La voix du Président américain (1934-2014)

Jacques Savoy

Université de Neuchâtel (Suisse) – Jacques.Savoy@unine.ch

Abstract

This paper describes a lexical study over the State of the Union addresses from 1934 until 2014. This corpus contains 81 governmental speeches uttered by thirteen presidents. This study shows that considering the most frequent lemmas does not provide useful and pertinent information. However when analyzing the part-of-speech (POS) distribution according to each president, we can see that some presidents such as Eisenhower or Kennedy are using more frequently noun phrases while others (e.g., Obama) prefer using more verbs. When observing the sentence length, we notice that the mean sentence tends to be shorter over the years. Based on an intertextual distance, this study demonstrates that speeches given by the same president tend to be very similar. This is not strong pattern and, for example, some of Reagan or Bush's (father) speeches tend to cluster with other interventions. Using a topic model (*latent Dirichlet allocation*), we found that some presidents (e.g., Nixon, Bush (son), Obama) tend to concentrate on a single and distinctive topic while speeches given by other presidents tend to cover different topics (e.g., Kennedy).

Résumé

Dans cette communication, nous présentons une analyse lexicale d'un corpus composé des discours sur l'état de l'Union de 1934 à 2014. Ce corpus couvre environ 80 ans de vie gouvernementale américaine avec les allocutions tenues par treize présidents. Cette étude indique que les lemmes les plus fréquents n'apportent pas d'information très pertinente. Par contre, en observant la distribution des catégories grammaticales, nous constatons que Eisenhower ou Kennedy recourent de manière plus fréquente aux groupes nominaux tandis que Obama tend à favoriser les verbes. Avec les années, on constate une légère préférence pour des phrases plus courtes. En s'appuyant sur une distance intertextuelle, nous remarquons que les allocutions tenues par le même président tendent habituellement à se regrouper entre elles. Cette tendance n'est pas générale et certains discours de Reagan ou Bush (père) ont tendance à se regrouper avec d'autres allocutions. En appliquant un modèle à thèmes (*topic model*), nous constatons que quelques présidences se concentrent sur un thème distinctif (par exemple, Nixon, Bush (son), ou Obama) tandis que d'autres abordent plusieurs sujets (par exemple, Kennedy).

Mots-clés : analyse du discours, discours politique, comparaison lexicale

1. Introduction

La numérisation des imprimés du projet Google Books (Delahaye et Gauvrit, 2013) a mis à notre disposition un volume considérable d'information dont l'accès aisé et gratuit devrait favoriser des analyses plus quantitatives. Cette tendance vers un accroissement du volume ne s'accompagne pas toujours d'une vérification de la qualité des données (par exemple, *Google Ngram Viewer* indique que le terme *internet* apparaît dans des livres français dès 1800 !). Désirant analyser des données textuelles de bonne qualité, nous nous sommes tournés vers le discours politique qui possède plusieurs autres avantages. D'abord, les documents sont facilement accessibles et sans droit d'auteur. De plus, ils peuvent couvrir des périodes relativement longues. Enfin, leur interprétation s'avère plus aisée que d'autres sources documentaires comme, par exemple, un domaine scientifique spécifique.

Ayant collecté un corpus, comment peut-on en déceler les grands thèmes ? Comment extraire les mots ou expressions significatifs d'une partie par rapport à l'ensemble ? La génération automatique d'un *nuage de termes* a été proposée comme synthèse compacte mais sans

pouvoir la mettre en relation avec d'autres textes. Dans le cadre de cette communication, nous souhaitons proposer quelques approches et évaluer empiriquement leur qualité.

Dans ce but, nous avons repris les discours sur l'état de l'Union durant les 80 dernières années. A priori, ce corpus d'allocutions devrait faire ressortir les affinités entre présidents. De tels rapprochements devraient exister entre présidents appartenant au même parti, donnant ainsi naissance à deux identités linguistiques distinctes, l'une républicaine et l'autre démocrate. En effet, nous pourrions nous attendre à ce que les démocrates s'expriment plutôt sur l'éducation, la famille et la santé tandis que les républicains devraient axer leurs interventions sur la libre entreprise, les valeurs morales strictes et sur une réduction des dépenses. Est-ce qu'une analyse lexicale peut confirmer cette hypothèse ? Quels outils sont les plus aptes à décrire les diverses tendances, affinités ou oppositions entre présidents ?

2. Travaux reliés

Dans l'analyse lexicographique et comparative des discours politiques, nous pouvons mentionner les travaux de (Labbé et Monière, 2003, 2008) qui comparent trois sources de discours gouvernementaux, soit le discours du Trône (Canada), le discours inaugural (Québec) et les déclarations de politique générale (France). Les avantages de cette étude tiennent au fait que cet ensemble de discours est rédigé dans la même langue et couvre une période relativement longue (de 1945 à 2000). On peut également souligner que cette analyse repose sur trois régimes parlementaires. De plus, même si les discours gouvernementaux expriment les idées de différents partis politiques, ils ont tendance à être plus similaires à ce que l'on pouvait s'y attendre. Les contraintes institutionnelles ne sont pas étrangères à ce phénomène. Par exemple, la continuité de l'exercice du pouvoir tend à gommer le clivage des partis. Les auteurs soulignent toutefois des modifications temporelles comme la tendance à disposer de discours plus longs au fil des années (plus grande complexité des questions abordées), avec une augmentation sensible de la longueur entre les discours de la IV^{e} et ceux de la V^{e} République. Une conclusion similaire a été établie pour l'Italie (Pauli et Tuzzi, 2009).

Afin d'extraire automatiquement un ensemble de thèmes d'un corpus, nous pouvons également nous appuyer sur des modèles à thèmes (ou allocation latente de Dirichlet (Blei et al., 2003)). Dans ce cadre, le corpus est analysé selon un modèle probabiliste de génération de documents avec chaque texte pouvant aborder plusieurs sujets. Chaque document se modélise comme une distribution de différents thèmes, avec chaque thème représentant une distribution spécifique de mots (l'ordre de ces derniers n'ayant pas d'importance, l'hypothèse du *sac de mots* est donc admise).

