Application de la méthode ALCESTE aux « gros » corpus et stabilité des « mondes lexicaux » : analyse du « CableGate » avec IRaMuTeQ

Pierre Ratinaud – ratinaud@univ-tlse2.fr
Pascal Marchand – pascal.marchand@iut-tlse3.fr

Laboratoire LERASS

Université de Toulouse

Application de la méthode ALCESTE aux « gros » corpus et stabilité des « mondes lexicaux » : analyse du « CableGate » avec IRaMuTeQ

- Présentation de la méthode ALCESTE
- Les adaptions de la C.H.D. réalisées dans IraMuTeQ
- Analyses sur le corpus du CableGate
  - Présentation du corpus
  - Hypothèse et traitements
  - Résultats
- Conclusions

Présentation de la méthode ALCESTE

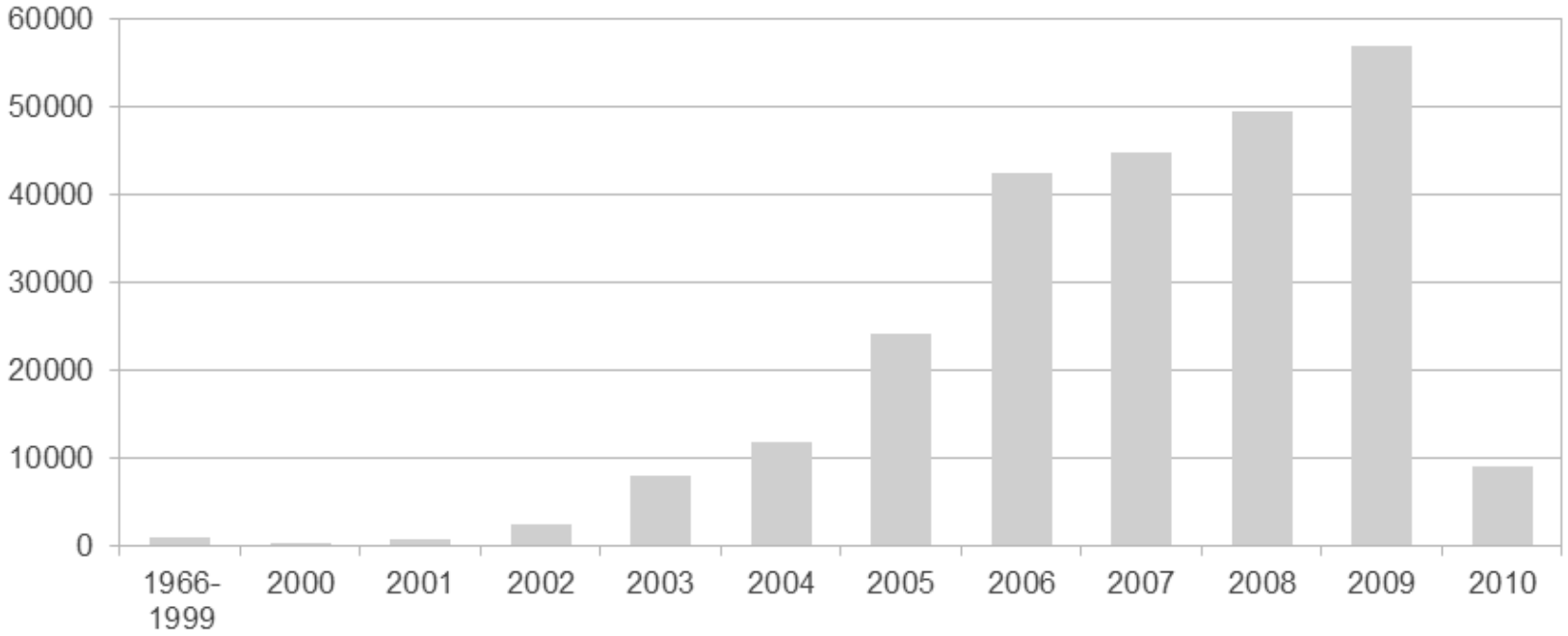
- Implémenté d'abord dans le logiciel ALCESTE Reinert (1983, 1990)
- puis dans IRaMuTeQ (Ratinaud & Déjean, 2009)
- Quelques particularités :
  - Découpage des unités du corpus (u.c.i.) en segments de texte (u.c.e.)
  - Sélection des formes « pleines »
  - Une Classification Hiérarchique Descendante originale
    - Série de bi-partitions en trois étapes
      - Analyse Factorielle des Correspondances
      - Permutation de toutes les unités jusqu'à la maximisation de l’inertie inter-classe
      - Élimination des formes significativement absentes des classes *

