SéminDoc 06/05/2009
LIRMM – Montpellier

Visualiser un texte par un nuage arboré

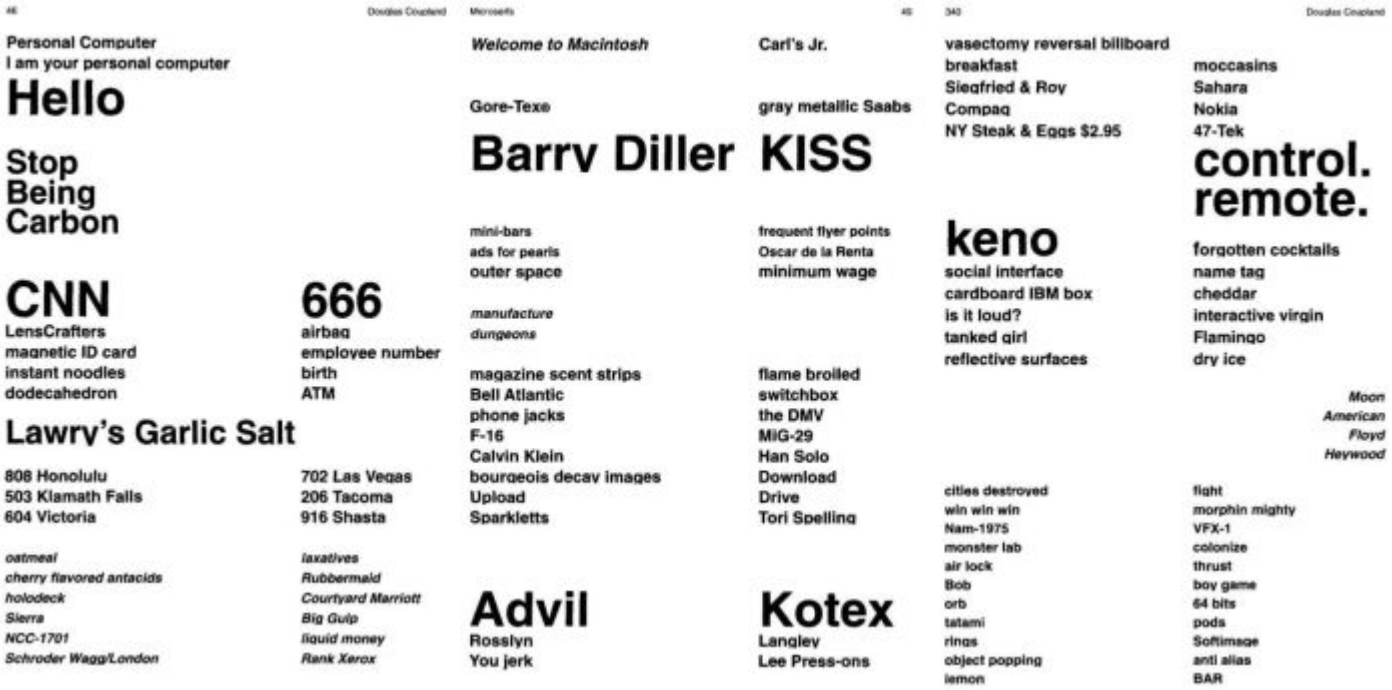
Philippe Gambette (équipes MAB/AIGco)
en collaboration avec Jean Véronis (Aix-en-Provence)

Plan

- Nuages de tags et de mots
- Nuages de tags améliorés
- Nuages arborés
- Etapes de construction
- Contrôle qualité
- Choix des paramètres
- Limites de la méthode

Nuages de tags

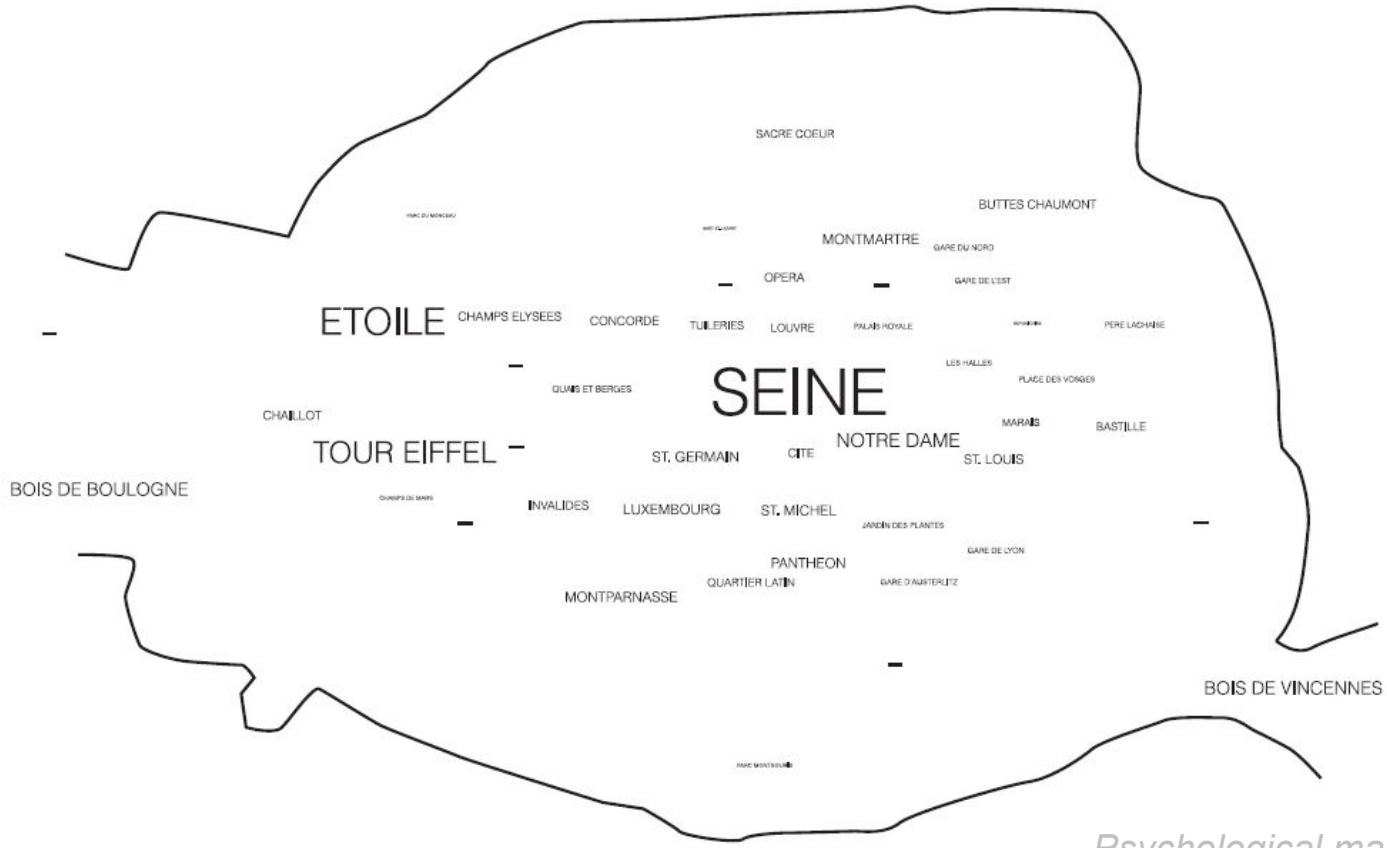
- Construits depuis un ensemble de tags
- Taille de police liée à la fréquence



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

Nuages de tags

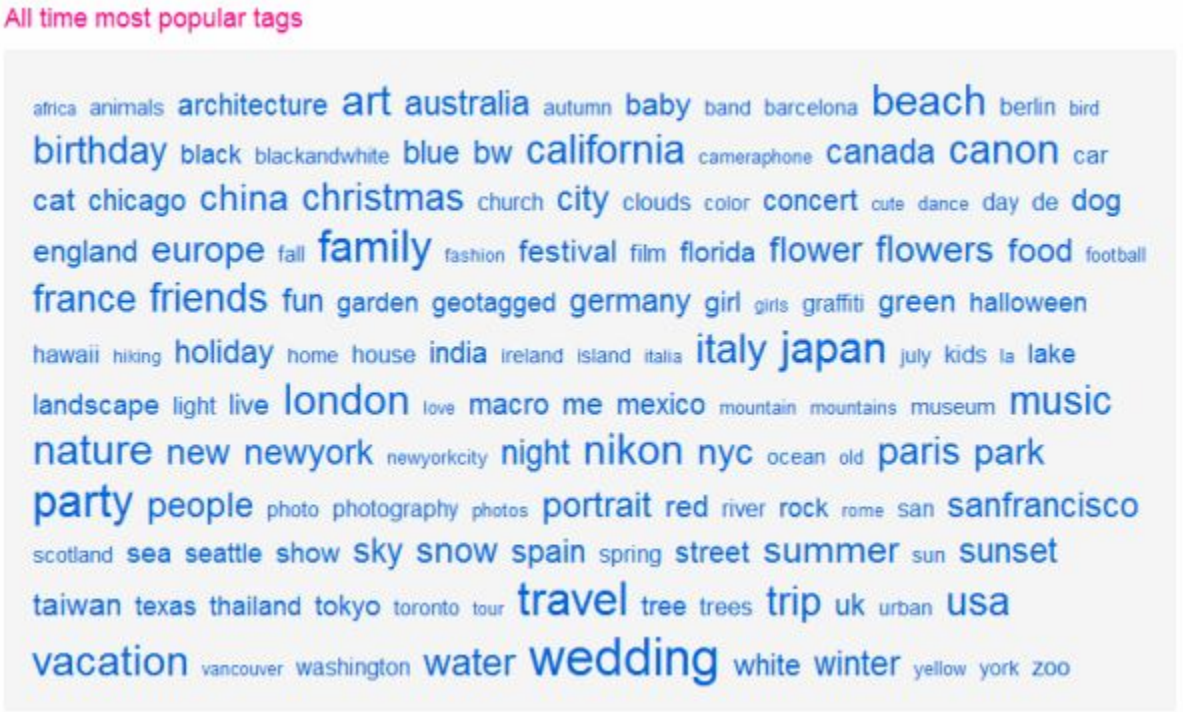
- Construits depuis un ensemble de tags
- Taille de police liée à la fréquence



Psychological maps of Paris
de Milgram et Jodelet
dans *Environmental Psychology*,
104-124, 1976

Nuages de tags

- Construits depuis un ensemble de tags
- Taille de police liée à la fréquence
- Se sont popularisés avec Flickr



Le nuage des tags les plus populaires dans tout Flickr

Nuages de mots

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

Nuages Wordle des mots les plus utilisés début 2009 dans les blogs des Top 100 politique et high-tech de Wikio

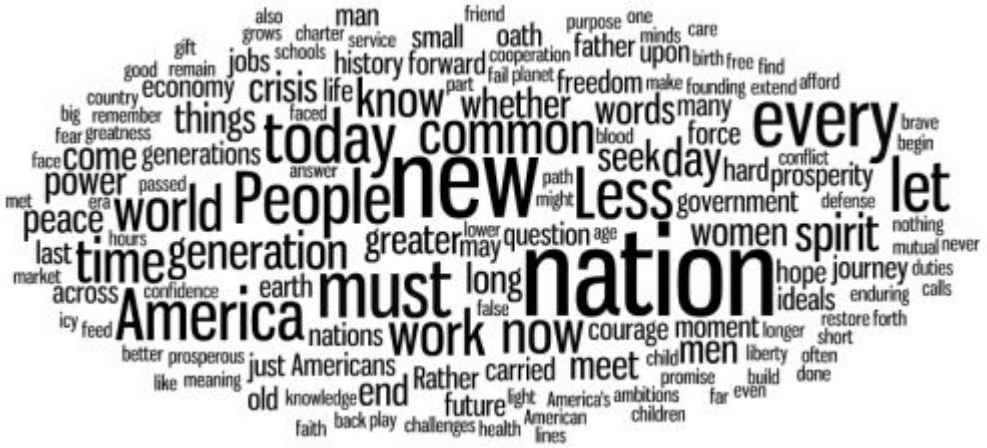
http://aixtal.blogspot.com/2009/04/web-de-quoi-parlent-les-blogs.html

Nuages de mots

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

Nuages de mots

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



GoogleImage(obama inaugural address wordle)

Nuages de mots

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

GoogleImage(obama inaugural address wordle)

Nuages de mots

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

GoogleImage(obama inaugural address wordle)

Nuages de tags/mots améliorés

Ajouter de l'information extraite du texte :

- pâleur pour exprimer la désuétude dans Amazon
- tags partagés en rouge dans del.icio.us
- regrouper les tags cooccurrents sur la même ligne

  Hassan-Montero & Herrero-Solana, InScit'06
- optimiser l'espace vide et la proximité sémantique

  Kaser & Lemire, WWW'07
- “topigraphy”: placement 2D d'après la cooccurrence

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

Nuages de tags/mots améliorés

Ajouter de l'information extraite du texte :

- pâleur pour exprimer la désuétude dans Amazon
- tags partagés en rouge dans del.icio.us
- regrouper les tags cooccurrents sur la même ligne

Hassan-Montero & Herrero-Solana, InScit'06

- optimiser l'espace vide et la proximité sémantique

Kaser & Lemire, WWW'07

- “topigraphy”: placement 2D d'après la cooccurrence

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

Nuages de tags/mots améliorés

Ajouter de l'information extraite du texte :

- pâleur pour exprimer la désuétude dans Amazon
- tags partagés en rouge dans del.icio.us
- regrouper les tags cooccurrents sur la même ligne

Hassan-Montero & Herrero-Solana, InScit'06

- optimiser l'espace vide et la proximité sémantique

Kaser & Lemire, WWW'07

- “topigraphy”: placement 2D d'après la cooccurrence

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

Nuages de tags/mots améliorés

Ajouter de l'information extraite du texte :

- pâleur pour exprimer la désuétude dans Amazon
- tags partagés en rouge dans del.icio.us
- regrouper les tags cooccurrents sur la même ligne

Hassan-Montero & Herrero-Solana, InScit'06

- optimiser l'espace vide et la proximité sémantique

Kaser & Lemire, WWW'07

- “topigraphy”: placement 2D d'après la cooccurrence

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

Nuages de tags/mots améliorés

Ajouter de l'information extraite du texte :

- pâleur pour exprimer la désuétude dans Amazon
- tags partagés en rouge dans del.icio.us
- regrouper les tags cooccurrents sur la même ligne