Afin de décrire les relations entre documents, nous pourrions également nous appuyer sur une mesure de distance lexicale (Labbé, 2007) entre deux discours, deux ensembles de discours ou entre quelques leaders politiques (Labbé et Monière, 2003), (Mayaffre, 2004) pour déterminer leur position relative au moyen d'une carte. Ainsi, avec les premiers ministres québécois, la distinction de genre entre discours oral et écrit semble plus forte que la distinction entre auteurs, avec quelques exceptions (les discours de R. Lévesque restent très similaires entre eux, qu'ils soient oraux ou écrits) (Monière et Labbé, 2006). En France, les discours gouvernementaux peuvent également faire l'objet d'une telle représentation permettant de voir la dynamique de rapprochement ou d'éloignement entre figures politiques (Mayaffre, 2004).

3. Les discours sur l'état de l'Union et traitements préalables

Afin de créer notre corpus, nous avons téléchargé[1] les discours sur l'état de l'Union (*State of the Union address*). Ce corpus se subdivise en 81 allocutions que l'on peut regrouper en 13 parties, chacune correspondant à un président. Dans notre corpus, le premier discours a été prononcé par Roosevelt (3 janvier 1934) et le dernier par Obama (28 janvier 2014). Nous avons éliminé deux discours, à savoir celui de 1946 (Truman) et de 1981 (Carter), deux discours seulement écrits et de taille nettement supérieure (de sept à dix fois plus long) aux autres allocutions des mêmes auteurs.

Afin d'analyser son contenu lexical, nous pouvons travailler sur les formes fléchies (e.g., *is*, *was*, *been, be, were* ou *wars*, *war*) d'une part et, d'autre part, sur les lemmes. Dans ce dernier cas, les formes appartenant à la même entrée dans le dictionnaire sont regroupées (e.g., *be* ou *war* de notre exemple précédent). Dans cette communication, nous avons retenu uniquement les lemmes possédant l'avantage de supprimer les effets liés à la syntaxe. Des conclusions très similaires auraient pu être extraites sur la base des formes fléchies.

Afin de déterminer les catégories grammaticales, nous avons recouru au logiciel d'étiquetage syntaxique automatique de l'Université de Stanford (Toutanova et al., 2003). Ce dernier attribue à chaque mot une étiquette syntaxique et des étiquettes morphologiques dérivées du corpus de Brown. Par exemple, depuis la phrase *Dangerous problems remain from Cuba to the South China Sea*, le système répondra par *Dangerous*/JJ *problem*/NN *remain*/VBP *from*/IN *Cuba*/NNP *to*/TO *the*/DT *South*/NNP *China*/NNP *Sea*/NNP ./. On y retrouve les étiquettes attachées au nom (NN, nom commun au singulier, NNS nom commun au pluriel, NNP nom propre au singulier), verbe (VB, entrée dans le dictionnaire, VBP présent non 3e personne, VBN participe passé), adjectif (JJ), pronom personnel (PRP), préposition (IN), adverbe (RB). Depuis chaque mot accompagné de son analyse morphologique, nous pouvons retrouver l'entrée dans le dictionnaire, essentiellement pour les noms par suppression du pluriel (e.g., *jobs*/NNS → *job*/NN) et par substitution de la forme fléchie des verbes (e.g., *argues*/VBZ → *argue*/VB). Lors cette transformation, les verbes irréguliers et quelques noms (e.g., *children*/NNS → *child*/NN) ne suivant pas la règle habituelle et subissent donc un traitement particulier.

Ce traitement automatique n'est pas exempt d'erreurs ou de syntagmes dont l'étiquetage proposé reste sujet à discussion, comme par exemple pour le groupe nominal S*enate Foreign Relations Committee*. Une première solution consiste à donner la même partie du discours aux quatre éléments (*Senate*/NNP *Foreign*/NNP *Relations*/NNP *Committee*/NNP). Comme alternative, on peut attribuer l'étiquette nom propre au pluriel au mot *Relations*/NNPS.

Ce processus d'étiquetage effectué, notre corpus gouvernemental comprend un total de 477 402 *tokens* (ou mots) et 11 160 lemmes distincts (taille du vocabulaire). Dans cet ensemble, on rencontre 4 021 *hapax* (mot ayant une seule occurrence, soit 36 % du vocabulaire) et 1 555 *dis legomena* (mot ayant exactement deux occurrences, correspondant à 14 % du vocabulaire). L'article défini (*the*, 24 785 occurrences) est le lemme le plus fréquent, suivi de *we* (16 601), *of* (15 574), *and* (15 512), *to* (14 535) et *be* (13 389).

Comme pré-traitement, nous avons remplacé certains codes UTF-8 par leurs équivalents ASCII (e.g., “ ” en ") ainsi que pour les lettres accentuées (e.g., *naïve* en *naive*). Quelques signes graphiques ont été supprimés ainsi que quelques fautes d'orthographe. Nous avons remplacé les lettres majuscules par des minuscules si cette dernière forme se retrouvait dans le

1 Depuis le site www.presidency.ucsb.edu. Le site www.WhiteHouse.gov propose aussi une version de ces allocations tandis que `www.c-span.org` en propose une version audio et vidéo.

corpus (e.g., *We* et *we*). A contrario, cette approche nous permet de conserver les noms propres (e.g., *Marshall*, *July*) ou des formes apparaissant toujours sous la même graphie (e.g., *Speaker*, *I*). Enfin, nous avons essayé de normaliser les formes différentes pouvant désigner la même entité comme, par exemple, avec *US*, *U.S.*, *U.S.A.*, *United States*. Parfois ces formes correspondent à différentes orthographes (*Viet Nam* ou *Vietnam*). Des graphies distinctes ont été conservées lorsque les entités désignées devaient être distinguées (*Soviet Union* et *Russia*).