Les adaptions de la C.H.D. réalisées dans IraMuTeQ

- Des trois étapes de la CHD, c'est la première qui impose les limites des tableaux
- Les adaptations
  - Matrices pleines → Matrices creuses
  - Svd dans R →svdlibc (Rhode, 2001)

Analyses sur le corpus du CableGate
Présentation du corpus

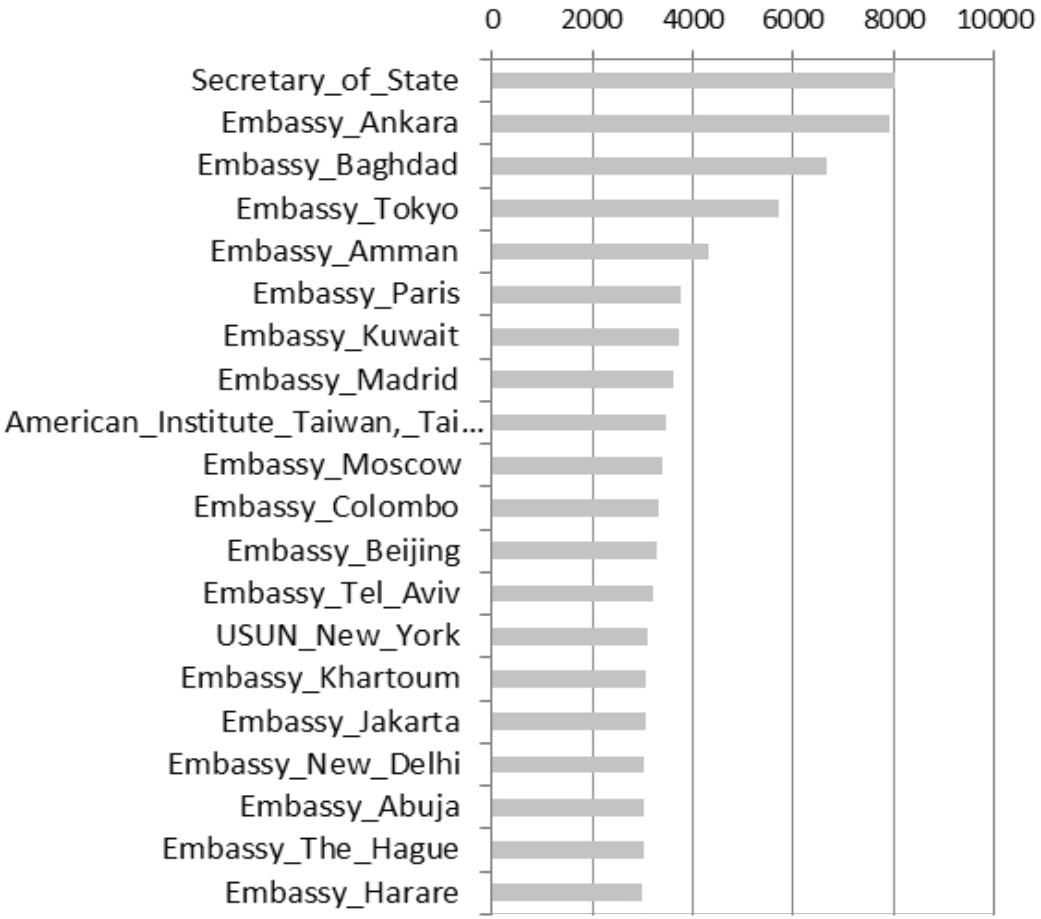
- Base de données PostgreSQL de 1,7 Go
- 251 287 câbles de la diplomatie américaine



