Journée d'études
« Outils de traitement de corpus textuels développés à Paris-Est »
18/06/2014 – Créteil

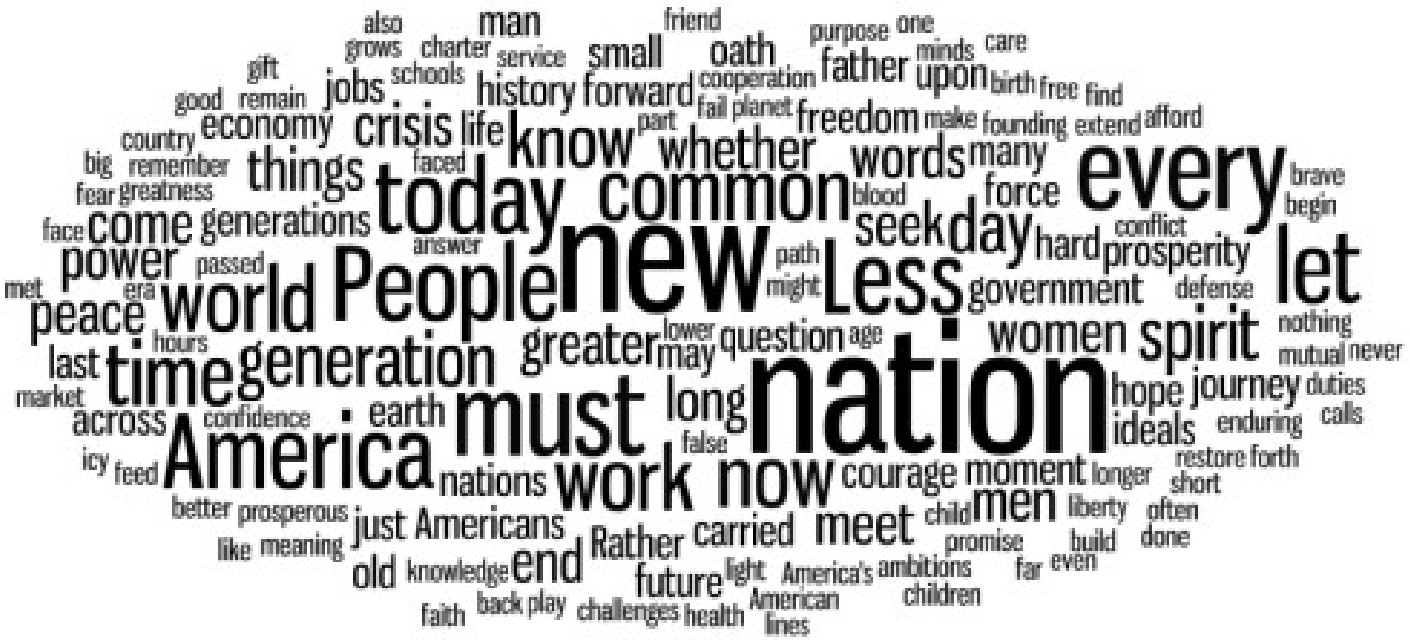
Visualisation et analyse de textes à partir d'arbres de mots avec TreeCloud

Philippe Gambette

LIGM
Université Paris-Est
Marne-la-Vallée

Le « nuage arboré », une information double

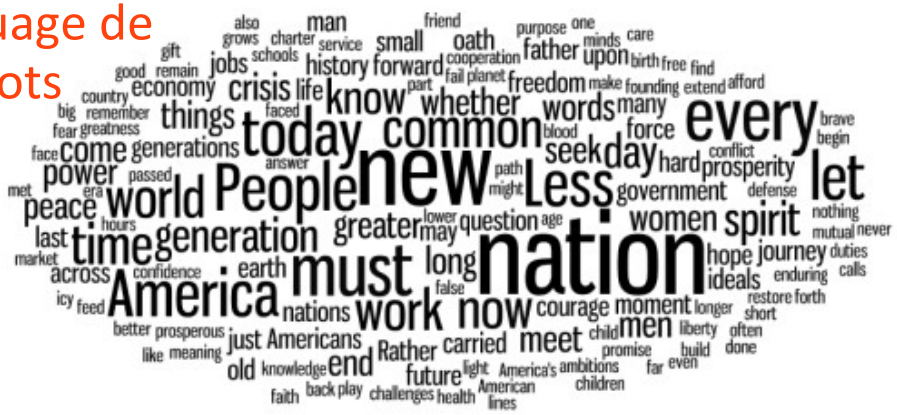
nuage de mots



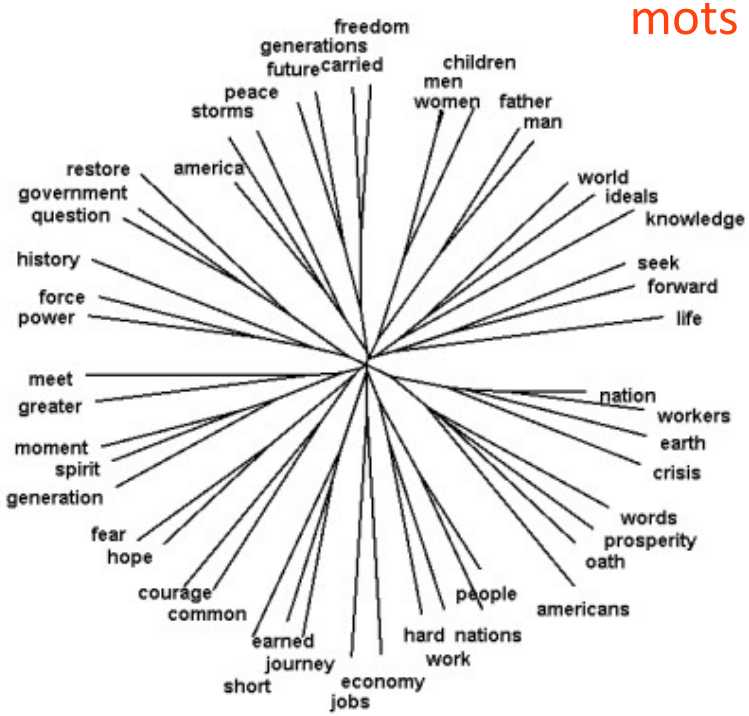
Discours inaugural de Barack Obama en 2008, Wordle

Le « nuage arboré », une information double

nuage de mots



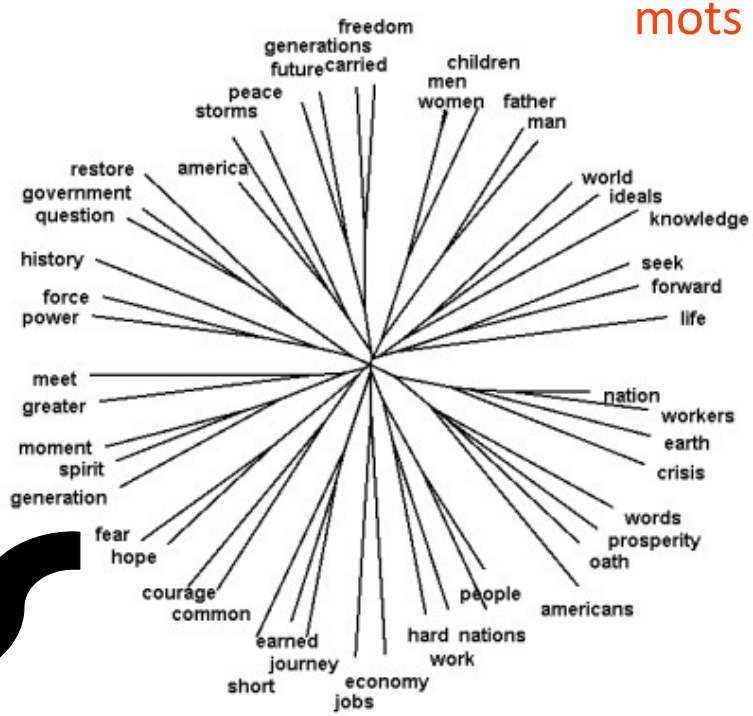
arbre de mots

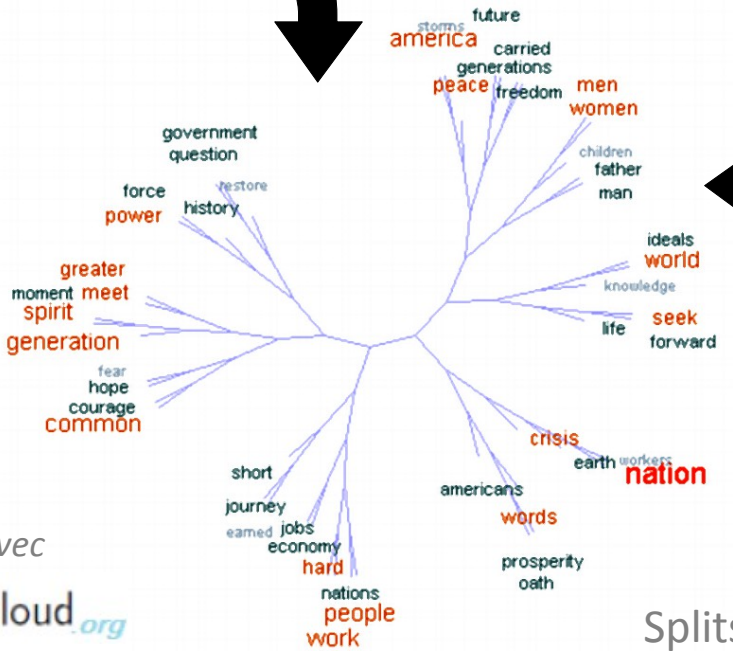


Le « nuage arboré », une information double

nuage de mots

arbre de mots





Discours inaugural de Barack Obama

construit avec

TreeCloud.org

SplitsTree4

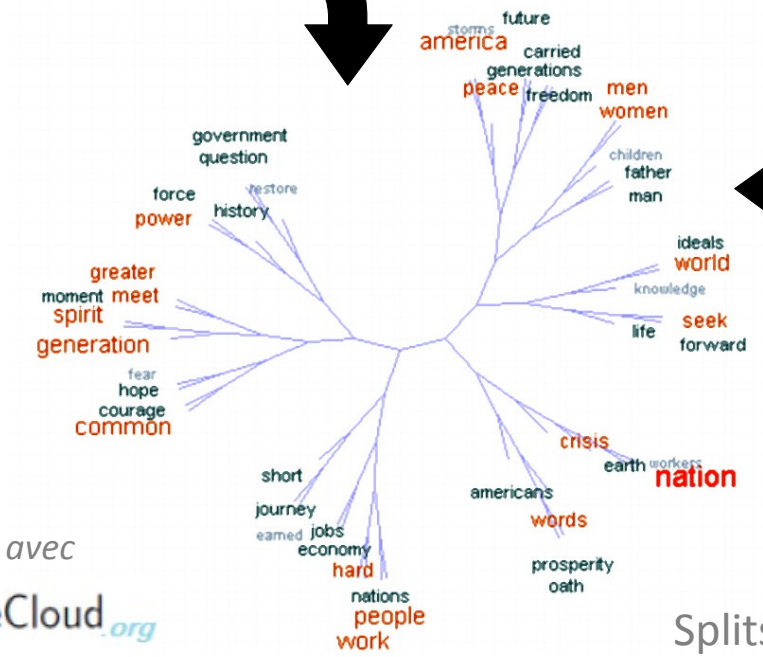
SplitsTree : Huson & Bryant, *Bioinformatics*, 2006