Rang	Années	Parti	N discours	Longueur moyenne	Lemmes distincts
F. D. Roosevelt	1934-45	Démocrate	12	4 422	1 018
H. Truman	1947-53	Démocrate	7	6 190	1 159
D. Eisenhower	1953-61	Républicain	9	6 729	1 428
J. F. Kennedy	1961-63	Démocrate	3	6 540	**1 449**
L. B. Johnson	1964-69	Démocrate	6	5 408	1 113
R. Nixon	1970-74	Républicain	5	4 419	**887**
G. Ford	1975-77	Républicain	3	5 202	1 153
J. Carter	1978-80	Démocrate	3	**4 260**	941
R. Reagan	1982-88	Républicain	7	5 307	1 190
G. H. Bush	1989-92	Républicain	4	5 033	1 075
W. Clinton	1993-00	Démocrate	8	**8 424**	1 391
G. Bush	2001-08	Républicain	8	5 743	1 234
B. Obama	2009-14	Démocrate	6	7 810	1 434

Tableau 1. Répartition des discours par président accompagnée de leur longueur moyenne

Afin d'avoir une idée du volume traité, nous avons présenté quelques statistiques dans le tableau 1. Pour chaque président, on indique le nombre de discours ainsi que le nombre moyen de lemmes et de lemmes distincts par discours. Sur l'ensemble de notre corpus, la longueur moyenne d'une allocution s'élève à 5 864,6 *tokens* (pour 1 194 lemmes distincts). Ce tableau signale que les discours de Carter sont, en moyenne, les plus brefs tandis que ceux de Clinton s'avèrent les plus longs. Au niveau du vocabulaire, Kennedy présente la plus grande richesse, en moyenne, avec 1 449 lemmes distincts par discours, et Nixon le vocabulaire le plus restreint (887 lemmes / discours). Enfin, nous désignerons par *H. Bush* George H. W. Bush (père) tandis que le nom *Bush* indiquera George W. Bush (fils).

4. Analyse macroscopique des discours sur l'état de l'Union

Afin de dégager quelques grandes tendances de ce corpus gouvernemental, nous pouvons nous appuyer, dans une première étape, sur les lemmes les plus fréquents (section 4.1) ou la distribution des catégories grammaticales et la longueur des phrases (section 4.2). Afin de visualiser les affinités et oppositions entre discours ou président, nous pouvons définir une distance intertextuelle et générer une arborescence par application d'une classification automatique (section 4.3). En recourant à un modèle à thèmes (*topic model*), nous pouvons disposer d'une analyse thématique globale complémentaire (section 4.4).

4.1. Les lemmes les plus fréquents

Si l'on consulte les lemmes les plus fréquents, on retrouve une centaine de mots fonctionnels (articles, prépositions, conjonctions, signes de ponctuation, verbes auxiliaires) sans intérêt majeur pour notre analyse. Ces formes ignorées, les lemmes les plus fréquents sont : *government* (6 879 occurrences), *State* (6 670), *Congress* (5 285), *United* (5 036), *country* (4 571), *would* (4 092), *people* (4 088), *nation* (3 682), *law* (3 572), *time* (3 335), *power* (2 902), *war* (2 410), …, *American* (adjectif, 2 171), …, *America* (nom, 1 680). Le discours sur l'état de l'Union concerne en premier lieu les Etats-Unis (*America*, *American*) et s'adresse à la *nation*, *country*, *people* et au *Congress*. Ces multiples occurrences correspondent aussi à

la mise en forme du discours. De plus, le lemme *world* constitue un thème récurrent dans les discours présidentiels car on le retrouve aussi fréquemment en Italie (Pauli et Tuzzi, 2009).

Cette allocution présente bien le programme annuel pour le gouvernement dont les thèmes récurrents se retrouvent dans les noms les plus fréquents comme les questions fiscales (*tax*, 1 152), *peace* (1 869), *child* (824), *job* (656) ou *economy* (915). Au niveau des verbes les plus fréquents (en ignorant les auxiliaires), on rencontre *provide* (1 782), *increase* (1 636), *continue* (1 499), *give* (1 300), *require* (1 251) et *exist* (1 162). Pour les adjectifs, la liste comprend *last* (2 701), *new* (2 545), *American* (2 171), *own* (1 871), *national* (1 770), *present* (1 732), *foreign* (1 594), *good* (1 380) et *important* (1 283).

Comme autre élément pertinent, nous pouvons observer les lemmes géographiques les plus fréquents. En première place, on trouve évidemment 1 680 lemmes *America* suivi de *Mexico* (858 occurrences), *Indian* (604), *Britain* (584), *Europe* (527), *Spain* (499), *France* (381), *China* (371), *Cuba* (371), (puis *Texas* (277), *Japan* (232), *Russia* (207), *Columbia* (196), *Panama* (195), *Soviet* (194), *Mississippi* (183) et *Philippine* (181)).

Comme notre corpus est diachronique, nous avons repris dans le tableau 2 les cinq noms géographiques les plus fréquents par président. Cette information donne une idée de l'intérêt dominant porté par chaque présidence ainsi que l'évolution durant les 80 dernières années. Ainsi si l'*Europe* était un centre d'intérêt essentiel dans les années 50 à 60, son attrait régresse au profit de l'*Asia* ou de la *China*. A l'exception des deux dernières présidences, la *Soviet Union* ou *Russia* occupe une place récurrente. On constate également que la région de l'*Afghanistan*, *Iran* et *Iraq* constitue un thème important depuis Carter.

Roo	**Tru**	**Eis**	**Ken**	**Joh**	**Nix**	**For**	**Car**	**Rea**	**HBus**	**Cli**	**Bush**	**Oba**
Europe	Europe	Europe	Europe	Vietnam	Washin.	Asia	Soviet	Soviet	Europe	China	Iraq	Washin.
Japan	Soviet	Soviet	Soviet	Europe	Soviet	Europe	Afghan.	Washin.	Soviet	Russia	Afghan.	Afghan.
China	Korea	Korea	Latin	Soviet	Vietnam	Japan	Israel	Nicaragua	Gulf	Europe	Mid. East	Iran
Germany	China	Columbia	Atlantic	Asia	Asia	Soviet	Asia	Afghan.	Iraq	Asia	Iran	Iraq
France	Japan	China	Cuba	Korea	China	China	Iran	Mid. East	Panama	Washin.	Washin.	China

Tableau 2. Les cinq noms géographiques les plus fréquents par chaque président

4.2. Répartition par catégories grammaticales

Notre étiquetage nous permet de connaître la répartition des diverses catégories grammaticales selon chacun locuteur. Le tableau 3 indique la répartition des diverses catégories grammaticales par président. Finalement, les valeurs les plus élevées sont indiquées en gras, les plus faibles en italique.

