Arbre et co-occurrences
Nouvel outil logométrique sur le net.
Application au discours de François Hollande

Laurent Vanni[1], Xuan Luong[2], Damon Mayaffre[3]

[1] UMR 7320, Bases, Corpus, Langage - CNRS - laurent.vanni@unice.fr
[2] UMR 7320, Bases, Corpus, Langage - CNRS - xuan.luong@unice.fr
[3] UMR 7320, Bases, Corpus, Langage - CNRS - damon.mayaffre@unice.fr

Abstract

Starting with 9 matrices representing presidential use of shared vocabulary (co-occurrences analysis), we produce a dissimilarities matrix of distances between presidents of the fifth republic. We give a tree analysis of this matrix and we improve the performances due a new topological approach. We present a new software that provides a graphical visualization with this tree analysis. Finally we suggest a first socio-linguistic description of the François Hollande's discourse which, according to political reporters, remains still difficult to define in the presidential history.

Résumé

A partir de 9 matrices mots x mots ou matrices co-occurrentielles (une par mandat présidentiel depuis 1958), nous produisons une matrice de dissimilarité consignant les distances entre les présidents de la Vème République. On donne une représentation arborée de cette matrice et on améliore ici les performances de la représentation grâce à une nouvelle approche topologique. On présente alors l'outil logiciel qui permet de tracer le graphe et au terme du parcours méthodologique, on produit une première description socio-linguistique du discours de François Hollande qui reste selon les observateurs politiques encore difficile à définir dans l'histoire présidentielle française.

Mots-clés : arborée, co-occurrence, discours, politique, représentation, distances, hyperbase

1. Introduction

La classification des textes fut (Mosteller et Wallace 1964) et reste (Brunet 2014 - sous presse) un objectif majeur de l'ADT. En l'occurrence : comment le discours de François Hollande, encore énigmatique, se situe-t-il par rapport aux discours élyséens antérieurs (de Gaulle, Pompidou, Giscard, Mitterrand, Chirac, Sarkozy) ?

Pour des raisons sémantiques développées ci-dessous, nous souhaitons calculer et représenter la distance entre textes (après de nombreuses études de Guiraud, Muller, Evrad, etc., cf. Corpus 2002) non pas sur la base de leurs occurrences lexicales, mais sur la base de leurs co-occurrences, c'est-à-dire de paires de mots constituées dont on mesure la distribution dans les différents textes du corpus.

A partir de 9 matrices mots x mots ou matrices co-occurrentielles (une par mandat présidentiel depuis 1958), classiques en ADT, nous produisons une matrice de dissimilarité consignant les distances entre les présidents de la Vème République sur la base de leurs

discours grand public (allocutions, interviews, discours de tribune majeurs)[1]. A la suite de nos premiers travaux (Luong, 1988; Bathélemy et Luong, 1998), on donne une représentation arborée de cette matrice et on améliore ici les performances de la représentation grâce à une nouvelle approche topologique. On présente alors l'outil logiciel qui permet de tracer le graphe dans le cadre d'une plateforme logométrique disponible sur le Web et qui vise, à terme, aussi bien l'édition que le traitement statistique des textes : Hyperbase Web Edition (http://hyperbase.unice.fr/). Enfin, au terme du parcours méthodologique, on produit une première description sociolinguistique du discours de François Hollande qui reste selon les observateurs politiques encore difficile à définir dans l'histoire présidentielle française.

2. Préalables linguistiques

Le passage d'une ADT occurrentielle à une ADT co-occurrentielle permet d'opérer un saut qualitatif décisif d'un point de vue linguistique ; sans qu'il soit insurmontable méthodologiquement.

En effet, en traitant des unités du corpus (les formes, les lemmes, etc.) de manière isolée, l'ADT peut être accusée de commettre un acte linguistiquement destructeur, dirimant à toute analyse ultérieure. De fait, Saussure comme Harris, Firth comme Guiraud, Halliday comme Tournier établissent les relations entre les unités - et non leur atomisation - comme condition du fonctionnement de la langue et propriété des textes; récemment, la sémantique de corpus (Rastier, 2011) pose quant à elle la contextualisation comme principe cardinal.

Précisément, nous percevons fondamentalement l'approche co-occurrencielle comme un effort de contextualisation des unités : la statistique établit, dans une fenêtre déterminée, la relation matérielle (ou coprésence significative) de deux unités. Traiter d'une co-occurrence n'est dès lors plus seulement analyser un *token* (jeton) du texte, sans valeur sémantique dans son isolement, mais une unité linguistique complexe, déjà signifiante car relationnelle : la co-occurrence comme forme minimale du contexte et unité constitutive de la textualité (Mayaffre, 2008).

