L'analyse de similitude appliquée aux corpus textuels : les primaires socialistes pour l'élection présidentielle française (septembre-octobre 2011)

Pascal Marchand[1], Pierre Ratinaud[2]

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

Abstract

The analysis of similarity (ADS) is a technique based on graph theory, conventionally used to describe the social representations using survey questionnaires. We integrated the analysis of similarity of a textual matrix to the software *Iramuteq* (P. Ratinaud).
The results can show, in a single graph, both common elements (usually absent of specific research, analysis of lexical correspondences or classifications), and the varying elements of variables related to the corpus.
The corpus analyzed here as an example is the discussion of socialist primary for the presidential election of 2012.

Résumé

L'analyse de similitude (ADS) est une technique, reposant sur la théorie des graphes, classiquement utilisée pour décrire des représentations sociales, sur la base de questionnaires d'enquête. Nous avons intégré au logiciel *Iramuteq* (P. Ratinaud) l'analyse de similitude d'une matrice textuelle.
Les analyses permettent de montrer, en un seul graphique, à la fois les éléments communs (généralement absents des recherches de spécificités, analyses des correspondances ou classifications lexicales), mais également les éléments différentiés en fonction de variables liées au corpus.
Le corpus analysé ici à titre d'exemple est constitué des débats des primaires socialistes pour l'élection présidentielle de 2012.

Mots-clés : Analyse de similitude ; Discours politique ; *Iramuteq*.

1. Introduction

On a parfois l'impression, après une analyse lexicométrique, que le monde lexical est bien partagé et que nos variables délimitent des territoires lexicaux bien tracés. Cette impression vient surtout du fait que le tableau lexical est, le plus souvent, partitionné selon des hypothèses plus ou moins clairement explicitées. La recherche de spécificités lexicales, l'analyse des correspondances, voire même la CDH (avec l'attention apportée aux éléments illustratifs), accentuent alors les différences et minimisent les ressemblances entre les colonnes du tableau. Nous montrerons que l'analyse de similitude (ADS) permet de représenter graphiquement la

structure d'un corpus, en distinguant également les parties communes et les spécificités des variables codées.

Nous proposerons de l'illustrer sur le corpus des « primaires socialistes », qui ont posé d'intéressantes questions d'analyse : les candidats devaient se différencier les uns des autres tout en préservant l'unité du parti qu'ils seraient amenés à défendre ensemble.

Les questions que nous pouvons nous poser sont les suivantes :

Quels sont les mots, les phrases et les relations lexicales qui peuvent caractériser chacun des débatteurs ?

Les trois débats ont-ils été équivalents ?

- Chacun des candidats a-t-il été constant dans les trois débats ou peut-on observer des changements ou des évolutions ?

2. Principes généraux de l'analyse de similitude (ADS)

L'ADS est une technique, reposant sur la théorie des graphes, classiquement utilisée pour décrire des représentations sociales, sur la base de questionnaires d'enquête (Flament, 1962 ; Flament, 1981 ; Vergès & Bouriche, 2001).

L'objectif de l'ADS est d'étudier la proximité et les relations entre les éléments d'un ensemble, sous forme d'arbres maximum : le nombre de liens entre deux items évoluant « comme le carré du nombre de sommets » (Flament & Rouquette, 2003 : 88), l'ADS cherche à réduire le nombre de ces liens pour aboutir à « un graphe connexe et sans cycle » (Degenne & Vergès, 1973 : 473).

Les bases théoriques de cette technique sont résumées dans un exemple développé par Flament & Rouquette (2003, o.c.). Dans la figure suivante, le graphique de gauche montre tous les liens possibles entre chaque item.

