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

Pierre Ratinaud[1], Pascal Marchand[2]

[1] Université Toulouse – ratinaud@univ-tlse2.fr
[2] Université de Toulouse – pascal.marchand@iut-tlse3.fr

Abstract

We propose an analysis of the entire 251,287 telegrams of the "CableGate". This corpus consists of 238 116 128 occurrences. After a short description of the adaptations of the ALCESTE procedure implemented in the software IRAMUTEQ that were necessary, we will present a comparison of two treatments: a clustering of a table that cross the 251 287 documents and the 5002 first "full" words (frequency minimum: 2536) and a clustering of a the table that cross these documents and the following 5000 "full" words (frequencies between 2534 and 781). The two tests, performed with the same parameters, produce 62 final clusters for the first and 69 for the second. The intersection between these two classifications shows clearly that they are not independent. In other words, a clustering conducted on the most common words of a text leads to a distribution of documents that has similarities with a clustering conducted on words much less frequent. These results help to maintain the hypothesis of «stabilized lexical worlds» (Reinert, 2008) and also allow to consider various options for ALCESTE analysis on large corpora.

Résumé

Nous proposons une analyse portant sur l'intégralité des 251287 télégrammes du « CableGate ». Ce corpus se compose de 238 116 128 occurrences. Après avoir rapidement décrit les adaptations de la procédure d'analyse ALCESTE implémentée dans le logiciel IRAMUTEQ qui ont été nécessaires, nous présenterons une comparaison entre deux traitements : une classification sur un tableau croisant les 251287 documents et les 5002 formes pleines les plus fréquentes (fréquence minimum : 2536) et une classification sur un tableau croisant ces mêmes documents et les 5000 formes pleines suivantes (fréquences comprises entre 2534 et 781). Les deux analyses, réalisées avec les mêmes paramètres, produisent 62 classes terminales pour la première et 69 pour la seconde. Le croisement entre ces deux classifications montre clairement qu'elles ne sont pas indépendantes. Autrement dit, une classification menée sur les formes les plus fréquentes de ce corpus conduit à une répartition des documents qui présente des similarités certaines avec une classification menée sur des formes beaucoup moins fréquentes. Ces résultats permettent de maintenir l'hypothèse des « mondes lexicaux stabilisés » (Reinert, 2008) et autorisent également, en tenant compte de leurs limites, d'envisager différentes options pour les analyses de type ALCESTE sur de gros corpus.

Mots-clés : classification, méthode ALCESTE, mondes lexicaux, gros corpus

1. Introduction

A partir du 02 novembre 2010, cinq grands quotidiens (The New York Times, The Guardian, Der Spiegel, Le Monde et El País) commencent à diffuser et à commenter des câbles de la diplomatie américaines mis en ligne par le site WikiLeaks[1]. Moins d'an an plus tard, dans la nuit du 1er au 2 septembre 2011, après la divulgation d'une première version complète de ces documents par un concurrent, ce même site publiera l'intégralité des 251287 télégrammes qui forment ce que l'on appelle le « CableGate ». Ces textes sont des communications entre les ambassades américaines du monde entier et Washington qui ont embarrassé et embarrasseront encore longtemps les diplomaties de la plupart des pays. Nous utiliserons ce corpus dans une démarche expérimentale qui se fixait comme premier objectif d'adapter l'implémentation de la méthode ALCESTE (Reinert, 1983, 1990) disponible dans IRAMUTEQ[2] (Ratinaud, 2009 ; Ratinaud & Déjean, 2009) à de « gros[3] » corpus. Après avoir présenté les modifications de l'algorithme de classification hiérarchique descendante nécessaires pour analyser un corpus de 238 116 128 occurrences, nous proposerons une analyse qui consiste à comparer une classification menée sur les 5002 formes pleines les plus fréquentes avec une classification sur les 5000 formes pleines suivantes.