Ainsi pour caractériser un texte, le calcul des spécificités, indice majeur de la lexicométrie (Lafon, 1980), gagnera à être établi non plus seulement sur les formes simples comme c'est le cas dans tous les logiciels, mais aussi sur des paires co-occurrentielles ; ainsi, autre exemple, une AFC gagnera à traiter non plus seulement des mots mais des couples ou des binômes. Ce sont là des implémentations fortes, encore peu exploitées, proposées par Etienne Brunet dans son logiciel Hyperbase depuis 2011 (Brunet, 2011).

Concrètement, nous avons montré ailleurs que le discours présidentiel, pour des raisons génériques ou institutionnelles, était obligé, quel que soit le président, de puiser dans un stock lexical imposé (« France », « pays », « politique », « gouvernement », etc.) (Mayaffre, 2012 : 62-63). Dès lors, c'est moins la fréquence prévisible de ces mots pris individuellement dans le corpus qui importe, que leur agencement ou combinaison co-occurrentiels par lesquels s'exprimera l'identité discursive de De Gaulle ou de Sarkozy, de Giscard ou de Hollande.

[1] Le corpus présidentiel constitué depuis plusieurs années compte aujourd'hui 573 discours de De Gaulle, Pompidou, Giscard, Mitterrand, Chirac, Sarkozy et Hollande équivalents à 2.824.973 occurrences.

3. Matrices et distances co-occurrentielles

A la suite de (Massonie, 1986) et de (Viprey, 1997), les matrices co-occurrentielles mots X mots se sont imposées en ADT sans doute car elles donnent la meilleure approximation mathématique de ce qu'est un texte étymologiquement : un tissu ou un tissage où chaque cellule du tableau constitue une maille de signification.

Nous sélectionnons ici 75 substantifs[2], très fréquents dans le corpus et partagés par les présidents, et les croisons[3] entre eux pour quantifier leurs relations ou entrelacements. (Tableau 1).

	avenir	*besoin*	*cas*	*cause*	*choix*	*chose*	*compte*	*condition*	*confiance*	*crise*	…
avenir		0	0	0	0	2	0	0	8	0	…
besoin	0		0	0	0	0	0	0	0	2	…
cas	0	0		0	0	1	0	0	0	0	…
cause	0	0	0		0	0	0	0	0	0	…
choix	0	0	1	0		0	0	0	0	0	…
chose	2	0	0	0	0		0	0	0	0	…
compte	0	0	0	0	0	0		6	2	0	…
condition	0	0	0	0	0	0	6		6	2	…
confiance	0	0	0	0	0	0	2	6		0	…
crise	0	2	0	0	0	0	0	2	0		…
…	...	…	…	…	…	…	…	…	…	…	…

Tableau 1. Matrice co-occurrentielle

Chaque cellule du tableau recueille le nombre de rencontres entre un mot en ligne et un mot en colonne. Une ligne dans son ensemble représente le profil co-occurrentiel exhaustif d'un mot. Pour les 9 textes du corpus, l'opération est reconduite, et nous nous proposons de mesurer la distance entre les 9 matrices co-occurrentielles produites : De Gaulle (1958-1969), Pompidou (1969-1974), Giscard (1974-1981), Mitterrand1 (1981-1988), Mitterrand2 (1988-1995), Chirac1 (1995-2002), Chirac2 (2002-2007), Sarkozy (2007-2012), Hollande (2012-2013). Mises bout à bout, les lignes des tableaux représentent le profil complet du locuteur. Dès lors, le calcul du Khi2 compare les 9 profils et nous permet de produire une matrice de

[2] 75 lemmes sélectionnés : avenir, besoin, cas, cause, choix, chose, compte, condition, confiance, crise, développement, difficulté, droit, élection, enfant, esprit, Europe, façon, fait, fin, fois, fonction, force, formation, France, guerre, heure, histoire, homme, idée, intérêt, jour, justice, liberté, lieu, loi, marché, mesure, mois, moment, monde, moyen, nombre, œuvre, part, parti, partie, pays, personne, peuple, place, point, pouvoir, premier, principe, problème, projet, question, raison, rapport, réalité, république, résultat, rôle, sécurité, sens, situation, société, sorte, système, terme, union, vie, volonté.