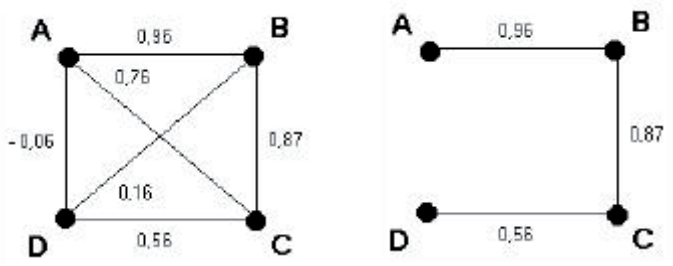

Figure 1 : Exemple de calcul de l'arbre maximum (ADS)

A partir de ces liens, on va chercher à représenter un arbre sans cycle, dit « arbre maximum », créé par les arêtes les plus fortes du graphique. C'est l'arbre le plus simple que l'on peut obtenir, mais c'est aussi le plus lourd (en termes d'information). A partir de l'exemple précédent : on considère la « clique » ABCA et on élimine le lien le plus faible (entre A et C). On considère ensuite la « clique » BCDB et on élimine le lien le plus faible (entre B et D). Et ainsi de suite pour toutes les « cliques » possibles. Le graphique de droite sur la figure 1 représente l'arbre maximum, sans cycle, du graphique de similitude de gauche.

L'analyse de similitude d'une matrice textuelle a été intégrée au logiciel *IRaMuTeQ* (développé par Pierre Ratinaud) et permet de décrire des classes lexicales, des profils de spécificités ou même des corpus entiers.

3. Le corpus

Les primaires socialistes pour l'élection présidentielle française de 2012 se sont déroulées en deux tours. Au premier tour s'affrontaient six candidats : Martine Aubry, Jean-Michel Baylet, François Hollande, Arnaud Montebourg, Ségolène Royal et Manuel Valls. Trois débats ont d'abord été organisés et diffusés sur des chaînes de radio et de télévision :

Jeudi 15 septembre 2011 (2h50 sur France 2, Le Monde)

Mercredi 28 septembre 2011 (2h30 sur i-Télé, Europe 1, Le Parisien, LCP-Assemblée Nationale)

Mercredi 05 octobre 2011 (2h20 sur BFM, RMC, Le Point, Public Sénat).

La retranscription des trois débats permet de dresser les tableaux suivants :

nombre d'uci :	295 (tours de parole)
nombre d'occurrences :	71913
nombre de formes :	5265
moyenne d'occurrences par forme :	18.96
nombre d'hapax :	1472 (2.05% des occurrences - 27.96% des formes)
moyenne d'occurrences par uci :	243.77

Tableau 1 : Caractéristiques générales (corpus lemmatisé)

Le vote du 9 octobre 2011 a permis de dégager les deux finalistes : Martine Aubry et François Hollande se sont affrontés le 16 octobre 2011.

Partie	occurrences	formes	hapax	Fréq. Max	Forme
Aubry1	4990	817	408	165	être
Aubry2	4956	830	412	187	être
Aubry3	4374	774	382	147	avoir
Aubry4	10543	1189	510	402	être
Baylet1	4160	750	367	198	être
Baylet2	4036	781	407	183	être
Baylet3	3368	742	402	149	être
Hollande1	4519	807	382	180	être
Hollande2	4096	772	373	179	être
Hollande3	3518	730	372	164	être
Hollande4	9352	1200	531	439	être
Montebourg1	3920	844	449	161	de
Montebourg2	3821	910	519	161	de

Montebourg3	3592	870	505	154	de
Royal1	4093	835	436	159	la
Royal2	4048	856	453	162	de
Royal3	3034	687	377	108	être
Valls1	4483	835	416	182	de
Valls2	4146	826	427	158	être
Valls3	3640	749	404	142	de

Tableau 2 : Principales caractéristiques lexicométriques (corpus lemmatisé)

Les premiers traitements du corpus suivent rigoureusement la méthode ALCESTE (Reinert, 1983, 1990) : reconnaissance et lemmatisation des formes, découpage en unités de contextes élémentaires (UCE), et création de la matrice habituellement soumise à une CDH « simple sur UCE ».

4. Analyses lexicométriques classiques

Classiquement, nous avons soumis le corpus à des analyses factorielles et classificatoires. Nous analysons ici le corpus partitionné selon les lignes du Tableau 2, c'est-à-dire en croisant les trois premiers débats avec les six locuteurs. Le tableau lexical comprend donc 18 colonnes et 1086 lignes (sélectionnées sur critère de fréquence).