TreeCloud : Gambette & Véronis, *IFCS'09*

Le « nuage arboré », une information double

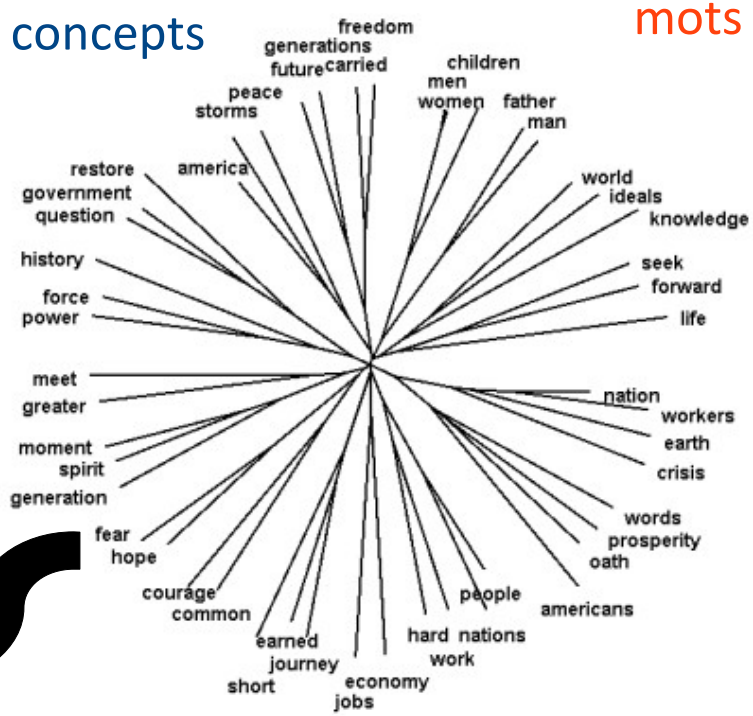
nuage de mots

hiérarchie des mots



hiérarchie des concepts

arbre de mots



Discours inaugural de Barack Obama

construit avec

TreeCloud.org

SplitsTree4

SplitsTree : Huson & Bryant, *Bioinformatics*, 2006

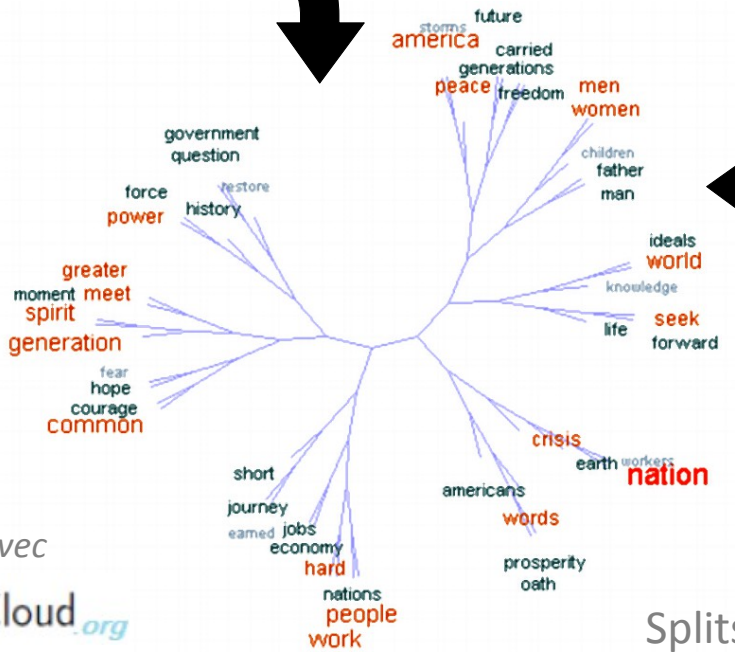
TreeCloud : Gambette & Véronis, *IFCS'09*

Le « nuage arboré », une information double

nuage de mots

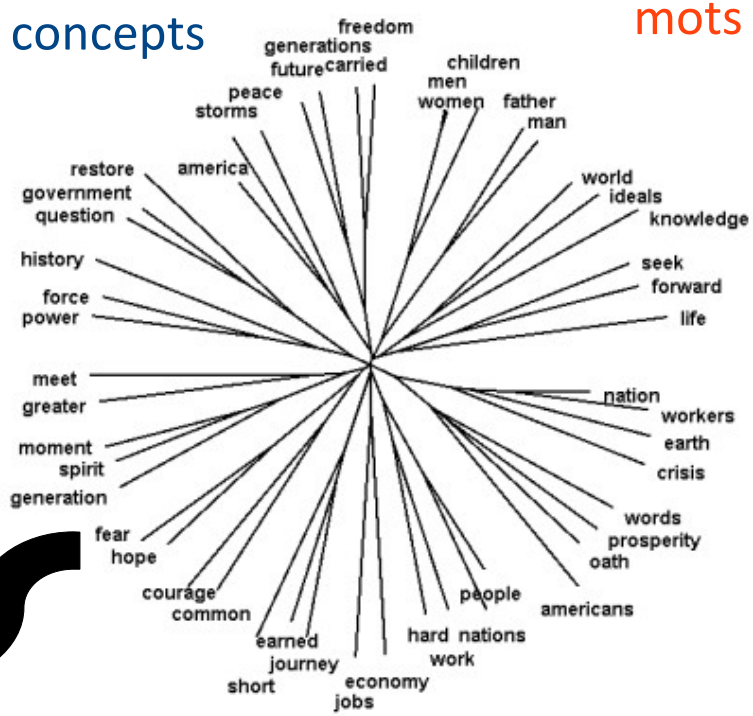
hiérarchie des mots

occurrences



hiérarchie des concepts

arbre de mots



cooccurrences

Discours inaugural de Barack Obama

construit avec

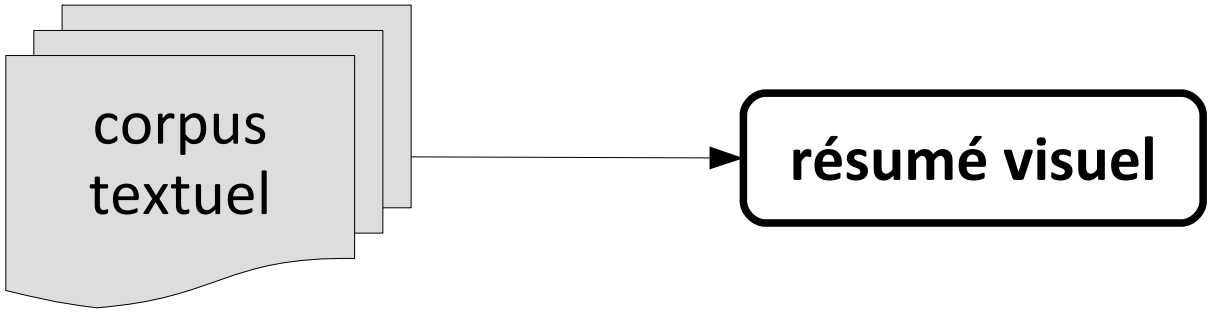
TreeCloud.org

SplitsTree4

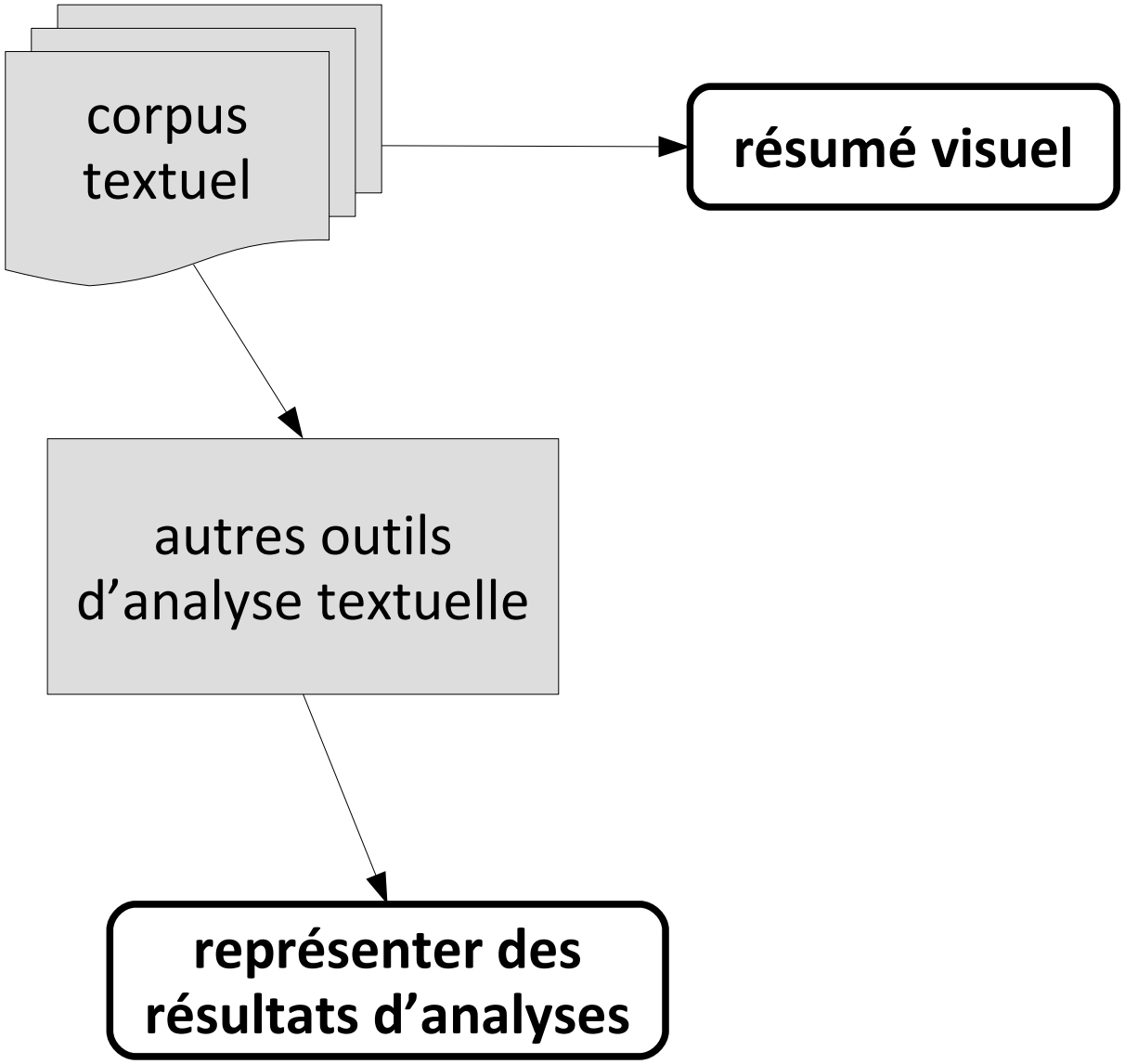
SplitsTree : Huson & Bryant, *Bioinformatics*, 2006

TreeCloud : Gambette & Véronis, *IFCS'09*

Le « nuage arboré », pour quoi faire ?

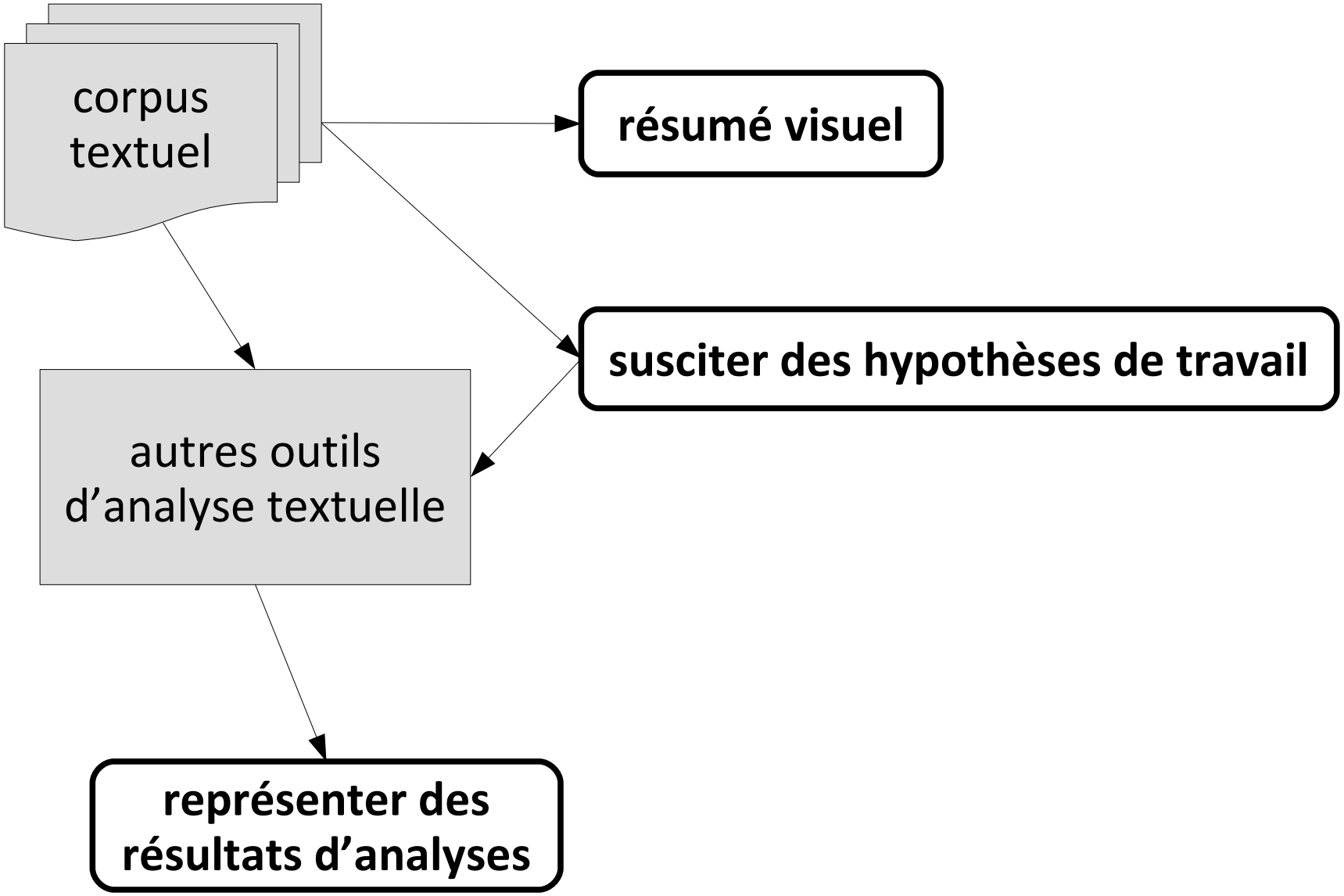


Le « nuage arboré », pour quoi faire ?