Répartition des télégrammes par date

Analyses sur le corpus du CableGate
Présentation du corpus

- Base de données PostgreSQL de 1,7 Go
- 251 287 câbles de la diplomatie américaine



Les 20 sources les plus fréquentes

Analyses sur le corpus du CableGate
Présentation du corpus

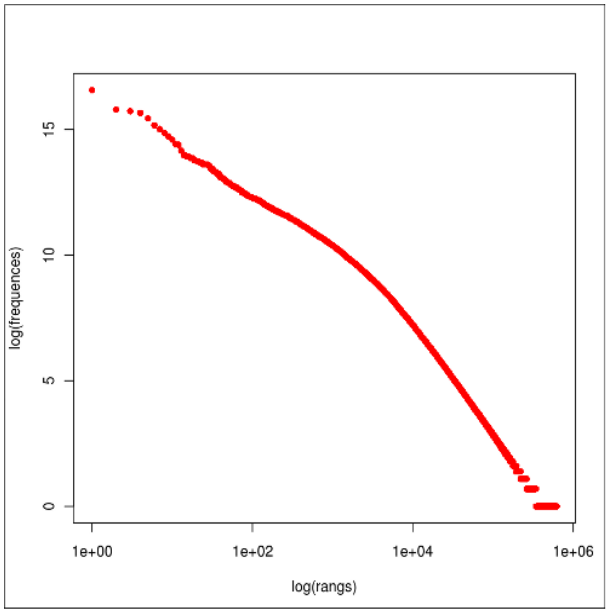
- Base de données PostgreSQL de 1,7 Go
- 251 287 câbles de la diplomatie américaine
- 238 116 128 occurrences après « nettoyage » (Fmax = 15 668 471, « the »)
- 624 202 formes (280 863 hapax - 44,9 % des formes - 0,11 % des occurrences)

UNCLAS STATE 204472
E.O. 12958: N/A
TAGS: PTER
SUBJECT: ANNUAL TERRORISM REPORT
(THIS CABLE HAS BEEN CLEARED BY M/P (SEP).)

. SUMMARY

THE DEPARTMENT IS REQUIRED BY LAW TO PROVIDE AN ANNUAL TERRORISM REPORT TO CONGRESS. THIS LAW REQUIRES THE REPORT BE A FULL AND COMPLETE FACTUAL RECORD OF TERRORISM-RELATED ACTIVITY IN ALL COUNTRIES THAT EXPERIENCED TERRORISM AND NOT BE TEMPERED BY CONCERNS ABOUT HOST GOVERNMENT

exemple de début de télégramme
en gras, les parties éliminées par le nettoyage

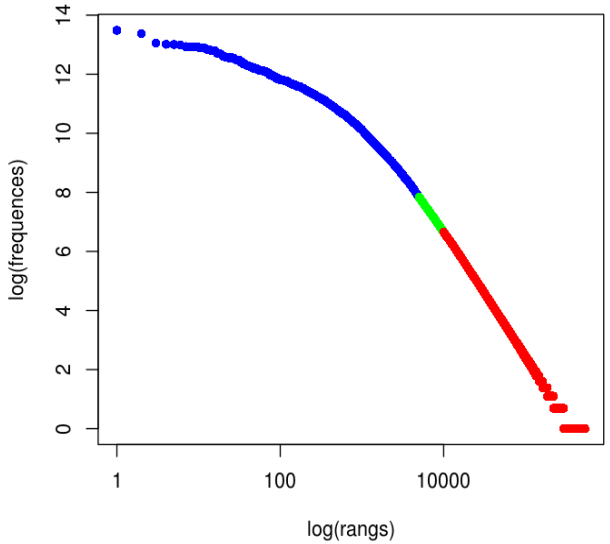


graphique rangs/fréquences des formes du corpus
(échelles logarithmiques)

