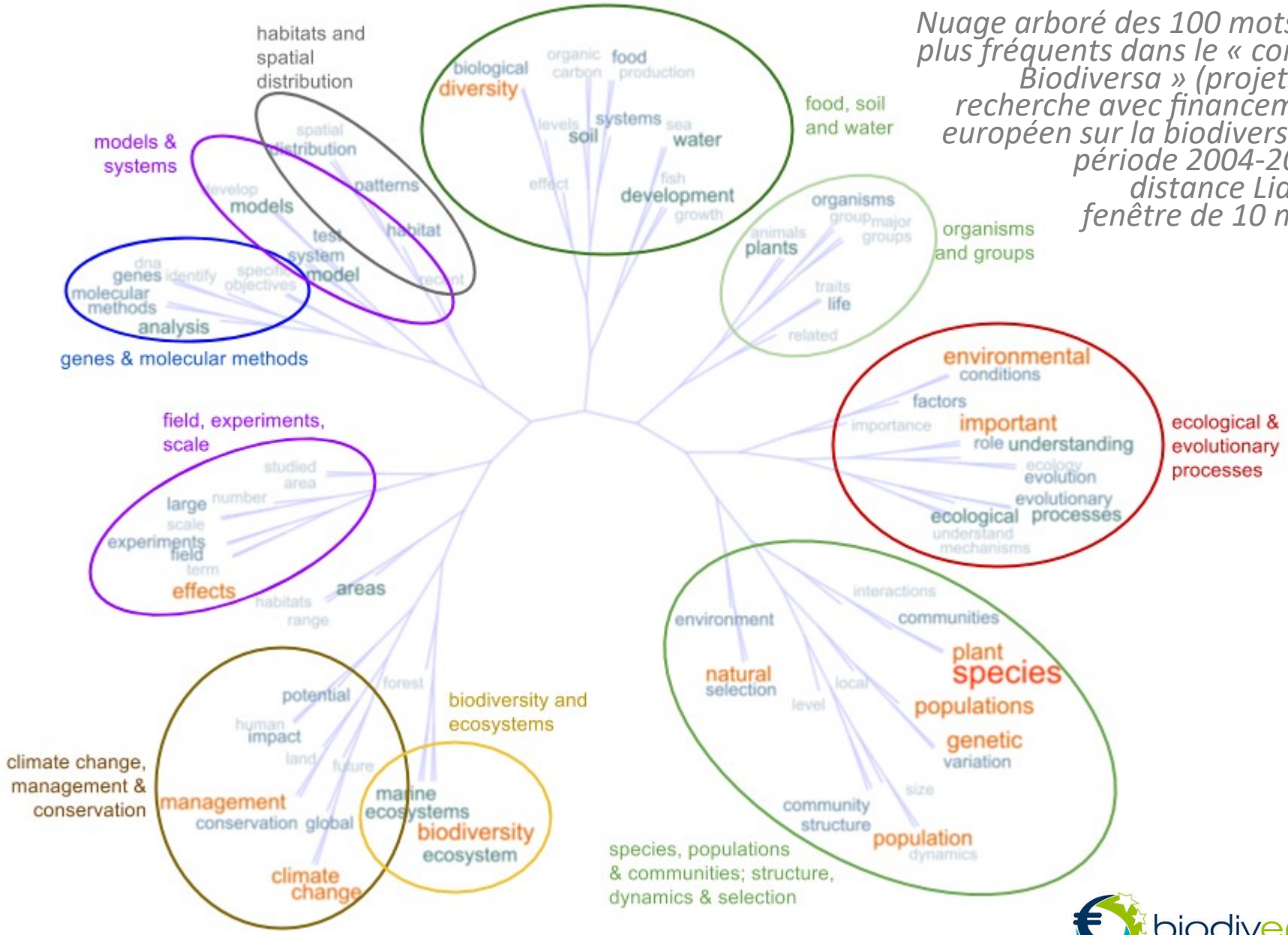
Si le corpus est assez grand :

- Identifier les mots plutôt au début ou à la fin du corpus
- Analyser le début de chaque phrase du corpus
- Représenter le contexte dans le corpus de certains mots ciblés

- Identifier des regroupements communs ou différents
- Comparer les voisinages d'un mot cible dans les deux corpus
- **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

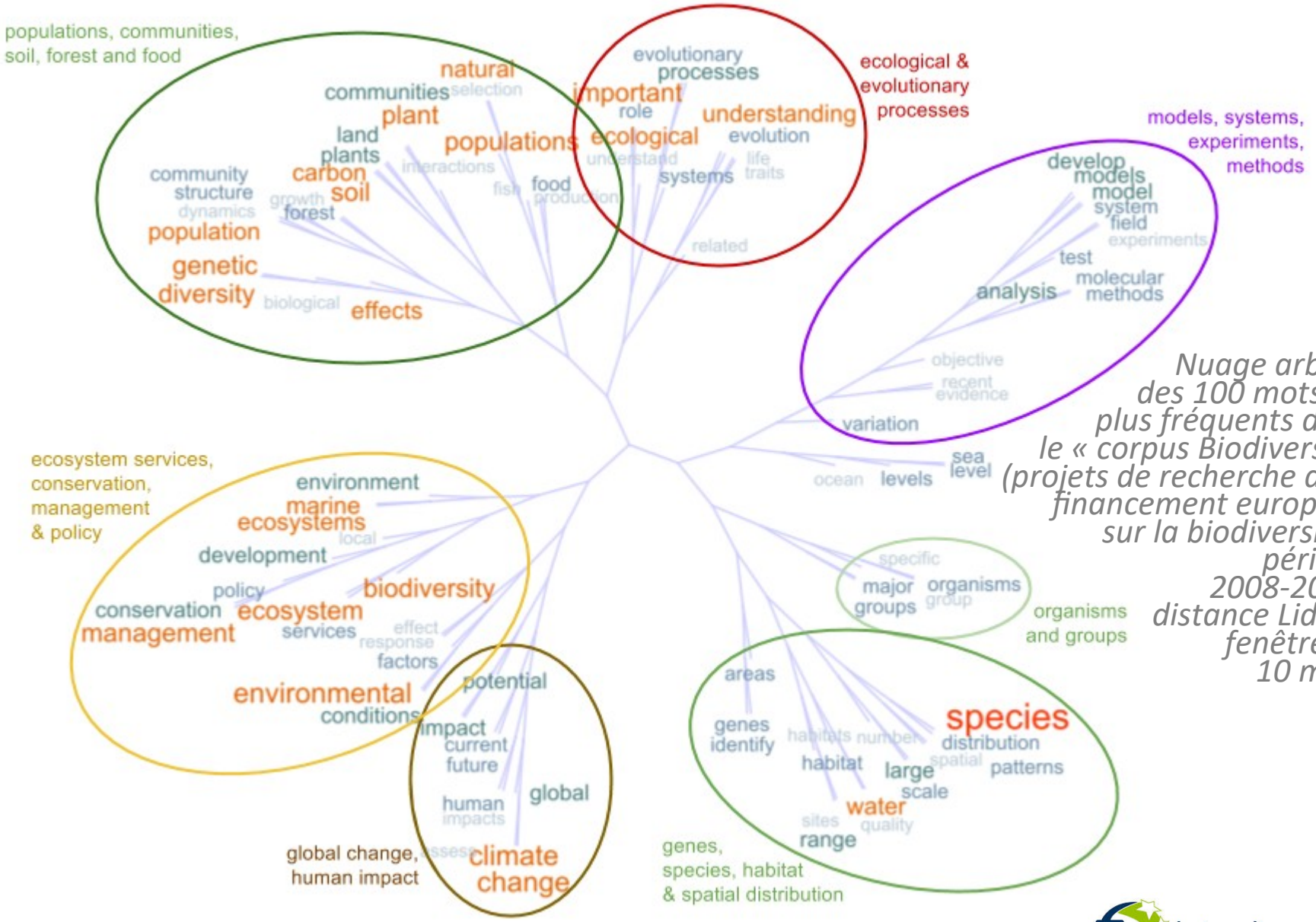
Méthode : comparaison de voisinages dans l'arbre



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

biodiversa

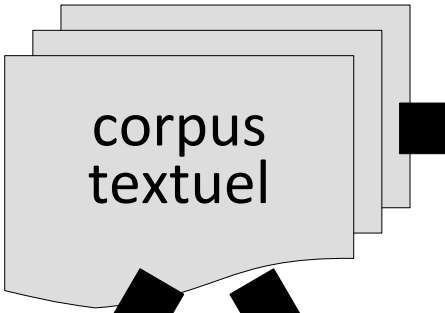
Méthode : comparaison de voisinages dans l'arbre