	Roo	**Tru**	**Eis**	**Ken**	**John**	**Nix**	**For**	**Car**	**Rea**	**HBus**	**Cli**	**Bush**	**Oba**
nom	20,1 %	20,3 %	**22,0 %**	20,9 %	*18,2 %*	18,8 %	20,0 %	19,3 %	19,6 %	18,7 %	19,0 %	20,2 %	18,9 %
npro.	*3,0 %*	3,0 %	3,7 %	3,4 %	3,8 %	3,4 %	4,0 %	3,4 %	3,8 %	3,6 %	3,9 %	**5,0 %**	3,5 %
verbe	13,2 %	13,9 %	12,9 %	*12,3 %*	14,2 %	13,7 %	13,5 %	13,2 %	14,2 %	14,7 %	15,1 %	15,4 %	**16,0 %**
conj.	4,0 %	3,5 %	3,6 %	4,6 %	4,1 %	*3,3 %*	3,7 %	**4,6 %**	4,0 %	4,4 %	3,7 %	4,6 %	3,8 %
prép.	**13,3 %**	12,1 %	12,4 %	11,4 %	11,7 %	12,8 %	11,2 %	10,7 %	10,7 %	*9,7 %*	9,9 %	10,0 %	10,0 %
article	11,3 %	10,8 %	10,3 %	10,1 %	10,2 %	**11,9 %**	9,7 %	10,0 %	9,0 %	9,7 %	9,0 %	*8,8 %*	9,0 %
adj.	8,1 %	8,0 %	**9,2 %**	8,1 %	6,4 %	7,1 %	8,7 %	8,5 %	7,1 %	6,2 %	6,4 %	6,7 %	*5,9 %*
pronom	6,1 %	6,8 %	*4,9 %*	6,0 %	7,8 %	6,9 %	6,6 %	7,1 %	7,2 %	8,1 %	**8,9 %**	7,2 %	8,1 %
adv.	7,7 %	7,3 %	7,2 %	7,6 %	7,8 %	8,3 %	7,4 %	7,9 %	8,8 %	8,3 %	8,5 %	*6,6 %*	**9,3 %**

Tableau 3. Pourcentage des différentes catégories grammaticales par président

Une première analyse révèle que les noms représentent un pourcentage significativement élevé pour Eisenhower et Kennedy. Cette importance accordée au groupe nominal peut s'expliquer par un besoin accru d'explication. Une fréquence plus faible caractérise le discours de Johnson. Pour les noms propres (deuxième ligne *npro*), Bush (fils) en fait un

usage important. La deuxième catégorie importante est le verbe qu'Obama et, dans une moindre mesure, Bush (fils) sur-emploient tandis que Kennedy le néglige. L'adjectif est favorisé dans les allocutions d'Eisenhower, tandis que Clinton opte volontiers pour les pronoms. Ceux-ci sont également abondamment usités par Obama qui se montre également un adepte des adverbes.

Si nous souhaitons présenter fidèlement la position des treize présidents, nous devons les placer dans un espace à neuf dimensions (nombre de catégories grammaticales considérées). Afin de respecter au mieux les vraies distances entre les présidents et de permettre une visualisation en deux dimensions, nous avons opté pour une analyse en composantes principales (ACP) (Lebart et al., 1998) sur la base du tableau 3. Cette opération a été réalisée au moyen du logiciel R (Baayen, 2008) et le résultat est présenté en figure 1.

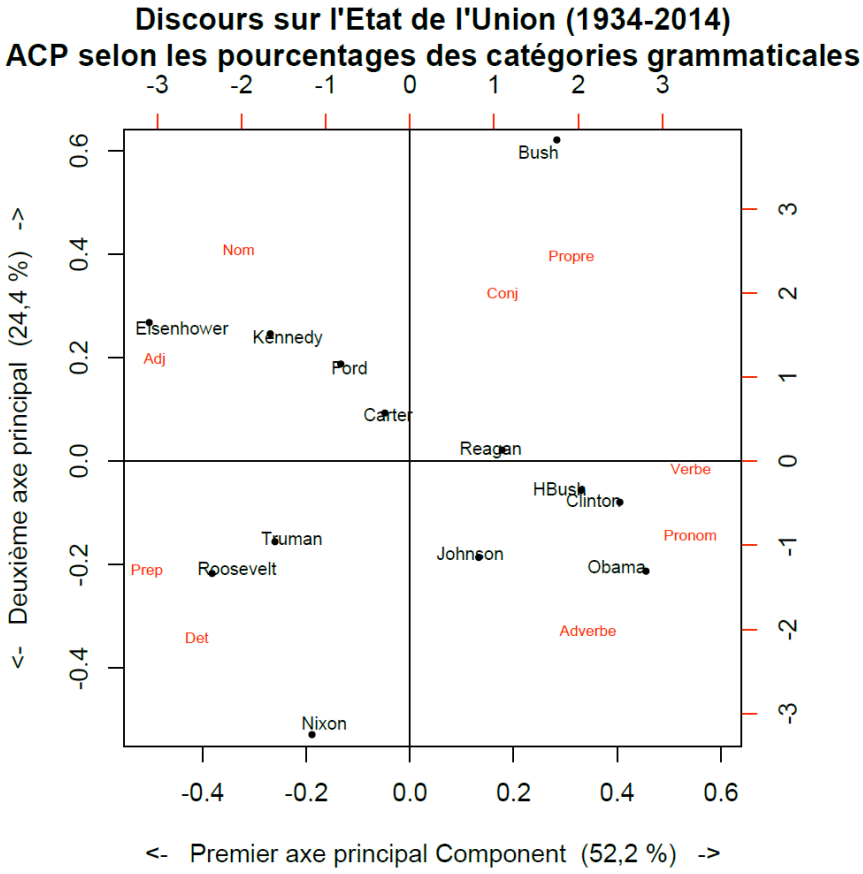

Figure 1. Représentation par l'analyse en composantes principales des catégories grammaticales des différents présidents

Dans cette figure, le premier axe principal souligne l'opposition entre le groupe nominal (noms communs, adjectifs, et prépositions) se situant à gauche et le groupe des verbes et pronoms qui se place à droite. Aux extrémités de cet axe, on retrouve l'opposition entre le binôme Eisenhower – Roosevelt d'une part et, d'autre part, Clinton – Obama. Le long du deuxième axe principal, on retrouve vers le haut le groupe des noms propres et conjonctions avec Bush (fils) comme figure emblématique. En bas de cet axe on retrouve les articles avec comme représentant typique Nixon. Le discours moyen serait plutôt l'apanage de Carter ou Reagan.