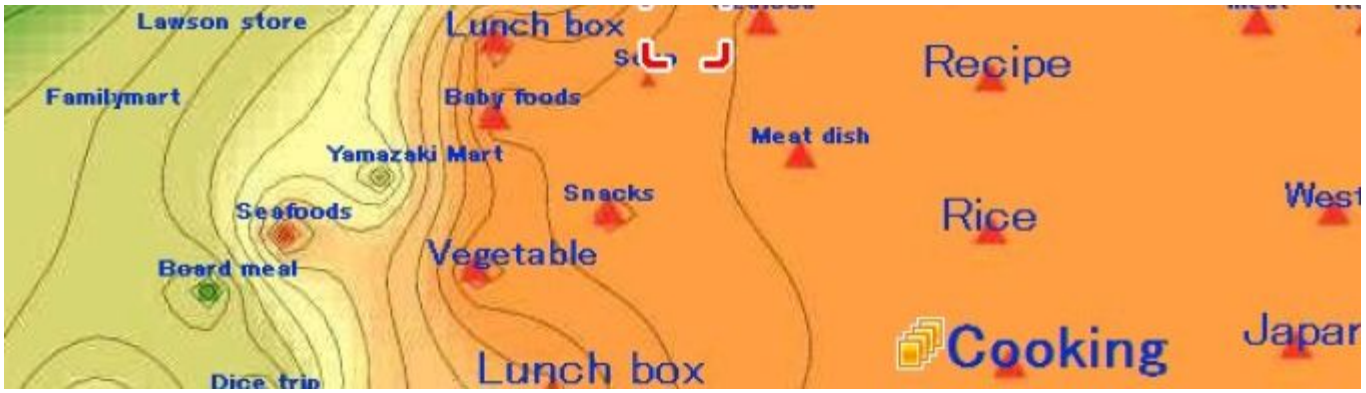
Hassan-Montero & Herrero-Solana, InScit'06

- optimiser l'espace vide et la proximité sémantique

Kaser & Lemire, WWW'07

- “topigraphy”: placement 2D d'après la cooccurrence

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



Extraire l'information sémantique d'un texte

- analyse littéraire :
approche philologique : se concentrer sur le texte

Brody

- analyse du discours :
analyse arborée, graphe de cooc., projection géodésique

Brunet (Hyperbase), Viprey (Astartex)

- fouille de texte :
graphe sémantique

Grimmer (Wordmapper), Viegas et al. (IBM Many Eyes)

- traitement des langues naturelles :
désambiguïsation

Véronis (Hyperlex)

Extraire l'information sémantique d'un texte

- analyse littéraire :
approche philologique : se concentrer sur le texte

Brody

- analyse du discours :
analyse arborée, graphe de cooc., projection géodésique

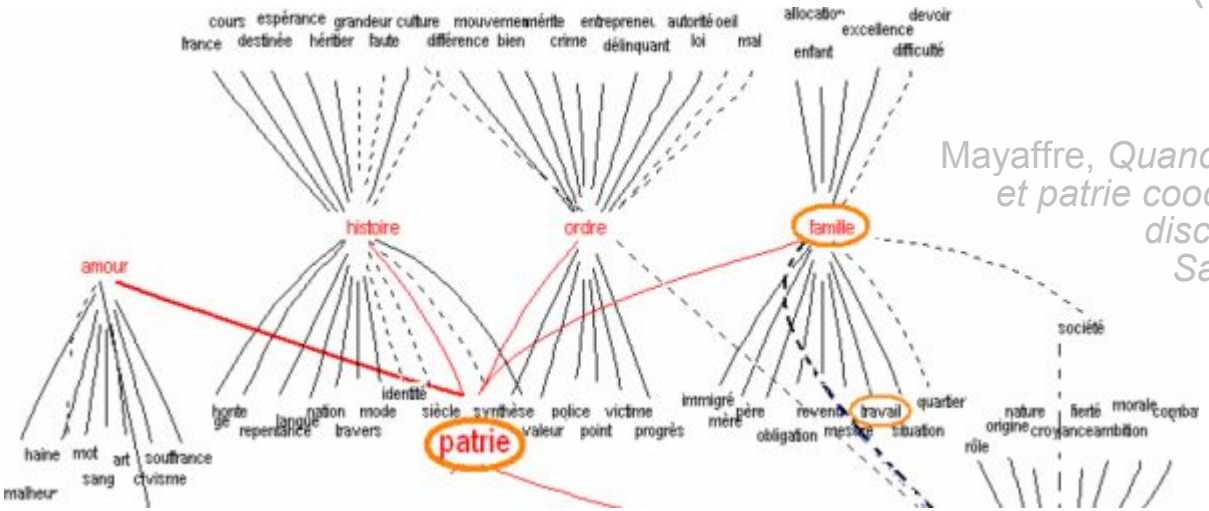
Brunet (Hyperbase), Viprey (Astartex)

- fouille de texte :
graphe sémantique

Grimmer (Wordmapper), Viegas et al. (IBM Many Eyes)

- traitement des langues naturelles :
désambiguïsation

Véronis (Hyperlex)



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

Extraire l'information sémantique d'un texte

- analyse littéraire :
approche philologique : se concentrer sur le texte

Brody

- analyse du discours :
analyse arborée, graphe de cooc., projection géodésique

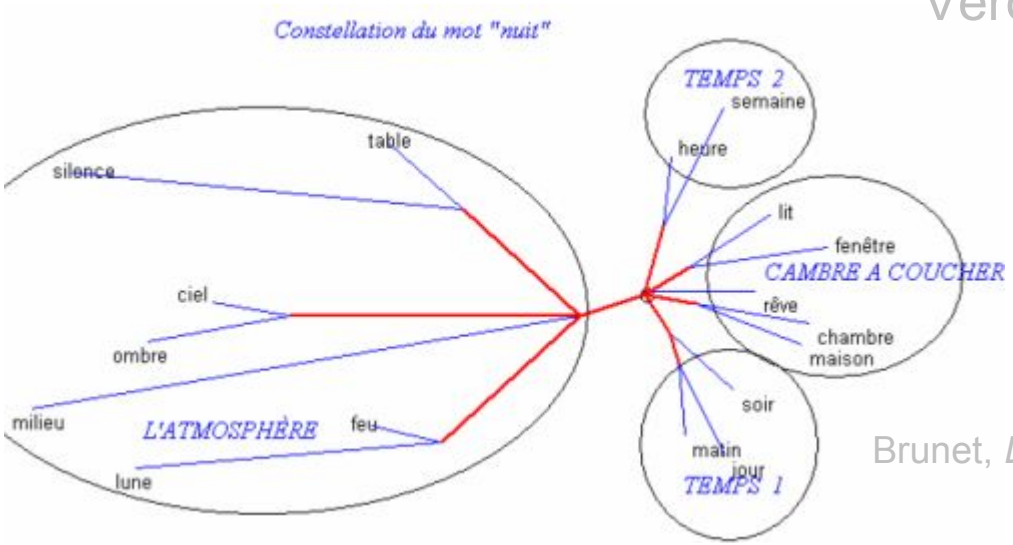
Brunet (Hyperbase), Viprey (Astartex)

- fouille de texte :
graphe sémantique

Grimmer (Wordmapper), Viegas et al. (IBM Many Eyes)

- traitement des langues naturelles :
désambiguïsation

Véronis (Hyperlex)



Brunet, *Les séquences (suite)*, JADT'08

Extraire l'information sémantique d'un texte

- analyse littéraire :
approche philologique : se concentrer sur le texte

Brody

- analyse du discours :
analyse arborée, graphe de cooc., projection géodésique

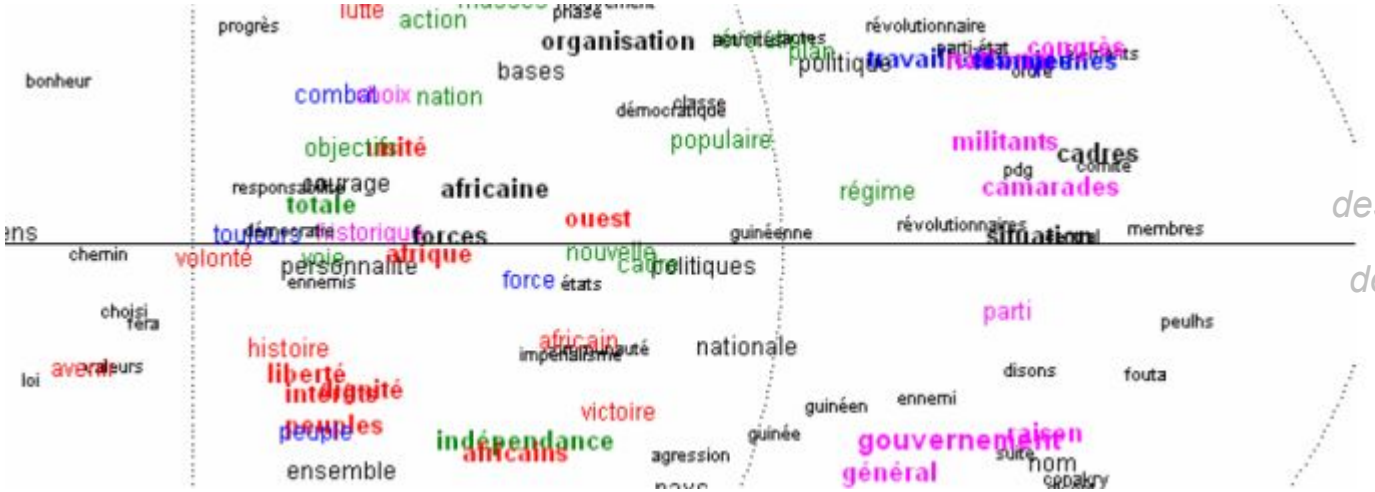
Brunet (Hyperbase), Viprey (Astartex)

- fouille de texte :
graphe sémantique

Grimmer (Wordmapper), Viegas et al. (IBM Many Eyes)

- traitement des langues naturelles :
désambiguïsation

Véronis (Hyperlex)



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

Extraire l'information sémantique d'un texte

- analyse littéraire :
approche philologique : se concentrer sur le texte

Brody

- analyse du discours :
analyse arborée, graphe de cooc., projection géodésique

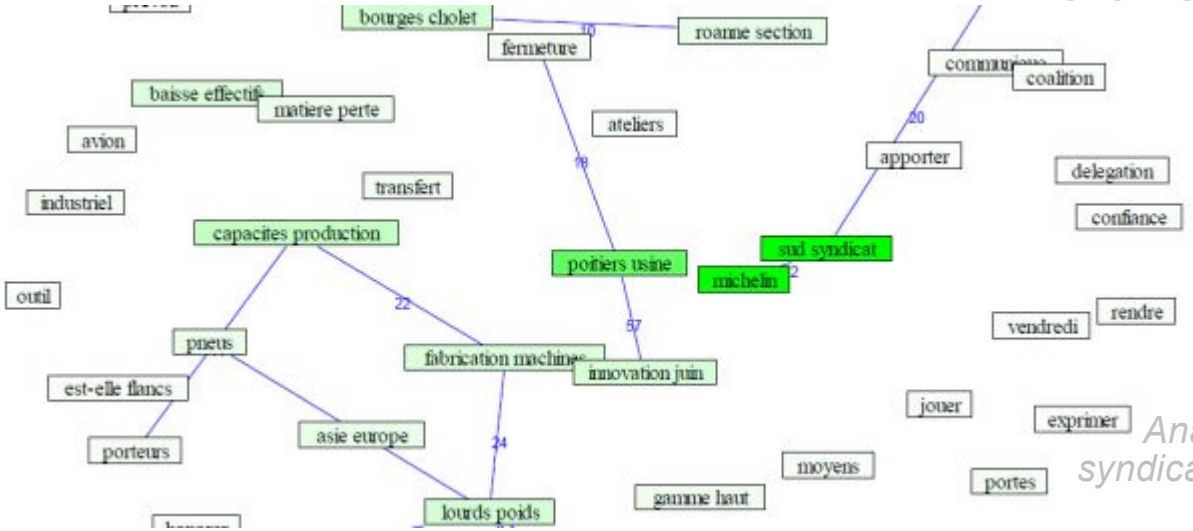
Brunet (Hyperbase), Viprey (Astartex)

- fouille de texte :
graphe sémantique

Grimmer (Wordmapper), Viegas et al. (IBM Many Eyes)

- traitement des langues naturelles :
désambiguïsation

Véronis (Hyperlex)



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

Extraire l'information sémantique d'un texte

- analyse littéraire :
approche philologique : se concentrer sur le texte

Brody

- analyse du discours :
analyse arborée, graphe de cooc., projection géodésique

Brunet (Hyperbase), Viprey (Astartex)

- fouille de texte :
graphe sémantique

Grimmer (Wordmapper), Viegas et al. (IBM Many Eyes)

- traitement des langues naturelles :
désambiguïsation

Véronis (Hyperlex)

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

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

Extraire l'information sémantique d'un texte

- analyse littéraire :
approche philologique : se concentrer sur le texte

Brody

- analyse du discours :
analyse arborée, graphe de cooc., projection géodésique

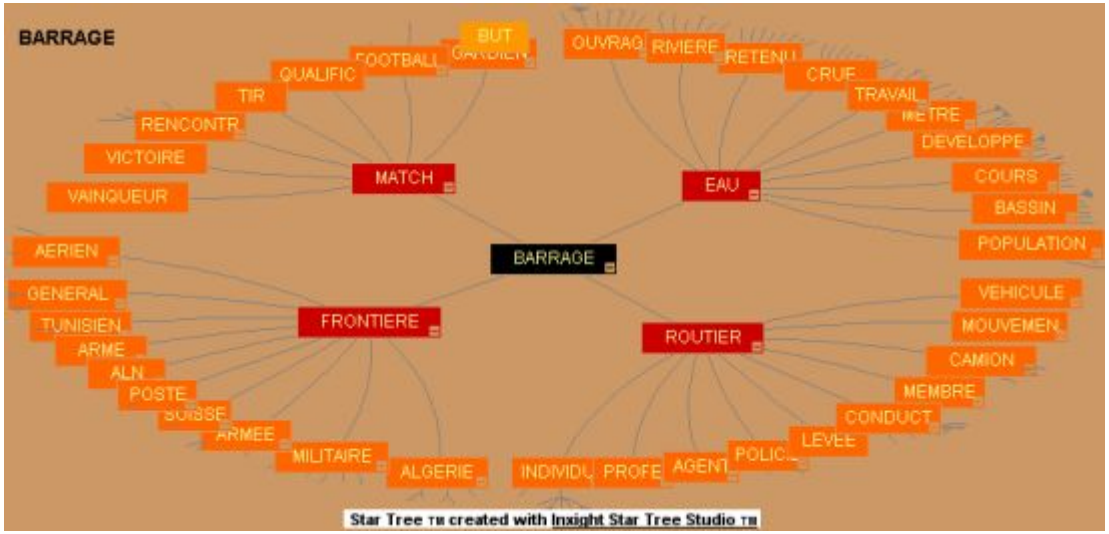
Brunet (Hyperbase), Viprey (Astartex)

- fouille de texte :
graphe sémantique

Grimmer (Wordmapper), Viegas et al. (IBM Many Eyes)