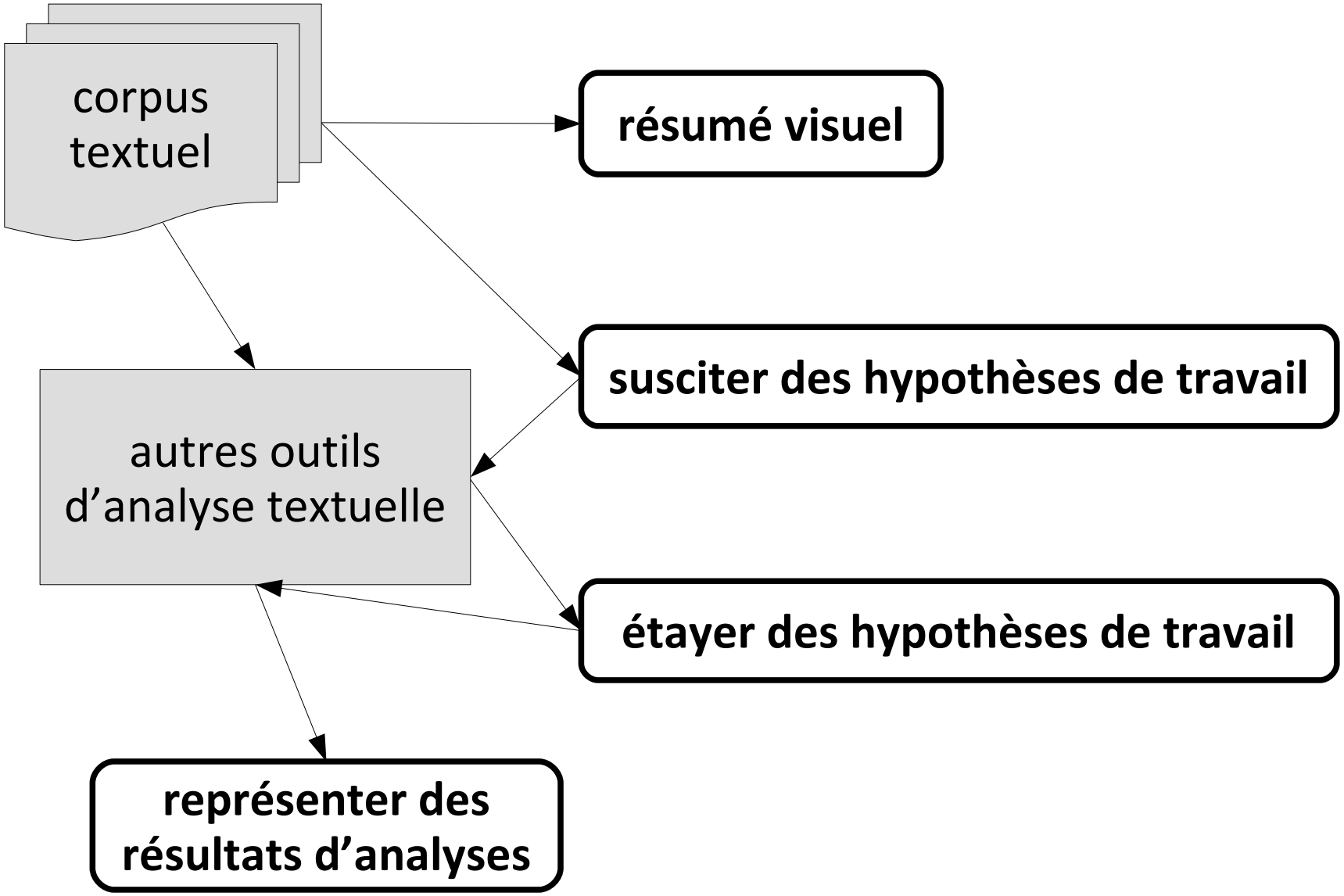
Amstutz & Gambette, *JADT 2010*

Le « nuage arboré », pour quoi faire ?



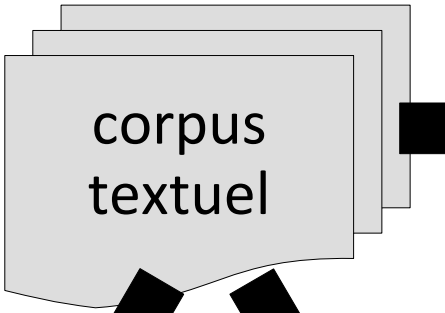
Amstutz & Gambette, *JADT 2010*

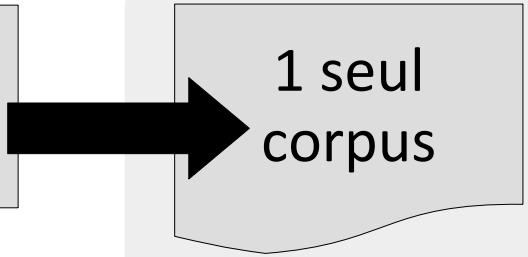
Le « nuage arboré », pour quoi faire ?



Amstutz & Gambette, *JADT 2010*

Exploration de corpus avec TreeCloud

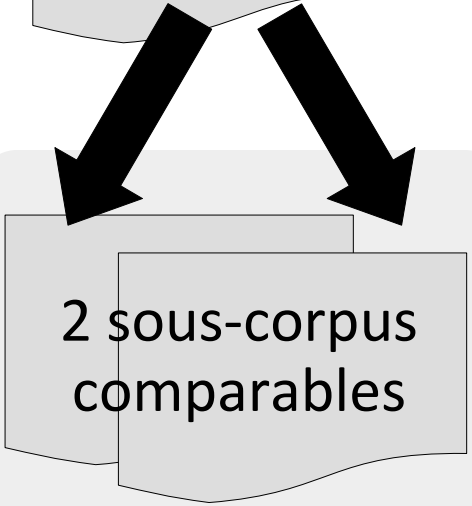




- Interpréter les regroupements
- Examiner les voisinages des mots fréquents dans l'arbre
- Examiner le voisinage de mots ciblés dans l'arbre («acteurs», verbes, etc.)

Si le corpus est assez grand :

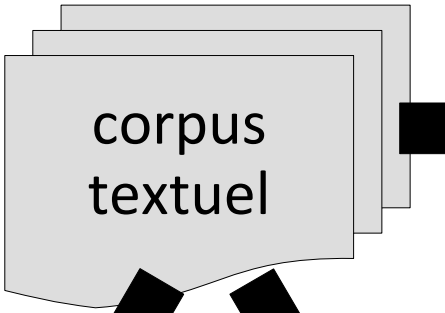
- Identifier les mots plutôt au début ou à la fin du corpus
- extraire le voisinage dans le corpus de certains mots ciblés et utiliser ces données comme corpus.

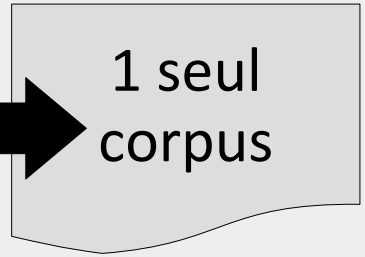


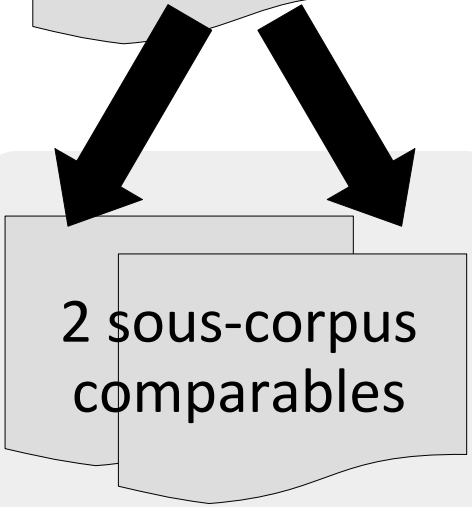
- Identifier des regroupements communs ou différents
- Comparer la fréquence des mots fréquents dans les deux arbres
- Comparer (quanti+quali) les regroupements dans les deux arbres
- Construire le nuage arboré des mots spécifiques de chaque sous-corpus

Travail en cours avec Edna Hernandez

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.

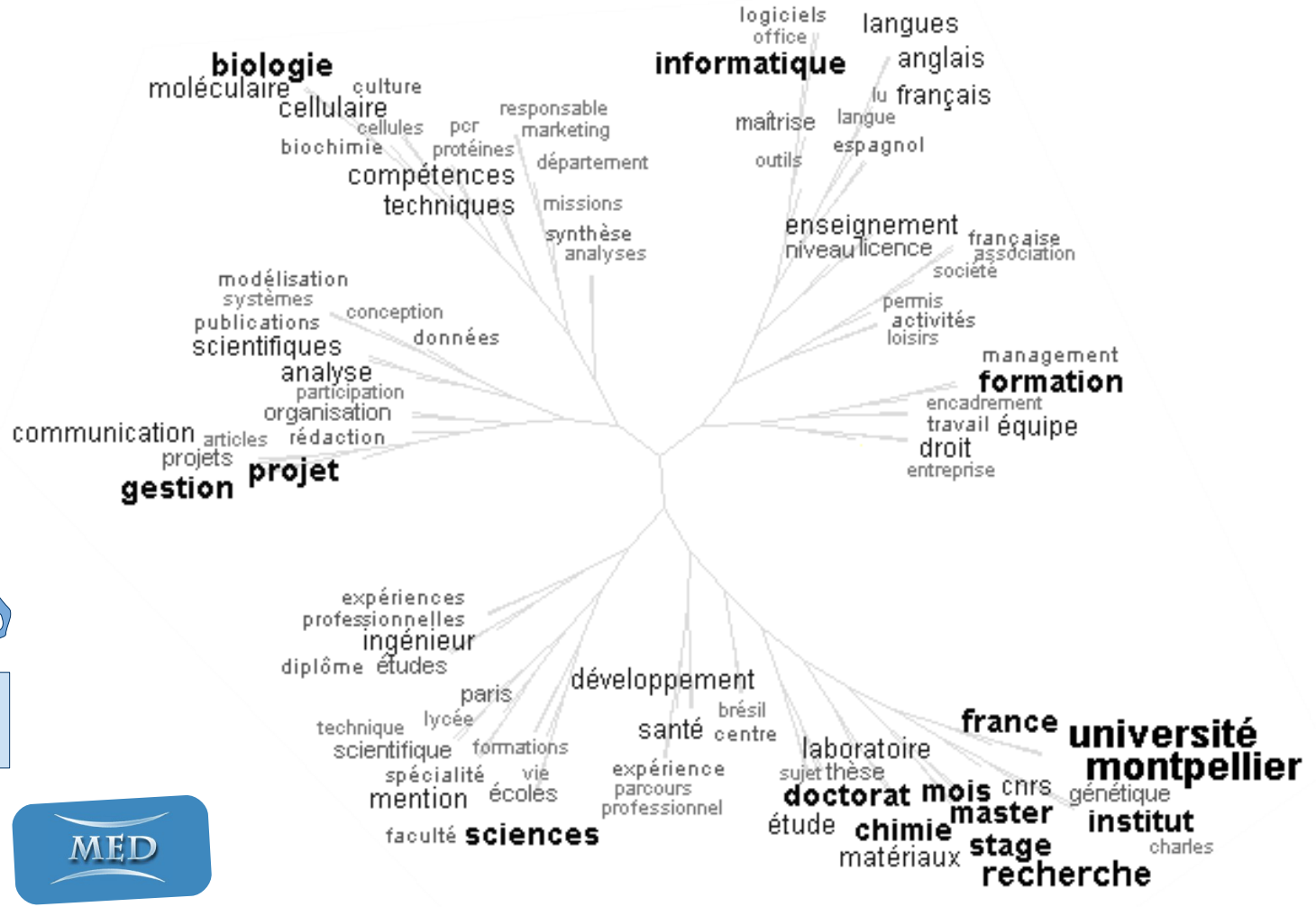
- Identifier des regroupements communs ou différents
- Comparer la fréquence des mots fréquents dans les deux arbres
- Comparer (quanti+quali) les regroupements dans les deux arbres
- Construire le nuage arboré des mots spécifiques de chaque sous-corpus

Travail en cours avec Edna Hernandez

Méthode : interpréter les regroupements

Dessiner des « patates »