[3] La fenêtre utilisée ici pour croiser les substantifs est le paragraphe.

dissimilarités (Tableau 2). C'est cette matrice que l'on se propose de représenter grâce à un arbre.

	DeGaulle	*Pompidou*	*Giscard*	*Mitterrand1*	*Mitterrand2*	*Chirac1*	*Chirac2*	*Sarkozy*	*Hollande*
DeGaulle	0	83	92	88	90	91	114	131	214
Pompidou	83	0	82	80	81	80	106	126	209
Giscard	92	82	0	80	82	80	107	126	206
Mitterrand1	88	80	80	0	71	76	101	124	206
Mitterrand2	90	81	82	71	0	77	104	120	208
Chirac1	91	80	80	76	77	0	99	119	208
Chirac2	114	106	107	101	104	99	0	138	216
Sarkozy	131	126	126	124	120	119	138	0	226
Hollande	214	209	207	206	208	208	216	226	0

Tableau 2. Matrice de dissimilarités

4. Représentation arborée

4.1. Principes et rappels

Comme nous l'avons rappelé aux JADT 1998 (Barthélemy et Luong, 1998), la représentation arborée d'une mesure de dissimilarité d sur un ensemble X à n éléments, consiste à déterminer un arbre dont les feuilles sont étiquetés par X et dont les arêtes sont munies de longueurs positives (ou nulles) de telle sorte que la somme des longueurs des arêtes qui constituent le chemin entre deux sommets réels x et y (distance lue sur l'arbre entre x et y) soit une "bonne approximation" de la dissimilarité d(x,y).

Les approches topologiques privilégient ainsi la recherche d'une structure d'arbre qui reflète "au mieux" les données. La plupart des algorithmes sont plus ou moins inspirés de ADTREE (Sattah et Tversky, 1977). Ils sont fondés sur l'observation que deux feuilles x et y d'un arbre sont adjacentes à un même nœud si et seulement si pour toute paire z, t de sommets réels distincts de x, y on a, pour la distance additive d de l'arbre :

$$d(x,y)+d(z,t) = \text{Min}(d(x,z)+d(y,t), d(x,t)+d(y,z), d(x,y)+d(z,t)) \quad (1)$$

On définit alors le score s d'une paire x,y de X comme le nombre de sommets réels z,t vérifiant l'égalité (1). On fusionne une paire de score maximal en un nœud, le considère comme un sommet réel et réévalue la dissimilarité du nœud ainsi formé à tous les autres sommets; on recalcule alors les nouveaux scores et ainsi de suite. A la fin de l'algorithme les longueurs d'arêtes de l'arbre ainsi créé sont réévaluées. La relation (1) peut s'exprimer par une propriété topologique : « les chemins joignant quatre sommets d'un arbre sont toujours dans la configuration d'un H ou d'une étoile.» (Figure 1).

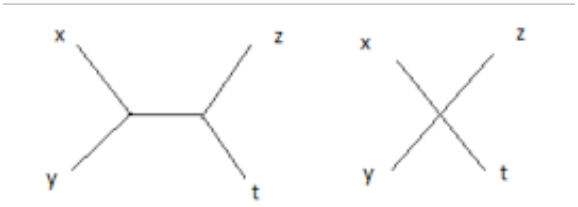

Figure 1. Configuration en H ou en étoile.

Dans un arbre, si deux feuilles x et y sont adjacentes à un sommet interne alors leur score est égal à ((n-1)(n-2))/2. Cette propriété permet de regrouper des ensembles de plus de deux sommets adjacents. Sous le nom de méthodes de groupement, nous avons construit un algorithme de reconstruction d'arbre qui généralise ADTREE (Luong 1988) et que la communauté ADT à souvent utilisé dans une version antérieure (par exemple (Mellet et Longrée, 2009)). A chaque itération, on calcule les scores pour en dégager les groupements, i.e. les feuilles qui sont adjacentes à un seul sommet intérieur, et on réévalue le nœud de chaque groupement. C'est parce qu'elle s'oppose, via l'égalité (1), à (presque) toutes les autres paires d'objets qu'une paire x, y va fusionner. Tout en conservant la nature additive de la distance obtenue et donc la possibilité de l'interpréter en termes d'intermédiarité, voire de filiation, les algorithmes de groupements, procédant par fusions successives, sont de nature classificatoire. C'est le contraire avec le monde de la reconstruction phylogénétique où on privilégie souvent la filiation, par exemple en biologie systématique ou en théorie de l'évolution (Saitou et Nei, 1987).