2. Présentation et adaptation de la C.H.D.[4] de la méthode ALCESTE

Rappelons que la méthode ALCESTE a été proposé par Reinert (1983, 1990) et a d'abord été implémentée dans le logiciel ALCESTE[5].

2.1. La méthode ALCESTE

Les particularités de cette technique d'analyse lexicale sont les suivantes :

- Un découpage des unités du corpus (nommée u.c.i.[6]) en segments de texte (nommée u.c.e.[7]) : dans les analyses que nous proposerons, cette phase ne sera pas traitée. Les unités classifiées ne seront pas des segments de texte, mais les télégrammes dans leur intégralité.
- Une sélection des formes « pleines » : dans la méthode ALCESTE, l'analyse ne porte que sur les formes dites « pleine » (les verbes, les noms, les adverbes, les adjectifs)[8] qui sont opposées aux formes supplémentaires (ou mots outils : les prépositions, les pronoms, les adjectifs possessifs, certains verbes et adverbes fréquents...).
- Une lemmatisation : par défaut, les formes sont lemmatisées.

---

1 http://wikileaks.org/

2 IRAMUTEQ : Interface de R pour les Analyses Multidimensionnelles de TExtes et de Questionnaires, http://www.iramuteq.org .

3 La notion de gros corpus (et l'ensemble des qualificatifs que l'on peut associer à « corpus ») est dépendante du champ d'application. Nous restreignons ici ce champ aux analyses utilisant la méthode ALCESTE.

4 Classification Hiérarchique Descendante

5 http://www.image-zafar.com/

6 Pour unité de contexte initiale

7 Pour unité de contexte élémentaire

8 C'est une description rapide des formes pleines, certaines exceptions existent.

- Une classification hiérarchique descendante : la technique de classification est sûrement la plus grande originalité de cette méthode. L'algorithme décrit par Reinert (1983) repose sur une série de bi-partitions construite sur la base d'une analyse factorielle des correspondances menée sur un tableau binaire (absence/présence) qui croise les unités textuelles choisies avec les formes pleines sélectionnées. C'est l'adaptation de cet algorithme que nous décrirons. Rappelons que chaque bi-partition est réalisée en trois étapes :

  1. Une analyse factorielle des correspondances (A.F.C.) est menée sur le tableau, puis, pour toutes les partitions possibles le long du 1er facteur de l'AFC, l'inertie inter-classe est calculée. Une première coupure intervient pour la partition qui maximise l'inertie inter-classe.
  2. Chaque unité du tableau est permutée d'une classe à l'autre et l'inertie inter-classe est recalculée. Si celle-ci est supérieure à l'inertie inter-classe précédente, la permutation est conservée. Cette partie de l'algorithme boucle jusqu'à ce qu'aucune permutation n'augmente l'inertie inter-classe.
  3. Les formes spécifiques d'une classe (au sens du chi2) sont retirées de l'autre.

2.2. Adaptation de la classification hiérarchique descendante de la méthode ALCESTE

Des trois étapes nécessaires à chaque bi-partition, c'est la première qui impose les limites des tableaux analysés. L'analyse factorielle des correspondances nécessite en effet le passage par une décomposition en valeurs singulières qui est une opération particulièrement lourde en terme de calcul. Dans les premières versions d'IRAMUTEQ, nous utilisions une adaptation de la librairie anacor (De Leeuw et Mair, 2009), qui, comme la plupart des librairies de R proposant ce type d'analyse, utilise la fonction svd9 pour réaliser les décompositions en valeurs singulières. Par ailleurs, nous utilisions des matrices pleines, ce qui n'est pas très pertinent pour les tableaux lexicaux sur lesquels nous travaillons, qui sont principalement composés de 0. Cet ensemble conduisait à une très importante consommation de mémoire lors de la classification, ce qui limitait les analyses à des tableaux « modestes » (de l'ordre de 2000 formes pour 80000 unités sur une machine disposant de 8 Go de RAM). Nous avons donc commencé par utiliser des matrices creuses, ce qui, dans un premier temps, ne changeait rien au problème, puisque la fonction svd les transforme en matrices pleines avant l'analyse. Notre recherche d'un algorithme de décomposition en valeurs singulières efficace et qui prendrait en entrée des matrices creuses nous a orienté vers la bibliothèque SVDLIBC (Rohde, 2011). Elle reprend l'algorithme las2 (Berry, 1992) de la bibliothèque SVDPACK (Berry *et al.*, 2011). Cet algorithme est particulièrement adapté à nos objectifs : il est optimisé pour les matrices creuses au prix d'une perte de précision sur les valeurs singulières de faible rang. Ce défaut ne nous concerne pas puisque seule la plus grande valeur est retenue dans l'analyse.