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



Rencontre Docteurs & Entreprises

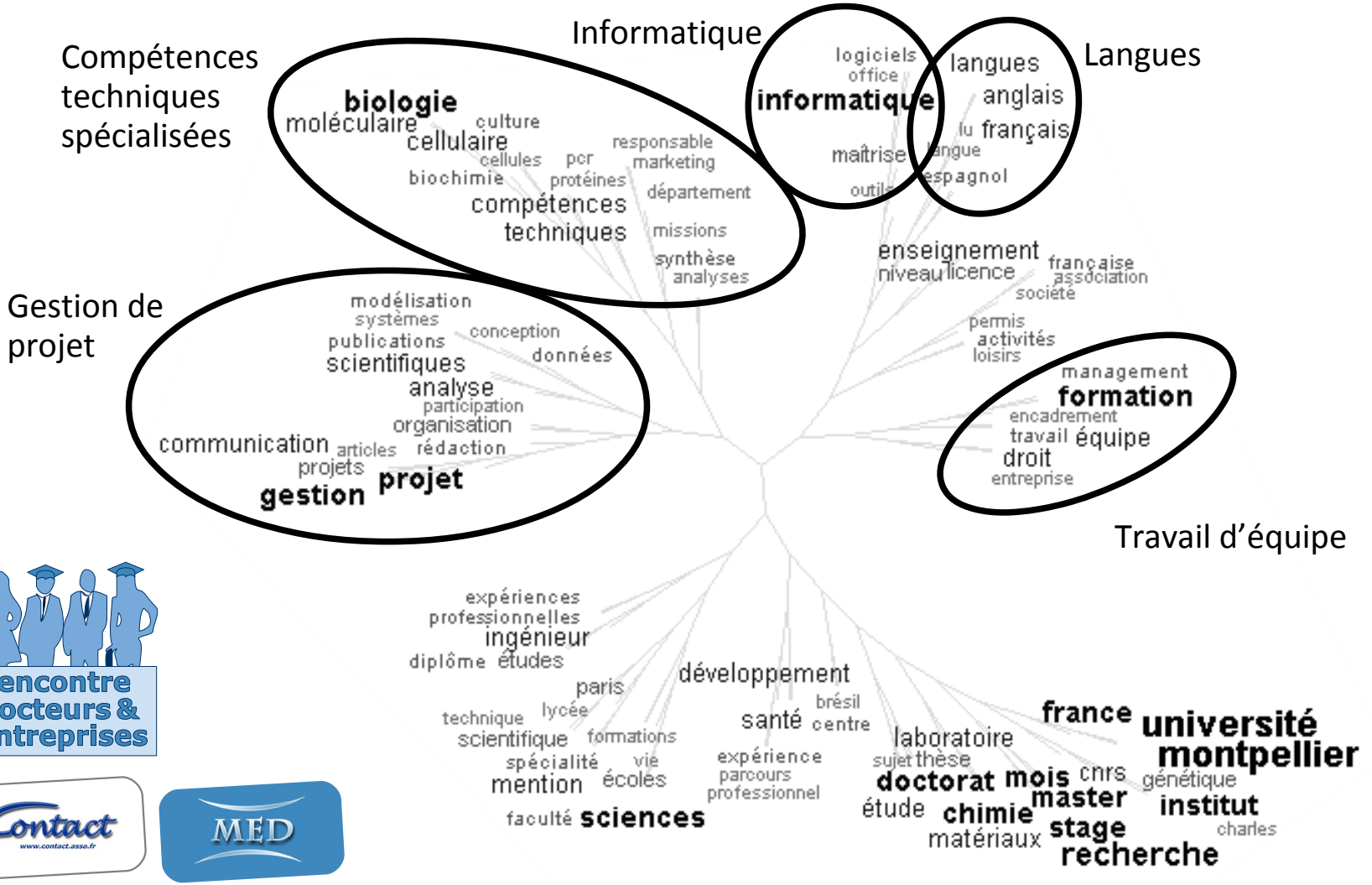
Contact
www.contact.asso.fr

MED

Méthode : interpréter les regroupements

Dessiner des « patates »

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

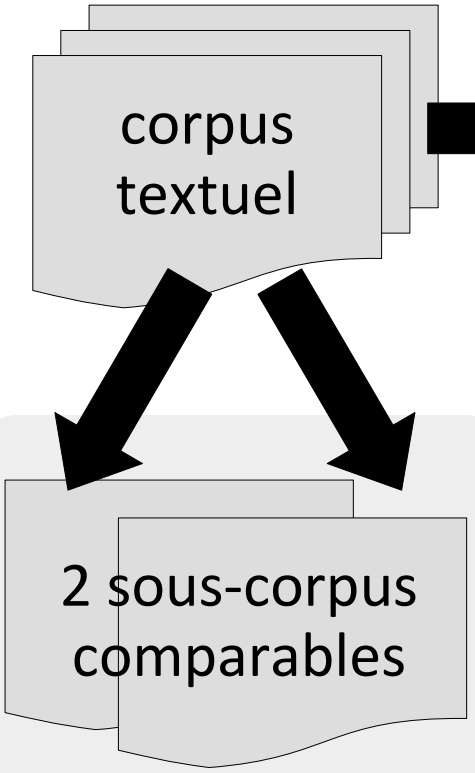


Rencontre Docteurs & Entreprises

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 :

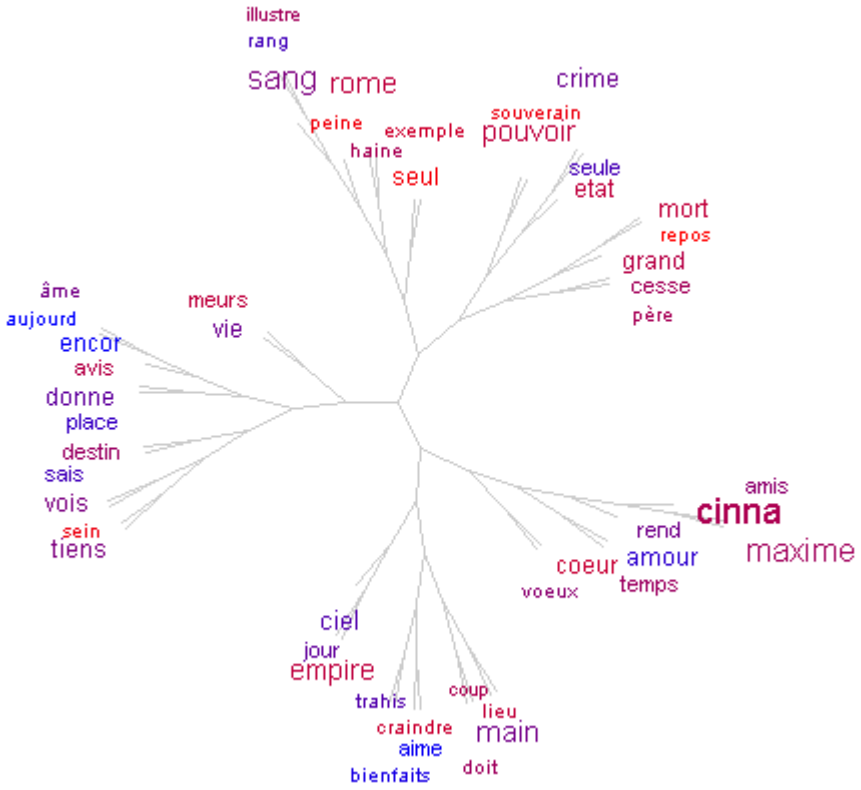
- Identifier les mots plutôt au début ou à la fin du corpus
- extraire le voisinage dans le corpus de certains mots ciblés et utiliser ces données comme corpus.

- 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 : voisinage des mots fréquents

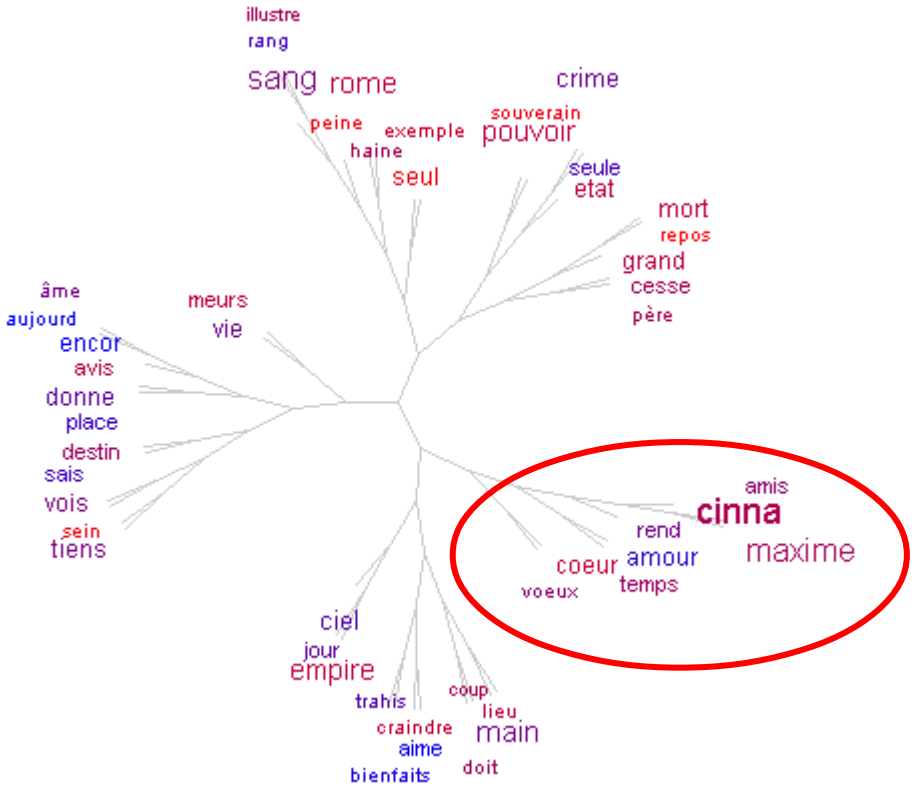
Amstutz & Gambette,
JADT 2010



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

Méthode : voisinage des mots fréquents

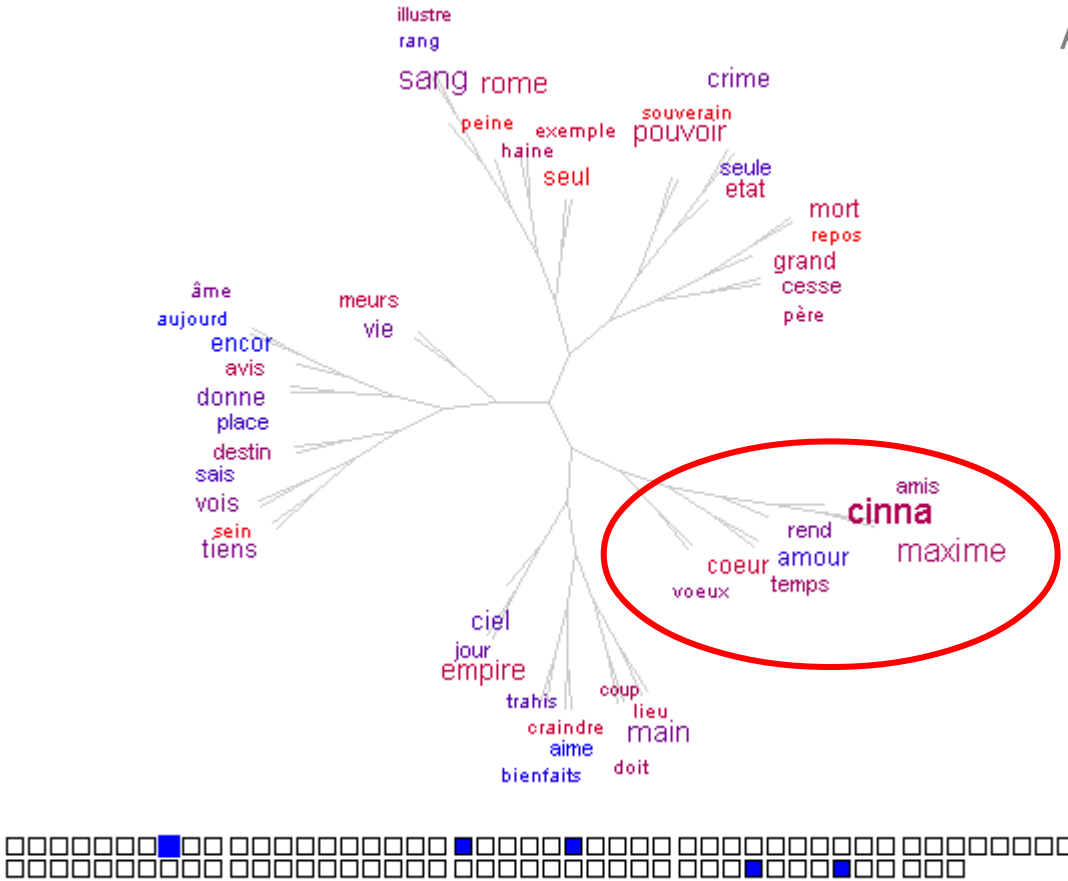
Amstutz & Gambette,
JADT 2010



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

Méthode : voisinage des mots fréquents

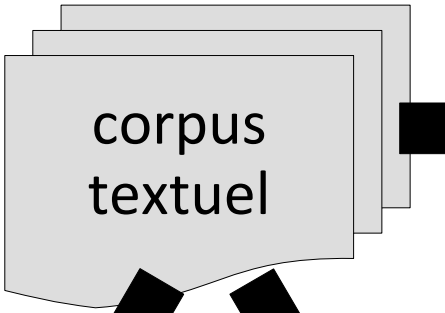
Amstutz & Gambette, JADT 2010

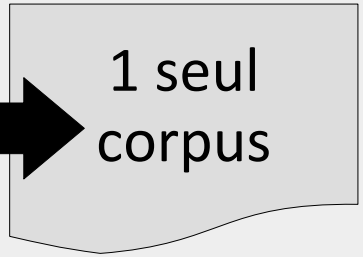


Carte des sections Lexico3 et contextes de « amis » dans les paroles d'Auguste dans Cinna.