Hypothèse et traitements

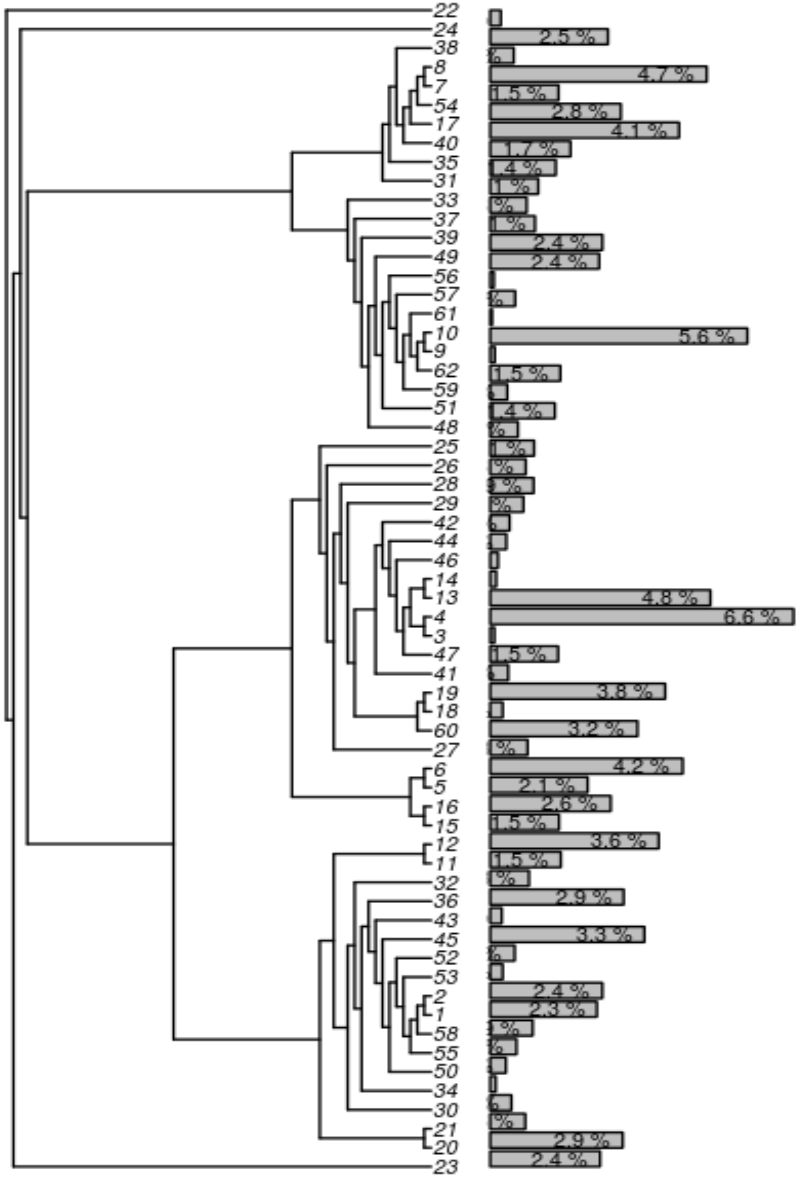
- Constats
  - La taille des unités classées a peu d'influence sur le mondes lexicaux mis à jour et sur leur organisation (Reinert, 1993, 1995)
  - Reinert (2008) constate des similarités entre les mondes lexicaux identifiés sur des corpus différents (mais qui entretiennent une proximité « thématique ») : « mondes lexicaux stabilisés »
  - « Il a bien fallu admettre l'existence d'un mode de fabrication de la différence relativement indépendant des domaines de connaissance dont relevaient ces corpus ! » (Reinert, 2008)
- Hypothèse
  - Nous devrions retrouver une certaine forme de stabilité dans l'organisation des documents mis à jour entre deux classifications portant sur des fenêtres de fréquences différentes.