- traitement des langues naturelles :
désambiguïsation

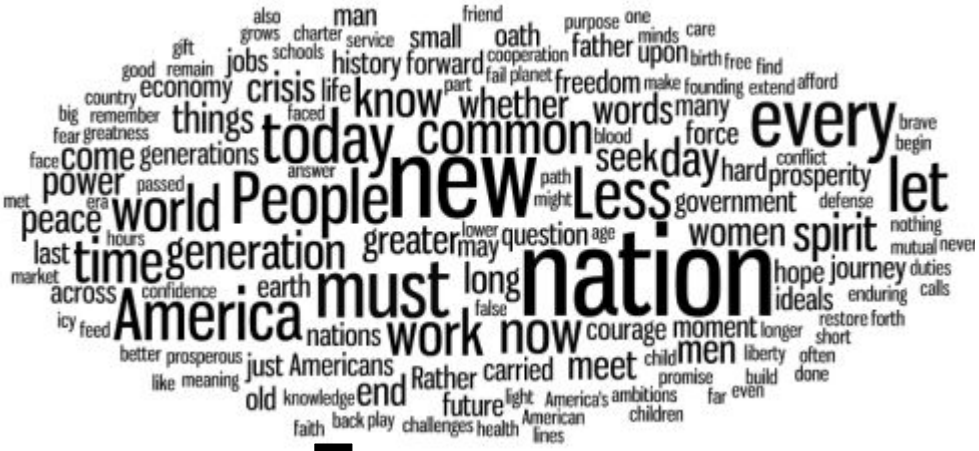
Véronis (Hyperlex)

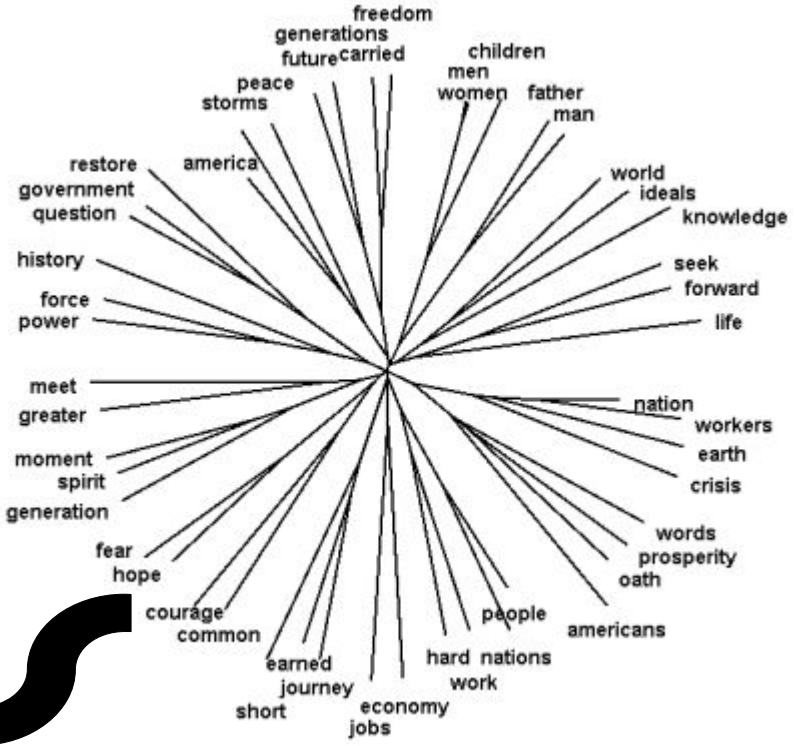


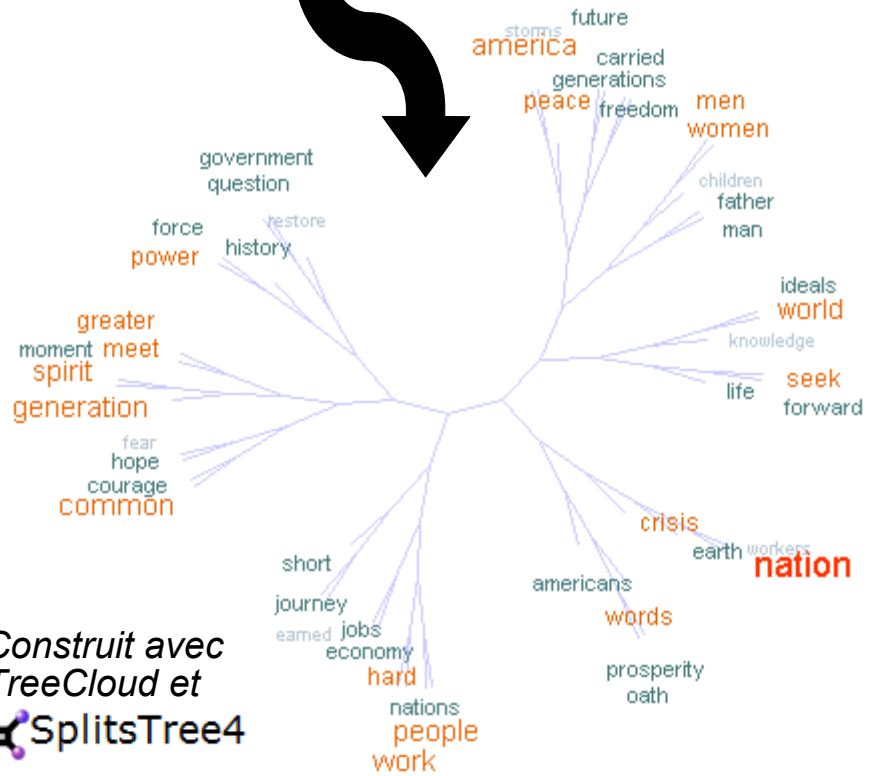
Désambiguïsation du mot “barrage”.

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

Nuage de tags + arbre = nuage arboré





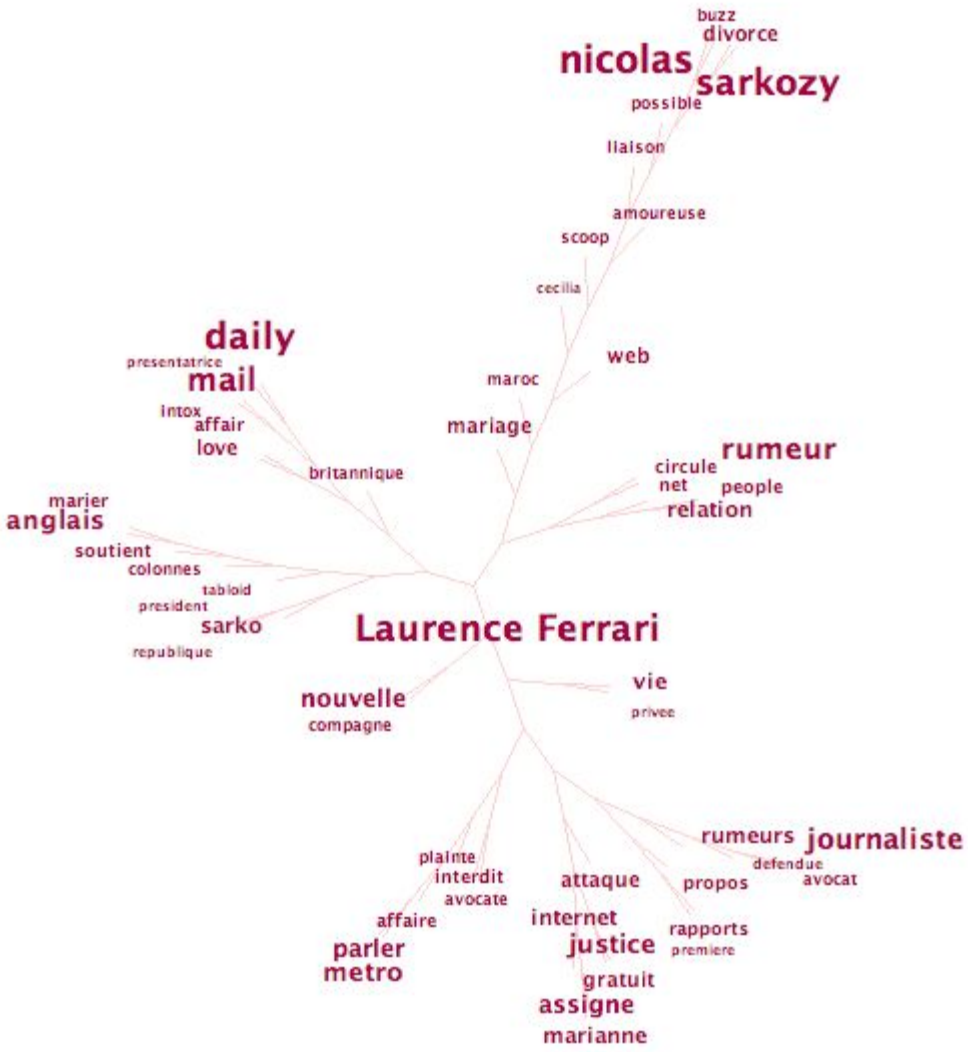


Construit avec TreeCloud et
SplitsTree4

SplitsTree : Huson 1998, Huson & Bryant 2006

TreeCloud en Python, license GPL
disponible sur http://www.treecloud.fr

Le premier nuage arboré



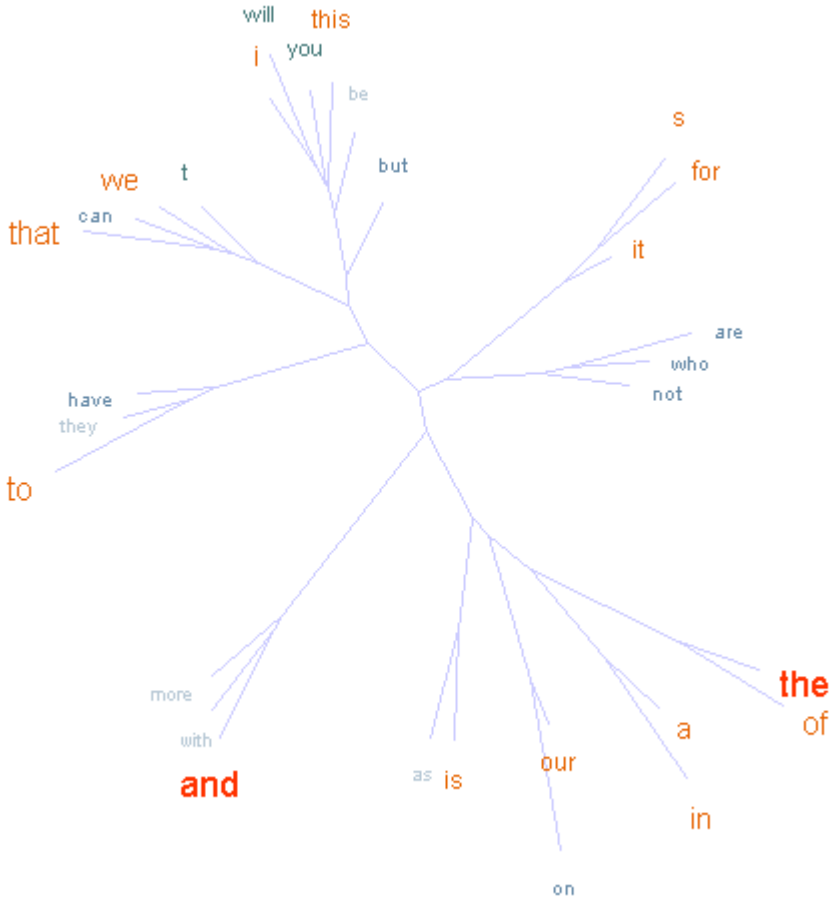
Nuage arboré des posts de blogs contenant “Laurence Ferrari”
du 25/11/2007 au 10/12/2007, par Jean Véronis
http://aixtal.blogspot.com/2007/12/actu-une-ferrari-dans-un-arbre.html

Construction – extraction des mots

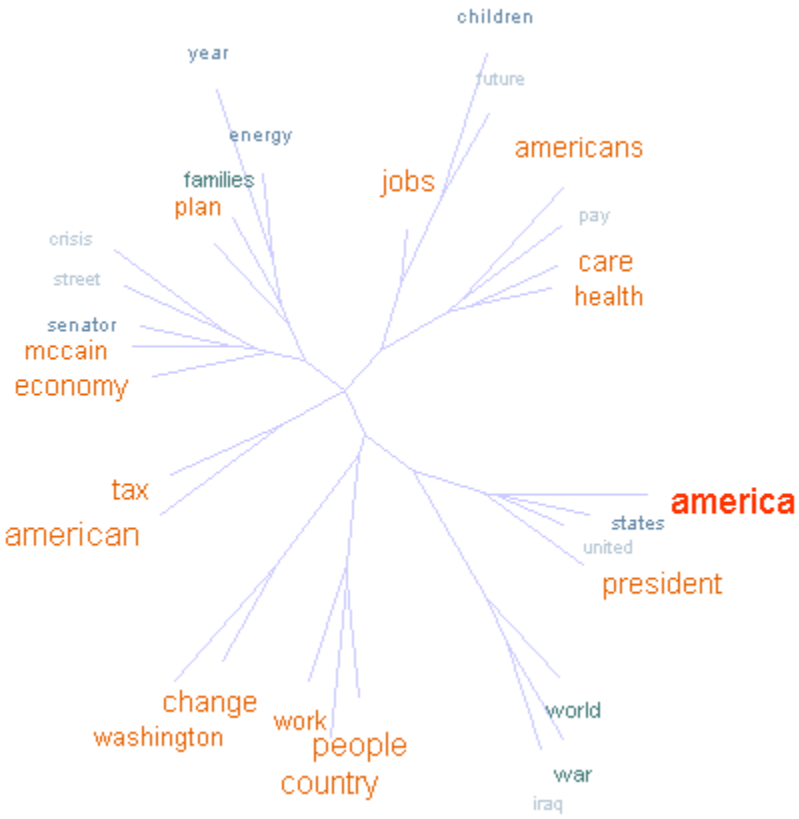
Extraire les mots avec leur fréquence :

- antidico ?