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

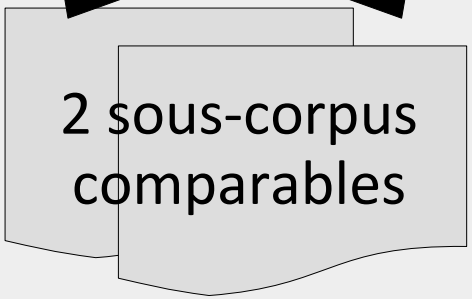




- Interpréter les regroupements
- Examiner les voisinages des mots fréquents dans l'arbre
- **Examiner le voisinage de mots ciblés dans l'arbre («acteurs», verbes, etc.)**

Si le corpus est assez grand :

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



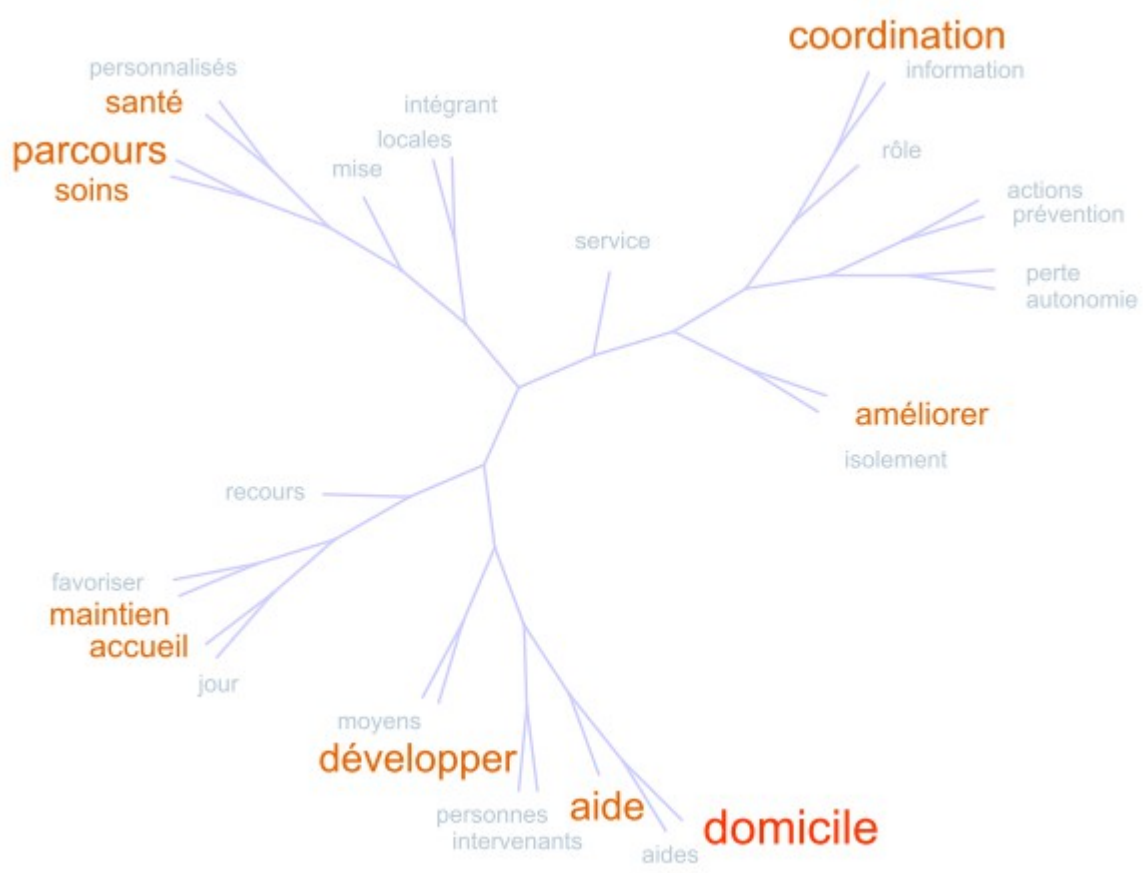
- Identifier des regroupements communs ou différents
- Comparer la fréquence des mots fréquents dans les deux arbres
- Comparer (quanti+quali) les regroupements dans les deux arbres
- Construire le nuage arboré des mots spécifiques de chaque sous-corpus

Travail en cours avec Edna Hernandez

Méthode : 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

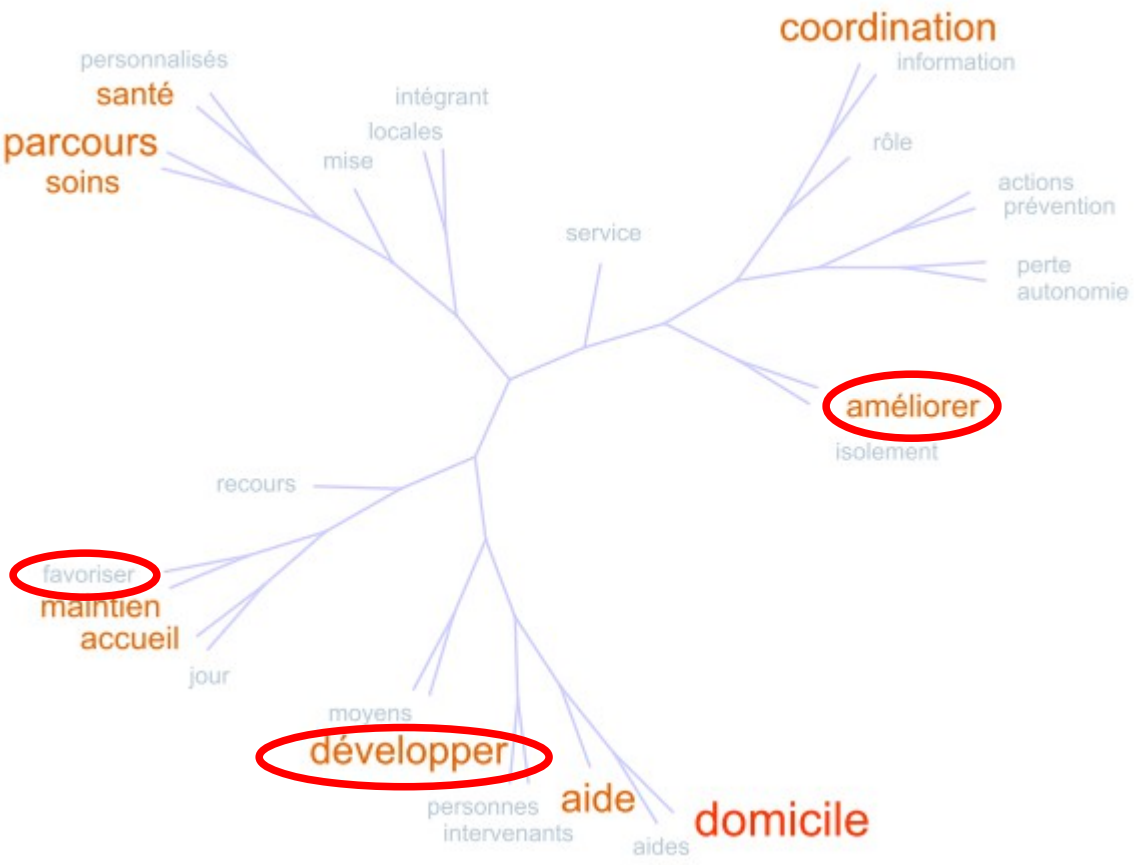


P. Grenier-Tisserand et
l'équipe projet de la DT 04

Méthode : voisinage des verbes

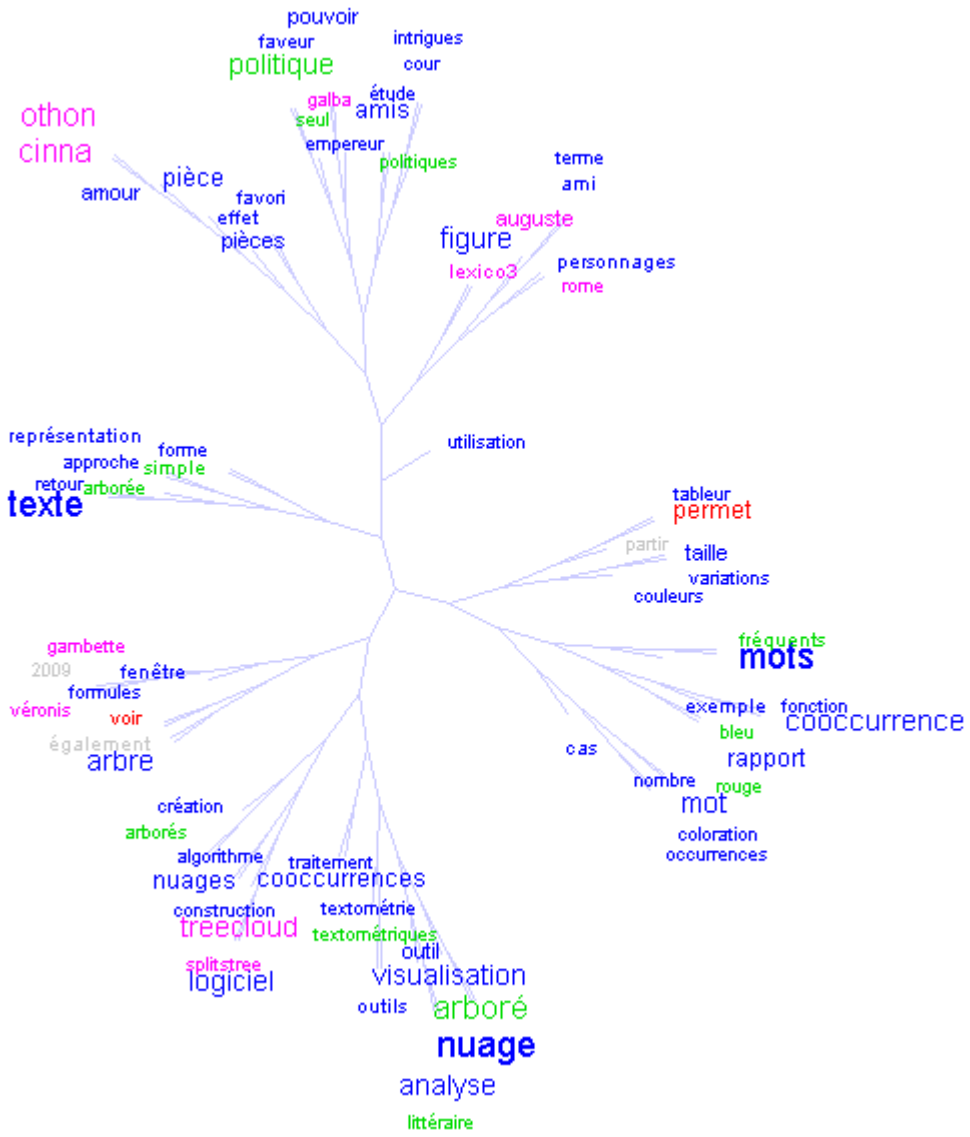
Corpus : réponses à des questions ouvertes à des professionnels de la santé sur le parcours de santé des personnes âgées dans les Alpes de Haute-Provence

Suggestions d'améliorations



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

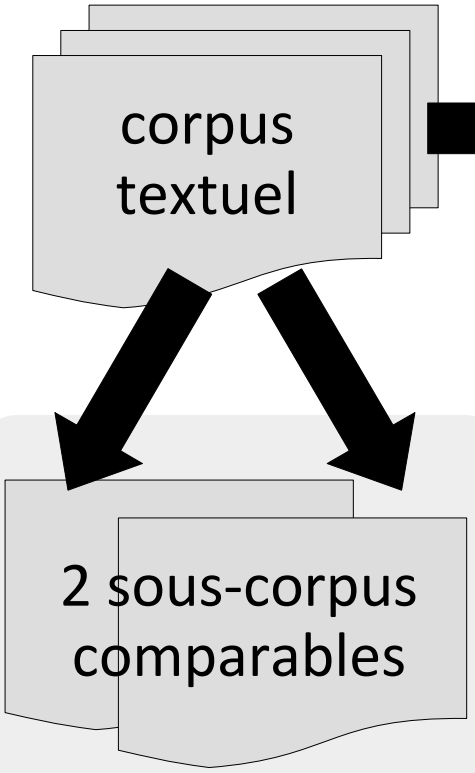
Perspective : coloration grammaticale



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.)