Hypothèse et traitements

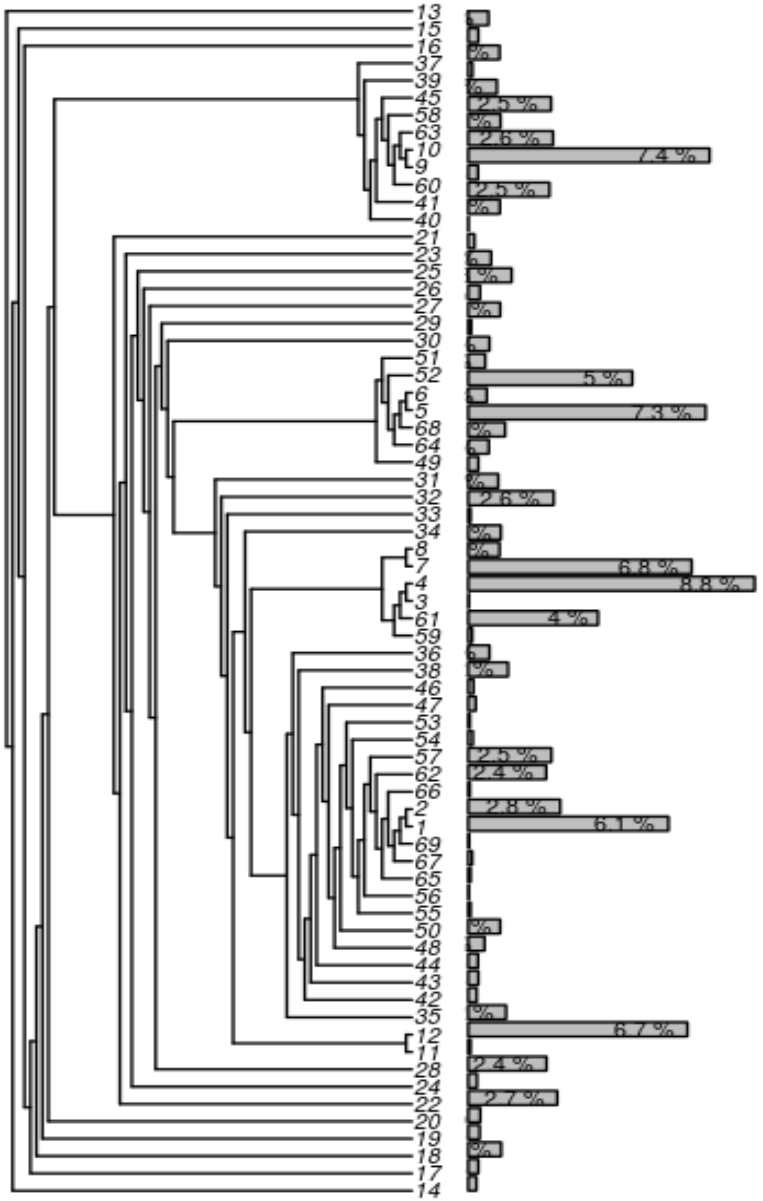


	Classification 1	Classification 2
Fréquence max. d'une forme pleine sélectionnée	720626	2534
Fréquence min. d'une forme pleine sélectionnée	2536	781
Nombre de formes pleines sélectionnées	5002	5000
Pourcentage de « 1 » dans la matrice	4,1%	0,33 %
Fréquence max. dans la matrice	173906 (end)	2396 (slug)
Fréquence min. dans la matrice	13 (aspirante)	17 (wof)
Nombre de lignes dans la matrice	251287	251287