4.1. Les scores revisités : un nouvel algorithme

Sur chaque sous-ensemble de X on peut calculer les scores de ses paires. On considère la topologie induite par la propriété : « deux éléments sont voisins si leur score est maximum ». Si X est un ensemble de distances arborées (« distances lues sur un arbre »), cette propriété permet de dégager un groupement, par exemple (a, b). On peut montrer que les sous-ensembles X-a et X-b sont équivalents au sens de cette topologie. Nous proposons un algorithme de reconstruction de l'arbre, en utilisant uniquement X et ses sous-ensembles pour dégager la structure de l'arbre, sans aucun autre calcul. On définit la notion de scores stricts s* : ce sont des scores qui ne se calculent que sur les figures en H. La propriété suivante caractérise les groupements :

Propriété 1. Soit x, y une paire de sommets réels issus d'un même groupement composé de k sommets réels, on a :

$$s^*(x, y) = ((n-k)(n-k-1)) / 2 \quad \text{et} \quad s(x, y) = ((n-2)(n-3)) / 2$$

Algorithm 1:

while *|X|>3* ***do***
Calcul des scores s et s . Déterminer les groupements. Déterminer le nœud de chaque groupement. Noter toutes les filiations de type feuille-nœud et leur distance. Gommer, de manière aléatoire, tous les arcs sauf un de chaque groupement*
end
if *Il reste 3 feuilles* ***then***
Avec les distances entre les feuilles on calcule le point O, centre de cette étoile à 3 feuilles, appelé « centre topologique » de l'arbre.
else
Il reste 2 feuilles, O est le milieu du chemin entre ces 2 feuilles.
end
Extraire le codage « père-fils » à partir des filiations. Construire graphiquement l'arbre à partir de ce codage.

Enonçons une propriété qui permet d'avoir un algorithme analogue au précédent pour les représentations arborées.

Propriété 2. Soit x et y deux éléments d'un groupement composé de k sommets et z un élément n'appartenant pas à ce groupement :

$$s(x, y) - s(x, z) \geq n - 3 \quad \text{et} \quad s^*(x, y) - s^*(x, z) \geq n - k - 1$$

On définit alors une notion de voisinage relatif à un ensemble de dissimilarités.

« x et y sont des pré-voisins » définit une relation binaire J° qui, en général n'est pas transitive; on note par J(∂) et on appelle équivalence de voisinage la fermeture transitive de J°(∂). J(∂) est une relation d'équivalence. Une classe d'équivalence modulo J(∂) est appelée un ∂-groupement. Si dans une itération on ne trouve aucun ∂-groupement défini précédemment, on prendra alors comme ∂-groupement une paire dont le score est maximum.

Algorithm 2:

while |X|>3 **do**
Calcul des scores s et s* . Déterminer les ∂-groupements.
if S'il en existe pas **then**
prendre alors une paire (x,y) de score maximum comme un ∂-groupement.
end
Déterminer le nœud de chaque ∂-groupement. Noter toutes les filiations de type feuille-nœud et leur distance. Gommer, de manière aléatoire, tous les arcs sauf un de chaque groupement
end
if *Il reste 3 feuilles* ***then***
Avec les distances entre les feuilles on calcule le point O, centre de cette étoile à 3 feuilles, appelé « centre topologique » de l'arbre.
else
Il reste 2 feuilles, O est le milieu du chemin entre ces 2 feuilles.
end
Extraire le codage « père-fils » à partir des filiations. Construire graphiquement l'arbre à partir de ce codage.

Remarques : Notre nouvel algorithme n'utilise que X et ses sous-ensembles pour avoir la structure de l'arbre, sans aucun calcul d'approximation, alors que les autres algorithmes des groupements évaluent les nœuds des ∂-groupements qui vont être utilisés par l'itération suivante. Ainsi les distances entre les éléments restants déterminés par un calcul sont de plus en plus petites, cela influe de manière notable les dernières itérations.

Quelques essais de comparaison entre des méthodes topologiques montrent que :
- Les coefficients de corrélation sont sensiblement améliorés.
- Le test sur l'inversion des quadruplets (cf (Barthélemy et Guénoche, 1991)) donne des résultats spectaculaires en faveur de notre algorithme.