4.1. Analyse des correspondances

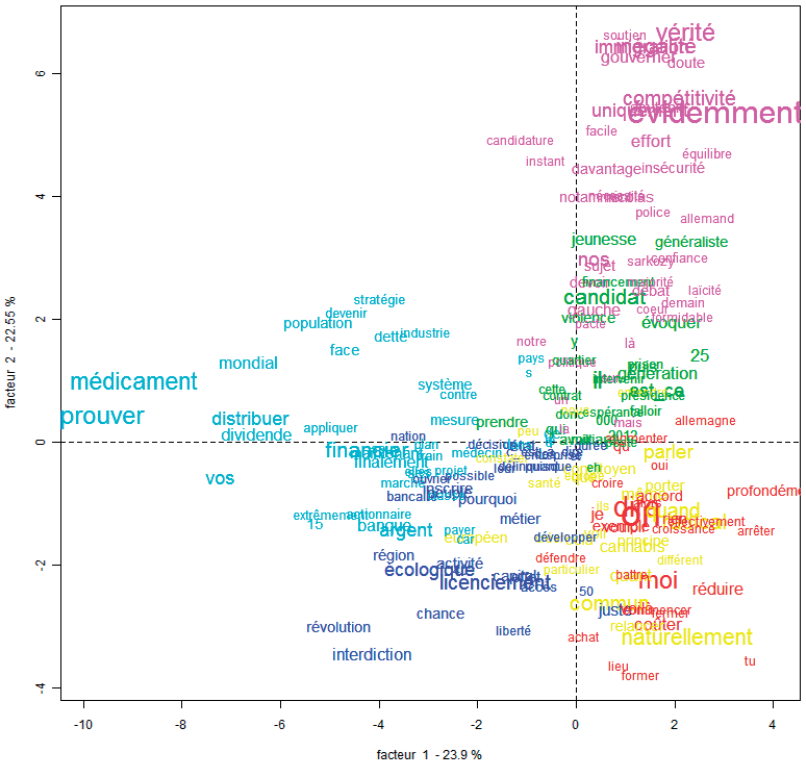

Figure 2 : AFC des formes lexicales pour les trois premiers débats

Le premier facteur oppose les formes : *inventer, imaginer, histoire, nouvelle, unir, vie, banque…*, aux formes : *priorité, sortir, falloir, recherche, justice, changer…*

Le deuxième facteur oppose les formes : *devoir, soutien, Nicolas Sarkozy, oublier, échec, doute, droite, étranger, drogue, effort, vérité, candidature…*, aux formes : *aider, an, je, Manuel, tenir, supprimer, centrale, nucléaire, absolument, bien sûr, moi, droit, ressource, smic, prix…*

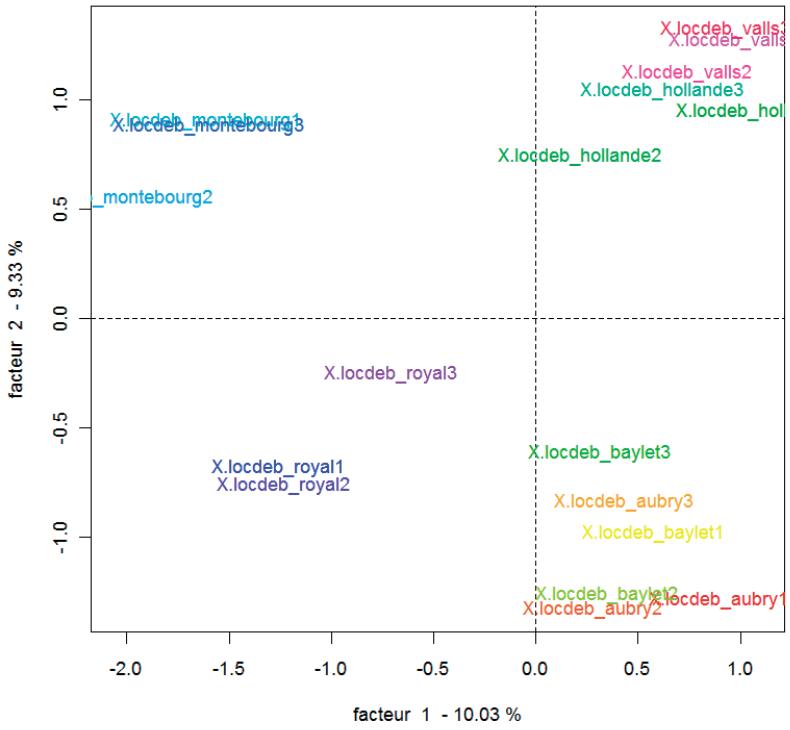

Figure 3 : AFC des six locuteurs dans les trois premiers débats (n_j=18)