Dans cette représentation, les affinités entre présidents ne semblent pas toujours suivre l'appartenance partisane (e.g., Roosevelt – Truman). Ainsi, on retrouve des paires formées par un représentant de chaque parti (H. Bush (père) – Clinton). Les regroupements s'opèrent plus

aisément en fonction de la proximité temporelle. Finalement, on notera également que Bush (fils) et Nixon sont assez éloignés du discours moyen.

Enfin, l'analyse de la longueur moyenne des phrases en nombre de *tokens* met en lumière de nouvelles distinctions. Sur l'ensemble de notre corpus (18 711 phrases), la longueur moyenne d'une phrase s'élève à 22,81 *tokens* (écart-type 13,8). Kennedy possède la moyenne la plus élevée tandis H. Bush (père) compose les phrases les plus brèves. Avec le temps, on perçoit une tendance vers une composition incluant des phrases plus courtes (par exemple, Roosevelt (26,7), Nixon (26,7), Clinton (22,4), Obama (20,8)).

	Roo	**Tru**	**Eis**	**Ken**	**Joh**	**Nix**	**For**	**Car**	**Rea**	**HBus**	**Cli**	**Bush**	**Oba**
moye	26,7	22,0	23,6	**27,9**	22,8	26,7	21,7	25,0	22,5	**19,0**	22,4	21,1	20,8
med.	24	20	22	23	20	**24**	20	20	20	*17*	19	20	19
N	1 964	1 946	2 535	691	1 424	822	709	576	1 625	1 046	2 968	2 153	2 216
min	2	4	2	4	2	3	2	3	2	3	2	2	2
max	146	105	148	148	129	111	93	148	102	115	136	81	97
écart	15,3	11,0	12,4	19,0	14,0	15,1	11,9	12,8	12,6	11,0	13,8	9,8	11,9

Tableau 4. Statistiques sur la longueur moyenne des phrases réparties selon les présidents

4.3. Distance intertextuelle et représentation graphique

Comme troisième analyse macroscopique, nous avons choisi de mesurer la distance entre chaque paire d'allocutions. Dans cette optique, nous avons adopté la distance intertextuelle proposée par Labbé (2007). Avec cette métrique, la valeur retournée varie entre 0 et 1 et dépend du taux de recouvrement des deux textes. Ainsi une distance nulle sépare deux textes identiques. Par contre, si ces derniers ne possèdent aucun mot en commun, la distance sera de 1. Entre ces deux valeurs extrêmes, la distance retournée sera fonction du nombre de mots communs aux deux textes ainsi que de leur fréquence.

En appliquant cette mesure de distance pour chaque paire de discours, nous obtenons une matrice symétrique (81 x 81 = 6 641 valeurs). Redonner simplement toutes ces valeurs ne présente pas un grand intérêt. Par contre, nous pouvons utiliser cette information pour opérer une classification automatique (Kaufman et Rousseeuw, 1990). Le résultat final nous permettra de découvrir les divers groupes formés de textes les plus similaires.

Afin d'atteindre cet objectif visuel, nous avons opté pour une représentation arborée, une technique classique en génomique (Baayen, 2008 ; Paradis, 2011). Le résultat obtenu est repris dans la figure 2. Son interprétation est similaire à celle que l'on retrouve dans d'autres applications (Mayaffre, 2004 ; Labbé, 2007).

En partant de la gauche de la figure et en suivant le sens des aiguilles d'une montre, nous rencontrons un groupe homogène formé par les discours tenus par Bush (fils) (de 2002 à 2008). Ces allocutions de Bush (fils) postérieures aux attentats du 11 septembre forment une classe assez homogène. Le président Bush semble avoir trouvé son style après ces événements tragiques. Après ce premier groupe, on voit un duo de discours de H. Bush (père) (BuH91, BuH92) et d'un groupe d'allocutions de Reagan (Rea84 à Rea88). Cet ensemble républicain se termine par trois discours un peu isolé, à savoir H. Bush (père) 1989 et 1992 et le premier de Bush (fils) 2001.

La suite comprend les deux derniers présidents démocrates débutant avec un ensemble correspondant à Clinton (formé de deux sous-groupes de Clinton 1997 à 2000, puis Clinton de 1994 à 1996, correspondant à ses deux mandats). Les discours d'Obama (de 2009 à 2014) forment un groupe moins soudé.

Sur la droite et vers le haut, on retrouve un groupe démocrate avec les trois allocutions de Kennedy et celle de Johnson en 1964. Ce groupe est bien distinct des autres groupes. Comme

Kennedy est assassiné le 22 novembre 1963, le temps a dû manquer au nouveau président pour imposer son style. On peut supposer que l'équipe Kennedy a donc écrit le premier discours de la présidence Johnson (discours très court, soit 3 629 termes).

Sur la droite, on repère trois groupes de discours correspondant à Roosevelt, à savoir les allocutions tenues de 1934 à 1939, le duo des années 1940 et 1941, puis le groupe des années de guerre (1942 à 1945). Ensuite, on rencontre un trio d'allocutions de Truman (1951 à 1953), et une longue séquence correspondant aux discours d'Eisenhower avec, inséré au milieu, deux duos de discours de Truman (1947-1950, puis 1948-1949). On constate que les discours des trois présidents Roosevelt, Truman et Eisenhower sont bien séparés du reste.