Résultats



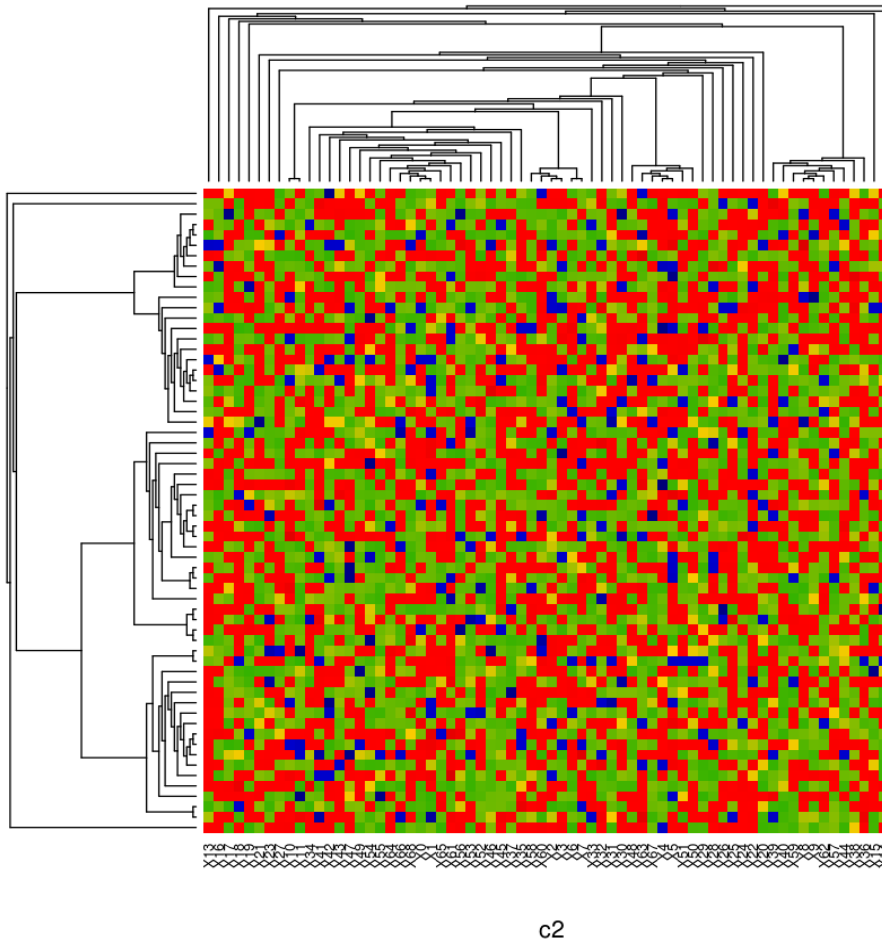
dendrogramme de la première classification