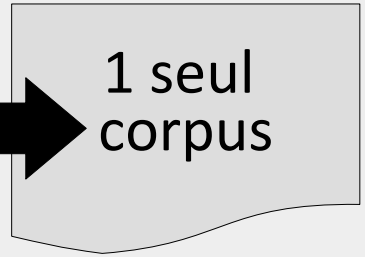


Nuage arboré des 100 mots les plus fréquents dans le « corpus Biodiversa » (projets de recherche avec financement européen sur la biodiversité), période 2008-2011, distance Liddel, fenêtre de 10 mots

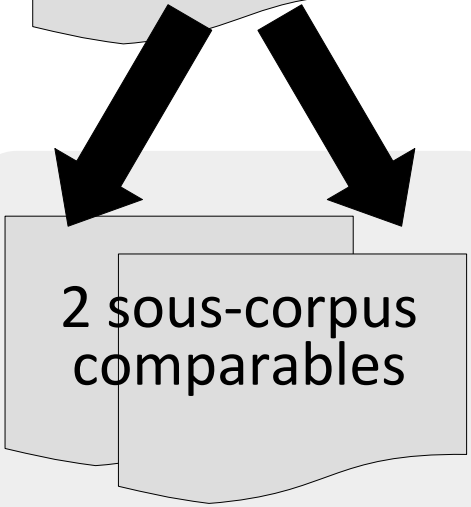
biodiversa

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 :

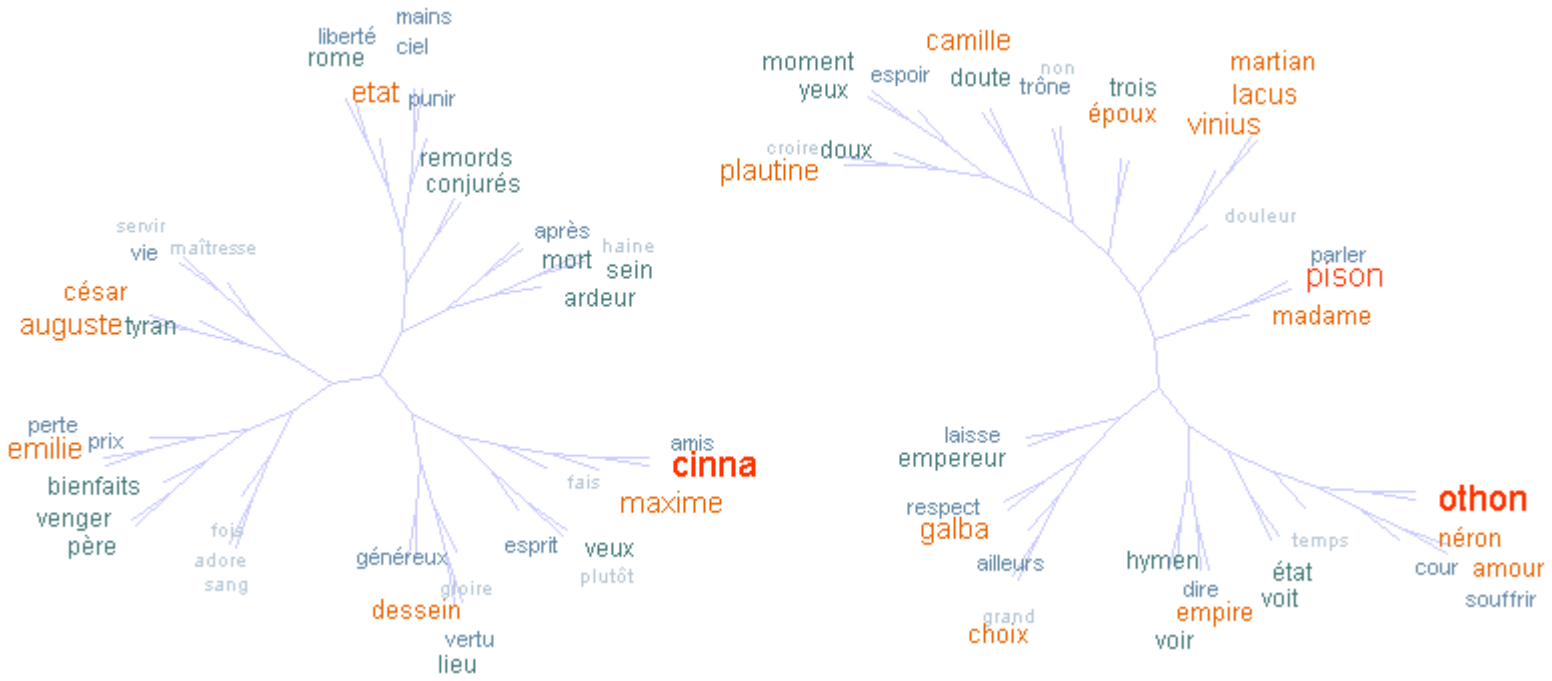
- Identifier les mots plutôt au début ou à la fin du corpus
- Analyser le début de chaque phrase du corpus
- Représenter le contexte dans le corpus de certains mots ciblés

- Identifier des regroupements communs ou différents
- Comparer les voisinages d'un mot cible dans les deux corpus
- 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

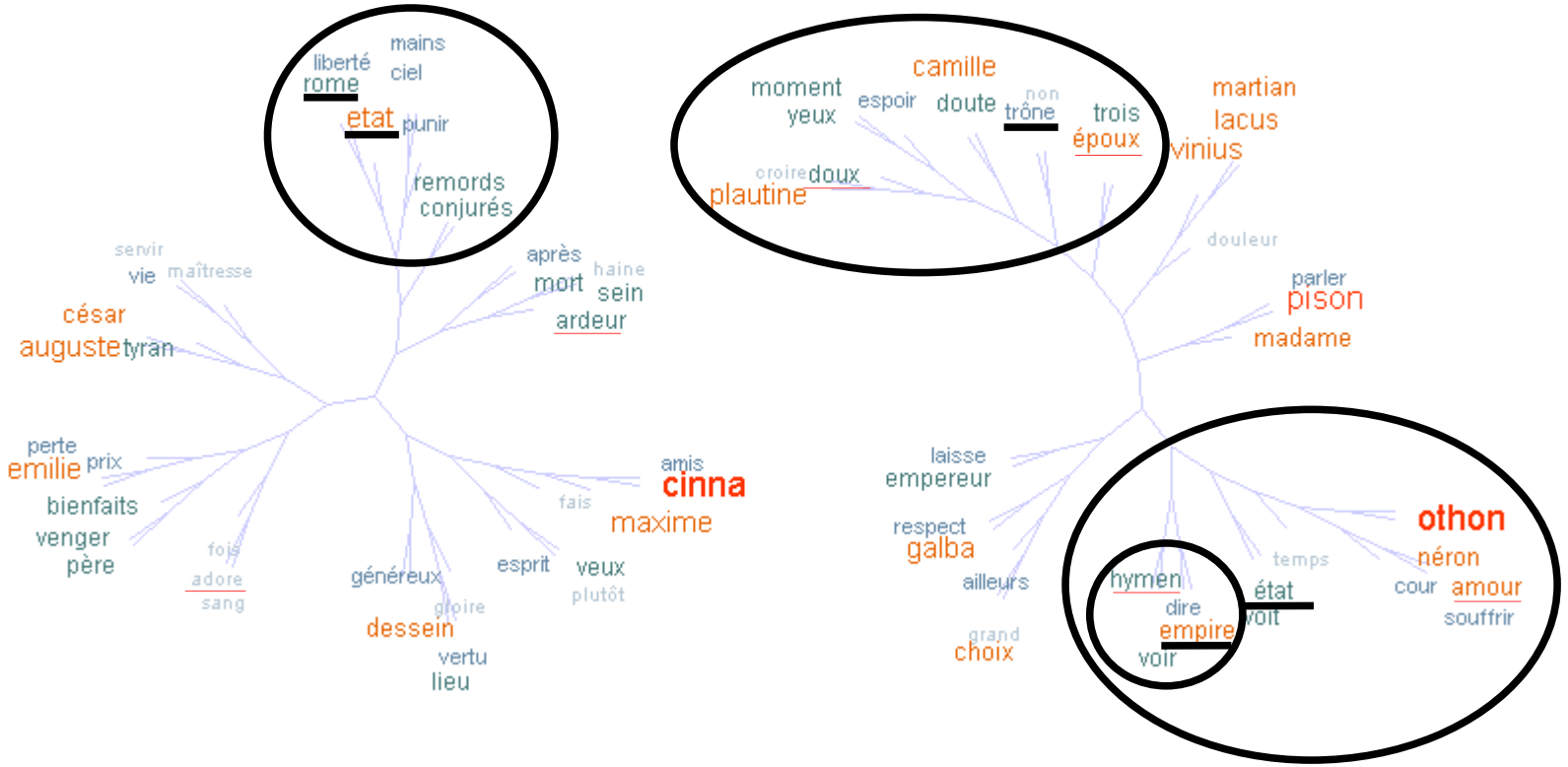


*Nuages arborés des **mots spécifiques** de Cinna et Othon, dimensionnés et colorés d'après leur spécificité calculée dans Lexico3.*

Quels moyens au service de la cause politique ?