On observe tout d'abord que les locuteurs restent sur des lexiques constants au long des trois débats. Les débats n'ont donc pas structuré le corpus aussi fortement que les débatteurs. On montre ensuite qu'Arnaud Montebourg s'oppose à quasiment tous les autres (1er facteur) et que Manuel Valls et François Hollande s'opposent à Martine Aubry, Jean-Michel Baylet et Ségolène Royal (2ème facteur). On recherche alors les spécificités des six débatteurs.

4.2. Spécificités (sur les trois premiers débats)

4.2.1. Martine Aubry

Sp+ : on, dire, moi, coûter, je, réduire, voilà, accord, exemple, profondément, qu, rien, vouloir, effectivement, oui, tu, achat, allemagne, alors, arrêter, augmenter, battre, commencer, croire, croissance, défendre, fermer, former, lieu, supprimer, sûr, taxe, ça…

Sp- : dans, de, y, un, une, nos, solution, être, dette, candidat, devoir, difficile, elles, il, entreprise, plus, évoquer, situation….

Uce caractéristiques : Et **moi**, **je** l'**ai dit**, **je** serai la **présidente** du redressement de la France, redressement économique… **mais non**, **mais je** vous **ai dit** que **je** ne répondrai pas…

Mais, **je le dis**, il faudra, **moi je** l'ai **dit** comme une priorité, **je** vais vous **dire** que **ça coûte**, il faut environ, il faut réorganiser la police, Manuel a totalement raison, ils font des tâches qui n'ont **rien** à voir avec ce **qu'on** leur demande, c'est-à-**dire** maintenir la sécurité pour les habitants.

4.2.2. Jean-Michel Baylet

Sp+ : naturellement, commun, radical, quand, parler, même, quant, cannabis, cela, porter, concitoyen, que, relancer, voir, européen, principe, santé, je, ils, différent, entendre, europe, nous, regarder, républicain, trop, constater, créer, particulier, peu, être…

Sp- : des, pays, qui, faire, exemple, augmenter, enfant, payer, pour, évidemment, financier, cette, notamment…

Uce caractéristiques : **Moi je** voudrais **quand même** en venir à l'**Europe** parce **que je vois que**… si, **je vois que** le temps tourne.

Je **ne suis** pas, encore une fois **je ne suis** pas pour les mesures coercitives.

4.2.3. François Hollande

Sp+ : il, candidat, est_ce, évoquer, génération, 25, jeunesse, y, puis, qui, prendre, violence, généraliste, avoir, milliard, prison, 0, 2012, financement, falloir, une, cette, donc, quartier, rapport, secteur, intervenir, poste, contrat, là, espérance, présidence, quinquennat, élection, eh, senior…

Sp- : je, cela, vouloir, moi, france, de, la, banque, contre, et, raison, européen, que, relancer, juste, chose, exemple, accord…

Uce caractéristiques : **Parce** que **cette jeunesse**, **qui a** des talents, mais **qui a aussi** des retards, **qui a aussi** des discriminations, **qui a aussi** des violences, **eh bien il** faut la faire espérer.

Et **puis, il y a ce** que j'ai appelé le **contrat** de **génération qui** servira **aussi**.

4.2.4. Arnaud Montebourg

Sp+ : approuver, médicament, financier, argent, vos, distribuer, mondial, banque, dividende, finalement, face, population, de, 15, système, ses, dette, maintenant, mesure, contre, européen, appliquer, le, plan, marché, sous, actionnaire, médecin, train, dans, devenir, s, stratégie…

Sp- : moi, falloir, quand, ça, juste, mais, dire, on, même, qu, je, vouloir, évidemment, français, parler, priorité, justice, sur, retraite, confiance, parce, jeune, là…

Uce caractéristiques : **Nous** n'avons, **dans notre pays**, **nous** n'avons pas **de** pénurie **de médecins**.