sans antidico



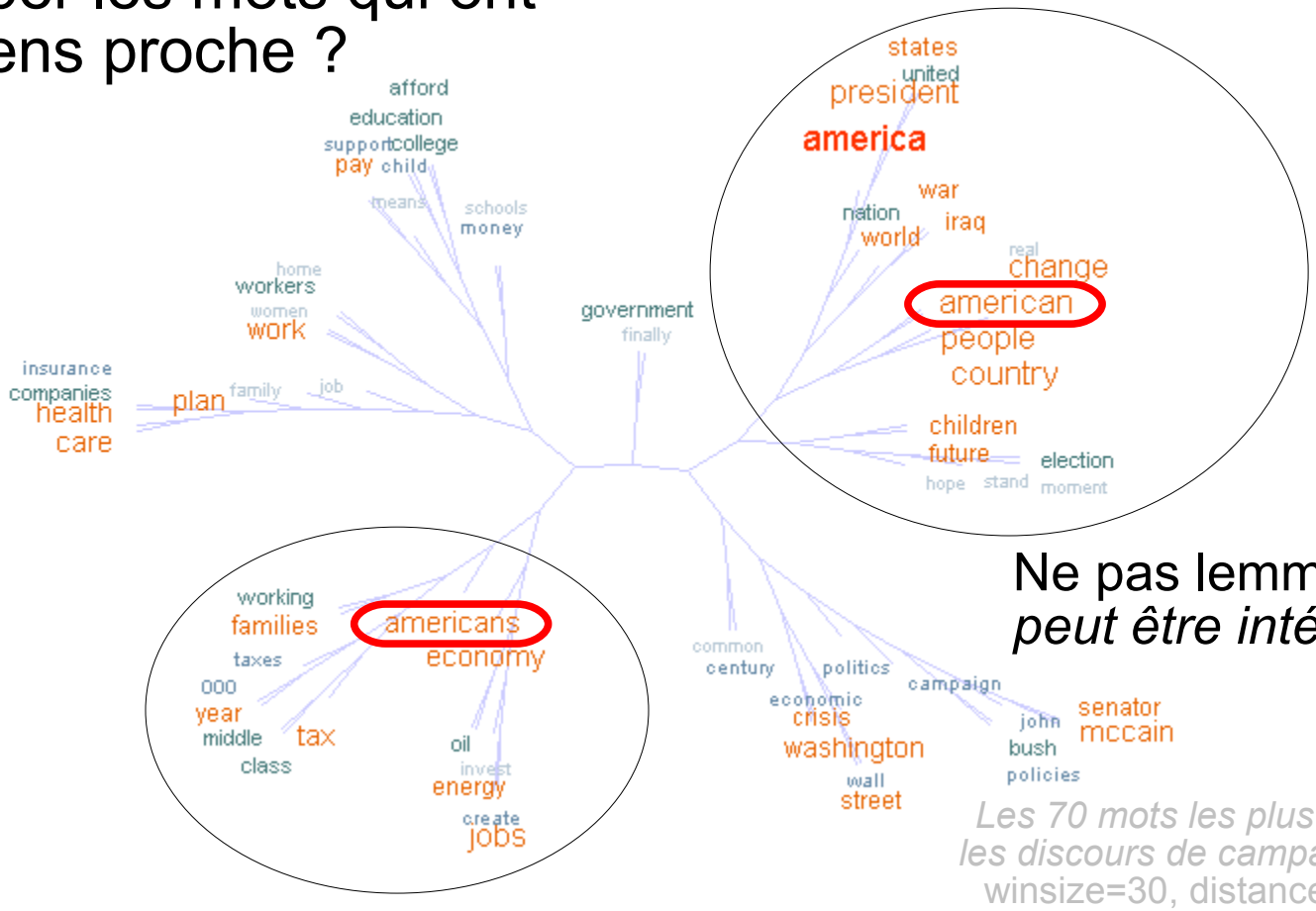
avec antidico



Construction – extraction des mots

Extraire les mots avec leur fréquence :

- lemmatisation?
- grouper les mots qui ont un sens proche ?



Ne pas lemmatiser...
peut être intéressant

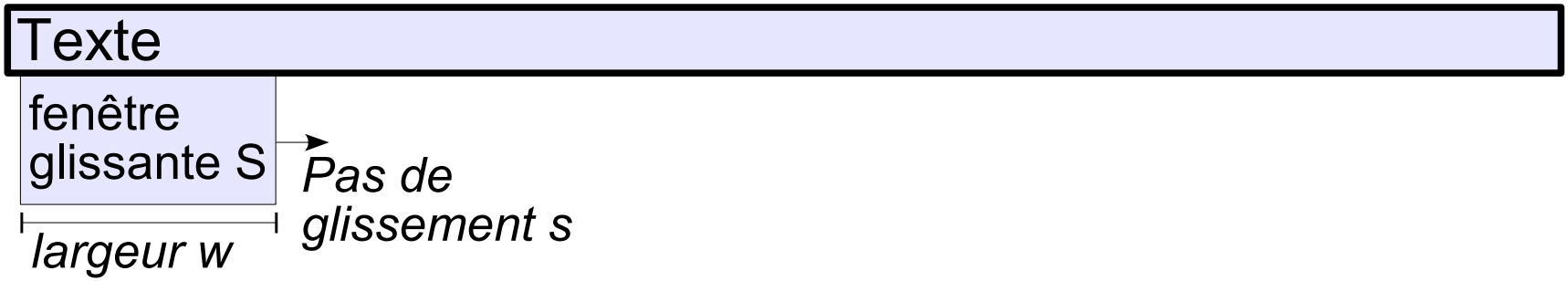
Les 70 mots les plus fréquents dans les discours de campagne d'Obama, winsize=30, distance=dice, NJ-tree.

Construction – matrice de dissimilarité

De nombreuses formules de distance sémantique utilisent la cooccurrence.

Construction – matrice de dissimilarité

De nombreuses formules de distance sémantique utilisent la cooccurrence.



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

	$v \in S$	$v \notin S$
$u \in S$	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

Construction – matrice de dissimilarité

Transformations à appliquer pour obtenir une dissimilarité :

- transformer la similarité en dissimilarité

- normalisation linéaire pour les matrices positives pour avoir des distances dans l'intervalle [0,1]

- normalisation affine pour les matrices avec des nombres positifs et négatifs, pour avoir des distances dans [α,1] (par exemple α=0.1)

Construction – construction de l'arbre

Plusieurs méthodes possibles :

- Neighbor-Joining

Saitou & Nei, 1987

SplitsTree4

- Variantes d'Addtree

Barthelemy & Luong, 1987

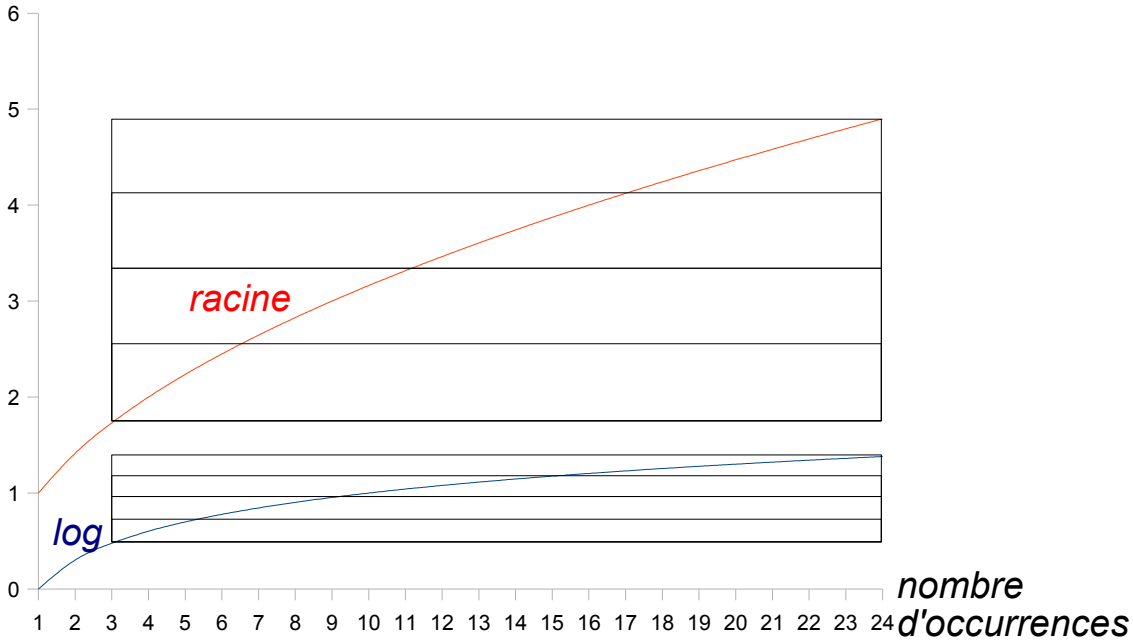
- Heuristique de quadruplets

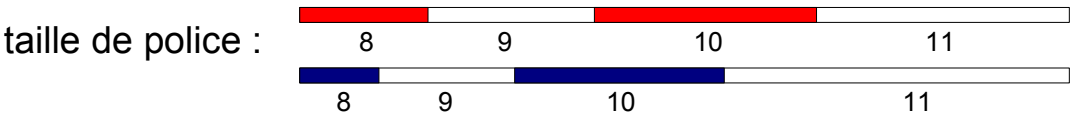
Cilibrasi & Vitanyi, 2007

Construction – décoration de l'arbre

Choix des tailles de polices :

- **fréquence** (appliquer un log !)

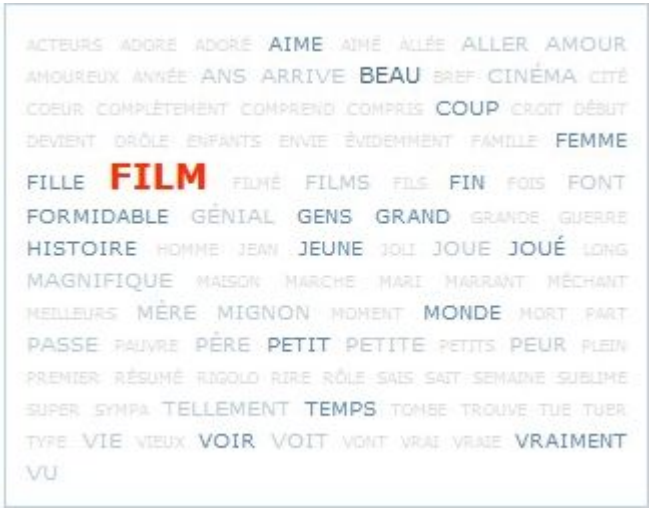




Construction – décoration de l'arbre

Choix des tailles de polices :

- **fréquence** (appliquer un log !)



racine

log

Merci, la loi de Zipf !

Construction – décoration de l'arbre

Choix des tailles de polices :

- **fréquence** (appliquer un log !)

ou

- **classement des fréquences** (distribution exponentielle)

ou

- **saillance** par rapport à un corpus de référence

ou

- **charge émotionnelle** calculée par l'entropie

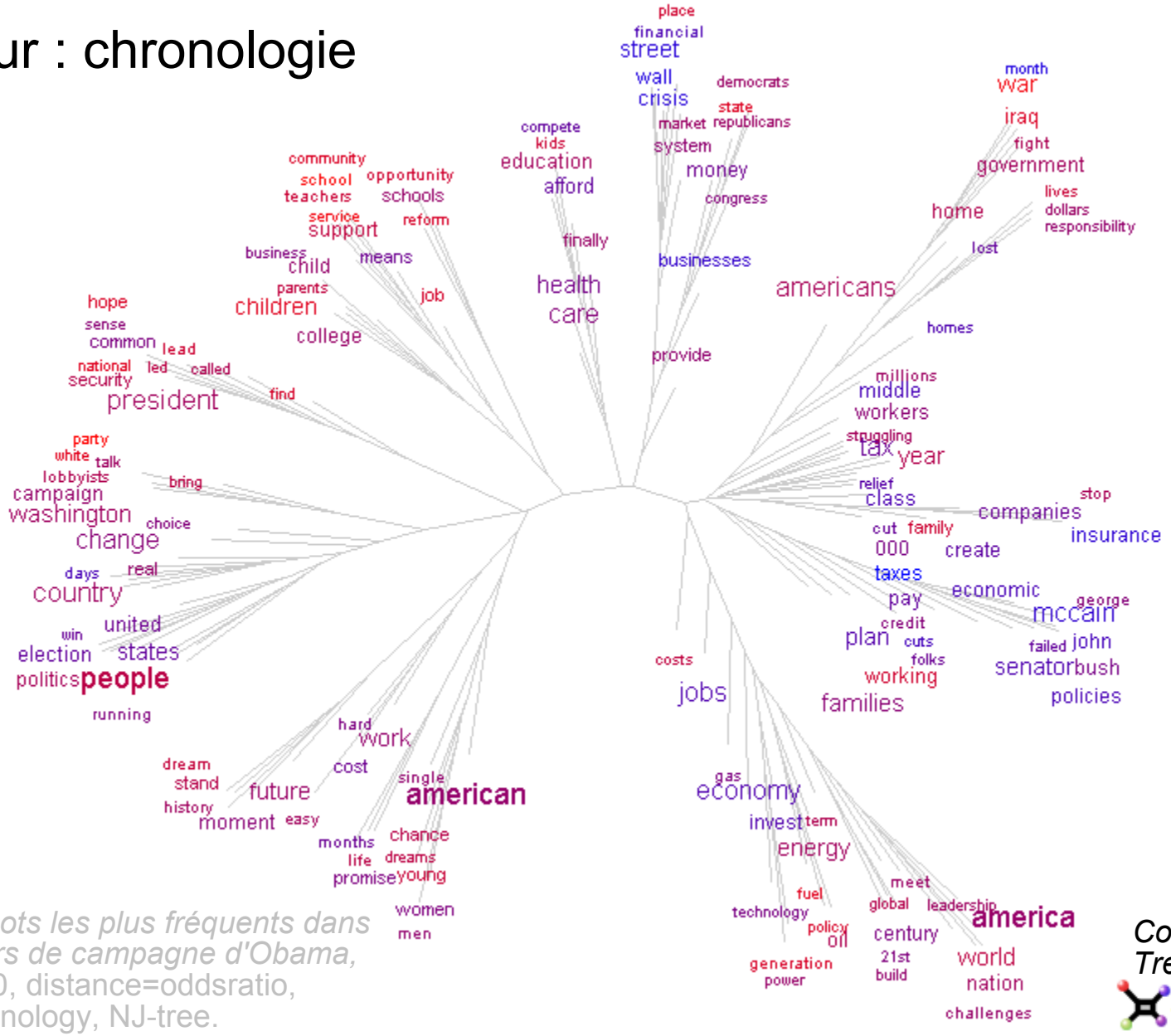
Eda, Uchiyama, Uchiyama, Yoshikawa, WWW 2009

Construction – décoration de l'arbre

couleur : chronologie

ancien

récent



Les 150 mots les plus fréquents dans les discours de campagne d'Obama, winsize=30, distance=oddsratio, color=chronology, NJ-tree.