Nous avons bien sûr dû réaliser un ensemble de modifications dans la façon de construire les matrices, mais cela relève plus du domaine de l'informatique que de la statistique textuelle10.

9 Elle même basée sur les routines de la librairie LAPACK (http://www.netlib.org/lapack).

10 A l'heure où nous écrivons ce document, toutes les étapes utilisées pour cette analyse ne sont pas encore présentes dans l'interface d'IRAMUTEQ. Toutefois, l'ensemble des procédures est disponible en ligne de commande et le code est accessible dans le dépôt subversion du logiciel (hhtp://www.netdig.org/svn-iramuteq/trunk)

3. Présentation du corpus

3.1. Dates, origines et classifications

Le corpus du CableGate se présente originellement sous la forme d'une base de données PostgreSQL de 1,7 Go. La base contient une table nommée « cable » composée de 9 champs : identifiant, date, référence, classification, origine, destination, en-tête et contenu. La figure 1 montre la fréquence d'apparition de chacune des dates présentes dans la base :

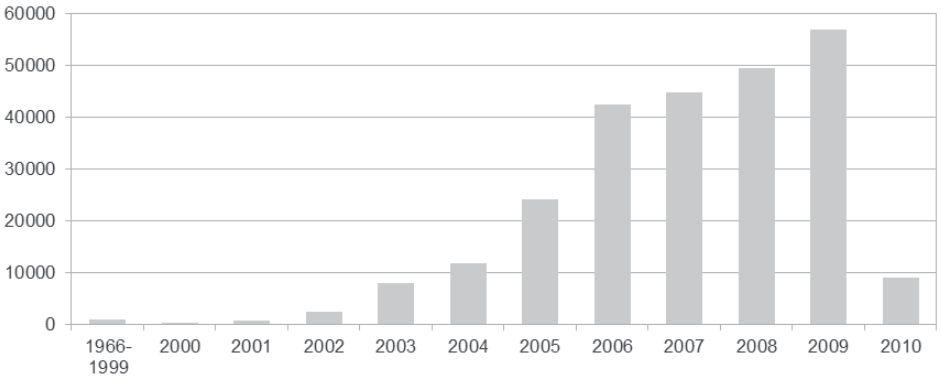

Illustration 1: Répartition des télégrammes par date - les années allant 1966 à 1999 ont été regroupées - N=251287

Bien que le plus ancien télégramme date de 1966, 99,6% des textes ont été écrits entre 2000 et 2010. Ces textes proviennent de 275 sources différentes. Une majorité (86 %) est issue des ambassades américaines du monde entier, 5% proviennent de Consulat et 3,1% proviennent du Secrétariat d'état américain, qui est le plus gros contributeur du corpus. La figure 2 montre la répartition des 20 plus gros contributeurs :

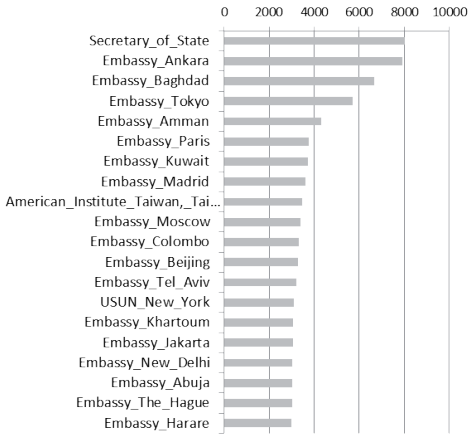

Illustration 2: Les 20 sources les plus fréquentes

6,2 % de ces documents sont classifiés secret[11], 40,5 % confidentiel[12]. Les 53,2% des documents restants ne sont pas classifiés.