Méthode : comparaison des spécifiques

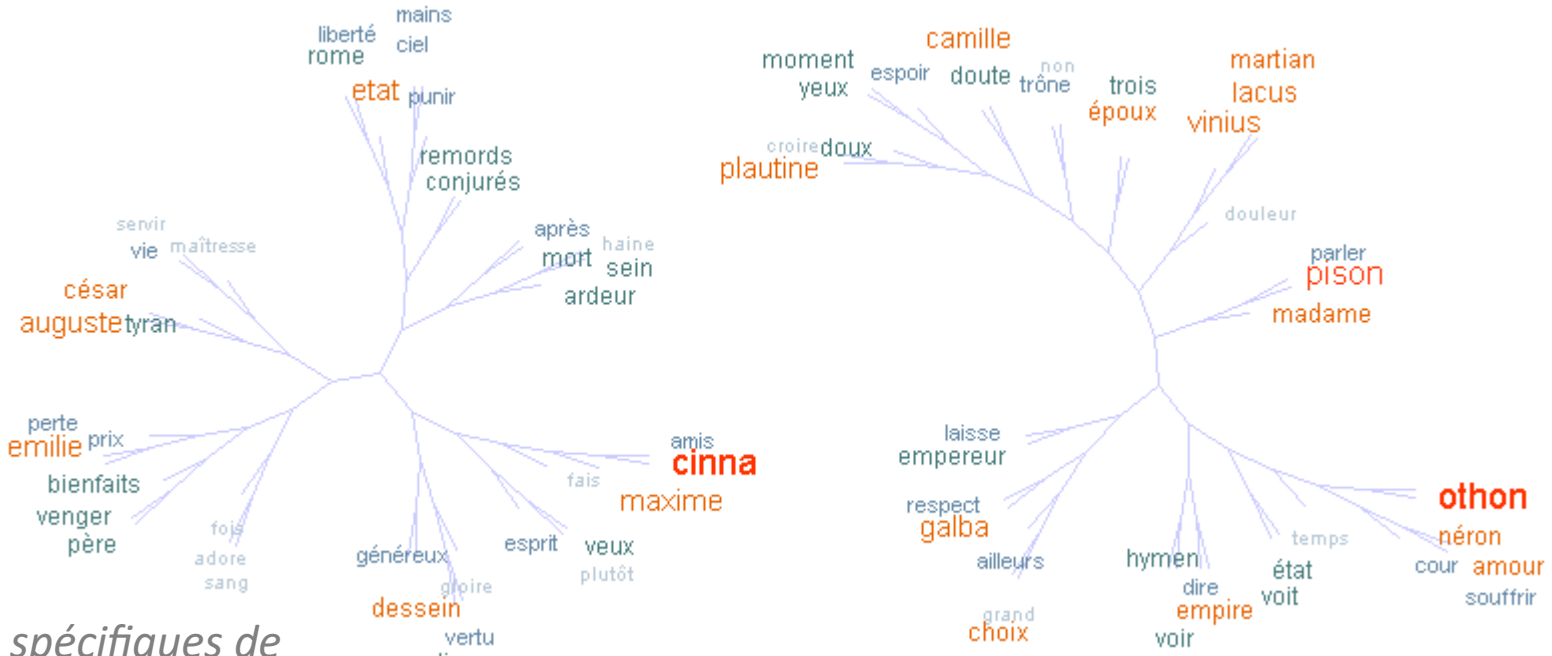
Amstutz & Gambette,
JADT 2010



*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

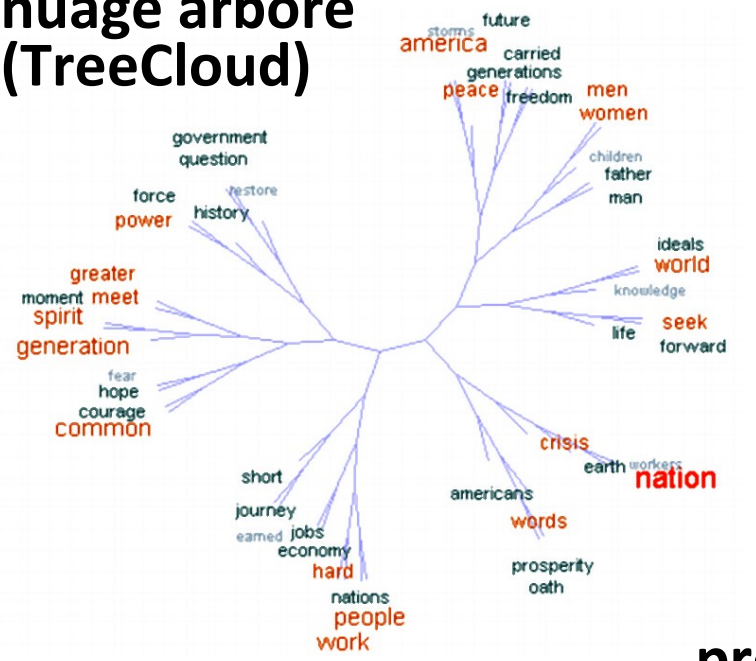


mots spécifiques de Cinna et Othon d'après Lexico3

	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

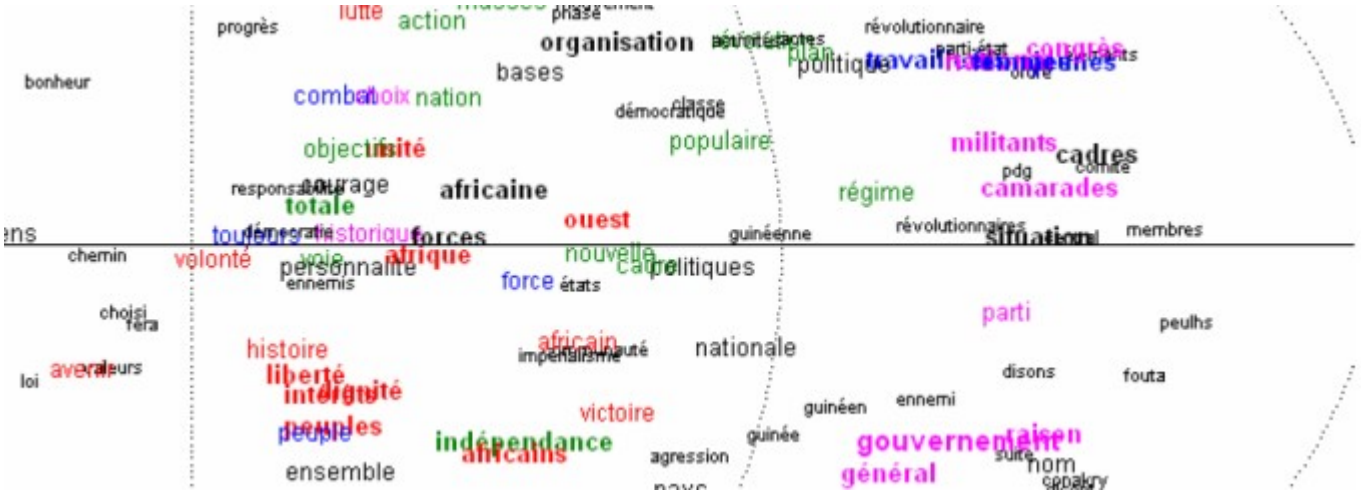
Comparaison avec d'autres visualisations

nuage arboré (TreeCloud)