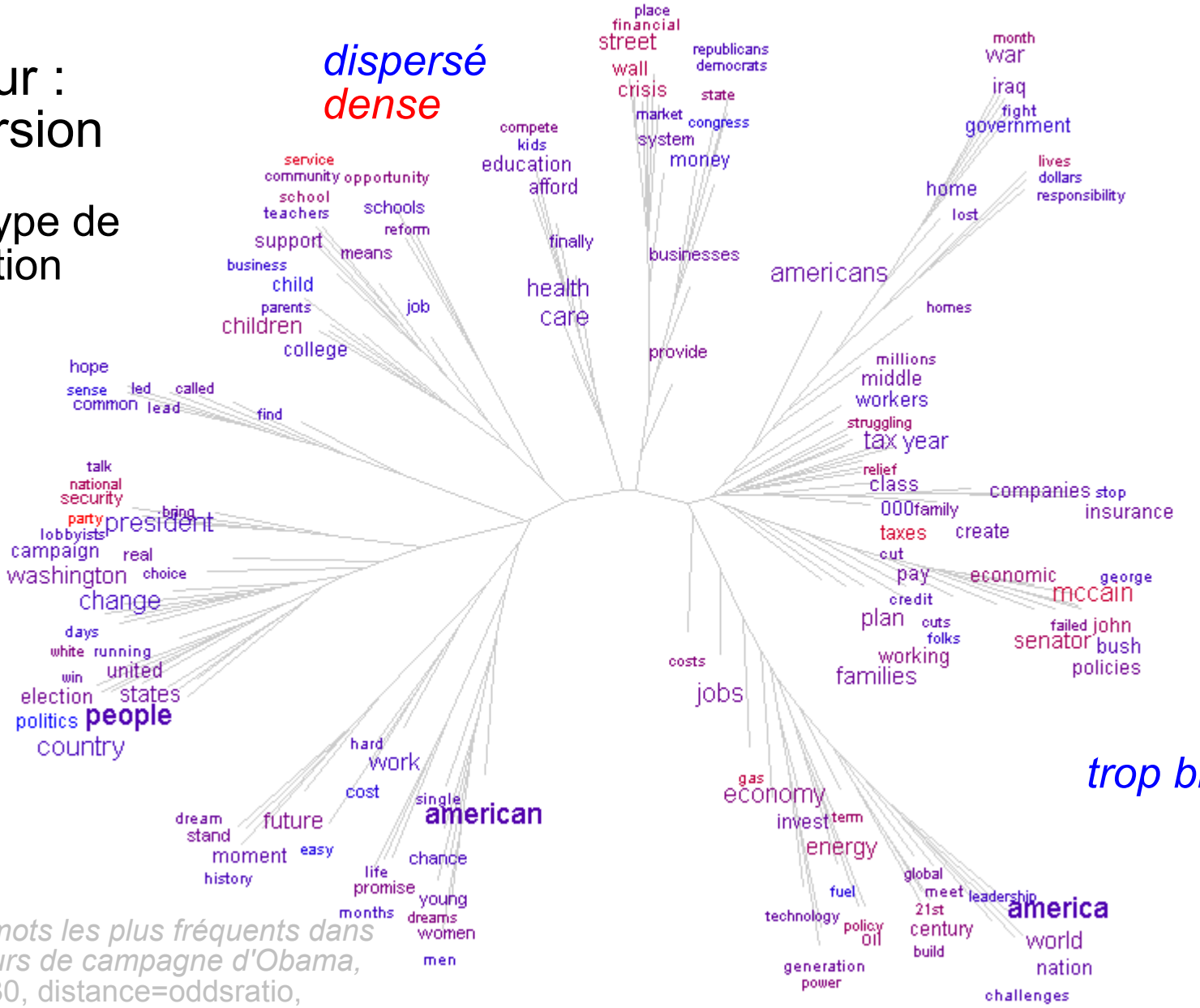
Construit avec TreeCloud et SplitsTree4

Construction – décoration de l'arbre

couleur :
dispersion

écart-type de
la position

dispersé
dense



trop bleu ?

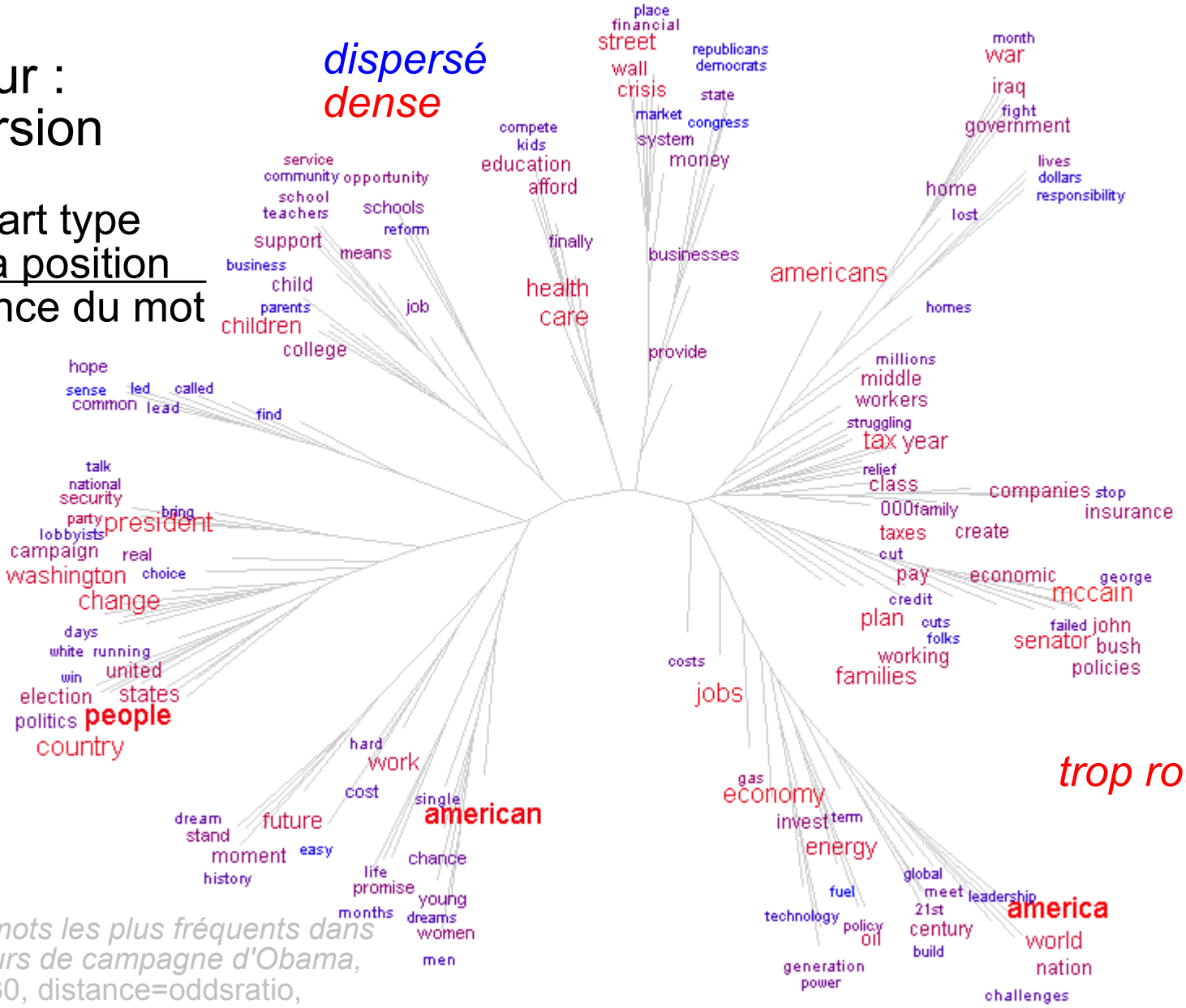
Les 150 mots les plus fréquents dans les discours de campagne d'Obama, winsize=30, distance=oddsratio, color=dispersion, NJ-tree.

Construction – décoration de l'arbre

couleur :
dispersion

écart type
de la position
fréquence du mot

dispersé
dense

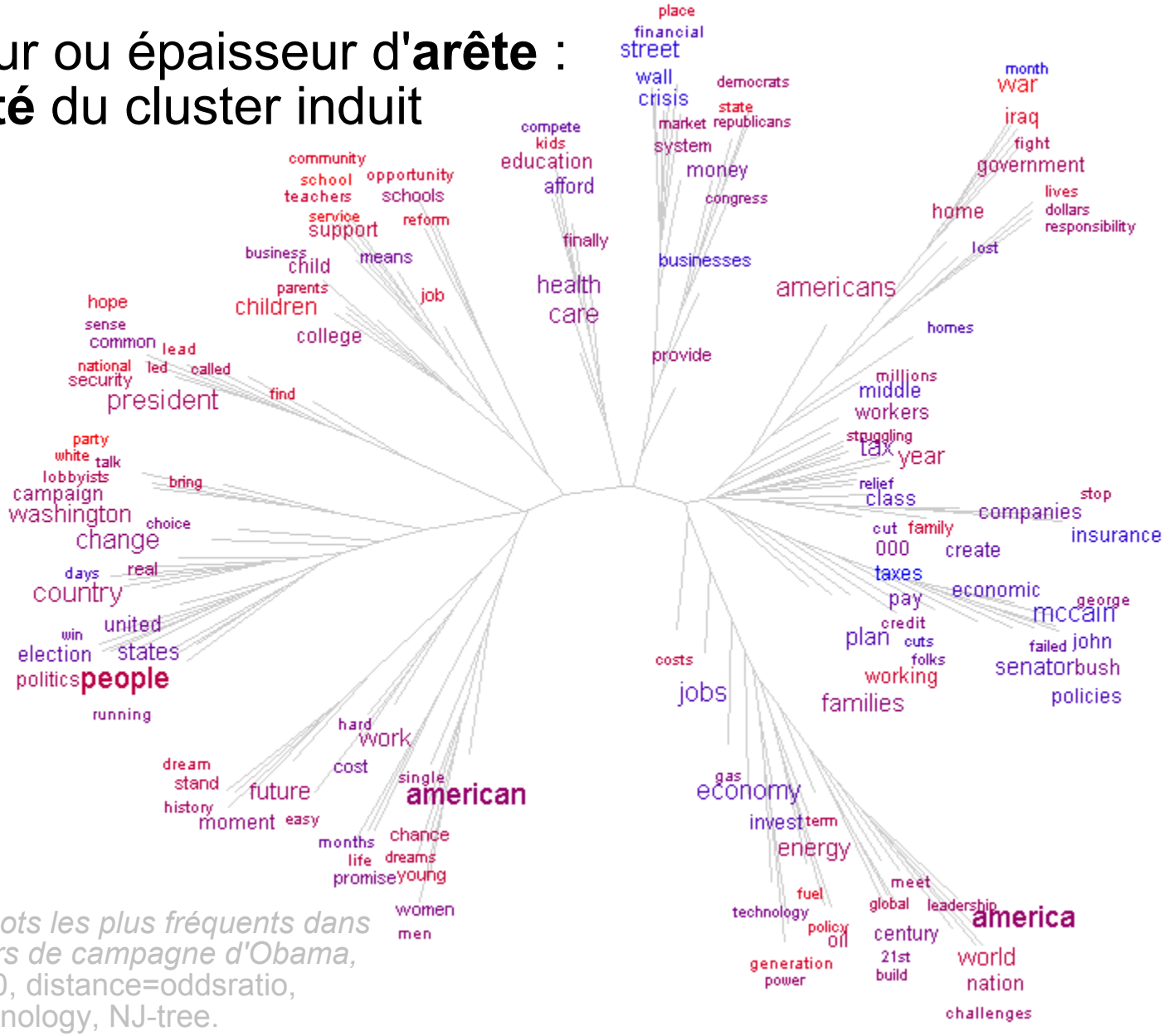


trop rouge ?

Les 150 mots les plus fréquents dans les discours de campagne d'Obama, winsize=30, distance=oddsratio, color=norm-dispersion, NJ-tree.

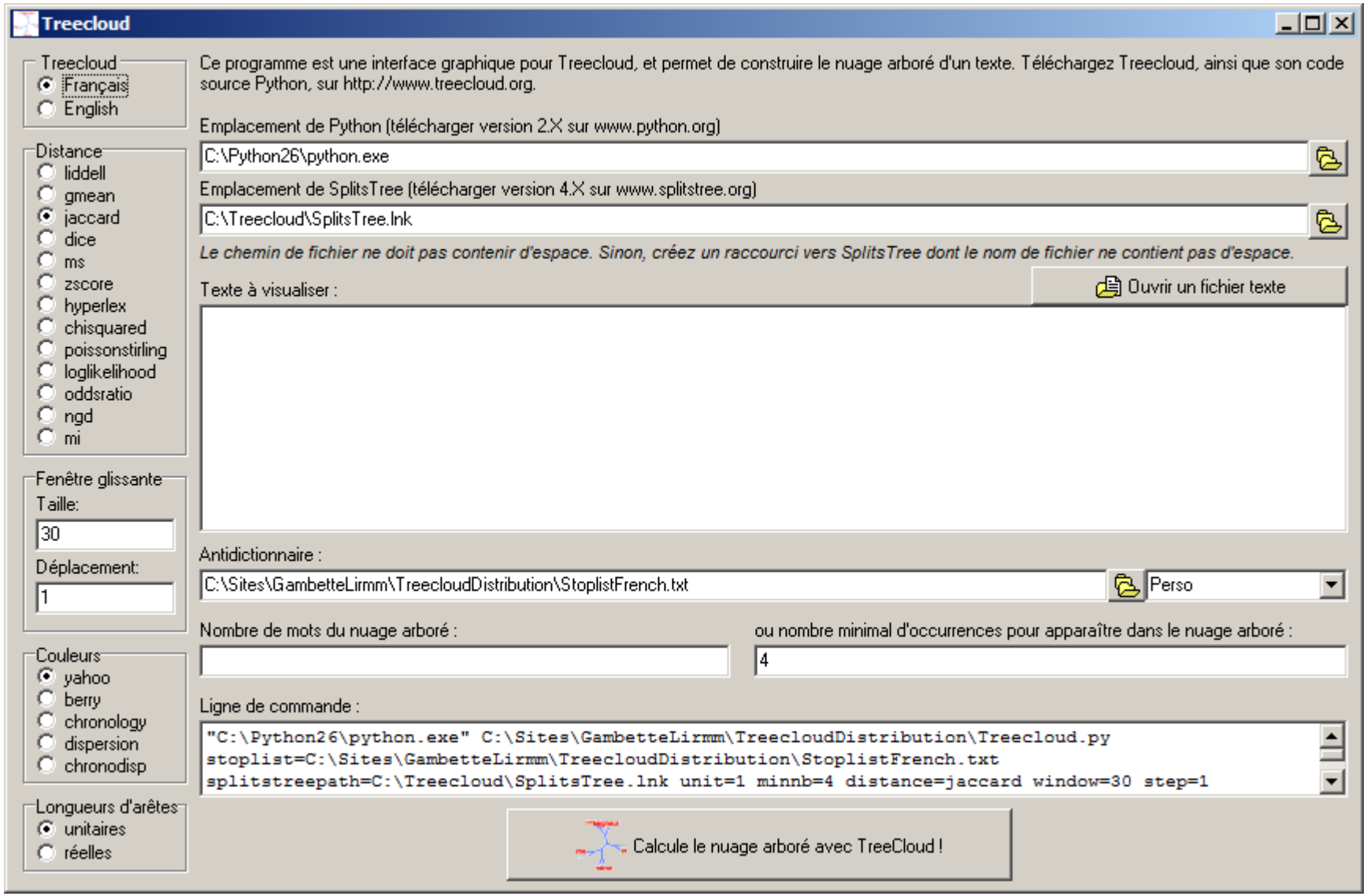
Construction – décoration de l'arbre

couleur ou épaisseur d'**arête** :
qualité du cluster induit



Les 150 mots les plus fréquents dans les discours de campagne d'Obama, winsize=30, distance=oddsratio, color=chronology, NJ-tree.

Démo



Interface graphique pour TreeCloud

Contrôle qualité

Peut-on mesurer objectivement la qualité des nuages arborés ?