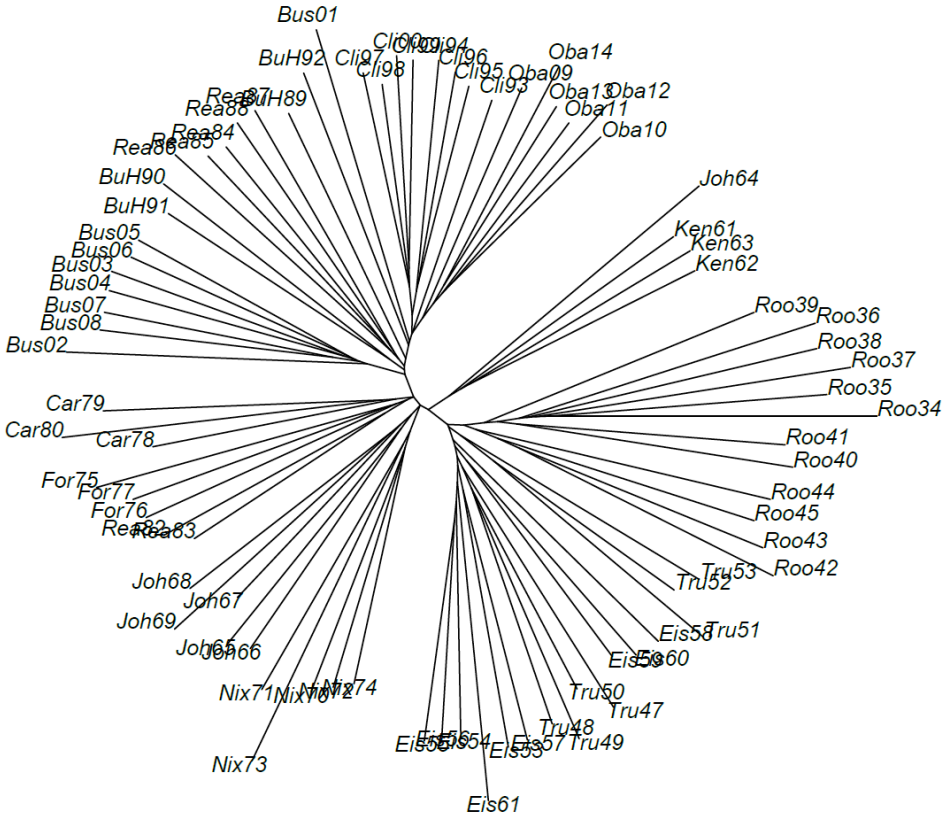

Figure 2. Représentation arborée des distances intertextuelles entre les discours

En bas, un peu sur la gauche, nous retrouvons les cinq discours de Nixon (1970 à 1974), et un deuxième groupe formé des cinq allocutions restants de Johnson (1965 à 1969). Distinct de ces deux entités, on rencontre ensuite un groupe du parti républicain formé de deux discours de Reagan (Rea82, Rea83) et trois de Ford (For76, For77, For75). Finalement, les trois discours de Carter (1978 à 1980) apparaissent.

Parmi ces distances entre discours, la plus grande (0,466) sépare l'allocution de Roosevelt en 1934 (Roo34) de celle d'Obama en 2014 (Oba14). La seconde plus grande distance (0,465) sépare Roosevelt 1937 de Bush (fils) 2001 (Bus01) ou Bush (fils) 2002 (Bus02). La plus petite distance (0,22) relie Clinton en 1998 de 1999. La deuxième plus petite distance (0,226) distingue le discours de Clinton en 1995 de celui prononcé en 1996.

Si l'on assemble toutes les allocutions prononcées par le même chef de l'exécutif, on peut construire une image moins détaillée mais révélant les affinités et oppositions entre les présidents. La figure 3 montre le résultat obtenu. En suivant un mouvement vertical du bas vers le haut, on retrouve les présidents presque dans l'ordre chronologique. De plus, on

retrouve les affinités fortes entre Truman et Eisenhower ou entre Obama et Clinton. Enfin, un rapprochement se dessine entre Nixon et Johnson, de même qu'entre Carter et Ford.

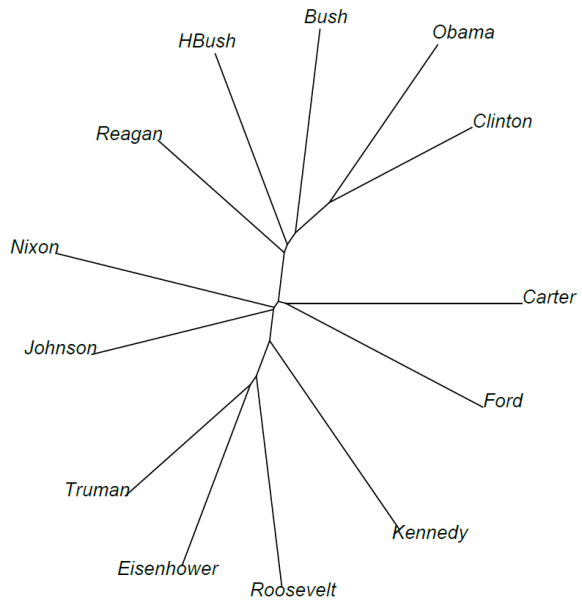

Figure 3. Représentation arborée des distances intertextuelles entre présidents

Dans cette comparaison comprenant les treize présidents, la distance la plus faible (0,185) se situe entre Clinton et Obama. La deuxième distance minimale (0,189) sépare Truman de Eisenhower tandis que la troisième (0,195) se situe entre Reagan et H. Bush (père). La plus grande distance (0,342) distingue Eisenhower et Obama (et la seconde plus importante (0,319) relie Clinton d'Eisenhower, puis on trouve Obama et Roosevelt (distance 0,317)). L'influence de la chronologie serait donc plus forte que celle des appartenances à un parti.

4.4. A la recherche des thèmes

La technique d'allocation latente de Dirichlet (ou modèle à thèmes, *topic model*) (Blei et al., 2003) correspond à un modèle probabiliste de génération de documents abordant plusieurs sujets. Dans ce cadre, chaque document d'un corpus se modélise comme une distribution sur différents thèmes. Dans ce contexte, un thème ne possède pas directement une sémantique (e.g., liste de vedettes-matière) mais correspond à une distribution spécifique des lemmes (l'ordre de ces derniers n'ayant pas d'importance, l'hypothèse du *sac de mots* est donc admise).

Dans ce modèle, un texte peut couvrir un seul thème, mais ceci constitue l'exception et non la norme. Ainsi, une première allocution peut traiter essentiellement du premier thème, un peu du second et marginalement du troisième (et ignorer tous les autres). Un mélange équilibré des deux premiers thèmes peut se retrouver dans le deuxième discours, et ainsi de suite. Chaque lemme peut apparaître sous plusieurs thématiques afin d'indiquer sa polysémie (comme, par exemple, *nuclear* ou *free*), son importance dans la forme (*program, nation* ou *Congress*) ou le fait qu'il corresponde à un mot fonctionnel présent dans presque toutes les allocutions (*the*, *in*, *have*, *be*, *and*).