5. Tracé de l'arbre : Algorithme et solutions techniques

Le logiciel présenté ici est une mise à jour du logiciel ARBOLING initialement développé par Xuan Luong et prévu pour fonctionner sur Mac OS 9. Son portage vers un langage moderne lui permet aujourd'hui de fonctionner directement sur le web et d'intégrer de nouvelles fonctionnalités et une représentation optimisée, basée sur l'algorithme 2 présenté précédemment. Ce logiciel est composé de deux parties, la première intègre le calcul du score et détermine le codage père/fils, et la deuxième se concentre sur le traçage de l'arbre. Nous allons nous concentrer sur cette deuxième partie qui propose une méthode simple qui optimise le rendu visuel de l'arbre proposé.

Pour comprendre l'algorithme il est nécessaire de revenir sur le codage père/fils obtenu après calcul du score et l'identification des groupes. La Figure 2 propose une représentation du codage père/fils qui permet de mettre en évidence les feuilles de l'arbre (les 9 présidents) ainsi que la totalité des nœuds intermédiaires et la racine topologique.

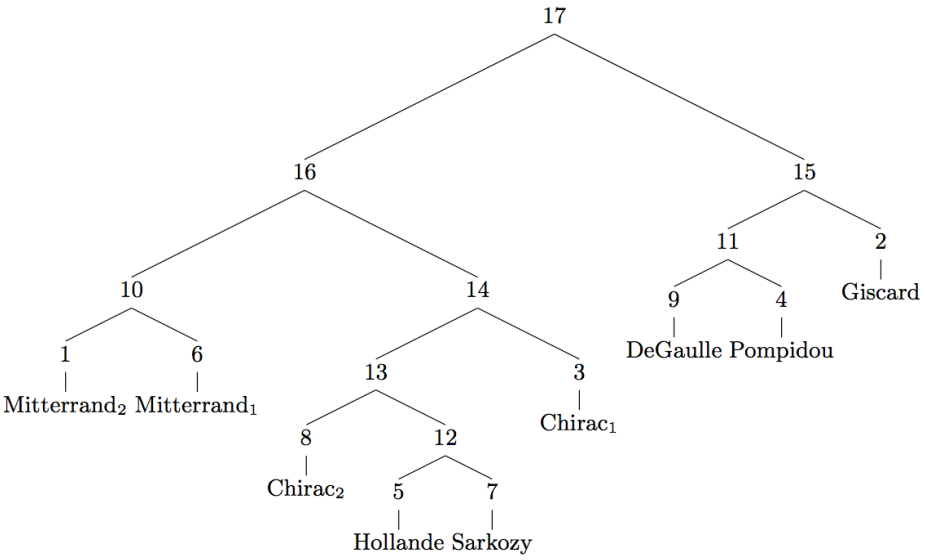

Figure 2. Représentation du codage pères/fils

Pour optimiser la surface utile occupée par le graphe, le logiciel trace un cercle qui occupe tout l'espace disponible et dont le centre représente le centre topologique de l'arbre. Ce cercle est ensuite découpé en autant de zones qu'il y a de feuilles dans l'arbre. Chaque zone de ce cercle forme un angle α qui est notre unité de base pour calculer l'espace disponible pour chaque sous-arbre[4] (Figure 3).

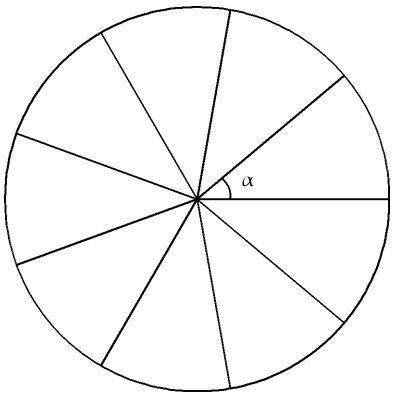

Figure 3. Découpe uniforme du cercle circonscrit de la représentions arborée

Une fois l'angle α calculé, l'arbre est parcouru dans l'ordre de plus profonde descente (*depth first search*), et pour chaque nœud un cône d'angle (nombre$_{\text{feuilles}}$ − 1) ∗ α est déterminé pour définir l'espace occupé par le sous-arbre.