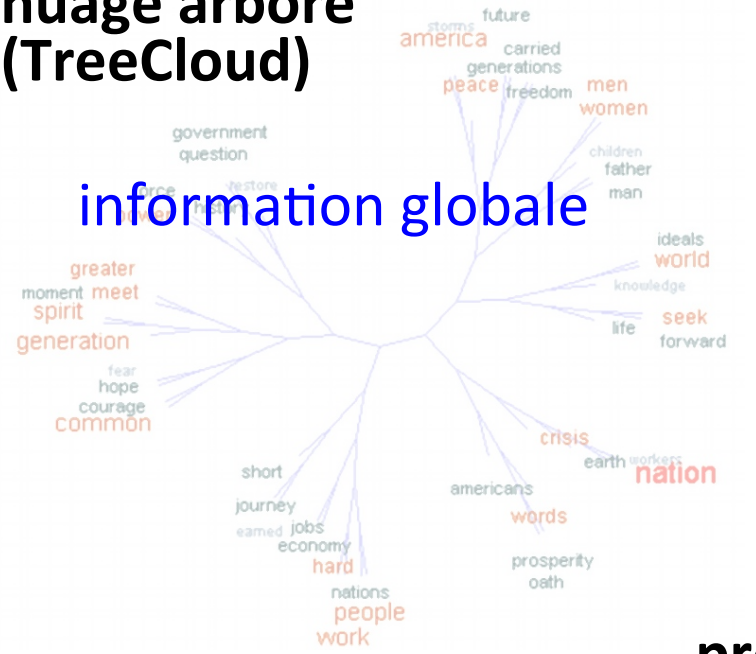
réseau de mots (PhraseNet d'IBM ManyEyes)

projection des mots (Astartex)



Comparaison avec d'autres visualisations

nuage arboré (TreeCloud)

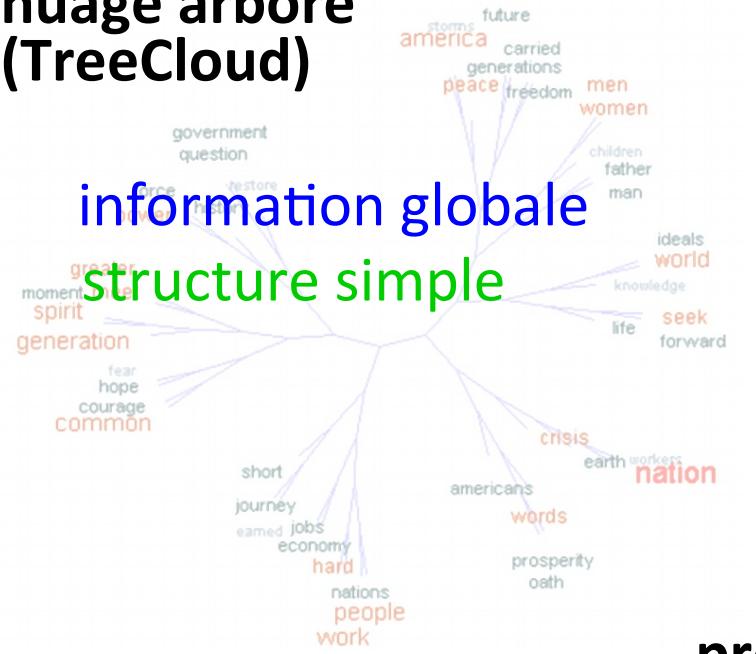


réseau de mots (PhraseNet d'IBM ManyEyes)

projection des mots (Astartex)

Comparaison avec d'autres visualisations

nuage arboré (TreeCloud)



information globale

structure simple

réseau de mots (PhraseNet d'IBM ManyEyes)

information locale

structure complexe

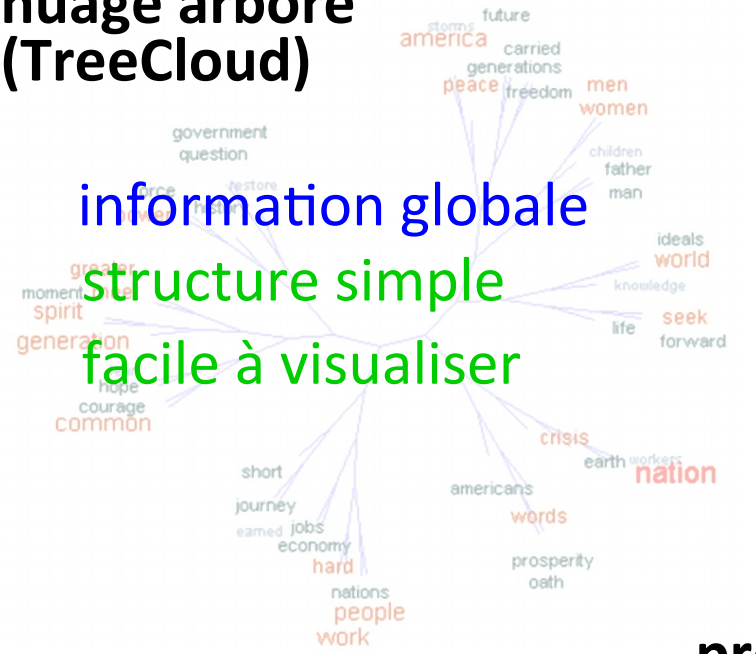
projection des mots (Astartex)

information globale

structure complexe

Comparaison avec d'autres visualisations

nuage arboré (TreeCloud)



information globale

structure simple

facile à visualiser

réseau de mots (PhraseNet d'IBM ManyEyes)

information locale

structure complexe

difficile à visualiser

projection des mots (Astartex)

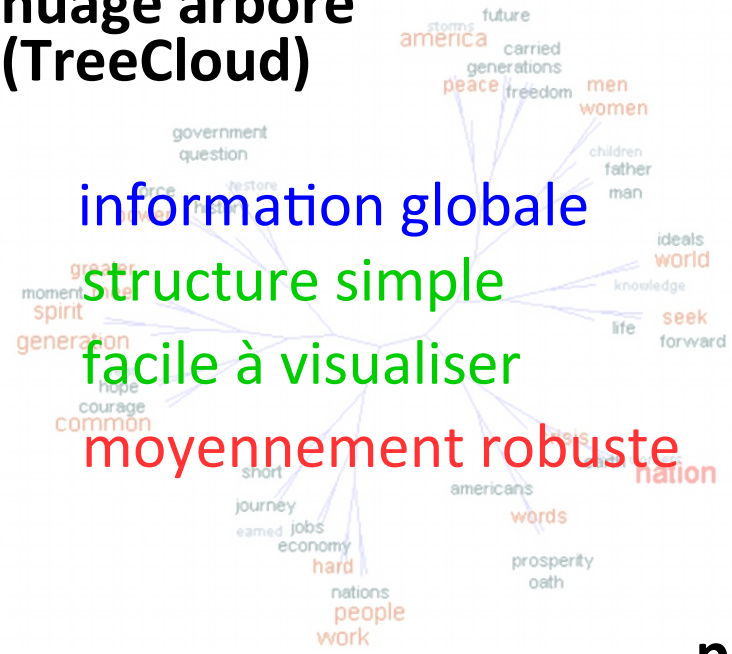
information globale

structure complexe

facile à visualiser mais chevauchements

Comparaison avec d'autres visualisations

nuage arboré (TreeCloud)



information globale

structure simple

facile à visualiser

moyennement robuste

réseau de mots (PhraseNet d'IBM ManyEyes)

information locale

structure complexe

difficile à visualiser

robuste

projection des mots (Astartex)

information globale

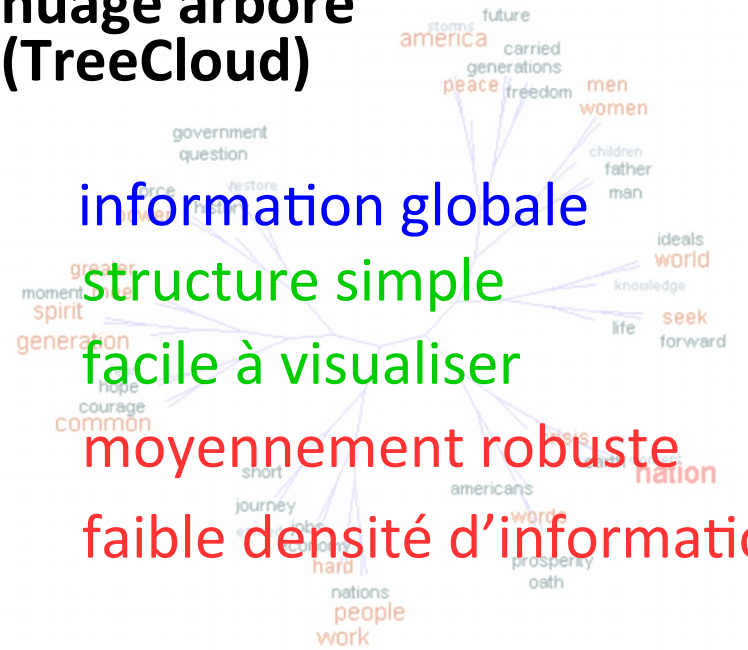
structure complexe

facile à visualiser mais chevauchements

robuste

Comparaison avec d'autres visualisations

nuage arboré (TreeCloud)