Dans notre application, nous avons fixé le nombre total des thèmes à dix, une valeur correspond approximativement au nombre de présidents de notre corpus. De plus, nous avons éliminé 658 termes fonctionnels afin de nous concentrer sur le contenu sémantique et moins sur la forme et le style. Le tableau 5 illustre ces dix thèmes par leurs lemmes les plus probables.

Si aucun sujet précis ne se dessine sous le premier thème, le deuxième fait apparaître en début de liste le lemme *peace*, un sujet lié à la guerre au Vietnam. Si pour Kennedy, il s'agissait de stopper l'avance du bloc communiste, Johnson veut gagner la guerre et Nixon cherchera à la finir. Le troisième thème regroupe des formulations plutôt républicaines avec les impôts (*tax*) et un *budget* à réduire tout en respectant les valeurs de la *family*, et de la liberté (*freedom*).

Th1	Th2	Th3	Th4	Th5	Th6	Th7	Th8	Th9	Th10
$	State	tax	work	national	work	economic	free	war	job
tax	great	budget	security	State	child	federal	Soviet	fight	work
billion	peace	work	Iraq	work	school	national	war	force	business
increase	President	family	child	peace	commun.	State	defense	production	tax
federal	federal	future	tax	man	family	security	peace	man	energy
State	energy	child	State	seek	challeng.	military	freedom	peace	cut
continue	tonight	tonight	freedom	power	support	increase	United	enemy	give
million	Union	freedom	health	great	century	policy	effort	plan	economy
act	increase	federal	terrorist	war	tonight	peace	economic	great	health
percent	goal	rate	economy	United	care	free	power	work	reform

Tableau 5. Distribution des lemmes les plus fréquents sur les dix thèmes

Le quatrième thème s'avère fortement lié aux mots *security*, *Iraq*, *freedom* et *terrorist* qui nous laissent entrevoir des sujets associés à la guerre et à la recherche de la sécurité. De plus, on y rencontre également le sujet des impôts (*tax*) et de l'assurance maladie (*health*).

Dans la cinquième liste de vedette-matières, on observe le mot *national* puis *peace*, *man*, *seek*, ou *war*. Une association directe avec un groupe de sujets n'est pas évidente. Le sixième thème possède une saveur démocrate avec une attention portée à l'éducation (*school*), à la famille (*child*, *family*) et aux défis (*challenge*) du siècle (*century*). En filigrane, on voit également le problème de l'assurance maladie (*health*).

Le huitième thème tend à être associé aux problèmes de la guerre froide (*Soviet*, *force*), la défense du monde libre (*defense*, *free*) et à la limitation des armes pour un effort vers la paix (*effort*, *peace*). Le neuvième semble assez similaire. Enfin, sous le dernier thème, on retrouve des lemmes liés à la politique économique (*job*, *business*), aux impôts (*tax*, *cut*) et à la santé (*health*).

Nous constatons que l'identification d'un thème (ou un groupe de thèmes) n'est pas toujours évidente avec cette approche. Les études précédentes (Chang et al., 2009), (Blei, 2012) indiquaient qu'une telle reconnaissance ne soulevait pas de problème important ou particulier. Notre expérience ne corrobore pas ce point de vue.

Comme deuxième sortie de l'allocation latente de Dirichlet (implémentation écrite en C par D.M. Blei), nous obtenons la répartition des thèmes dans les discours tenus par les treize présidents. Le tableau 6 indique les répartitions estimées.

Avec cet outil d'analyse, on peut voir apparaître quelques présidents qui concentrent leurs allocutions sous un thème. Par exemple, Nixon s'avère fortement lié (94 %) au deuxième thème (*peace*), Bush (fils) au quatrième (88,2 %, *Iraq, security*), Clinton au sixième (68 %, *community, school, health*), et Obama au dernier (83,4 %, *job, tax, health*).

On peut également voir une association entre un thème et les idées principales d'un parti, comme par exemple avec le troisième thème et les présidents Reagan et H. Bush (père), le binôme démocrate Clinton – Obama avec le sixième thème, ou Truman, Kennedy et Carter avec le huitième thème. Dans ce dernier cas, les sujets sous-jacents (*free*, *Soviet*, *defense*, *peace*, *economic*) ne s'associent pas de manière évidente au parti démocrate. Toutefois, on

peut se rappeler que le président est également le *Commander-in-Chief*, que celui-ci soit républicain ou démocrate.

	Th1	Th2	Th3	Th4	Th5	Th6	Th7	Th8	Th9	Th10
Roosevelt	0,0 %	0,0 %	0,0 %	0,0 %	**66,2 %**	0,0 %	0,4 %	2,0 %	31,3 %	0,0 %
Truman	11,3 %	0,1 %	0,0 %	0,0 %	10,8 %	0,0 %	33,9 %	33,1 %	10,5 %	0,2 %
Eisenhower	20,8 %	0,5 %	0,2 %	0,0 %	2,4 %	0,0 %	**63,3 %**	11,8 %	1,1 %	0,0 %
Kennedy	20,5 %	24,2 %	2,7 %	0,2 %	4,5 %	1,9 %	23,7 %	19,3 %	1,2 %	1,8 %
Johnson	40,7 %	36,5 %	1,5 %	0,2 %	1,3 %	8,0 %	0,6 %	7,1 %	2,7 %	1,4 %
Nixon	0,0 %	**94,0 %**	0,0 %	0,0 %	4,7 %	0,4 %	0,6 %	0,0 %	0,2 %	0,0 %
Ford	28,2 %	**53,5 %**	10,3 %	0,1 %	0,1 %	0,3 %	1,2 %	4,7 %	0,5 %	1,2 %
Carter	10,6 %	29,0 %	13,2 %	0,3 %	1,5 %	0,4 %	1,0 %	40,8 %	0,0 %	3,2 %
Reagan	2,1 %	9,4 %	**74,0 %**	0,5 %	0,5 %	0,2 %	2,4 %	10,2 %	0,0 %	0,6 %
HBush	0,0 %	3,9 %	**57,3 %**	5,6 %	0,6 %	4,9 %	1,2 %	17,2 %	0,6 %	8,9 %
Clinton	0,9 %	0,1 %	16,4 %	0,5 %	0,1 %	**68,0 %**	0,0 %	0,0 %	0,0 %	14,1 %
Bush	1,0 %	0,1 %	5,9 %	**88,2 %**	0,0 %	3,0 %	0,0 %	0,4 %	1,2 %	0,2 %
Obama	0,0 %	0,9 %	1,7 %	1,0 %	0,0 %	13,0 %	0,0 %	0,0 %	0,0 %	**83,4 %**

Tableau 6. Répartition en pourcentage des dix thèmes selon chaque président

5. Conclusion

En prenant comme domaine d'application l'ensemble des discours sur l'état de l'Union (1934-2014), notre étude propose d'en synthétiser les grandes tendances au moyen de quatre approches. En premier, nous pouvons retenir les lemmes les plus fréquents en ignorant les termes fonctionnels. Cette technique permet de dresser à grands traits les thèmes du corpus (e.g., *peace, job, economy, work, child*) ainsi que les formulations récurrentes (*America, Congress, government, people, country*). Les termes débutant par une majuscule permettent de situer le corpus dans l'espace (*Soviet, Europe, China*). Il s'avère par contre difficile de bien distinguer les particularités de chaque locuteur ainsi que leurs possibles affinités.