Contrôle qualité

Peut-on mesurer objectivement la qualité des nuages arborés ?

Quelle est la meilleure méthode pour construire un nuage arboré à partir de mes données ?

Contrôle qualité

Peut-on mesurer objectivement la qualité des nuages arborés ?

Quelle est la meilleure méthode pour construire un nuage arboré à partir de mes données ?

Variations du nuage arboré en cas de petits changements ?

➡ **bootstrap** pour évaluer :

- **stabilité du résultat**
- **robustesse de la méthode**

Contrôle qualité

Peut-on mesurer objectivement la qualité des nuages arborés ?

Quelle est la meilleure méthode pour construire un nuage arboré à partir de mes données ?

Variations du nuage arboré en cas de petits changements ?

➡ **bootstrap** pour évaluer :

- **stabilité du résultat**
- **robustesse de la méthode**

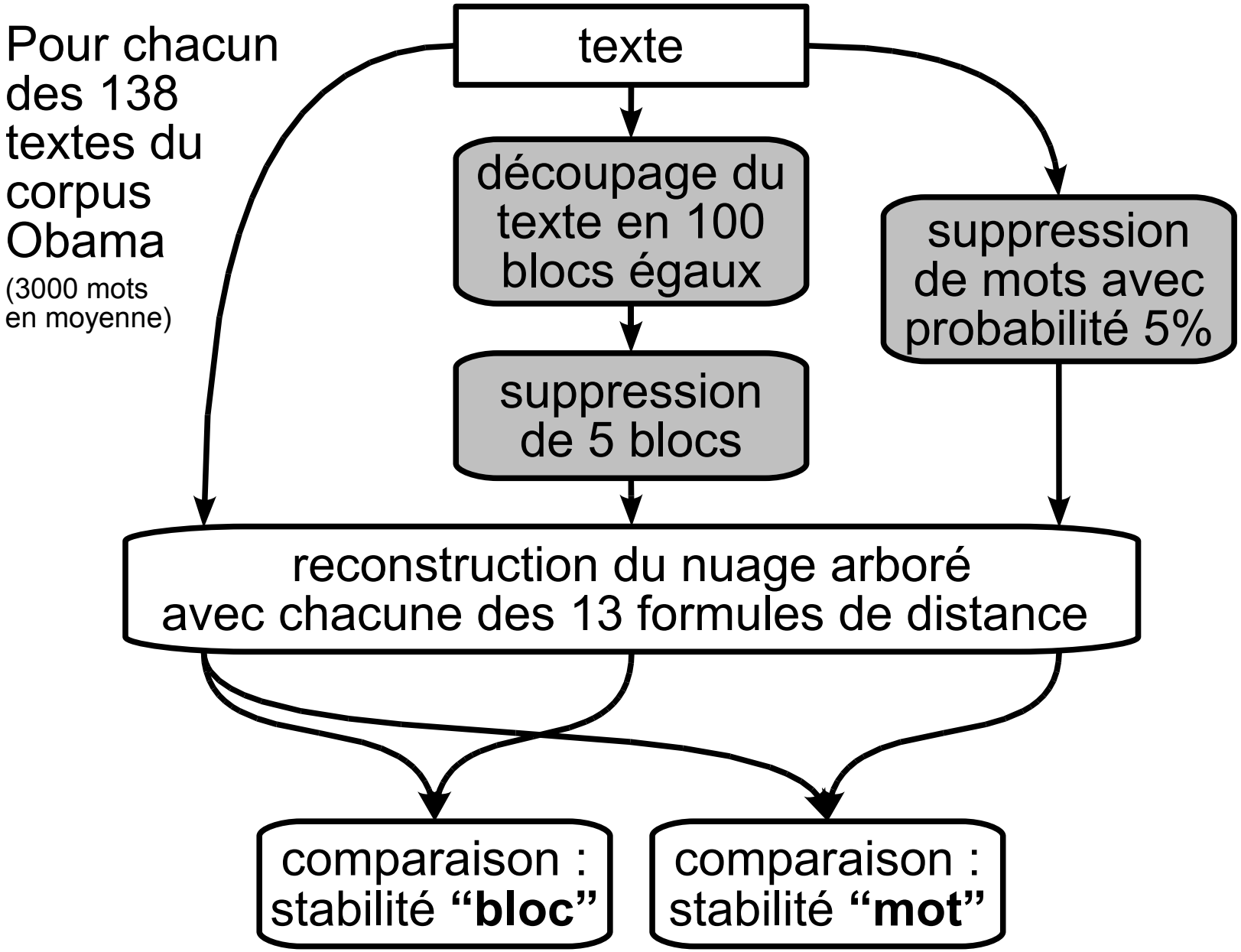
Y a-t-il une méthode plus directe ?

➡ l'**arboricité** montre à quel point la matrice de distance correspond à un arbre

➡ **implique stabilité** ?

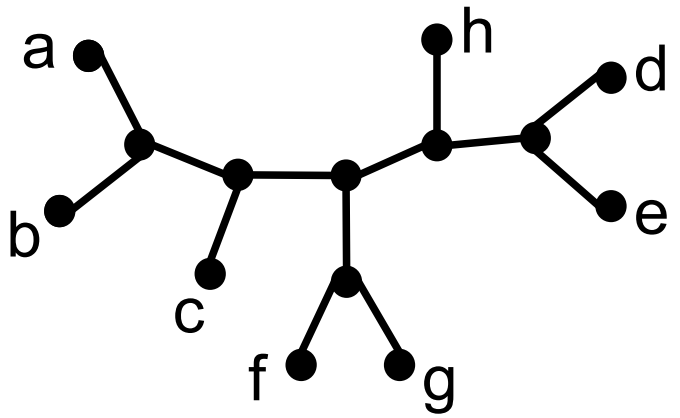
Guénoche & Garreta, 2001
Guénoche & Darlu, 2009

Contrôle qualité – bootstrap

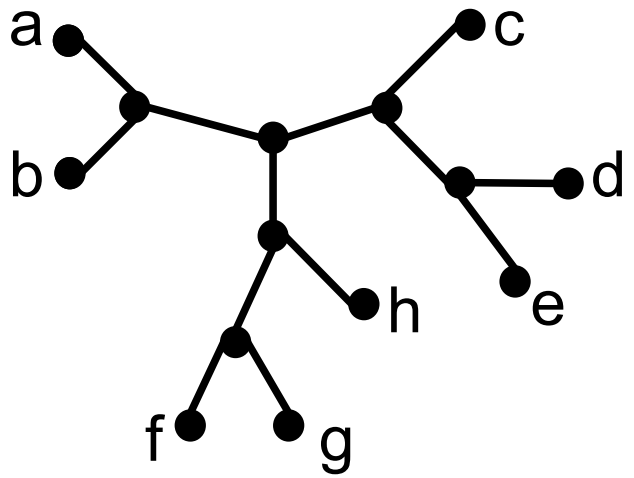


Contrôle qualité – similarité d'arbres

arbre obtenu sur le texte



arbre obtenu sur le texte altéré

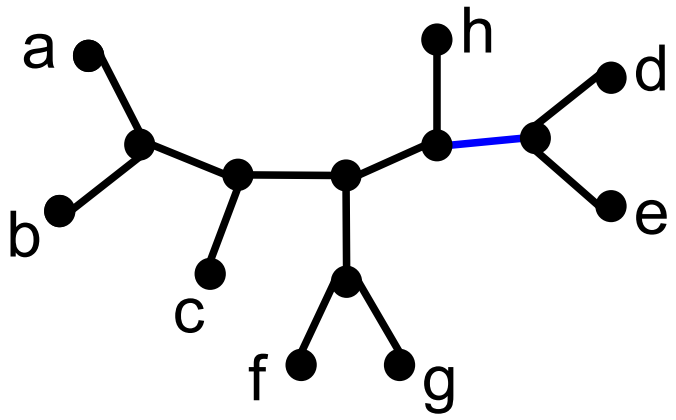


Distance de Robinson-Foulds :
Nombre de splits différents

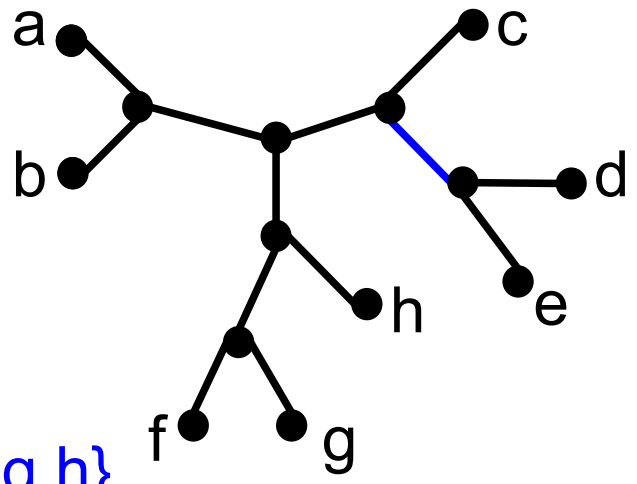
Similarité :
Pourcentage de splits non triviaux identiques

Contrôle qualité – similarité d'arbres

arbre obtenu sur le texte



arbre obtenu sur le texte altéré



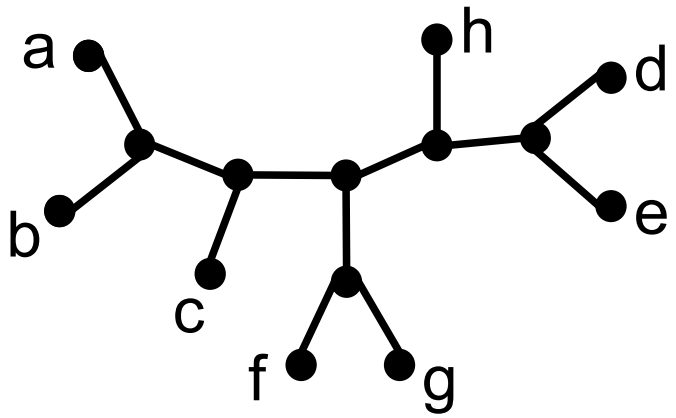
split {d,e} séparé de {a,b,c,f,g,h}

Distance de Robinson-Foulds :
Nombre de splits différents

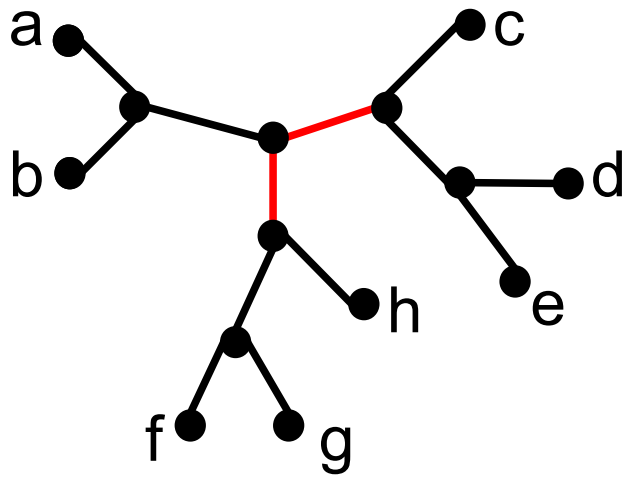
Similarité :
Pourcentage de splits non triviaux identiques

Contrôle qualité – similarité d'arbres

arbre obtenu sur le texte



arbre obtenu sur le texte altéré

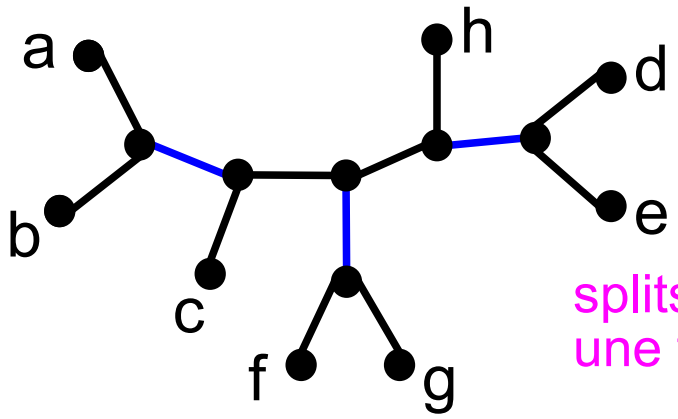


Distance de Robinson-Foulds :
Nombre de splits différents : 2

Similarité :
Pourcentage de splits non triviaux identiques

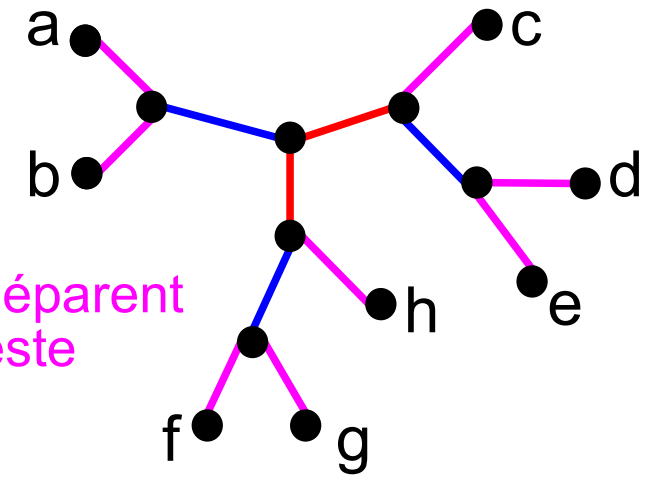
Contrôle qualité – similarité d'arbres

arbre obtenu sur le texte



splits *triviaux* : séparent une feuille du reste

arbre obtenu sur le texte altéré



Distance de Robinson-Foulds :
Nombre de splits différents : 2

Similarité :
Pourcentage de splits non triviaux identiques : 60%

Contrôle qualité – arboricité

Arboricité "discrète" d'une matrice symétrique M :

$$\mathrm{Arb}_d(M) = \frac{1}{C_n^4} \left| \{ \{i,j,k,l\} \text{ tels que Smax-Smed<Smed-Smin} \} \right|$$

où Smin, Smed, Smax sont les trois sommes $M_{i,j}+M_{k,l}$, $M_{i,k}+M_{j,l}$ et $M_{i,l}+M_{j,k}$ rangées dans l'ordre croissant.

Guénoche & Garreta, Jobim 2001

Arboricité "continue" :

$$\mathrm{Arb}_d(M) = \frac{1}{C_n^4} \sum_{i<j<k<l} \frac{\text{Smed - Smin}}{\text{Smax - Smed}}$$

Guénoche & Darlu, Alphy 2009

Contrôle qualité – arboricité

Arboricité "discrète" d'une matrice symétrique M :

$$\mathrm{Arb}_d(M) = \frac{1}{C_n^4} \left| \{ \{i,j,k,l\} \text{ tels que Smax-Smed<Smed-Smin} \} \right|$$

condition liée à la *condition des quatre points*

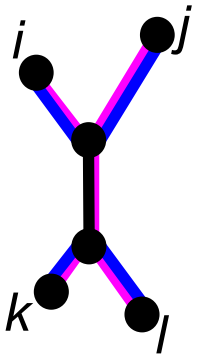
où Smin, Smed, Smax sont les trois sommes $M_{i,j}+M_{k,l}$, $M_{i,k}+M_{j,l}$ et $M_{i,l}+M_{j,k}$ rangées dans l'ordre croissant.