3.2. Description du contenu

Le contenu des télégrammes se présente sous des formes relativement stables. Une série de tags commence la plupart des messages (classification, destinataire, sujet, etc.). Ces tags sont nombreux et nous avons décidé, pour cette analyse, de supprimer tous ceux que nous avons pu détecter. Le texte suivant est un exemple de début de télégramme, nous avons mis en gras les parties éliminées :

UNCLAS STATE 204472
E.O. 12958: N/A
TAGS: PTER
SUBJECT: ANNUAL TERRORISM REPORT
(THIS CABLE HAS BEEN CLEARED BY M/P (SEP).)
---
1. 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

Illustration 3 : un exemple de début de câble

Le corpus a été passé en minuscule et tous les caractères en dehors d'une liste restreinte[13] ont été éliminés. Ainsi nettoyé, le corpus se compose de 238 116 128 occurrences (Fmax = 15 668 471, « the »). Il est constitué de 624 202 formes différentes, dont 280 863 hapax (44 ,9% des formes, 0,11% des occurrences). La figure 3 présente le graphique rangs/fréquences (sur des échelles logarithmiques) du corpus :

11 SECRET ou SECRET NOFORN

12 CONFIDENTIAL ou CONFIDENTIAL NOFORN

13 a-z0-9àâäáéèêëìïòôöùûüçñ

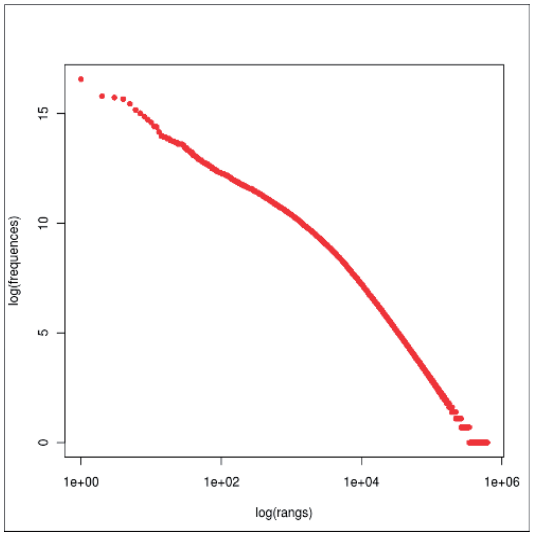

Illustration 4 : graphique rangs/fréquences des formes du corpus (échelles logarithmiques)

4. Hypothèse et Analyse

Parallèlement au développement de la méthode ALCESTE et du logiciel du même nom, Reinert a élaboré un modèle théorique autour de la notion de « monde lexicaux stabilisés » (Reinert, 2008). Ce modèle repose sur l'hypothèse que « dans l'activité langagière, les mots pleins constituent […] des traces possibles des contenus de nos activités. Ils ne sont pas les signifiants mais bien des traces possibles de ce contenu en acte. » (Reinert, 2008, p. 3). La stabilité des classifications sur les formes pleines d'un corpus lorsque l'on fait varier la taille des unités et les similarités constatées par Reinert entre certaines classifications sur des corpus différents participent à valider cette hypothèse.