En parcourant cet arbre (codage père/fils), le premier groupe que l'on rencontre est Mitterrand1 et Mitterrand2, dont le père est le nœud d'indice 10 (les indices des nœuds

[4] On appelle sous-arbre un arbre dont la racine est un fils de la racine topologique de l'arbre.

intermédiaires commencent à 10 car les 9 premiers indices sont occupés par les 9 feuilles de l'arbre). Si on remonte jusqu'à la racine de l'arbre on rencontre aussi le nœud 16 responsable de 6 feuilles, le nœud 14 responsable de 4 feuilles et le nœud 15 responsable de 3 feuilles. Pour chaque nœud intermédiaire on détermine l'espace qui lui est réservé sur le cercle en fonction du nombre de feuilles dont il est responsable. Par exemple, le nœud 10 va occuper un cône égal à $(\text{nombre\_feuilles} - 1) * \alpha$ c'est-à-dire un cône d'angle α. La figure 4 nous montre cette étape intermédiaire de la création de l'arbre, avec le positionnement des premières feuilles et des premiers nœuds intermédiaires. L'algorithme est ensuite répété jusqu'à obtenir la totalité des nœuds disposés sur le cercle.

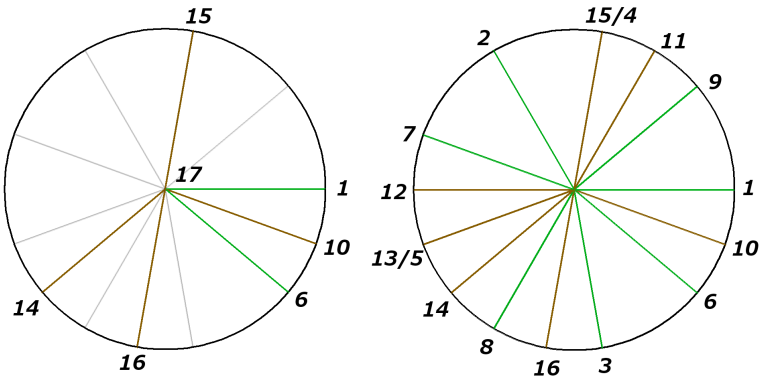

Figure 4. Positionnement des premiers nœuds (Cercle gauche) et des derniers nœuds (Cercle droite)

La dernière étape de cet algorithme consiste à appliquer les distances entre les nœuds (obtenues par le codage Pères/Fils, non visibles sur la figure 2) et à relier chaque nœud père avec l'ensemble de ses fils. Nous obtenons ainsi la représentation arborée attendue (Figure 5).

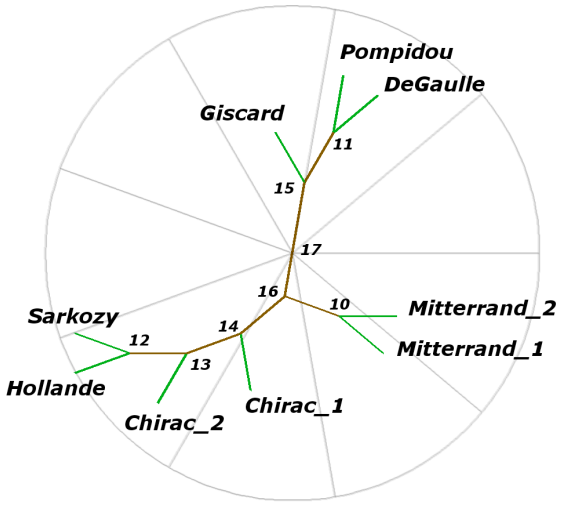

Figure 5. Représentation arborée finale

Ce logiciel, disponible en ligne sur la plateforme Hyperbase Web Edition, en cours de développement (Figure 6), va faire l'objet d'une nouvelle mise à jour visant à intégrer l'analyse arborée avec une suite d'outils documentaires et statistiques. Il sera notamment question d'utiliser les co-occurrences directement en manipulant les entrées et les sorties du logiciel sans demander à l'utilisateur de saisir à la main les matrices de dissimilarités comme proposé dans la version actuelle.

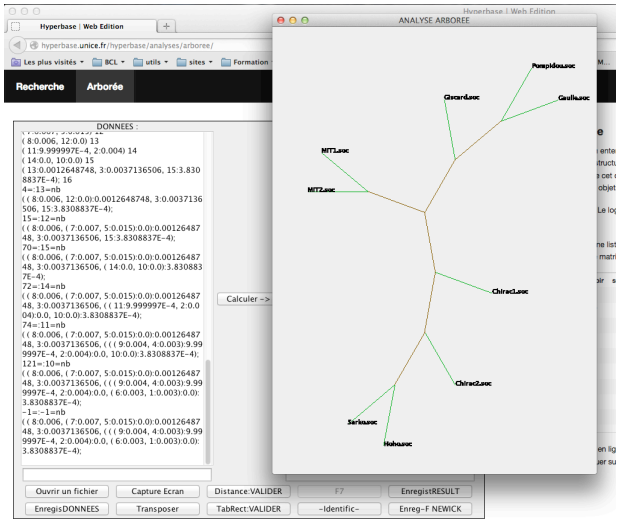

Figure 6. Logiciel en ligne : http://hyperbase.unice.fr

6. Conclusion : première interprétation socio-linguistique

Entre rupture et continuité, le discours présidentiel français évolue depuis l'avènement de la Vème République comme nous l'avons montré ailleurs (Mayaffre, 2012). L'accession de François Hollande à l'Elysée enrichit le corpus jusqu'ici dominé par les présidents de droite, et le prisme co-occurrentiel modifie le point de vue.