information globale

structure simple

facile à visualiser

moyennement robuste

faible densité d'information

réseau de mots (PhraseNet d'IBM ManyEyes, Tropes)

information locale

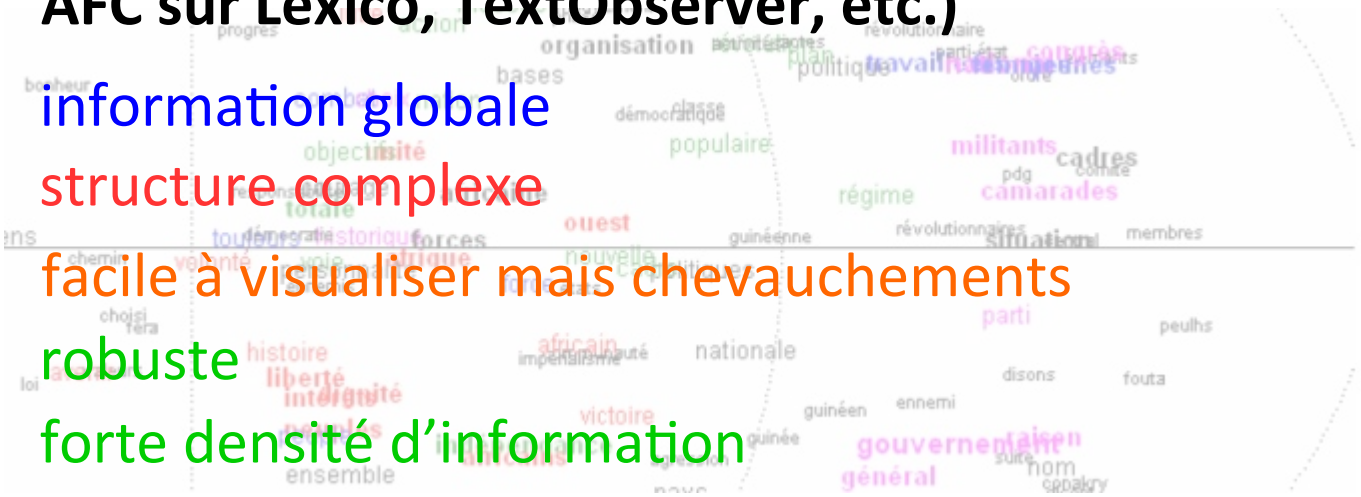
structure complexe

difficile à visualiser

robuste

forte densité d'information

projection des mots (Astartex, AFC sur Lexico, TextObserver, etc.)