Guénoche & Garreta, Jobim 2001

Arboricité "continue" :

$$\mathrm{Arb}_d(M) = \frac{1}{C_n^4} \sum_{i<j<k<l} \frac{\text{Smed - Smin}}{\text{Smax - Smed}}$$

Guénoche & Darlu, Alphy 2009



Condition des quatre points :
Smin ≤ min(Smed,Smax) pour tout quadruplet ssi M correspond à une distance d'arbre

Contrôle qualité – arboricité

Arboricité "discrète" d'une matrice symétrique M :

$$\mathrm{Arb}_d(M) = \frac{1}{C_n^4} \left| \{ \{i,j,k,l\} \text{ tels que Smax-Smed<Smed-Smin} \} \right|$$

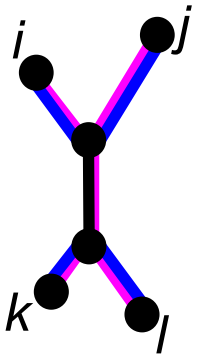
où Smin, Smed, Smax sont les trois sommes $M_{i,j}+M_{k,l}$, $M_{i,k}+M_{j,l}$ et $M_{i,l}+M_{j,k}$ rangées dans l'ordre croissant.

Guénoche & Garreta, Jobim 2001

Arboricité "continue" :

$$\mathrm{Arb}_d(M) = \frac{1}{C_n^4} \sum_{i<j<k<l} \frac{\text{Smed - Smin}}{\text{Smax - Smed}}$$

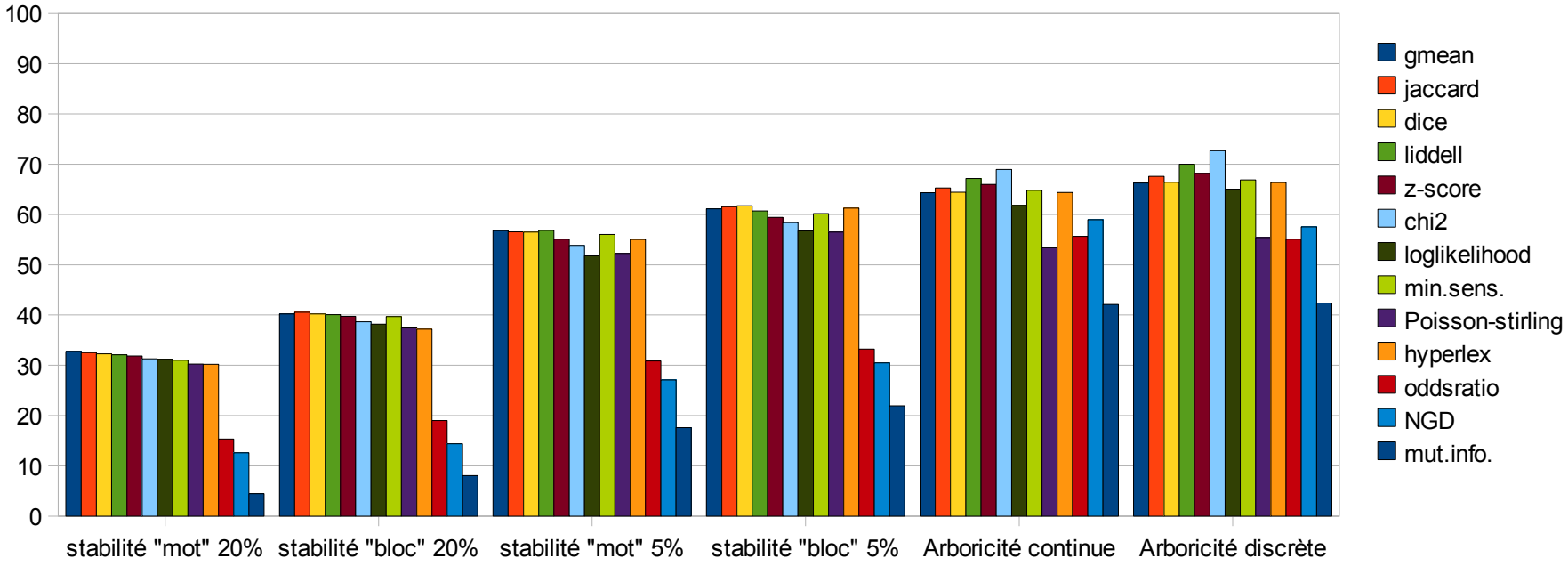
Guénoche & Darlu, Alphy 2009



Smin ≤ Smed = Smax
Une distance est **proche d'une distance d'arbre** si Smed est plus proche de Smax que de Smin.

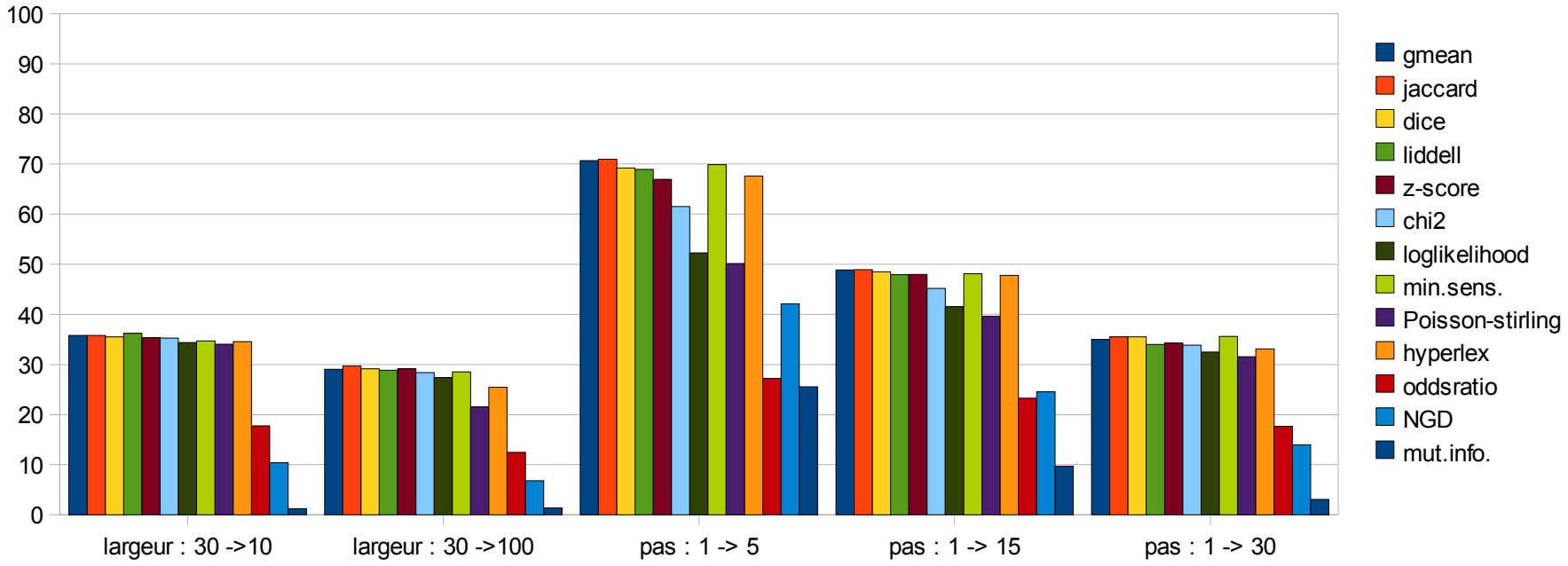
Contrôle qualité – résultats

Qualité de bootstrap et arboricité :



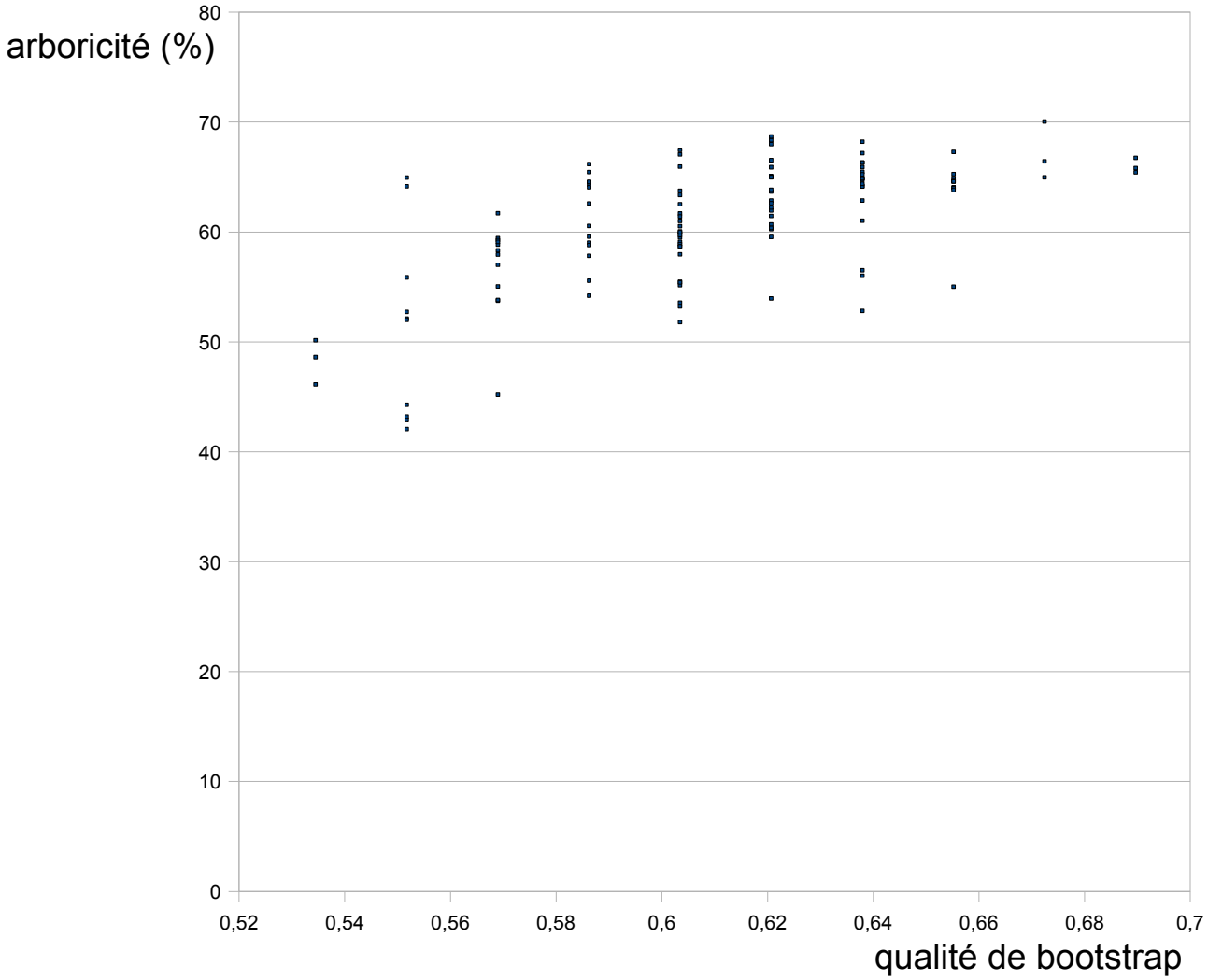
Contrôle qualité – résultats

Robustesse au changement de paramètres de la fenêtre glissante :



Contrôle qualité – arboricité

Relations entre “qualité de bootstrap” et arboricité :

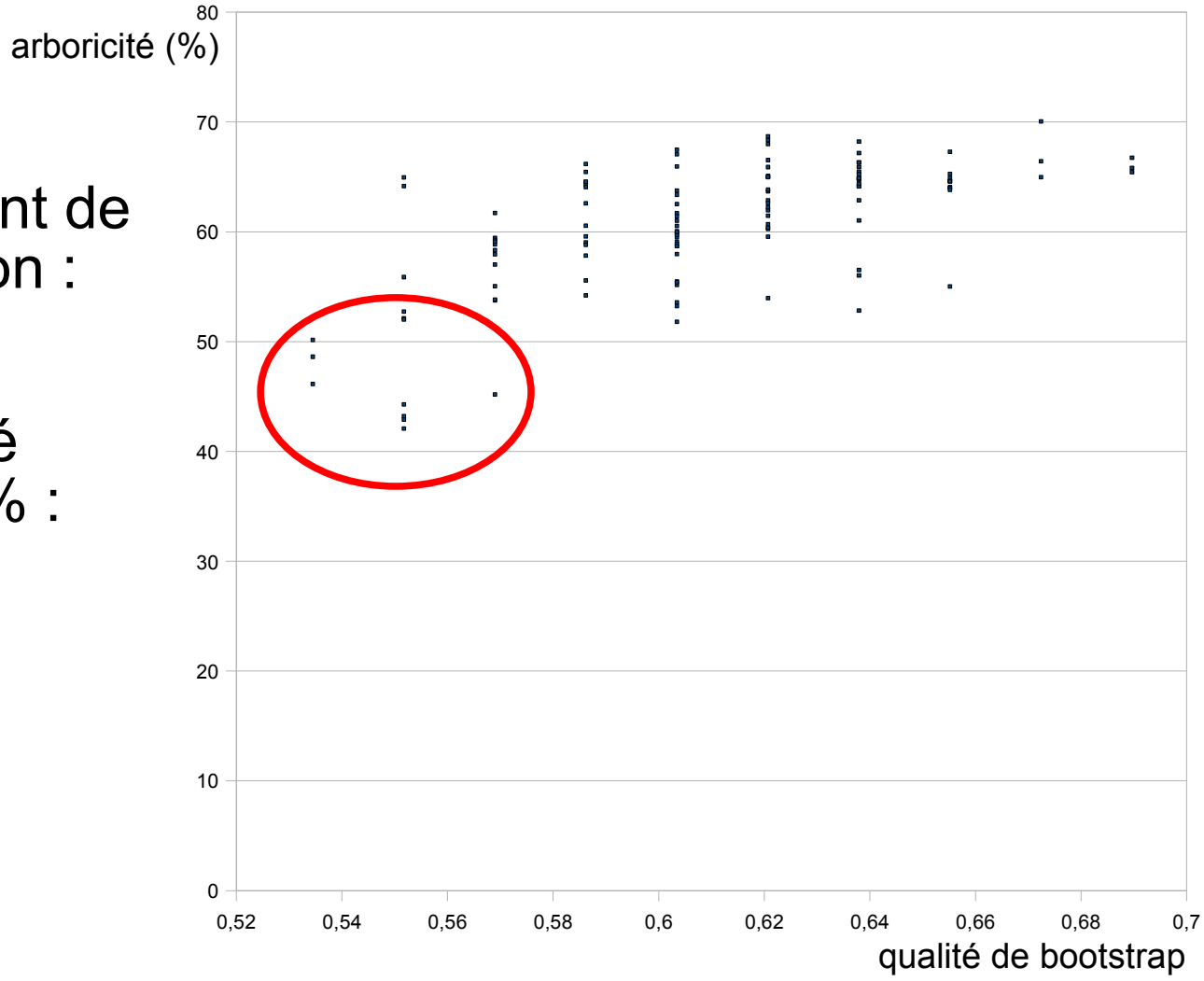


Contrôle qualité – arboricité

Relations entre “qualité de bootstrap” et arboricité :