De façon à vérifier que cette stabilité est également présente dans différentes fenêtres de fréquence des formes pleines, nous avons procédé à deux classifications des télégrammes du CableGate en retenant une fois les 5002[14] formes pleines les plus fréquentes et une fois les 5000 formes pleines suivantes. Le tableau 1 résume les deux analyses :

	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

Tableau 1 : résumé des analyses

14 Un changement de fréquence intervenant au rang 5003, nous avons préféré conserver les deux formes supplémentaires plutôt que de sélectionner arbitrairement des formes dans le rang précédant.

Nous avons bien sûr conservé le même paramétrage pour ces deux analyses en demandant 80 classes terminales et en ne conservant que les classes regroupant au moins 500 télégrammes.

5. Résultats

La première classification a conduit à 62 classes terminales et la seconde à 69. Les dendrogrammes suivants résument les deux classifications :

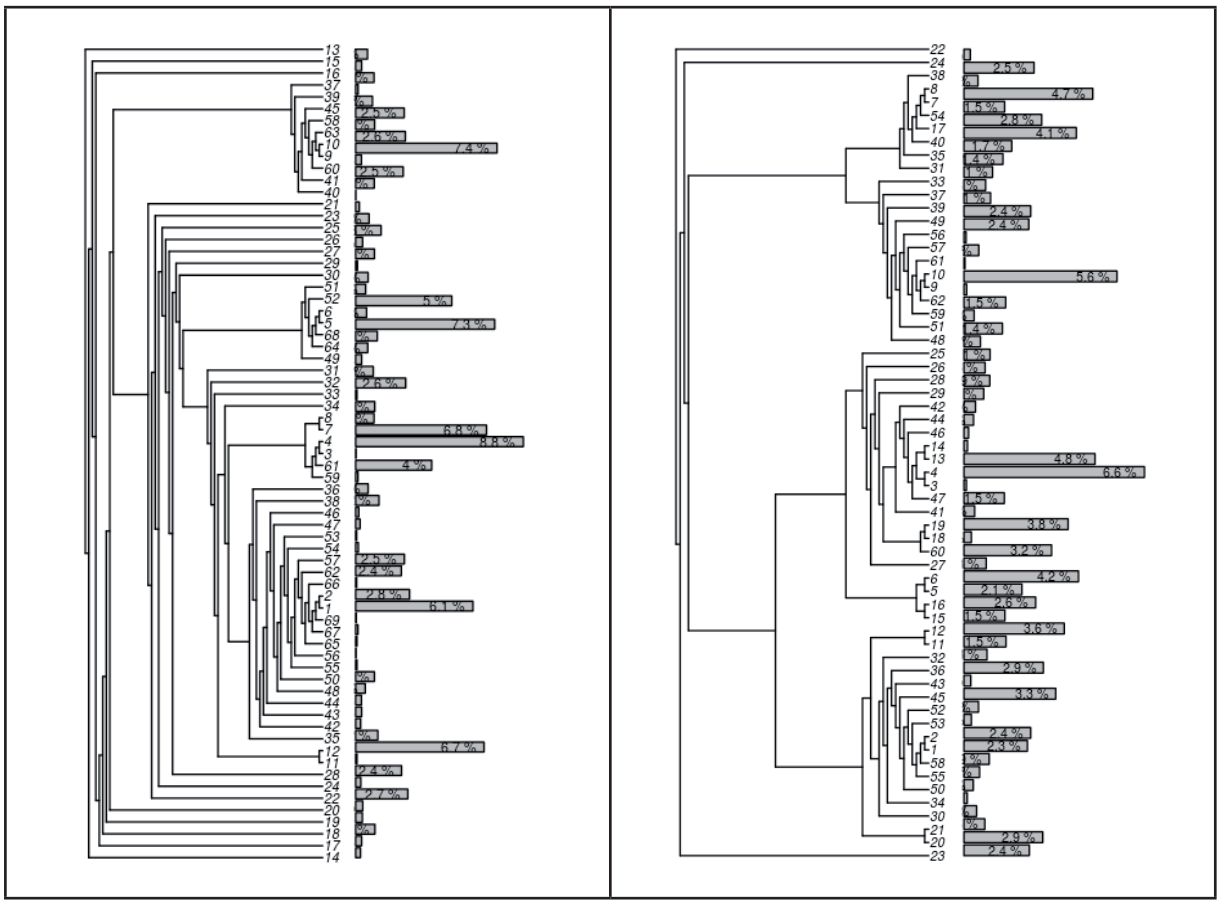

Illustration 5 : à gauche, dendrogramme de la première classification, à droite, dendrogramme de la seconde. Les barres indiquent la taille relative des classes.