information globale

structure complexe

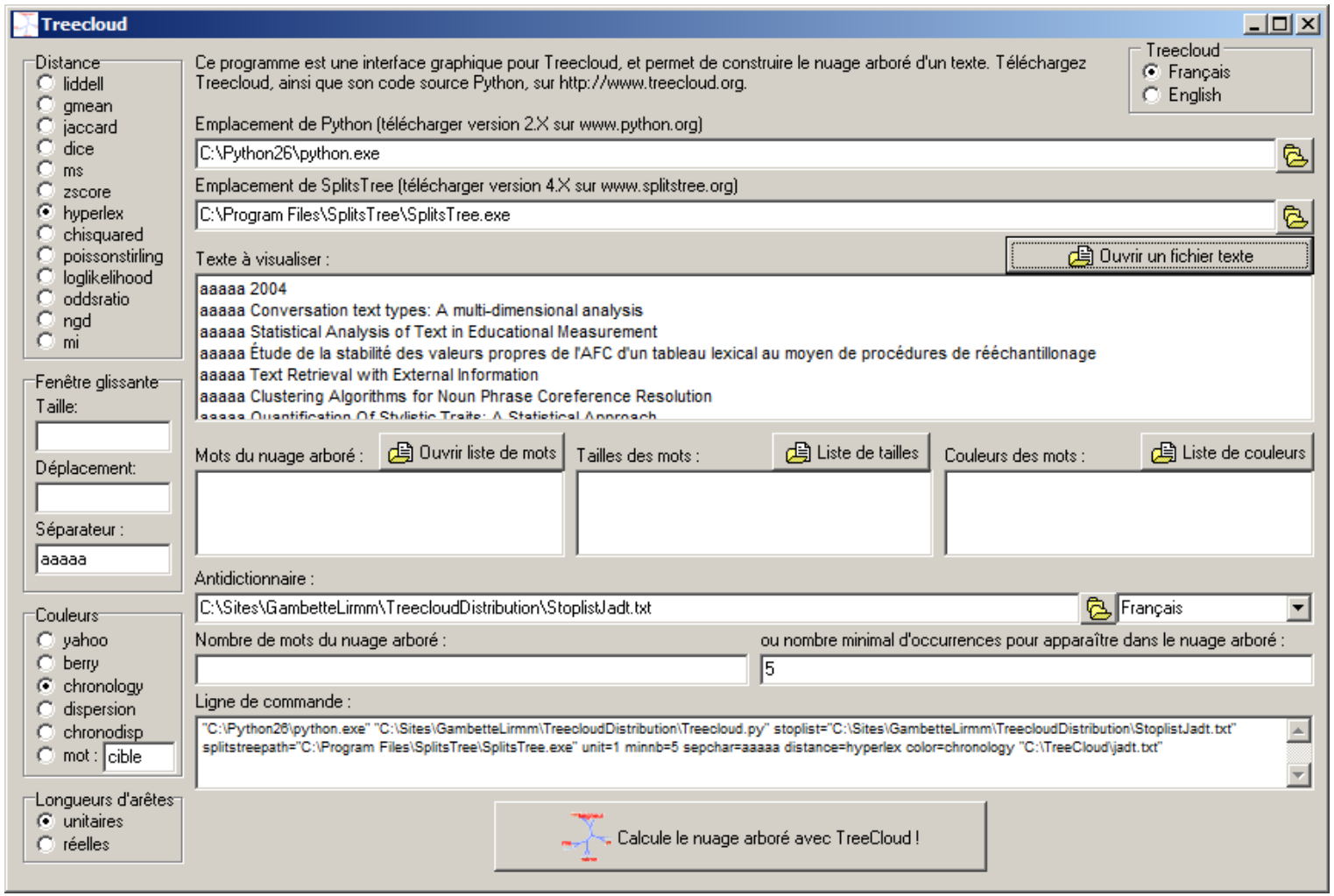
facile à visualiser mais chevauchements

robuste

forte densité d'information

Implémentations

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

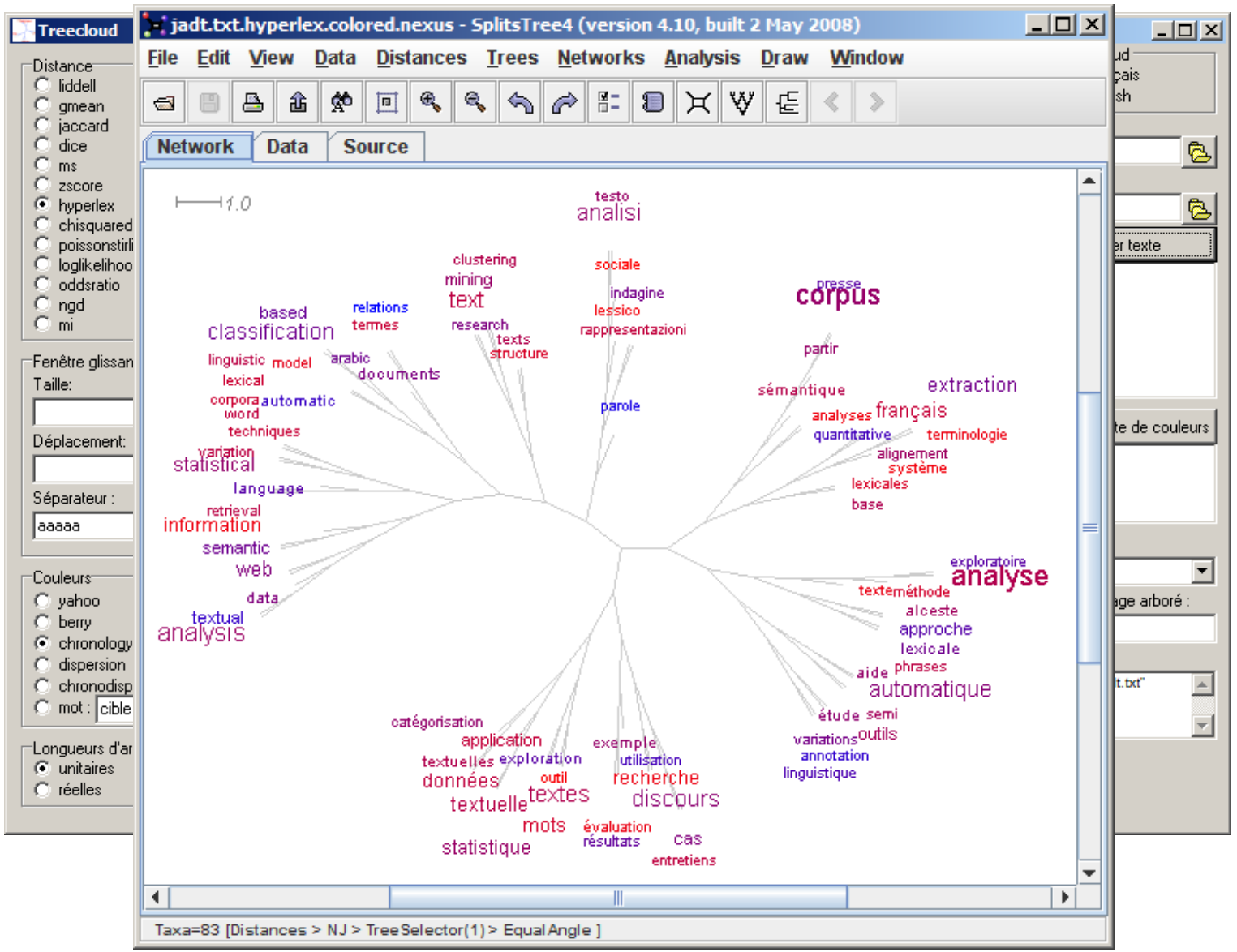


www.treecloud.org

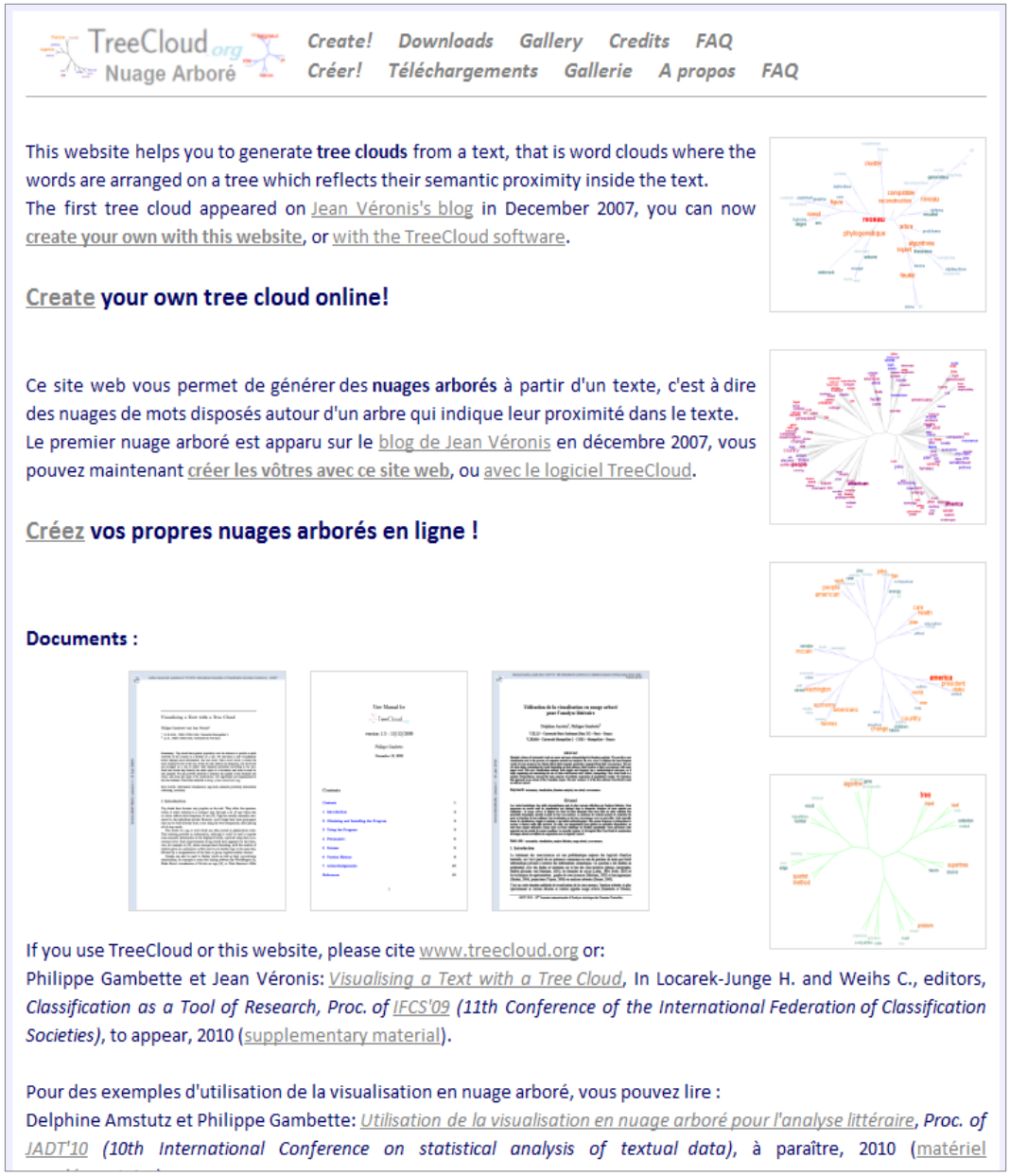
www.splitstree.org

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

TreeCloud.org
Nuage Arboré

Créer! Téléchargements Gallerie A propos FAQ

Créez vos propres nuages arborés !

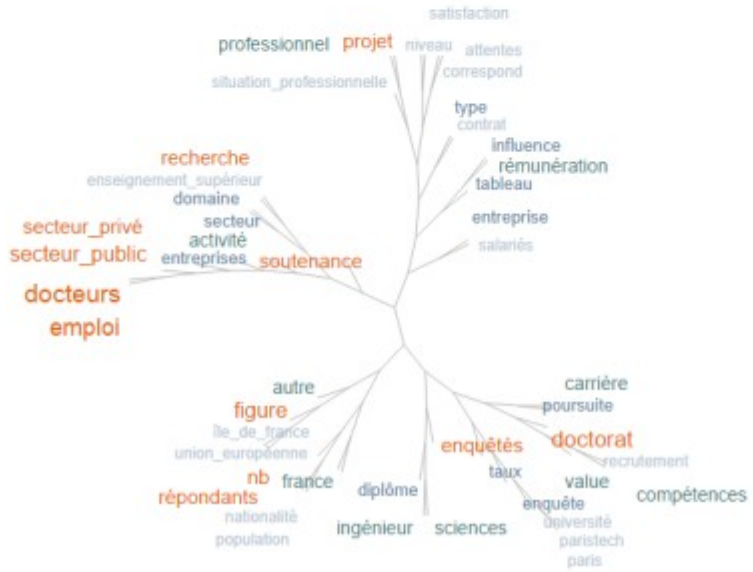
Collez votre texte dans le cadre ci-dessous, puis cliquez sur *Envoyer* ! Attention, l'utilisateur suivant verra votre texte quand il se connectera au site, si vous ne voulez pas faire apparaître vos textes, installez plutôt TreeCloud sur votre machine.

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

TreeCloud.org
Nuage Arboré

Créer! Téléchargements Gallerie A propos FAQ

Créez vos propres nuages arborés !