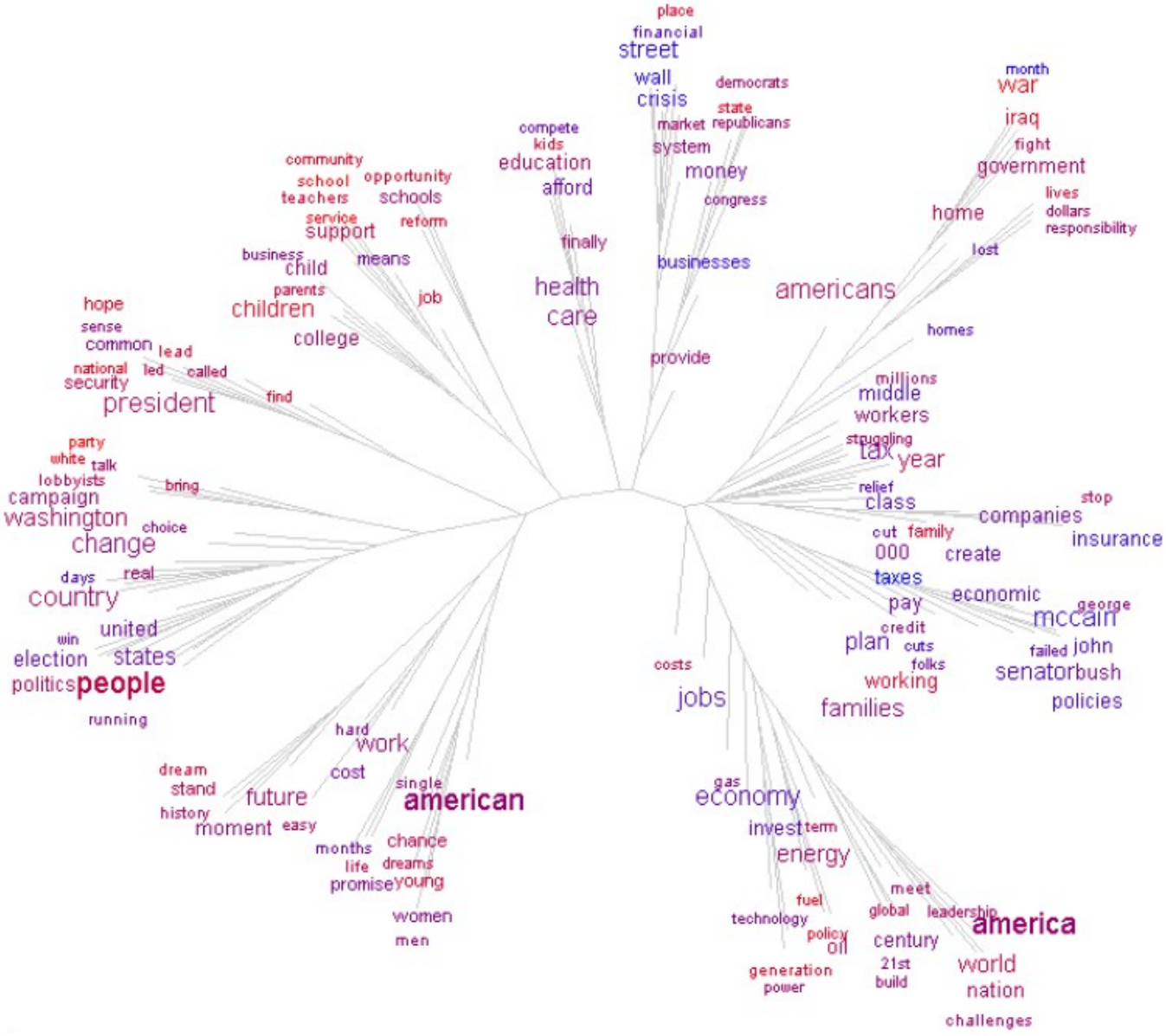
Si le corpus est assez grand :

- **Identifier les mots plutôt au début ou à la fin du corpus**
- 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 : mots au début ou à la fin

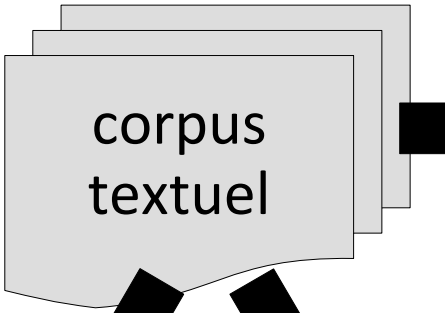


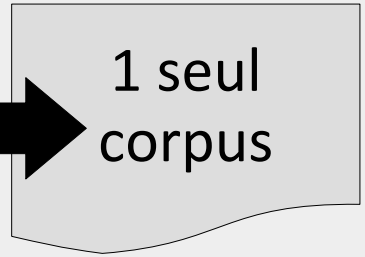
Nuage arboré de l'ensemble des discours de campagne de 2008 de Barack Obama, coloration chronologique

début de la campagne
fin de la campagne

Gambette & Véronis, IFCS 2009

Exploration de corpus avec TreeCloud

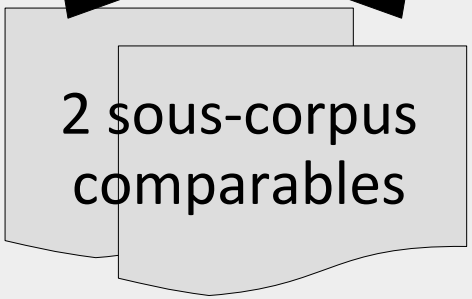




- Interpréter les regroupements
- Examiner les voisinages des mots fréquents dans l'arbre
- Examiner le voisinage de mots ciblés dans l'arbre («acteurs», verbes, etc.)

Si le corpus est assez grand :

- Identifier les mots plutôt au début ou à la fin du corpus
- extraire le 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

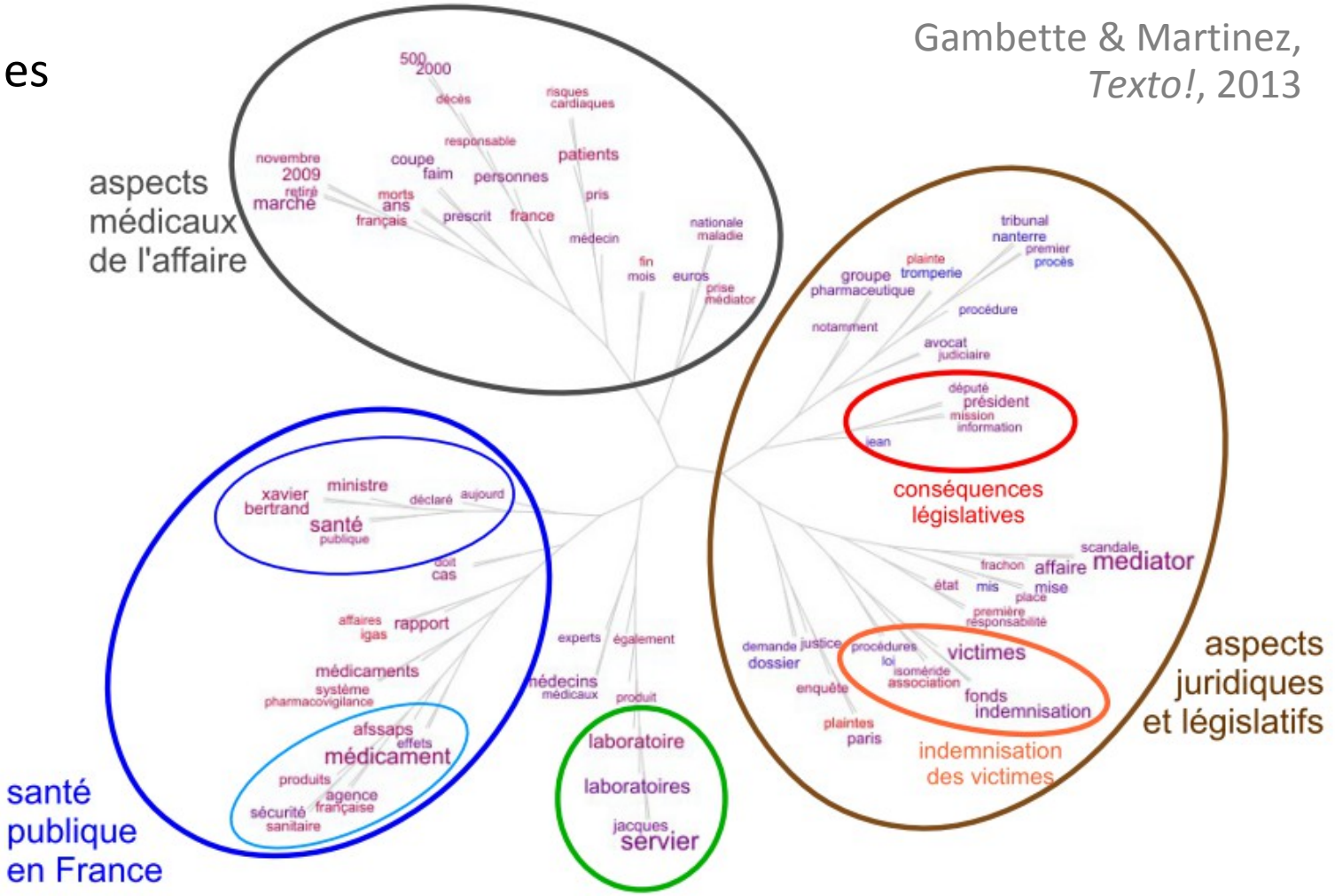
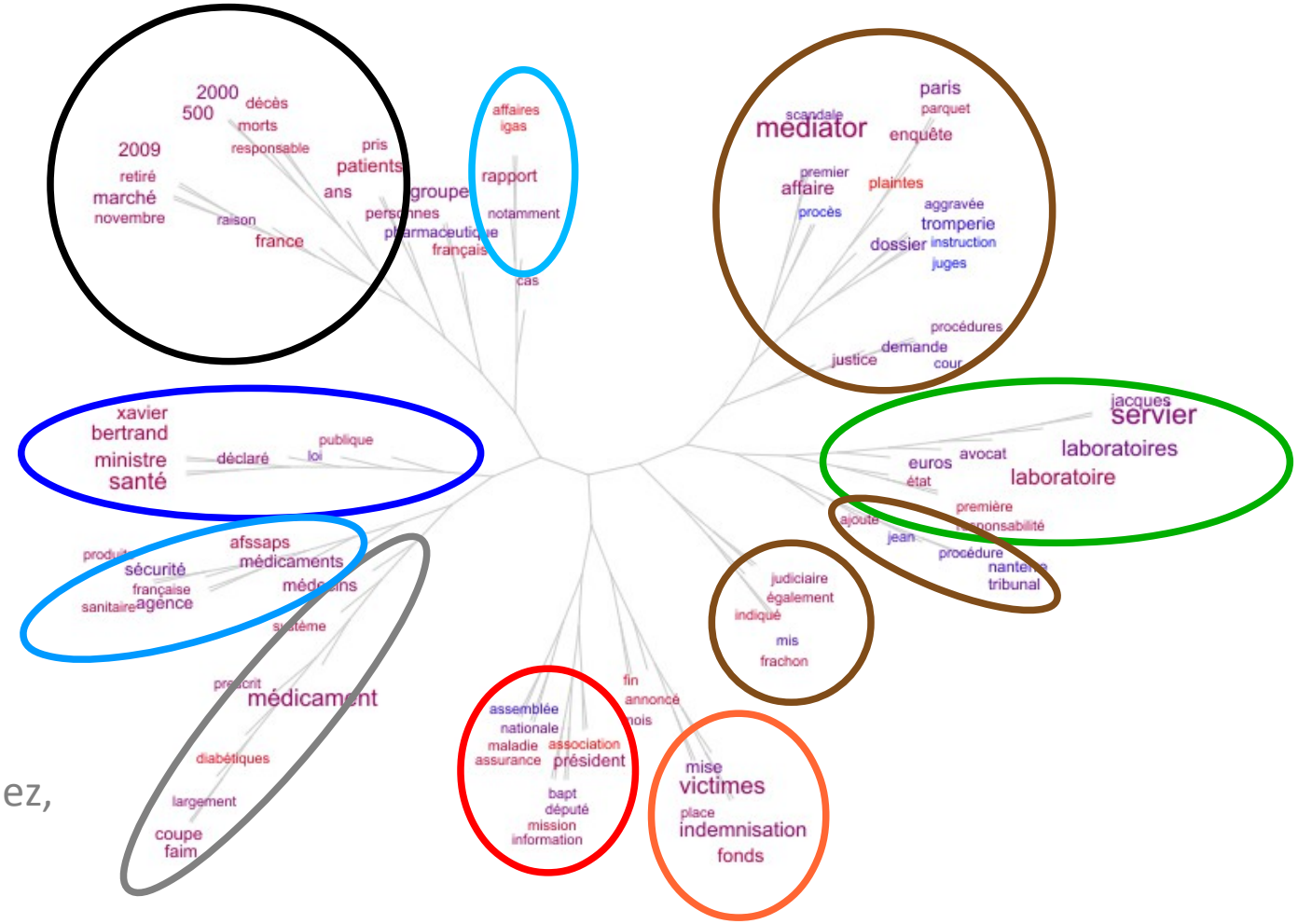