Ces premiers résultats nous indiquent que le paramétrage à 80 classes terminales n'était pas pertinent, les grandes classes apparaissant ici peuvent certainement être encore dissociées. De façon à pouvoir comparer ces classifications, nous avons construit le tableau de contingence qui les croise. Les graphiques suivants rendent compte de la dépendance entre ces deux classifications à partir de la comparaison entre les effectifs théoriques et les effectifs observés d'une part, et à partir de l'étude des résidus standardisés liés au chi2 de ce tableau d'autre part:

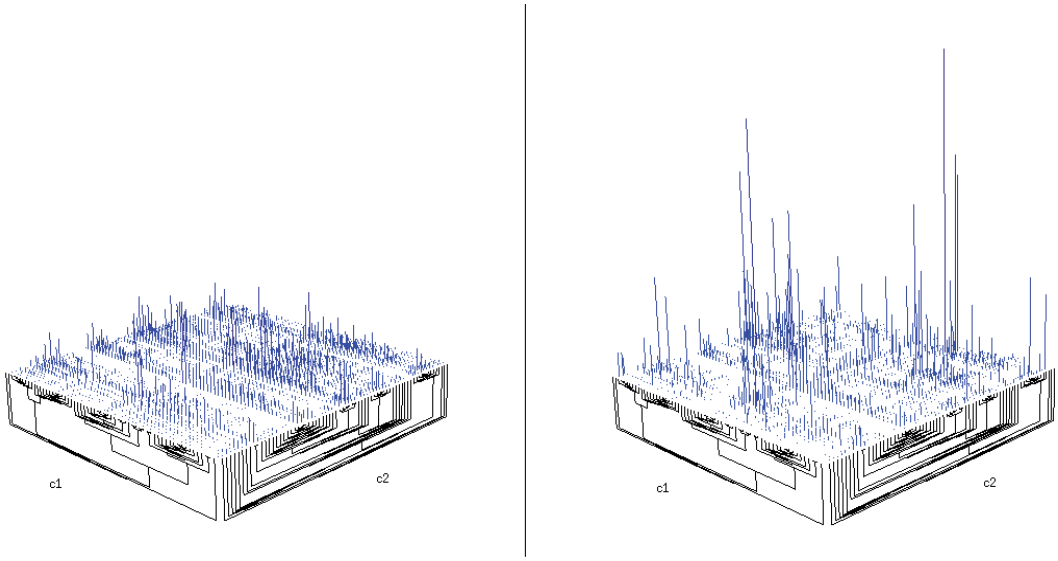

Illustration 6 : à gauche, les effectifs théoriques du tableau de contingence (en % du total), à droite, les effectifs observés (en % du total). Les classifications 1 et 2 apparaissent respectivement à gauche et à droite sur chaque graphique.