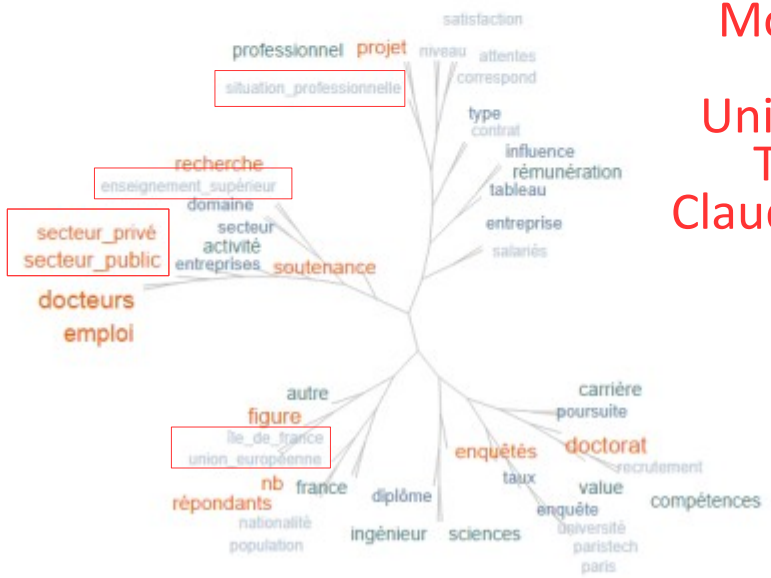
Collez votre texte dans le cadre ci-dessous, puis cliquez sur *Envoyer* ! Attention, l'utilisateur suivant verra votre texte quand il se connectera au site, si vous ne voulez pas faire apparaître vos textes, installez plutôt TreeCloud sur votre machine.

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

TreeCloud.org
Nuage Arboré

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

This website helps you to generate **tree clouds** from a text, that is word clouds where the words are arranged on a tree which reflects their sema
The first tree cloud appeared on Jean Véronis's blog
create your own with this website, or with the TreeClou

Create your own tree cloud online!

Ce site web vous permet de générer des **nuages arbo**
des nuages de mots disposés autour d'un arbre qui indi
Le premier nuage arboré est apparu sur le blog de Je
pouvez maintenant créer les vôtres avec ce site web, o

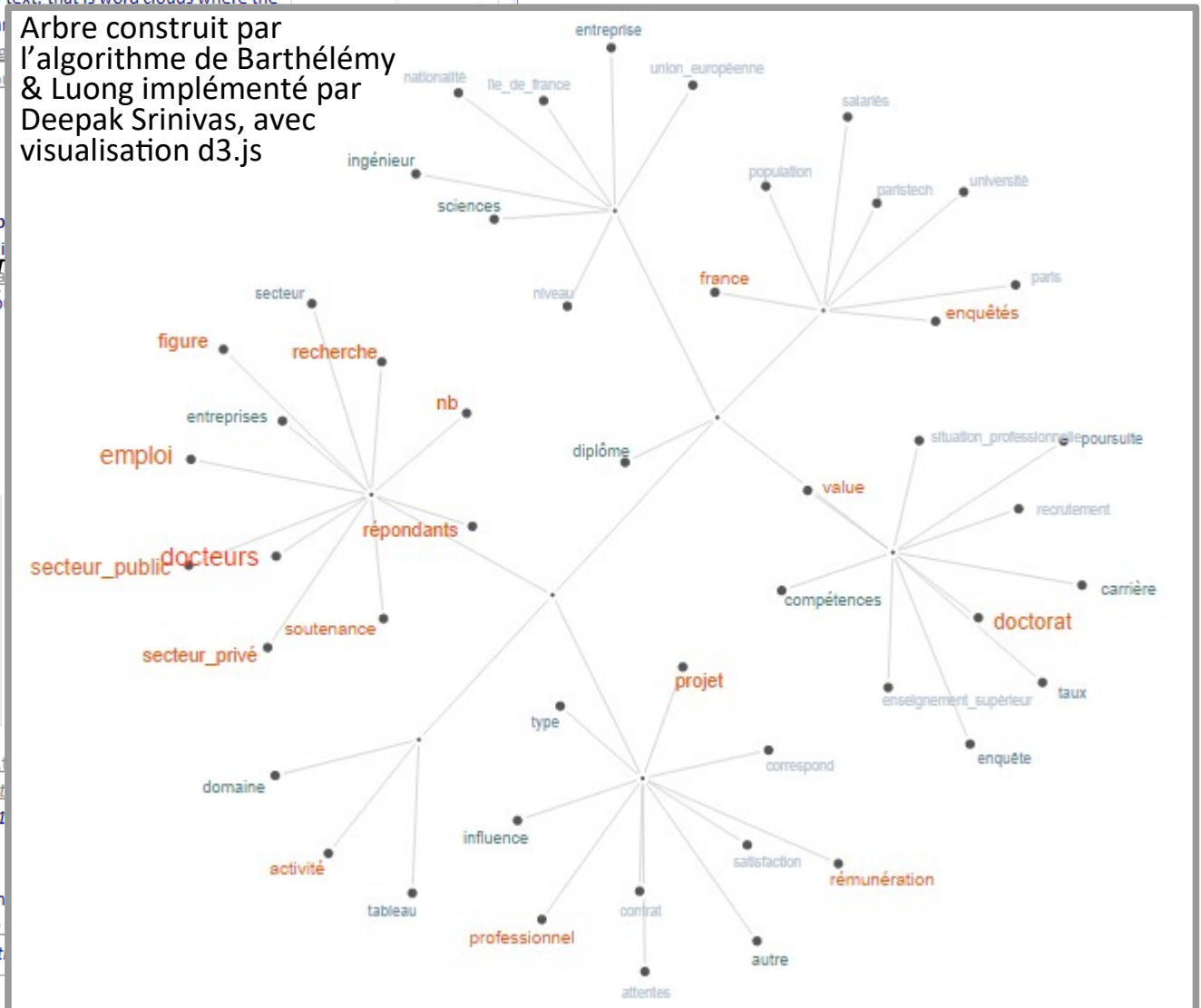
Créez vos propres nuages arborés en ligne

Documents :

If you use TreeCloud or this website, please cite www.t
Philippe Gambette et Jean Véronis: *Visualising a Text Classification as a Tool of Research, Proc. of IFCS'09 (1 Societies)*, to appear, 2010 (supplementary material).

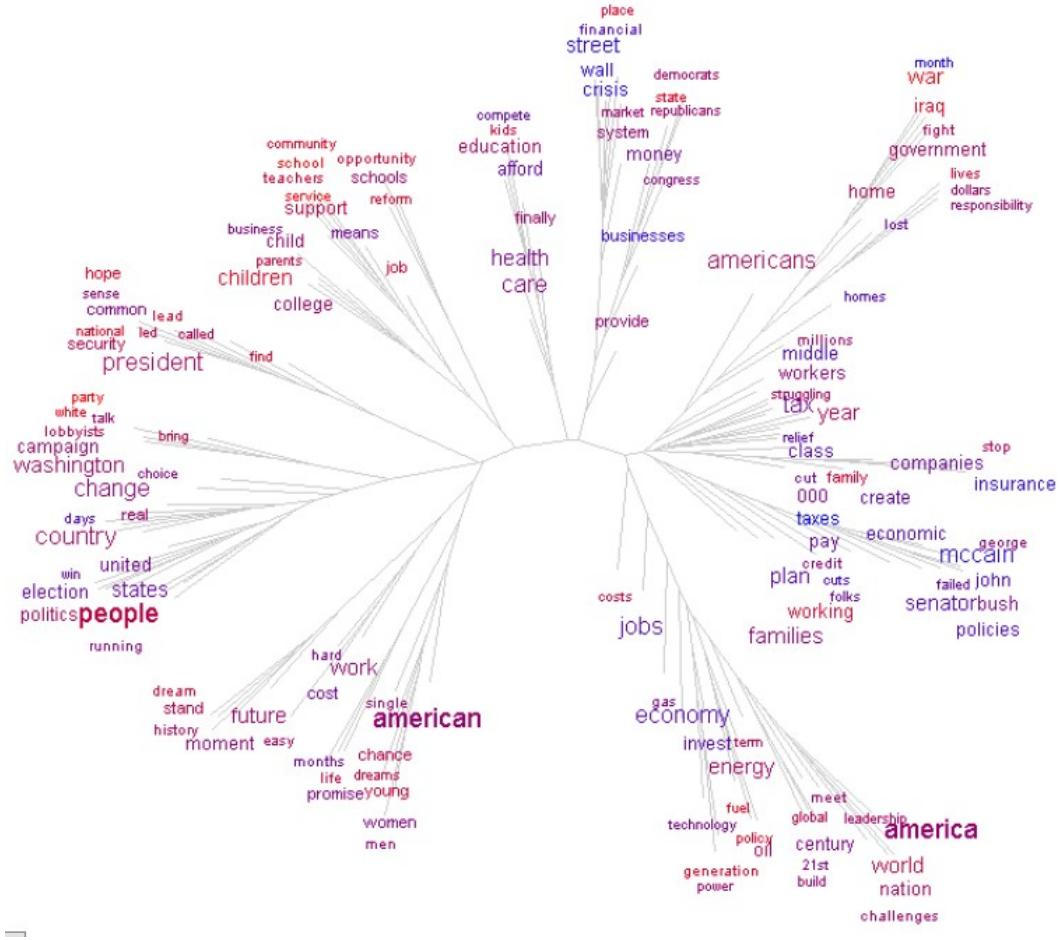
Pour des exemples d'utilisation de la visualisation en n
Delphine Amstutz et Philippe Gambette: *Utilisation de JADT'10 (10th International Conference on statist*