Il n'y aurait, je vous **le** dis, si **nous** avions mis si les **dirigeants** qui, aujourd'hui, **nous ont** précipité **dans** cette crise - **car** cette crise est la **conséquence de** l'incompétence **de** nos

dirigeants -, si **nous** avions mis en place ces **mesures**, **nous** n'aurions pas aujourd'hui **de** crise **de** la zone euro.

4.2.5. Ségolène Royal

Sp+ : licenciement, écologique, juste, interdiction, inscrire, capital, activité, pourquoi, état, peuple, métier, effet, région, chance, révolution, c_est_à_dire, 40, accès, ouvrier, bancaire, des, retraite, 50, durée, et, rentrer, liberté, équitable, développer, remettre, entreprise, décision, délinquant, travail, puisque, nation, possible, écouter, leur, banque…

Sp- : nous, il, falloir, nos, quand, on, parler, gauche, s, mais, évidemment, hôpital, sarkozy, nicolas, devoir, manière, aujourd, hui, président, avoir, y, notre, médecin…

Uce caractéristiques : Quelle **décision** ? celle que l'on a déjà entendue avant la crise de 2008, c'est l'**interdiction des** banques **et** ça ça doit être une **décision** européenne **et** même internationale, l'**interdiction des** banques de spéculer sur la dette **des états** c'est-à-dire sur la misère **des peuples**.

Et demain je **veux** que toutes les régions **puissent** entrer au **capital des entreprises stratégiques**.

4.2.6. Manuel Valls

Sp+ : évidemment, vérité, inégalité, compétitivité, nos, immigration, uniquement, effort, gouverner, gauche, nicolas, devoir, soutenir, débat, sujet, insécurité, notamment, doute, davantage, mais, sarkozy, notre, là, sur, confiance, demain, police, équilibre, soutien, facile, un…

Sp- : je, vous, payer, quand, on, avoir, milliard, finalement, voir, sûr, argent, moi, aider, voilà, an, prendre, déjà…

Uce caractéristiques : Si **nous** pensons **un** seul **instant** que **demain il y** a une **majorité** et que **nous** pourrons tout faire **uniquement sur notre** programme et que **nous** n'aurons pas entendu **la** voix des français, alors je ne donne pas le cher du temps que **nous** passerons dans **un** contrat de **confiance** avec les Français.

Donc, **il y** a d'abord **un** échec majeur de **Nicolas Sarkozy** et de **la** droite **sur** ce **sujet-là**.

4.3. ADS des sous-corpus

L'analyse de similitude est appliquée à chacun des sous-corpus définis par les locuteurs, après découpage en UCE, et création de la matrice *formes* * *UCE*. Si l'on retrouve, sur chacun des arbres de similitude ci-après, les spécificités définies ci-dessous pour chaque locuteur, des formes communes apparaissent également et avec un critère de centralité (*France, français, aller*…)

Figure 4 : *ADS du sous-corpus Aubry*

Figure 5 : *ADS du sous-corpus Baylet*

Figure 6 : *ADS du sous-corpus Hollande*

Figure 7 : *ADS du sous-corpus Montebourg*

Figure 8 : *ADS du sous-corpus Royal*

Figure 9 : *ADS du sous-corpus Valls*

Les méthodes classiques permettent de spécifier les colonnes du tableau lexical et rendent très bien compte des différences qui s'instaurent entre les six locuteurs, qui structurent davantage le lexique que les trois situations de débat. Quant à l'ADS de chacun des sous-corpus de locuteurs, elle fournit une représentation graphique indépendante des calculs factoriels et des spécificités, mais qui les confirme néanmoins, tout en restituant les usages communs.

5. ADS du corpus global

Après segmentation, reconnaissance et lemmatisation des formes, puis partition en UCE, la matrice du corpus global peut être représentée de diverses façons (arbres linéaires ou circulaires ; taille des formes proportionnelle à la fréquence ou à la liaison statistique…). On représente ici l'arbre des liaisons lexicales du corpus (calcul de cooccurrence et algorithme de Fruchterman-Reingold).