Comme deuxième approche, nous avons analysé la distribution des catégories grammaticales par président ainsi que la longueur moyenne des phrases. A ce niveau, nous avons le sentiment que la chronologie possède une grande influence comme l'ont démontré d'autres études linguistiques (Juola, 2003) (Hughes et al., 2012). Nous observons l'usage récurrent de phrases plus courtes, et le recours plus abondant aux pronoms. Une haute fréquence du groupe nominal tend à caractériser les discours de Roosevelt à Kennedy.

En recourant à une mesure de distance intertextuelle et à un outil associant classification et représentation graphique, nous pouvons détecter les relations plus étroites entre chaque allocution ou entre chaque président. De manière générale, les discours du même président ont tendance à se grouper entre eux. De plus, le facteur diachronique permet d'expliquer des rapprochements supplémentaires. Toutefois des exceptions demeurent comme les allocutions de Truman qui appartiennent à deux groupes distincts (correspondant à ses deux mandats : 1947 - 1950 et 1951 - 1953). Pour H. Bush (père), les quatre allocutions tendent à former un duo (1990-1991) et deux singletons (1989 et 1992) difficiles à attribuer à un groupe donné.

A un niveau plus élevé, la période 1934 - 2014 se subdivise en trois parties. En premier, on observe que les présidents Roosevelt, Truman et Eisenhower tendent à former un groupe séparé, et que Kennedy représente une présidence de transition. En deuxième, on retrouve les présidents Ford et Carter qui se rapprochent et, dans une moindre mesure, Nixon et Johnson qui suivent un mouvement similaire. Enfin, la dernière période débute avec Reagan et comprend des présidences relativement proches mais qui conservent leurs caractères distincts.

Enfin, nous avons retenu le modèle à thèmes (*topic model*) qui permet de synthétiser le corpus via un nombre donné de thèmes. Les grandes thématiques peuvent ainsi être révélées et, pour chaque document (ou groupe de documents), on peut estimer la distribution sur les divers thèmes. Cette approche permet de détecter des styles très personnels (comme ceux de Nixon, Obama ou Bush (fils)) ou, au contraire, ceux qui se retrouvent parmi deux ou plusieurs auteurs (Carter et Kennedy, ou Reagan et Bush H (père)).

Remerciements

Cette recherche a été financée par le Fonds national suisse pour la recherche scientifique (subside n^0 200020-129535). L'auteur tient à remercier D. Labbé et les relecteurs pour leurs commentaires pertinents sur une version préliminaire de cette communication.

Références

Baayen H.R. (2008). *Analysis Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge University Press, Cambridge.

Blei D.M. (2012). Probabilistic topic models. *Communications of the ACM*, 55, 77-84.

Blei D.M., Ng A.Y. et Jordan, M.I. (2003). Latent Dirichlet allocation. *Machine Learning Research*, 3, 993-1022.

Chang J., Boyd-Graber J., Gerrish S., Wang C., et Blei, D.M. (2009). Reading tea leaves: How humans interpret topic models. *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems*.

Delahaye J.-P. et Gauvrit N. (2013). *Culturomics. Le numérique et la culture*. Odile Jacob, Paris.

Hughes J.M., Foti N.J., Krakauer D.C. et Rockmore D.N. (2012). Quantitative Patterns of Stylistic Influence in the Evolution of Literature. *Proceedings of the PNAS,* 109(20), pp. 7682-7686.

Juola, P. (2003). The Time Course of Language Change. *Computers and the Humanities*, 37, 77-96.

Kaufman L. et Rousseeuw P.J. (1990). *Finding Groups in Data : An Introduction to Cluster Analysis*. Wiley Interscience, Hoboken.

Labbé D. (2007). Experiments on authorship attribution by intertextual distance in English. *Journal of Quantitative Linguistics*, 14(1), pp. 33-80.

Labbé D. et Monière, D. (2003). *Le discours gouvernemental. Canada, Québec, France (1945-2000)*. Honoré Champion, Paris.

Labbé D. et Monière D. (2008). *Les mots qui nous gouvernent. Le discours des premiers ministres québécois: 1960-2005*. Monière-Wollank, Montréal.

Labbé D. et Monière D. (2013). *La campagne présidentielle de 2012. Votez pour moi !* L'Harmatan, Paris.

Lebart L., Salem A. et Berry, L. (1998). *Exploring Textual Data.* Dordrecht, Kluwer.

Mayaffre D. (2004). Analyse logométrique de la cohabitation Chirac/Jospin (1997-2002). *Actes JADT 2004*, Louvain-La-Neuve, pp. 787-792.

Monière D. et Labbé D. (2006). L'influence des plumes de l'ombre sur les discours des politiciens. *Actes JADT 2006*, Besançon, pp. 687-696.

Paradis E. (2011). *Analysis of Phylogenetics and Evolution with R.* 2nd Ed., Springer, New York.

Pauli F. et Tuzzi, A. (2009). The end of year addresses of the Presidents of the Italian Republic (1948–2006): discoursal similarities and differences. *Glottometrics*, 18, pp. 40–51

Savoy J. (2012). Attribution d'auteur : Une approche basée sur l'allocation latente de Dirichet (LDA). *Actes JADT 2012*, Louvain-La-Neuve, pp. 897-909.

Toutanova K., Klein D., Manning C. et Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclid dependency network. *Proceedings of HLT-NAACL 2003*, pp. 252-259.