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)

Références (*treecloud.org*)

Philippe Gambette, Jean Véronis (2009)
Visualising a Text with a Tree Cloud, *IFCS'09, Studies in Classification, Data Analysis, and Knowledge Organization* 40, p. 561-570
http://igm.univ-mlv.fr/~gambette/Re20090317.pdf

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://igm.univ-mlv.fr/~gambette/Re20100611.pdf

Philippe Gambette, Nuria Gala & Alexis Nasr (2012)
Longueur de branches et arbres de mots, *Corpus* 11:129-146
http://igm.univ-mlv.fr/~gambette/Re20120209.pdf

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

Philippe Gambette, Hilde Eggermont & Xavier Le Roux (2014)
Temporal and geographical trends in the type of biodiversity research funded on a competitive basis in European countries, *rapport BiodivERsa*
http://www.biodiversa.org/700/download

Nadège Lechevrel & Philippe Gambette (2016)
Une approche textométrique pour étudier la transmission des savoirs biologiques au XIXe siècle, *Nouvelles perspectives en sciences sociales* 12(1):221-253
https://hal-upec-upem.archives-ouvertes.fr/hal-01408455

Philippe Gambette, Tita Kyriacopoulou, Nadège Lechevrel & Claude Martineau (2017)
Anatomie, animaux, vocabulaire de la vivisection : construire des ressources lexicales pour visualiser une thématique dans un corpus littéraire, *Colloque AnimalHumanité, Expérimentation et fiction : l'animalité au cœur du vivant*, décembre 2016
https://hal-upec-upem.archives-ouvertes.fr/hal-01609198

Claude Martineau (2017)
TreeCloud, Unitex: increased synergy, *ECLAVIT Workshop*, 24 novembre 2017
https://hal-upec-upem.archives-ouvertes.fr/hal-01702091

Atelier

Tutoriel :
https://docs.google.com/document/d/1OauE9EflJTyVR3gM7ZPc3cGJ3-N0Iq2ghPD5RNrb_YY/edit?usp=sharing

Possibilités de **réalisations numériques sur vos corpus** :

- en ajoutant des liens sur les mots du nuage arboré, avec *TreeCloud Linker* :
  - Mots-clés des publications de l'UPEM :
    http://treecloud.univ-mlv.fr/treecloud-linker/
  - Collections du musée Fragonard de l'école vétérinaire d'Alfort :
    http://treecloud.univ-mlv.fr/treecloud-linker/fragonard.html

- en les chargeant dans *TreeCloud Corpus* :
  - *Vœux présidentiels* rassemblés par Jean-Marc Leblanc, 1960-2018
    http://treecloud.univ-mlv.fr/treecloud-voeux/
  - *Lettres républicaines* de Daniel Stern (pseudonyme de Marie d'Agoult), 1848
    http://treecloud.univ-mlv.fr/treecloud-corpus/lettres-republicaines/

Prétraitements divers

Utilisation de formules de tableur pour pré-traiter des corpus :

- **Données d'enquête** :
  - ajouter “ a a a a a a a a a a ” à la fin de chaque réponse à une question ouverte, pour éviter que les mots d'une réponse soient considérés proches des mots de la réponse suivante (si *fenêtre glissante* paramétrée à 10 mots).
  - possibilité de filtrer les lignes pour sélectionner uniquement les réponses d'un échantillon donné.

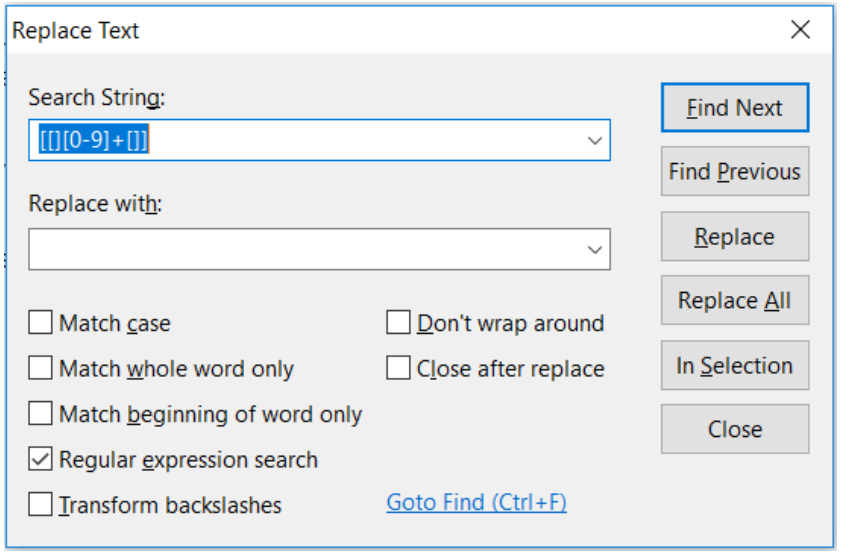
- **Données d'entretien** :
  - mettre le nom du locuteur sur la ligne précédant ses paroles
  - possibilité d'utiliser une formule pour créer une colonne avec le nom du locuteur sur chaque ligne
  - Filtre pour sélectionner seulement les paroles d'un locuteur

→ consulter et copier ou télécharger **ce document tableur partagé**

Prétraitements de textes obtenus par OCR

Rechercher/remplacer (par exemple avec Notepad2) :

- remplacer les apostrophes courbes : remplacer "’" par "'"
- supprimer les références aux notes de fin de texte, en utilisant les “expressions régulières”, remplacer "[[][0-9]+[]]" par "" : un caractère "[" suivi d'un chiffre, éventuellement répété, suivi d'un caractère "]".



Prétraitements de textes obtenus par OCR

Aide automatique à la détection des césures de fin de ligne :
coupeCésure
http://igm.univ-mlv.fr/~gambette/text-processing/coupeCesure/

Code couleur :

- mot dont la césure a été supprimée car trouvé en entier dans le dictionnaire
- mot pour lequel le trait d'union a été gardé

Texte obtenu après remplacements...

Quand vous y ferez quelque réflexion, je crois
que vous trouverez que j'ai raison, et que si je fusse retournée , je rendois mon voyage inutile par être trop court. Pour mon fils et sa femme, ils sont ravis de passer ici jusqu'au carême avec moi : en ce temps-là j'irai à Rennes par complaisance pour eux, parce que ce
temps est plus triste que l'hiver à la campagne : peut-être que ce projet changera, il ne faut point voir de si loin. Ge qui est sûr, ma fille, c'est que l'air d'ici est fort
bon; vous lui faites tort de le croire mauvais. Il fait
depuis plus de deux mois le plus beau temps du monde,
des chaleurs dans la canicule, un mois de septembre

Extraction de contextes avec VoyantTools

- Charger le corpus dans **http://voyant-tools.org**
(ex. : corpus des voeux présidentiels, de Jean-Marc Leblanc ; sélection de plusieurs fichiers txt sur le disque dur avec le bouton « Charger »)
- Charger les **contextes** de “religion”, par exemple, dans le cadre en bas à droite, puis exporter avec le bouton entouré en rouge :

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

→ Visualiser les fréquences d'une liste de mots **VisuLexique**