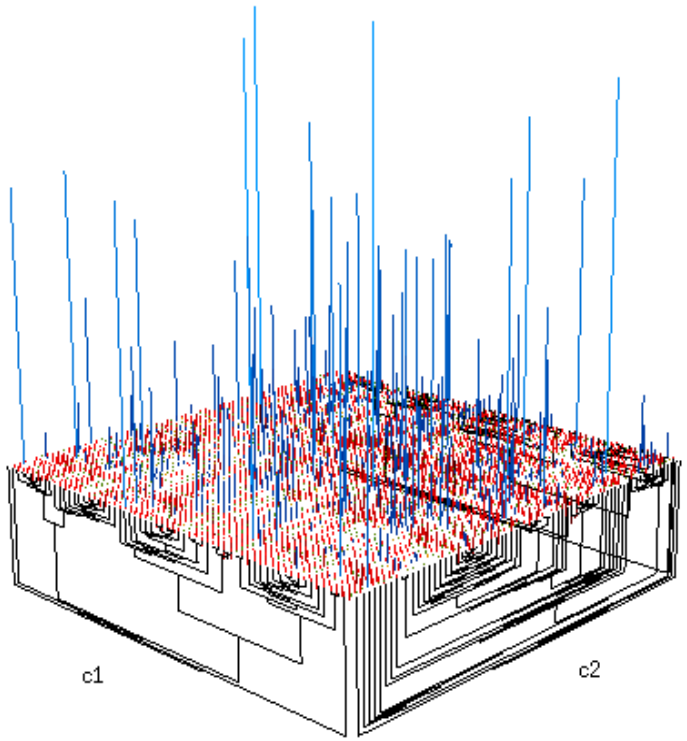


dendrogramme de la seconde classification

Les barres indiquent la taille relative des classes

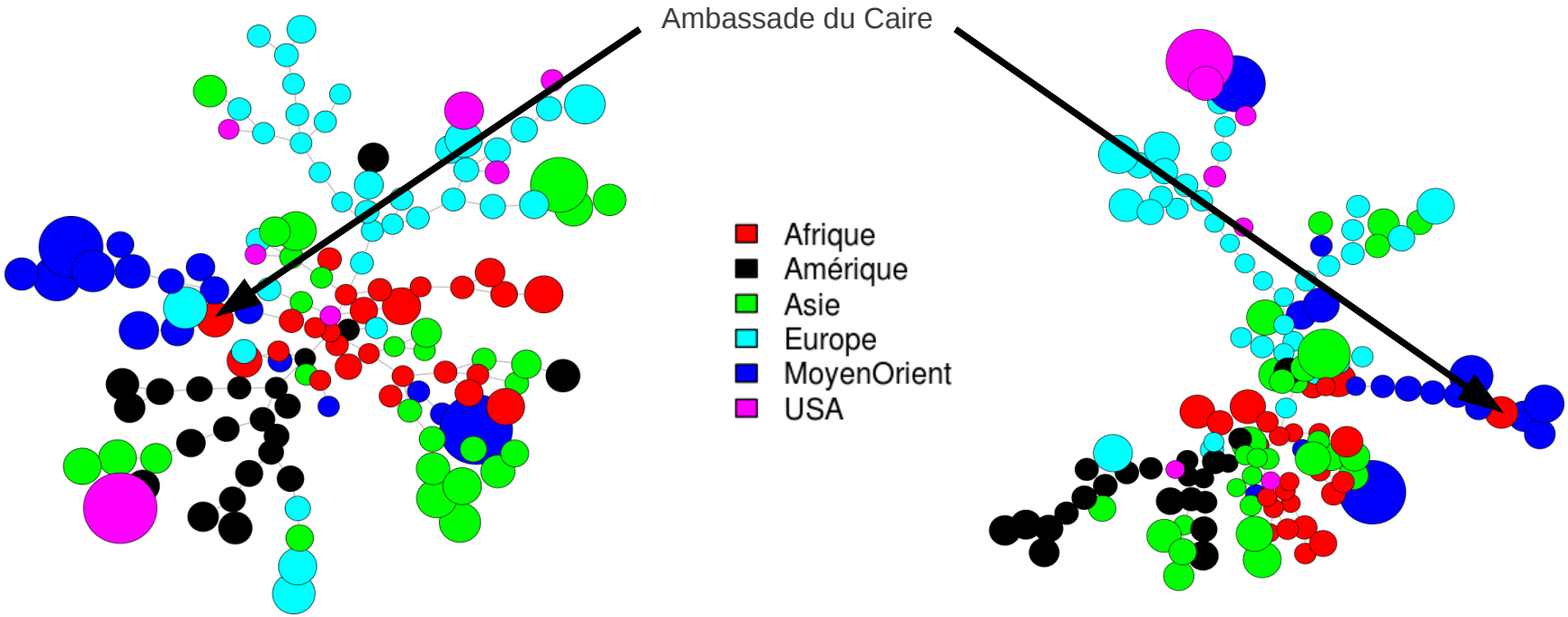
Résultats



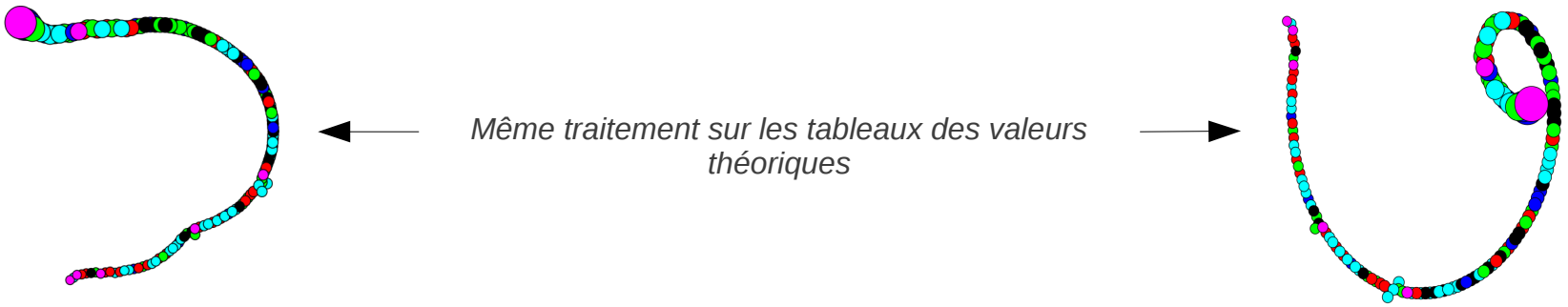


résidus standardisés du chi2 mené sur le tableau de contingence qui croise les deux classifications
les points bleus représentent les contributions significativement positives (>1,96)
les niveaux de vert représentent les contributions non-significatives
les niveaux de rouge représentent les contributions significativement négatives (< -1,96)

Résultats



Arbres minimum des graphes des matrices des distances euclidiennes des tableaux de contingence qui croisent les classes avec les sources apparaissant au moins 500 fois N=138