Illustration sur le corpus Mediator

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



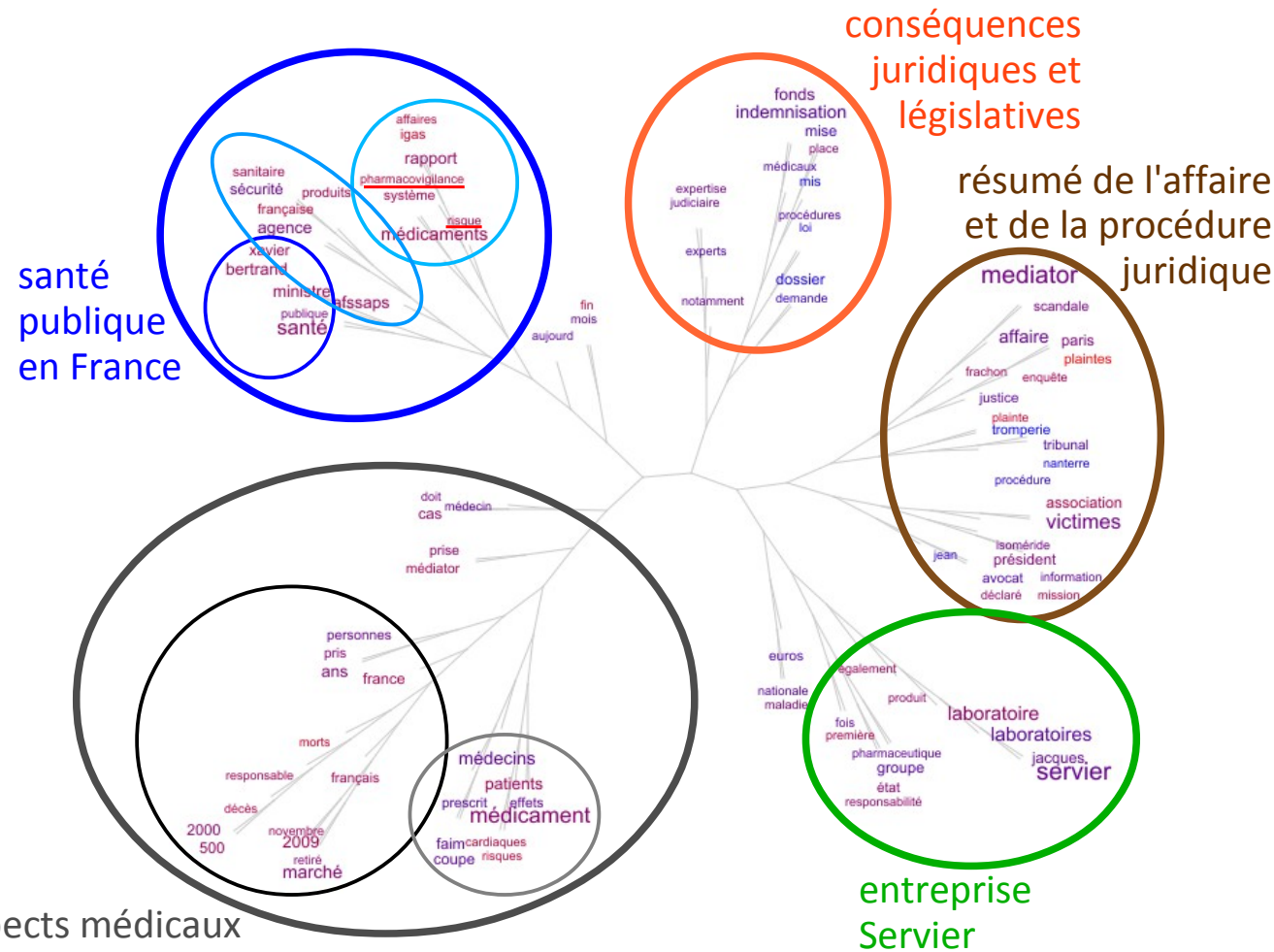
Gambette & Martinez, *Texto!*, 2013

Illustration sur le corpus Mediator

Comparer les articles d'agences et articles de journalistes

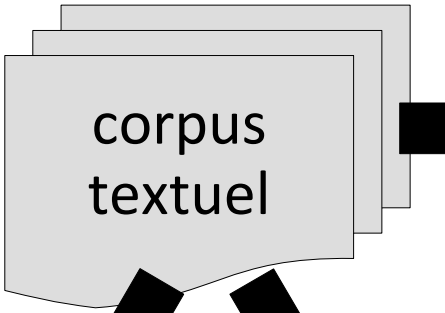
Corpus : 595 articles d'agences contre 1496 articles de journalistes de 2011 évoquant l'affaire du Mediator dans la presse française.

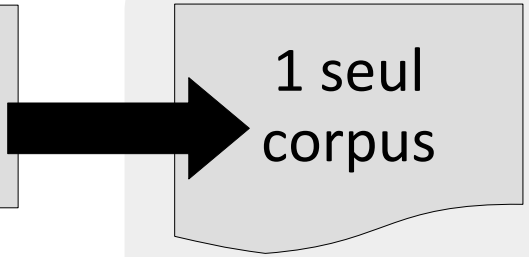
Articles
de journalistes



Gambette & Martinez,
Texto!, 2013

Exploration de corpus avec TreeCloud

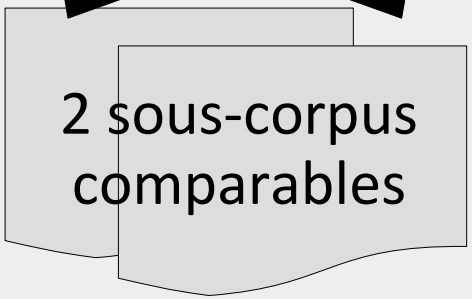




- Interpréter les regroupements
- Examiner les voisinages des mots fréquents dans l'arbre
- Examiner le voisinage de mots ciblés dans l'arbre («acteurs», verbes, etc.)

Si le corpus est assez grand :

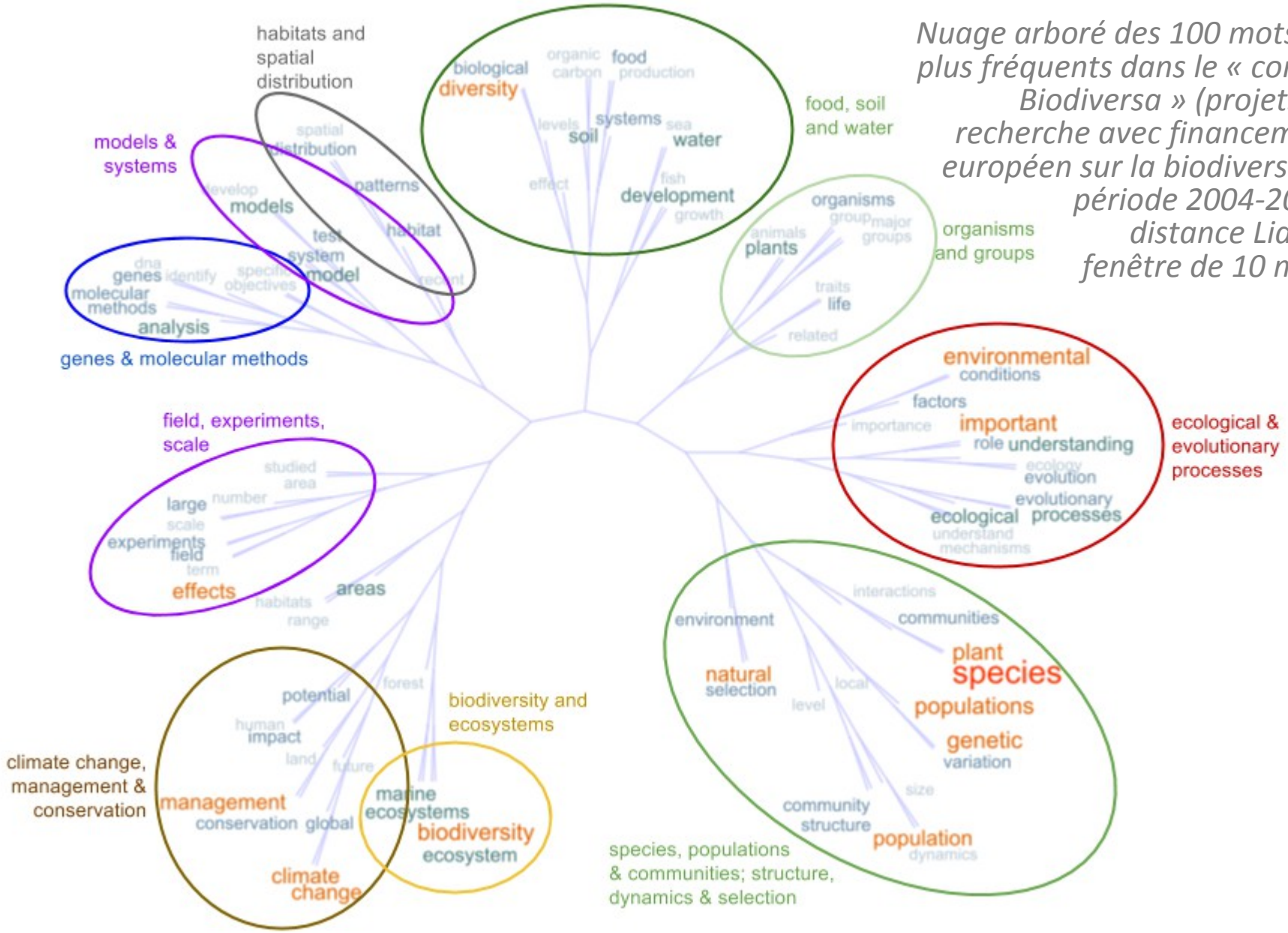
- Identifier les mots plutôt au début ou à la fin du corpus
- extraire le voisinage dans le corpus de certains mots ciblés et utiliser ces données comme corpus.



- 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

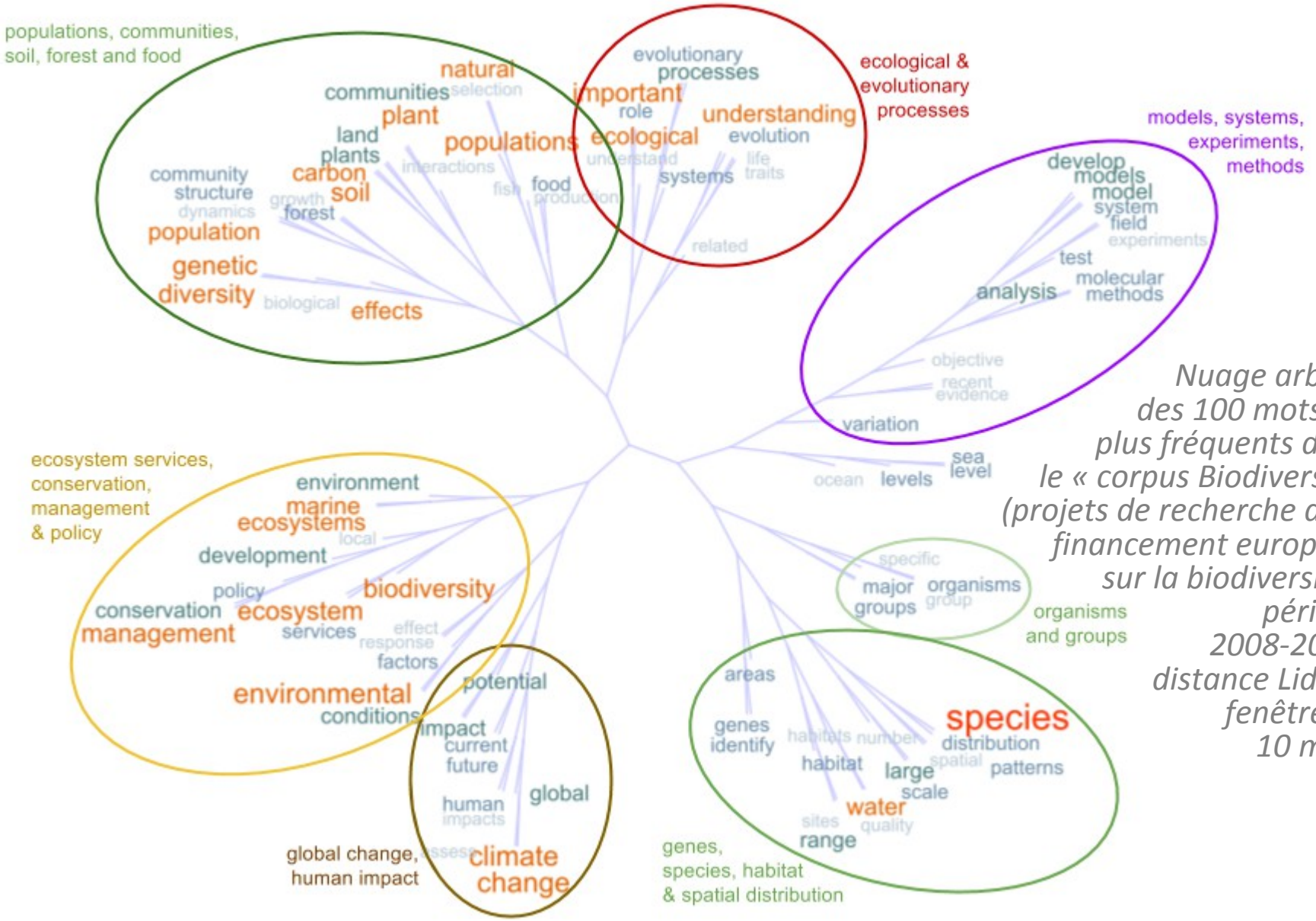
Travail en cours avec Edna Hernandez

Méthode : comparaison de voisinages dans l'arbre



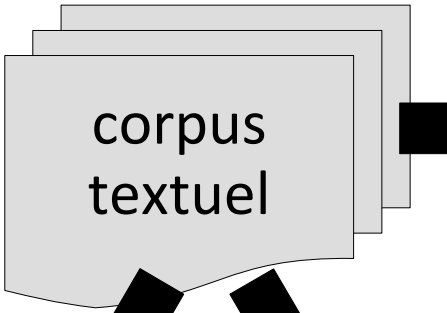
Nuage arboré des 100 mots les plus fréquents dans le « corpus Biodiversa » (projets de recherche avec financement européen sur la biodiversité), période 2004-2007, distance Liddel, fenêtre de 10 mots

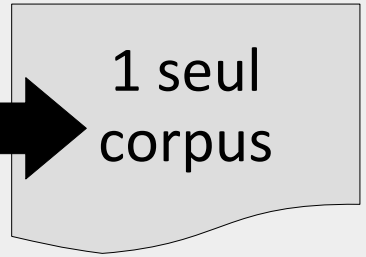
Méthode : comparaison de voisinages dans l'arbre



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

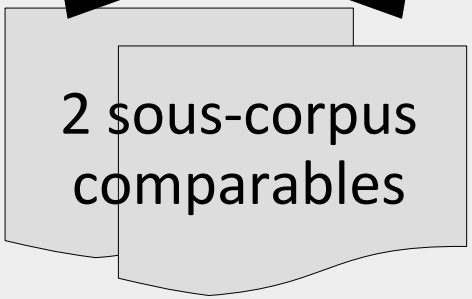




- Interpréter les regroupements
- Examiner les voisinages des mots fréquents dans l'arbre
- Examiner le voisinage de mots ciblés dans l'arbre («acteurs», verbes, etc.)