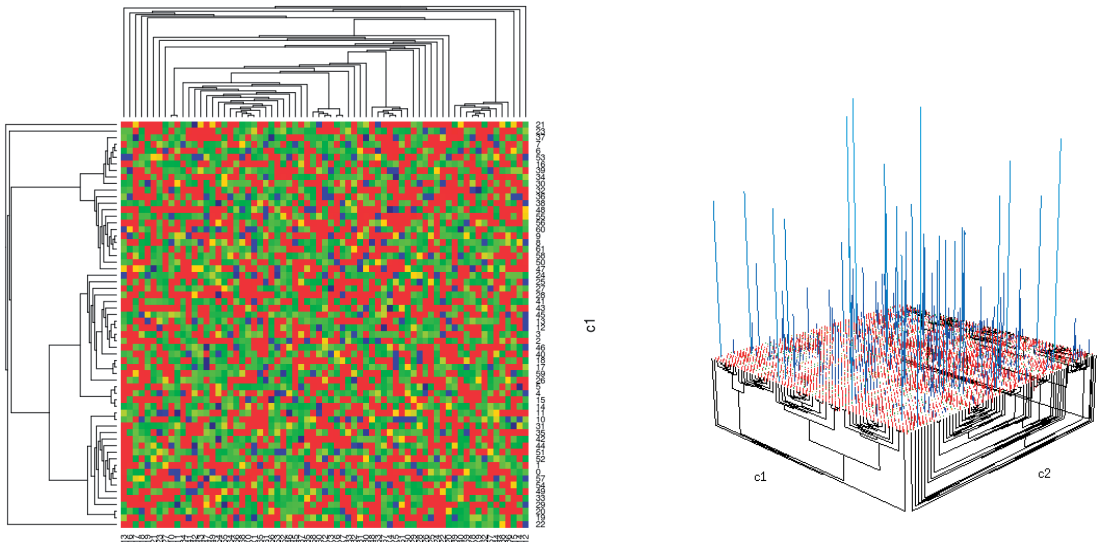

Illustration 7 : à gauche, résidus standardisés du chi2 mené sur le tableau de contingence (les points bleus représentent les contributions significativement positives (>1,96), les niveaux de vert représentent les contributions non-significatives, les niveaux de rouge représentent les contributions significativement négatives (< -1,96) ; à droite, la même représentation en trois dimensions.

Pour contrôler que ces classifications nous proposent bien une organisation globalement commune du corpus, nous avons construit les tableaux de contingence qui croisent chacune des deux classifications avec les sources des télégrammes. Pour ce traitement, nous n'avons conservé que les sources apparaissant au moins 500 fois. Les matrices des distances euclidiennes entre les sources ont été calculées pour chacun des deux tableaux de contingences obtenus. Nous

avons ensuite utilisé la librairie *igraph* (Csardi et Nepusz, 2006) de R pour tracer les arbres minimum de ces matrices en repérant les sources en fonction de leur situation géographique.