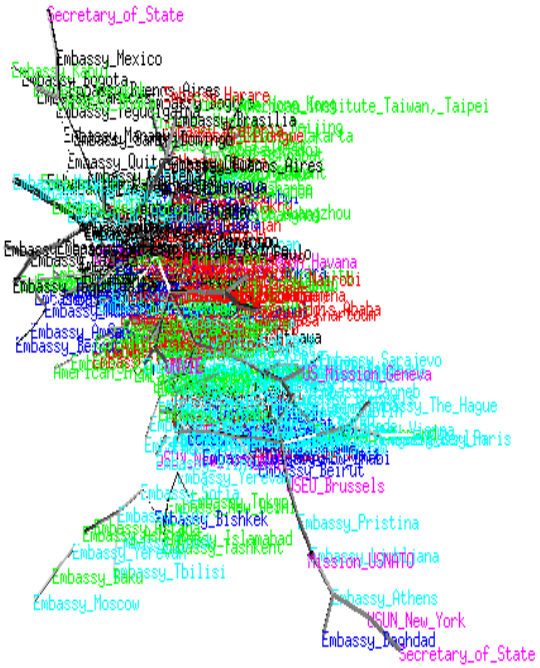


Même traitement sur les tableaux des valeurs théoriques

Conclusion

- On retrouve bien une forte similarité dans l'organisation des documents alors que les classifications portent sur des matrices très différentes :
  - aucune forme commune entre les matrices
  - la seconde matrice est plus de dix fois plus vide que la première
- Cette propriété peut être utilisée pour envisager différentes stratégies de traitements sur les gros corpus
- Les adaptations de l’algorithme permettent de produire des classifications doubles sur UC sur l'intégralité de gros corpus en maximisant le nombre de formes pleines prises en compte.

Merci



- Afrique
- Amérique
- Asie
- Europe
- MoyenOrient
- USA