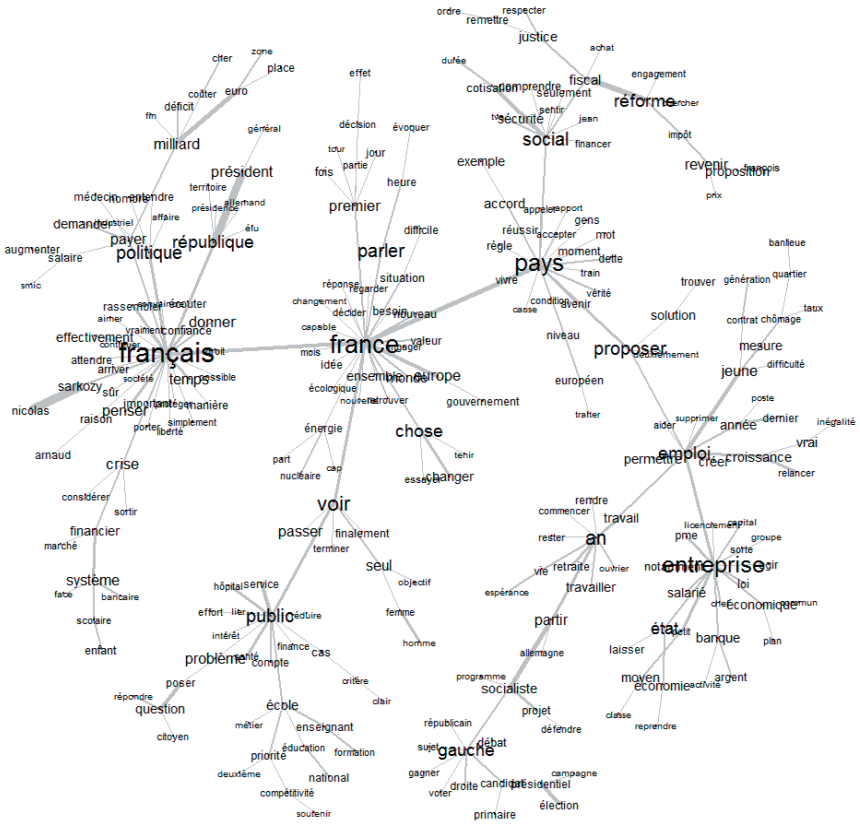

Figure 10 : Exemple d'ADS appliquée au corpus «primaires PS»

A partir de cette représentation, on peut mettre en évidence les spécificités des locuteurs (figure 11 : il est possible de colorier les formes lexicales en fonction des locuteurs et la taille de police est proportionnelle à la spécificité) [1].