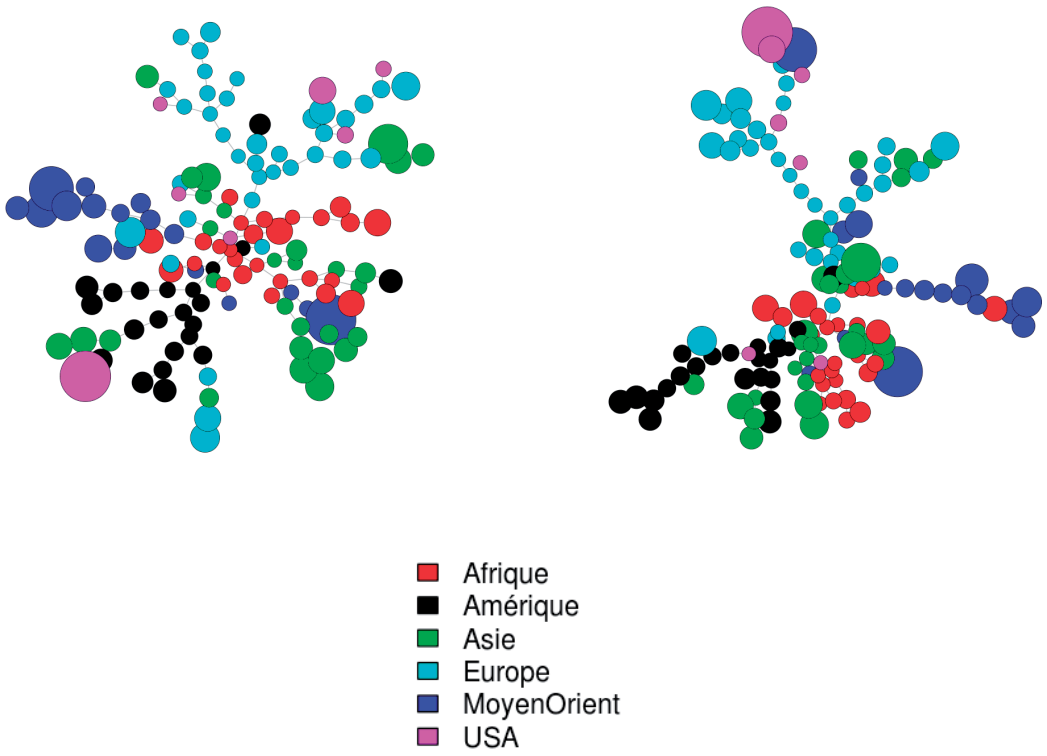

Illustration 8 : à gauche, arbre de la première classification ; à droite, arbre de la seconde classification

Bien qu'ils ne soient pas superposables, ces deux arbres nous montrent que les deux classifications ont abouti à un regroupement géographique des textes.

6. Conclusion

Les résultats que nous venons de présenter permettent, selon nous, de maintenir l'hypothèse des « mondes lexicaux stabilisés ». Les deux classifications que nous comparons proviennent de matrices très différentes, même si elles sont issues du même corpus. La seconde matrice est plus de dix fois plus vide que la première. Pourtant, elles mènent à des organisations des documents qui sont clairement dépendantes. Ces résultats attestent également de la pertinence de la méthode ALCESTE sur ce type de corpus.

Par ailleurs, les adaptations réalisées sur l'algorithme de classification rendent possible l'utilisation de la méthode ALCESTE classique (avec la double classification sur uc) sur des corpus de plusieurs dizaines de millions d'occurrences tout en travaillant sur un nombre important de formes pleines. La procédure que nous avons suivie pourrait également permettre d'améliorer

la recherche de stabilité dans l'analyse des gros corpus en utilisant les classifications sur les formes moins fréquentes pour préciser le contour des classes obtenues sur les classifications des formes fréquentes.

Références

Berry, M. (1992). Large Scale Singular Value Computations. *International Journal of Supercomputer Applications*. 6. (1). 13-49.

Berry, M. Do, T. O'Brien, G. Krishna, V. and Varadhan, S. (2011). SVDPACK. http://www.netlib.org/svdpack.

Csardi, G. et Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*. Complex Systems. (1695). http://igraph.sf.net.

De Leeuw, J. et Mair, P. (2009). Simple and Canonical Correspondence Analysis Using the R Package anacor. *Journal of Statistical Software*. 31. (5). 1-18.

Ratinaud, P. (2009). IRAMUTEQ : Interface de R pour les Analyses Multidimensionnelles de TExtes et de Questionnaires. http://www.iramuteq.org.

Ratinaud, P. and Déjean, S. (2009). IRaMuTeQ : implémentation de la méthode ALCESTE d'analyse de texte dans un logiciel libre. *Modélisation Appliquée aux Sciences Humaines et Sociales (MASHS2009)*. Toulouse - Le Mirail.

Reinert, M. (1983). Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. *Les cahiers de l'analyse des données*, VIII, (2), 187-198.

Reinert, M. (1990). ALCESTE : Une méthodologie d'analyse des données textuelles et une application : Aurélia de Gérard de Nerval. *Bulletin de méthodologie sociologique.* 26. 24-54.

Reinert, M. (2008). Mondes lexicaux stabilisés et analyse statistique de discours. *9èmes Journées internationales d'Analyse statistique des Données Textuelles*.

Rohde, D. (2011). SVDLIBC. http://tedlab.mit.edu/~dr/SVDLIBC.