L'arbre final (Figure 5) montre d'abord la prégnance de la chronologie sur le corpus. Au regard des relations co-occurrentielles entre substantifs étudiés - c'est-à-dire des noyaux de sens élémentaires ou encore noyaux thématiques - les présidents se distribuent grosso modo d'une extrémité à l'autre de l'arbre selon une logique historique qui part des années 1950 jusqu'à la période actuelle. Le discours de François Hollande se rapproche ainsi naturellement de celui de son immédiat prédécesseur - atténuant par-là, sans doute, l'idée de changement que portait la dernière élection. Le calcul des paires cooccurrentielles spécifiques de François Hollande (versus l'ensemble du corpus) laisse comprendre les pesanteurs de la conjoncture immédiate : « compétitivé-entreprise », « crédits-impôts », « avenir-emploi », « marché-travail ». Tout comme Nicolas Sarkozy (Mayaffre 2012-b), François Hollande s'exprime ainsi comme un président-Premier ministre en charge de la vie quotidienne des Français dans un contexte de crise économique aigüe. Depuis l'invention du quinquennat, la geste présidentielle gaullienne d'un président régalien au-dessus des affaires domestiques se dissipe pour laisser place, dans le discours, aux problèmes économiques et budgétaires.

Dans ce cadre, il conviendrait d'étudier plus avant la structure de l'arbre pour travailler la chronologie. Ici, sur un arbre composé seulement de 9 feuilles, le centre topologique de l'arbre (marqué par le numéro 17) semble confirmer que la césure principale intervient dans les années 1980 comme nous l'avons illustré par la structure nominale et référentielle (versus verbale et phatique) des discours de part et d'autre des mandats de François Mitterrand (Mayaffre 2012-a : 41 et ss).

Mais nous proposons pour conclure de complexifier le point de vue en ajoutant au corpus initial 3 nouveaux textes, ceux de Sarkozy durant les campagnes 2007 et 2012, et celui de Hollande durant la campagne 2012 ; le corpus de De Gaulle ayant quant à lui été divisé autour de l'élection de 1965 pour un total de 13 textes.

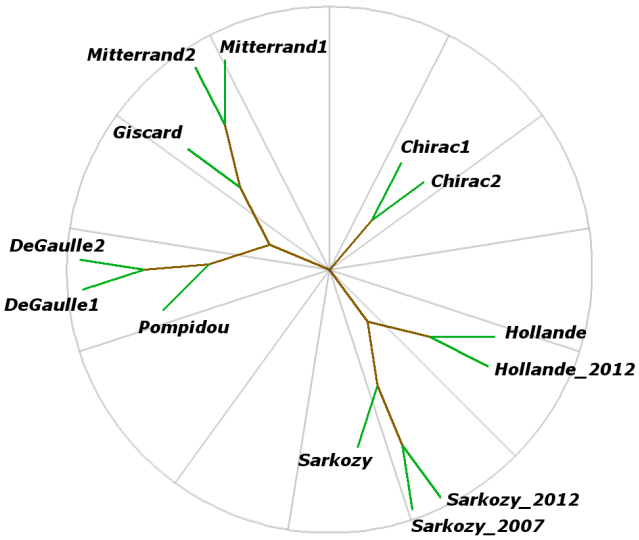

Figure 7. Représentation arborée finale

La classification arborée sur les co-occurrences (Figure 7) montre alors de manière éclatante, à côté d'une chronologie confirmée, la remarquable identité discursive des locuteurs. De Gaulle, Mitterrand ou Chirac ne modifient pas notablement leur discours au cours de leur double mandature. Surtout, Sarkozy et Hollande semblent s'exprimer de la même manière durant leur campagne (discours électoral) et durant leur présidence (discours institutionnel). La co-occurrence apparait comme un indice robuste - plus robuste que l'occurrence simple ? - pour déterminer la signature d'un discours.