1 On se reportera à la version électronique pour visualiser les couleurs.

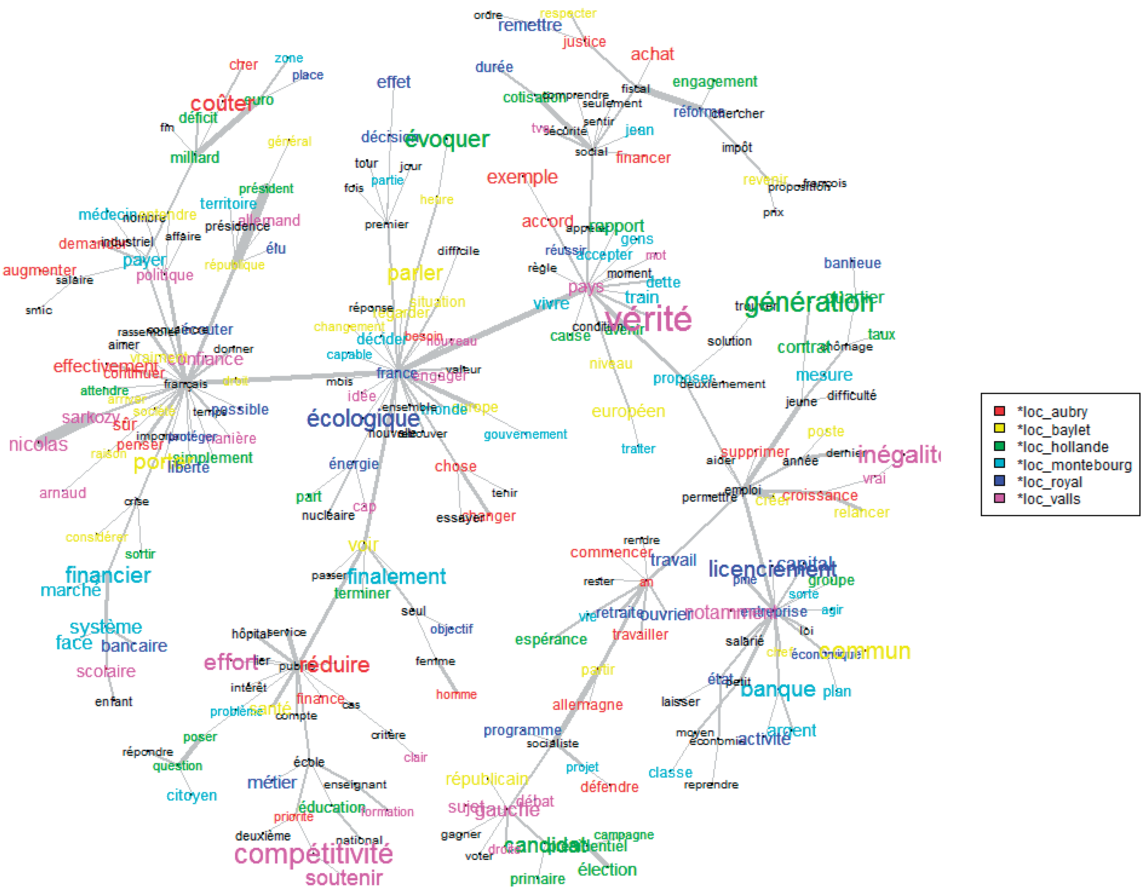

Figure 11 : Exemple d'ADS appliquée au corpus «primaires PS» (avec spésificités)

6. Conclusions

L'ADS permet de voir que les différences entre les modalités de variables (le plus souvent en colonnes du tableau lexical) ne sont pas aussi absolues que des méthodes plus « classiques » inciteraient parfois à le penser. L'intérêt de l'ADS est de rétablir la partie commune, en représentant les relations entre les formes lexicales dans un corpus non partitionné par des variables exogènes, mais uniquement en unités de contexte.

On repère alors les diverses thématiques qui structurent le corpus et on observe que les spécificités de chaque locuteur ne sont pas toutes regroupées dans le graphe des similitudes du corpus global. On peut donc en conclure que les locuteurs convergeaient sur les thématiques abordées, mais différaient dans la façon de les aborder.

7. Epilogue

A l'issue du premier débat, et sur la base de la Figure 3, on pouvait imaginer :

Qu'Arnaud Montebourg éprouverait des difficultés à prendre position pour l'un des deux finalistes ;

Que Manuel Valls se rallierait à François Hollande ;

Que Jean-Michel Baylet et Ségolène Royal se rallieraient à Martine Aubry.

Seules les deux premières hypothèses se sont vérifiées, indiquant que la proximité lexicale ne saurait expliquer toutes les stratégies électorales.

Si l'on introduit, dans le corpus, le débat du deuxième tour, on observe que les deux finalistes sont restés dans leur vocabulaire et n'ont fait aucun mouvement de rapprochement vers les lexiques des candidats éliminés.