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.

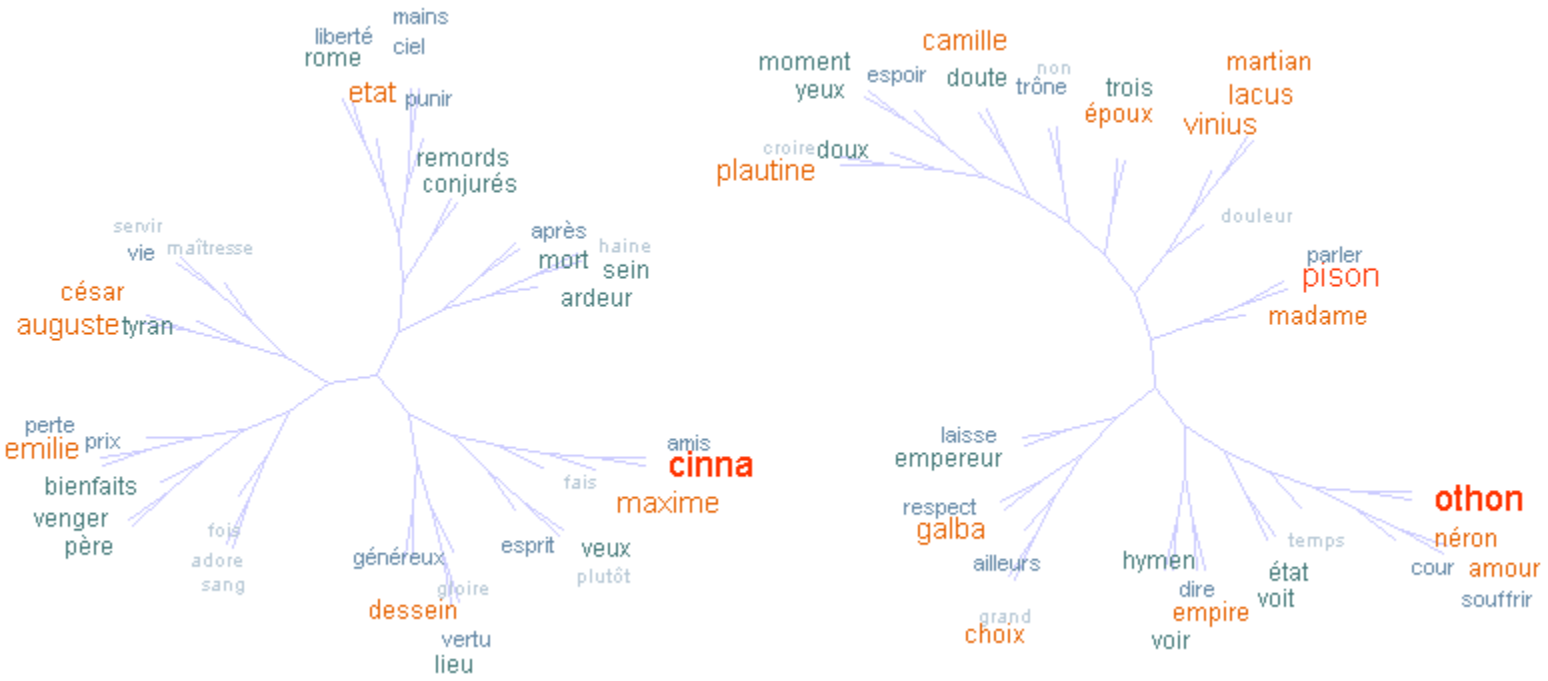


- Identifier des regroupements communs ou différents
- Comparer la fréquence des mots fréquents dans les deux arbres
- Comparer (quanti+quali) les regroupements dans les deux arbres
- **Construire le nuage arboré des mots spécifiques de chaque sous-corpus**

Travail en cours avec Edna Hernandez

Méthode : comparaison des spécifiques

Amstutz & Gambette,
JADT 2010

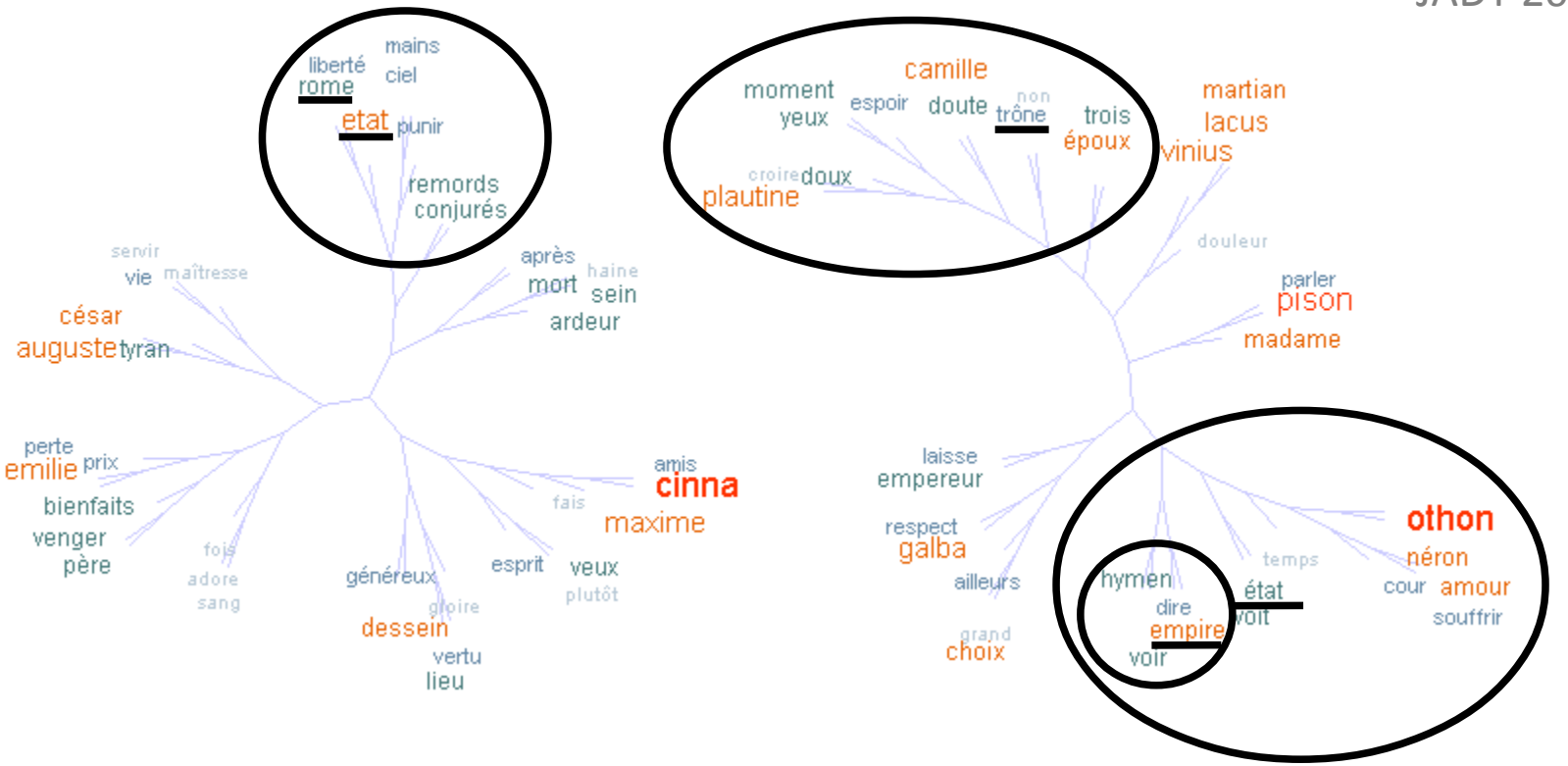


*Nuages arborés des **mots spécifiques** de Cinna et Othon, dimensionnés et colorés d'après leur spécificité calculée dans Lexico3.*

Quels moyens au service de la cause politique ?

Méthode : comparaison des spécifiques

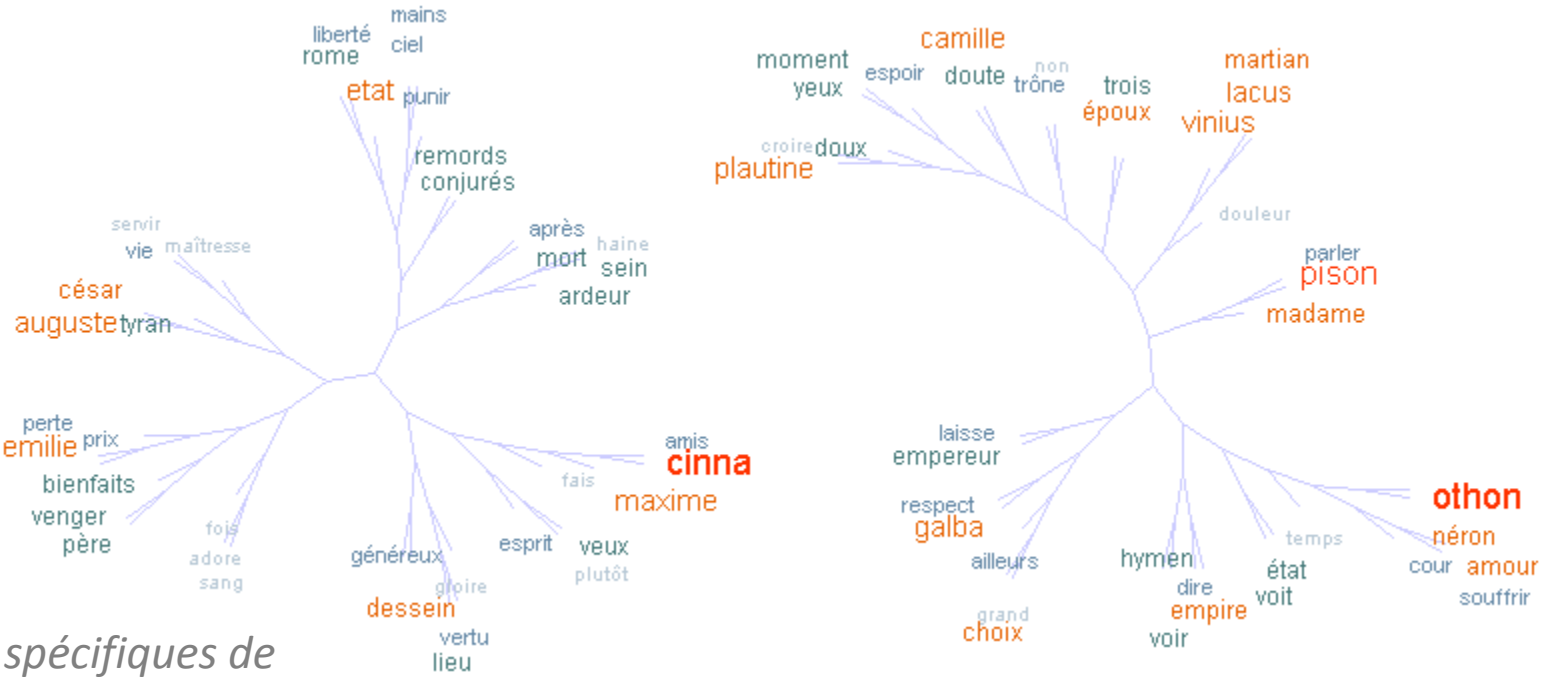
Amstutz & Gambette,
JADT 2010



*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

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
- Claude Martineau : intégration de prétraitements Unitex dans TreeCloud
- Xavier Le Roux, Hilde Eggermont : analyse du corpus de projets sur la biodiversité