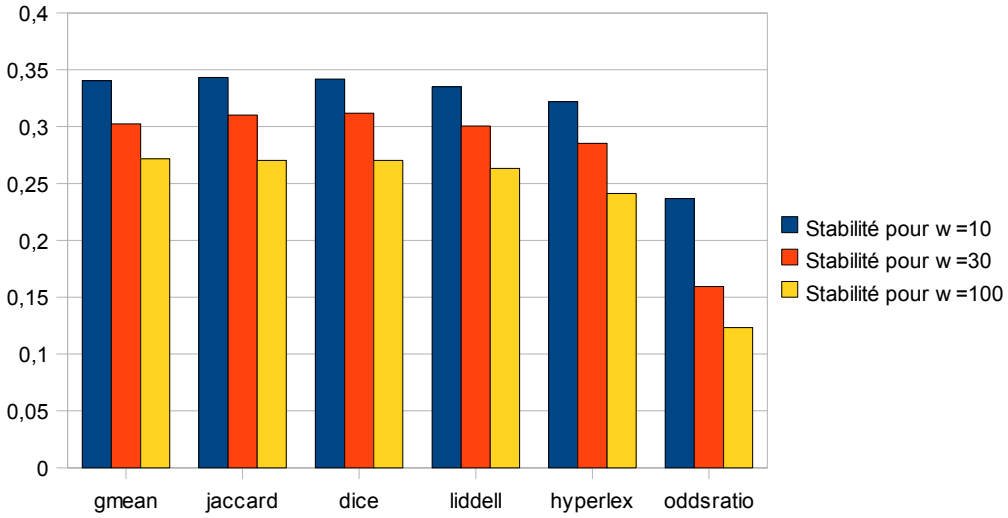
Coefficient de corrélation : 0.64

Arboricité sous 50% : danger !



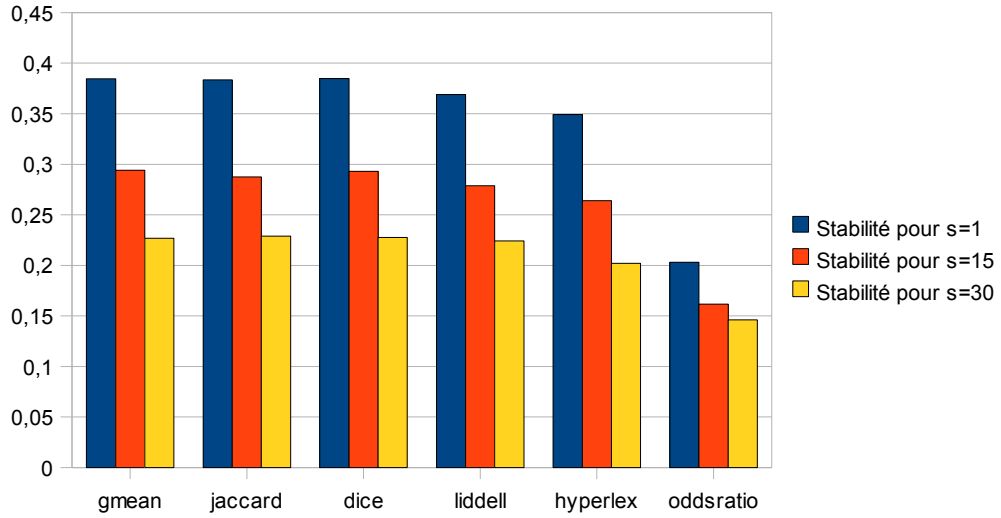
Choix des paramètres – fenêtre glissante

Choix de la largeur de la fenêtre glissante :



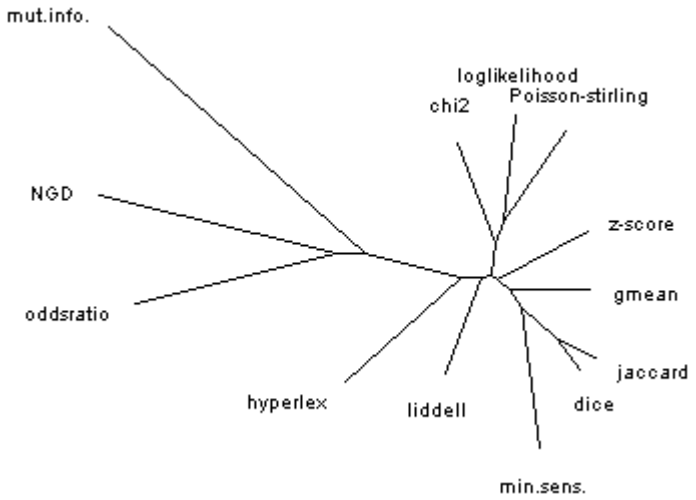
Choix des paramètres – fenêtre glissante

Choix du pas de glissement de la fenêtre glissante :



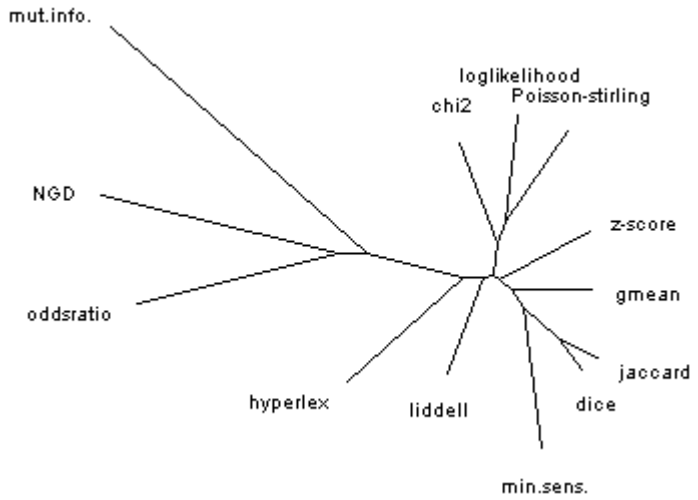
Choix des paramètres – distance

L'arbre des distances entre arbres de distances :



Choix des paramètres – distance

L'arbre des distances moyennes, sur les textes du corpus Obama, entre les arbres de mots, en fonction du choix de la formule de distances entre mots :



Limites de la visualisation

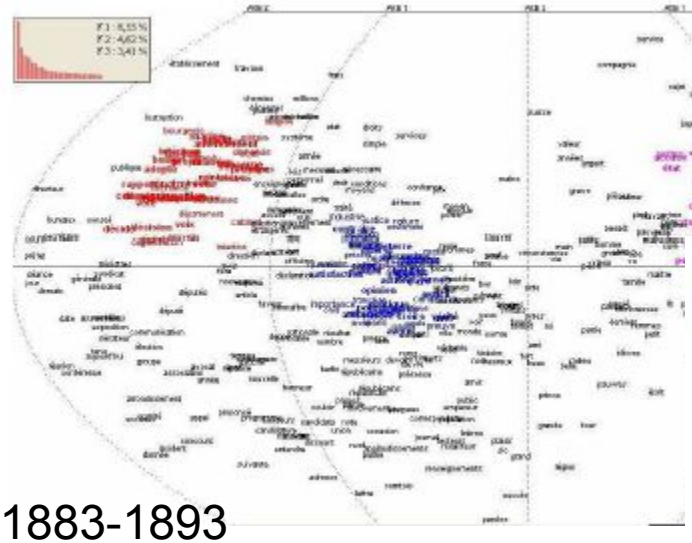
- problème de **place** :
  espace en $O(n^2)$ pour représenter $O(n)$ mots

- problème de **robustesse** :
  5% de modifications sur le texte induit 40% de modifications sur l'arbre

Limites de la visualisation

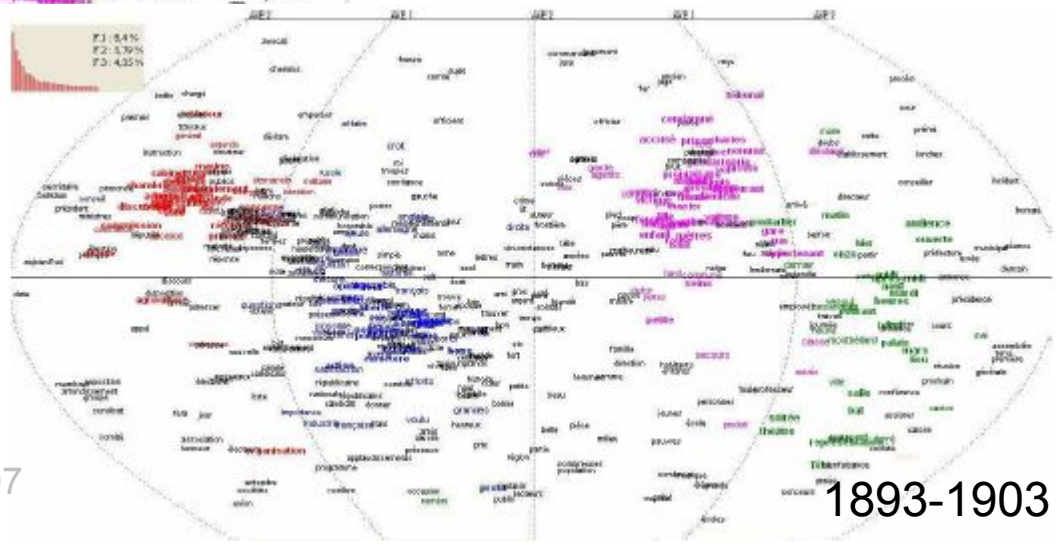
- problème de **robustesse** :
  5% de modifications sur le texte induit 40% de modifications sur l'arbre

Robustesse de l'AFC sur gros corpus :



1883-1893

Coloration des pôles isotropiques pour 1883-1893.



1893-1903

Viprey, Lethier
Lecture de l'archive et linguistique de corpus : presse écrite du XIX° siècle,
Journées de la linguistique de corpus 2007

Perspectives

- Créer une interface web
  SplitsTree : problème de droits. Scriptree ?

http://www.treecloud.fr
http://www.scriptree.org

- Tester d'autres méthodes de construction d'arbre

- Evaluer l'utilité des nuages arborés en analyse de textes

- Construire le nuage arboré quotidien des personnalités citées sur les blogs, avec WIKIO Labs

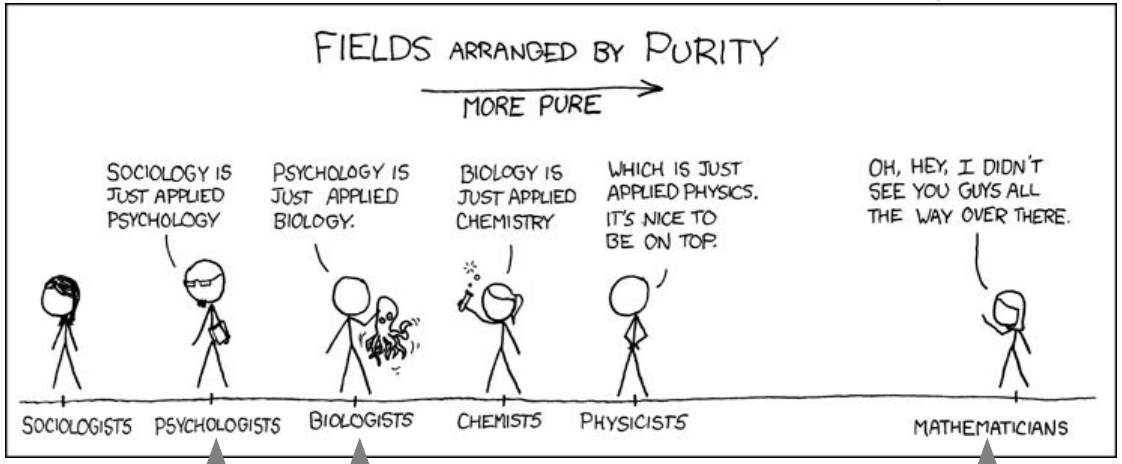
http://labs.wikio.net

Merci pour votre attention !

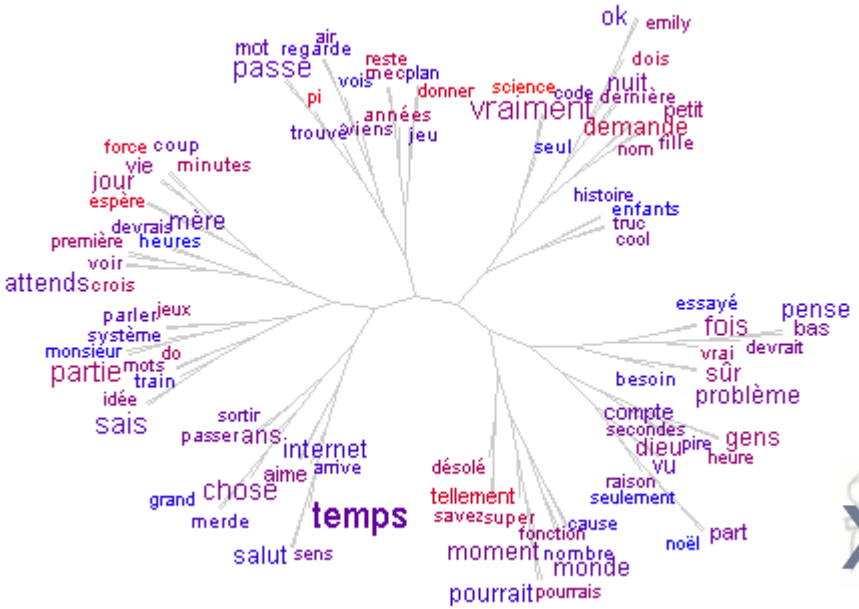
Merci pour votre attention !

http://xkcd.com/365

Domaines triés par pureté croissante :
Le sociologue
Le psychologue : la sociologie, c'est juste de la psychologie appliquée.
Le biologiste : la psychologie, c'est juste de la biologie appliquée.
Le chimiste : la biologie, c'est juste de la chimie appliquée.
Le physicien : ... qui est juste de la physique appliquée. C'est chouette de tout dominer.
Le mathématicien : oh, salut les gars, je vous avais pas vu d'aussi loin.



domaines mis en jeu dans la conception de TreeCloud



Nuage arboré de la traduction du webcomic xkcd

xkcd — *UNE BD EN LIGNE SUR FOND D'AMOUR, SARCASME, MATHS, ET LINGUISTIQUE.*

Xkcd en français : http://xkcd.free.fr (il reste 7% des planches à traduire...)