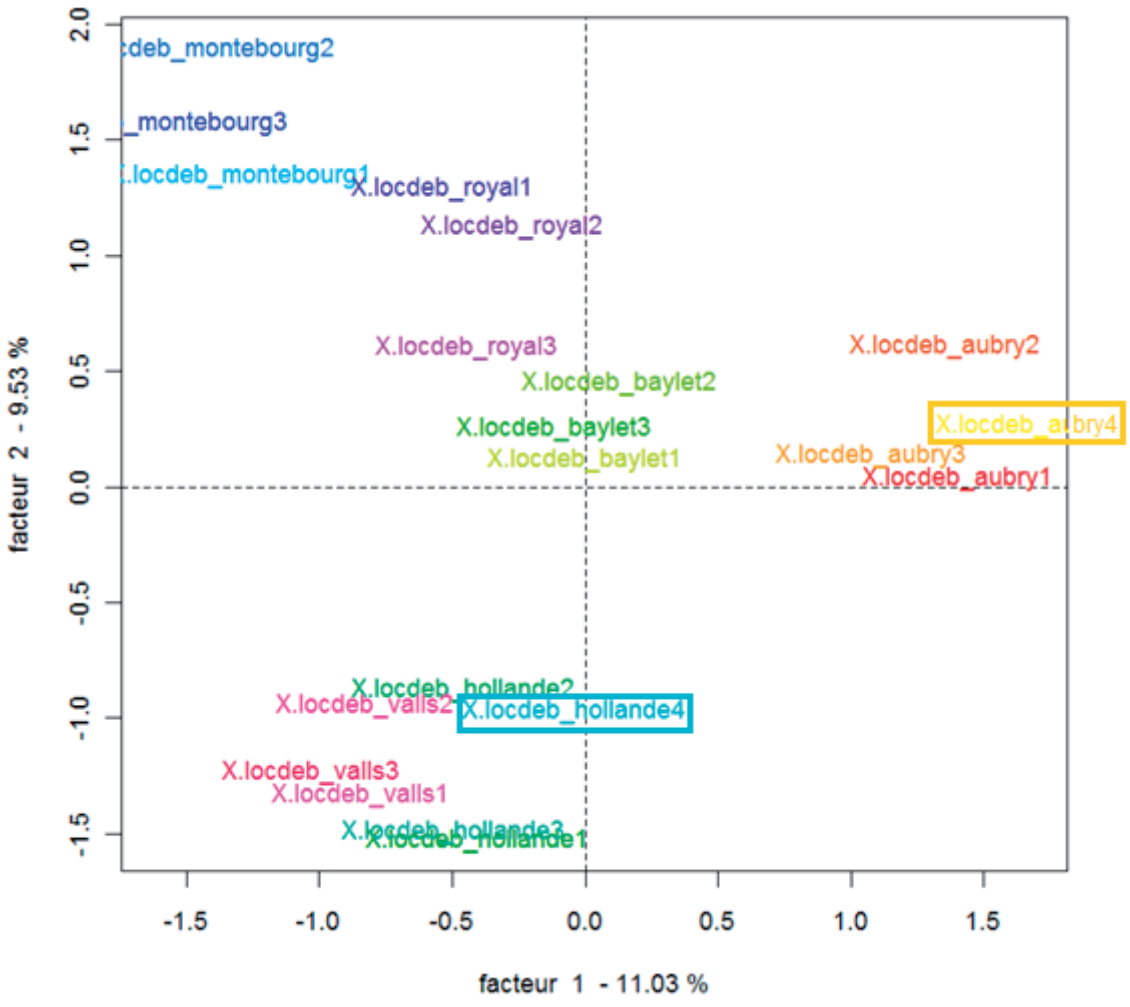

Figure 12 : AFC des six locuteurs dans les quatre débats (n_j=20)

Références

Degenne, A., Vergès, P. (1973). Introduction à l'analyse de similitude. *Revue française de sociologie, 14* (4), 471-511.

Flament, C. (1962). L'analyse de similitude. *Cahiers du centre de recherche opérationnelle*, 4, 63-97.

Flament, C. (1981). L'Analyse de Similitude, une Technique pour les Recherches sur les Représentations Sociales. *Cahiers de Psychologie Cognitive*, 1, 375- 395.

Flament, C., Rouquette, M.L. (2003). *Anatomie des idées ordinaires* : *comment étudier les représentations sociales*. Paris : Armand Colin.

Ratinaud, P. (2003). *Les professeurs et Internet : Contribution à la modélisation des pensées sociale et professionnelle par l'étude de la représentation professionnelle d'Internet d'enseignants du secondaire.* Thèse de l'Université de Toulouse 2 - Le Mirail, décembre 2003.

Ratinaud, P. (2009). *Iramuteq : Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires*. www.iramuteq.org

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.

Vergès, P. & Bouriche, B. (2001). L'analyse des données par les graphes de similitude. *Sciences Humaines* (en ligne : http://www.scienceshumaines.com/textesInedits/Bouriche.pdf).