Références

Barthélemy J.-P. et Luong N.X. et Mellet S. (2003). « Prenons nos distances pour comparer des textes, les analyser et les représenter » in CORPUS 2 *La distance intertextuelle* CNRS-Université de Nice.

Barthélemy J.-P. et Luong N.X *« Représenter les données textuelles par des arbres », in JADT 1998, Actes des 4e Journées Internationales d'analyse de données textuelles, Université de Nice, UMR 6039, 1998, p. 49-70.*

Barthélemy J.-P. et Guénoche A. (1991). *Trees and proximity representations*, New York : John Wiley et Sons (première édition française : *Les arbres et les représentations des proximités*, Paris : Masson 1988).

Barthélemy J.P. et Luong, N.X. (1988). "Sur la topologie d'un arbre phylogénétique : aspects théoriques, algorithmiques et applications à l'analyse des données textuelles", *Math. et Sciences Humaines*, Paris.

Brunet E. (2014-sous presse). *Au bout du compte Questions linguistiques*, textes édités par Bénédicte Pincemin, préface de François Rastier, Paris : Champion.

Brunet E. (2011). « Nouveau traitement des cooccurrences dans Hyperbase », Corpus, 12 (http://corpus.revues.org/2275).

Buneman P. (1971). "The recovery of trees from measures of dissimilarity", in *Mathematics in Archeological and Historical Sciences.* Hodson et al. Eds, Edinburgh University Press.

CORPUS 2, 2003, "La distance intertextuelle" (dir. Xuan Luong, Jean-Pierre Barthélémy et Sylvie Mellet).

Day W.H.E. (1987). "Computational complexity of inferring phylogenies from dissimilarity measures", *Bulletin of Mathematical Biology* 49, 461-467.

Firth J.R. (1957). *Papers in Linguistics 1934-51*. Oxford. Oxford University Press.

Gambette Ph., Gala N. et Nasr A. (2011), « Longueur de branches et arbres de mots », Corpus, 12 (http://corpus.revues.org/2245).

Lafon Pierre (1980), "Sur la variabilité de la fréquence des formes dans un corpus", Mots, 1, p 127-165.

Luong N. X. (1988). *Méthodes d'analyse arborée. Algorithmes. Applications*. Thèse de doctorat d'état. Université de Paris V.

Luong N.X. (ed.) (1989). *Analyse arborée des données textuelles. Tree Analysis of Textual Data,* CUMFID 16, Nice : CNRS - INLF.

Massonie J.-Ph. (1986). « Q-occurrences libres », in : Brunet E. (dir.), *Méthodes quantitatives et informatiques dans l'étude des textes*, Paris : Champion, p. 611-623.

Mayaffre D. (2008). « De l'occurrence à l'isotopie. Les co-occurrences en lexicométrie », Sémantique et Syntaxe, n°9, 2008, pp. 53-72. (Hal-Shs : http://hal.archives-ouvertes.fr/hal-00551114/fr/).

Mayaffre D. (2012). *Le discours présidentiel sous la Vème République. Chirac, Mitterrand, Giscard, Pompidou, de Gaulle*, Paris, Presses de Sciences Po.

Mayaffre D. (2012-b). *Mesure et démesure du discours. Nicolas Sarkozy 2007-2012*, Paris, Presses de Sciences Po.

Mellet S. et Longrée D. (2009). « Syntactical Motifs and Textual Structures. Considerations based on the Study of a Latin historical Corpus », in S. Mellet et D. Longrée, *New approaches in text linguistics*, Amsterdam, John Benjamins, pp. 161-173.

Mosteller Frederick et Wallace David L. (1964). *Inference and Disputed Authorship : The Federalist. Reading*, Addison-Wessley Publishing Company. Republié sous le titre *Applied Bayesian and Classical Inference : The Case of the "Federalist Papers"*. New York : Springer-Verlag, 1984.

Rastier F. (2011). *La mesure et le grain. Sémantique de corpus*. Paris, Champion.

Saitou N. et Nei, M. (1987). « The neighbor-joining method : a new method for reconstructing phylogenetic trees », *Molecular Biology Evolution*, 4, pp. 406-425.

Sattath S. et Tversky, A. (1977). « Additive similarity tree », *Psychometrika*, vol 42, 3, pp. 319-345.

Tournier M. (1980). « En souvenir de Lagado », *Mots*, 1, pp. 5-9.

Viprey J.-M. (1997). *Dynamique du vocabulaire des Fleurs du mal*, Paris, Champion.