Apports de la lexicométrie à l'analyse des entretiens. Le cas de l'enquête " Les Français et la politique ". Avec une note d'Etienne Schweisguth et une réponse de l'auteur.

Dominique Labbé

Colloque DIME-SHS
"Des instruments au service de la recherche en sciences sociales"
Paris – 28 septembre 2018

Apports de la lexicométrie à l'analyse des entretiens. Le cas de l'enquête « Les Français et la politique ». Avec une note d'Etienne Schweisguth et une réponse de l'auteur.

Communication à la Table-ronde
"Analyse textuelle en renfort de l'exploitation des données"

Dominique Labbé
PACTE
(CNRS Université Grenoble-Alpes)
dominique.labbe@umrpacte.fr

Résumé

Dépouillement lexicométrique des 64 entretiens de l'enquête « les Français et la politique » pilotée par Etienne Schweisguth en 1983 : balisage des textes, standardisation des graphies, étiquetage des mots, longueur des textes, principales caractéristiques de leurs vocabulaires décrites à partir du tableau lexical, des index hiérarchiques, des concordances et des syntagmes répétés. Le calcul des distances entre textes et leur classification révèlent l'influence de l'enquêteur sur les propos de l'enquêté. En conclusion, un appel à constituer une grande base de transcriptions d'entretiens. En annexe, une note d'E. Schweisguth et une réponse de l'auteur.

Au printemps 2018, le Réseau Quetelet et le CDSP (SciencesPo) nous ont remis 64 entretiens - menés entre mars et décembre 1983 auprès de Français en âge de voter - portant sur le thème « Les Français et la politique » (enquête FP dans la suite de cette communication). Pour une présentation de ces documents : Garcia 2013[1] et pour une première exploitation, les quatre publications d'Etienne Schweisguth (E. S. dans la suite) qui pilotait cette étude et a réalisé la majorité des entretiens.

Une question nous a été posée par les organisateurs de cette journée d'étude : « Que peut apprendre la statistique lexicale à propos de ce corpus ? » Et plus largement : « Quelle aide peuvent apporter les mathématiques appliquées, la lexicologie et l'informatique aux chercheurs en sciences sociales dans le dépouillement des textes et plus particulièrement les transcriptions d'entretiens ? »

Cette communication présente quelques éléments de réponse à ces questions. Remarquons au préalable que :

- l'intérêt de ces méthodes est maintenant assez bien connu (pour une synthèse : Grimmer & Stewart 2013) et son application aux données textuelles issues des enquêtes a fait l'objet de publications assez nombreuses (par exemple : Brugidou 2008). Nous nous concentrons ici sur les apports spécifiques de la lexicométrie.

- les 64 transcriptions – d'excellente qualité - comportent un numéro et un titre donné par E. S. (par exemple n°11 : « GaucheHuma1 »). L'annexe 1 donne la correspondance entre numéros et titres. Pour une analyse automatique sans aucune hypothèse préalable, tous les traitements ont été effectués sur les textes sans prendre en considération les autres renseignements. En quelque sorte, les programmes ont travaillé en « aveugle ».

- tous les traitements sont confiés à des automates ; les calculs suivent des procédures et des formules publiées, de telle sorte que les résultats présentés dans cette note sont contrôlables et reproductibles.

- ces méthodes ont déjà été appliquées à quelques collection d'entretiens, notamment : Pionchon 2001 ; Labbé & Labbé 2001 ; Labbé D. 2002.

Enfin, à part deux ou trois notations, nous nous contentons de présenter les données sans commentaire ou interprétation. Auparavant, voici quelques mots à propos des traitements opérés sur les textes

I. *Trois opérations préalables*

Une présentation de ces trois opérations, sur un corpus d'entretiens, est consultable en ligne (Labbé D. 2002). En résumé :

Balisage

En tête de chaque texte, on place des « balises » indiquant le titre de l'enquête, la date et un numéro qui renvoie à un autre fichier dans lequel figurent les caractéristiques de l'enquêté(e) :

[1] Voir les références placées à la fin de ce texte. La plupart des documents cités sont consultables en ligne.

sexe, âge, profession, situation familiale, niveau d'étude, commune de résidence, votes, etc. Naturellement, l'anonymat de l'enquêté est respecté et tout élément susceptible de l'identifier ont été effacés dans le corps du texte. Une autre balise indique la date des traitements lexicométriques effectués sur le texte et le nom de l'opérateur ayant supervisé ces traitements.

Dans le corps du texte des balises, reconnaissables par l'ordinateur, identifient les propos de l'enquêteur et ceux de l'enquêté. Dans la suite de cette note, seuls les propos des enquêtés seront analysés mais nous verrons que les propos de l'enquêteur ne manquent pas d'intérêt…

Standardisation des graphies

Dans tout texte imprimé, les variantes graphiques concernent plus d'un mot sur dix (sans compter les erreurs d'orthographe, les majuscules initiales de phrases, les noms communs affublés à tort d'une majuscule). Si l'on opère des calculs sur le texte « brut » sans aucune standardisation des graphies – comme le font les logiciels usuels d' « analyse des données textuelles » - cela signifie que les résultats obtenus sont affectés d'une incertitude d'au moins 10%... ou encore, que les analyses sur différents corpus sont difficilement comparables.

Naturellement, la quasi-totalité de cette standardisation est réalisée automatiquement, les interventions de l'opérateur étant limitées au maximum et strictement encadrées.

Une fois les textes balisés et les graphies standardisées, on procède à leur étiquetage.

Etiquetage

Chacun des mots du texte reçoit une étiquette comportant sa graphie standardisée, et son vocable (entrée de dictionnaire et catégorie grammaticale). Par exemple, en français, toutes les flexions d'un verbe (modes, temps, personnes) sont regroupées sous l'infinitif de ce verbe ; de même les substantifs sont identifiés par leur genre (président/présidente). Ainsi, "le politique" peut être un homme ou un concept mais pas une action ou une femme...

Autrement dit, cette étiquette replace le mot dans le lexique de la langue, selon les conventions généralement acceptées par les locuteurs de cette langue. Quand l'ordinateur a commencé à être utilisé pour le dépouillement des textes, il a été proposé d'utiliser cette nomenclature « lexicale », notamment par C. Muller dans un article de 1963 et dans son manuel de 1977. Telle est la raison pour laquelle nous parlons de « statistique lexicale », ou mieux de « lexicométrie ». Nous avons réalisé l'implémentation de ces normes lexicographiques dans des algorithmes et des programmes informatiques (présentation : Labbé D. 1990 ; Pibarot & Al. 1995).

Enfin, ces opérations n'altèrent en rien le texte original. Elles permettent de recueillir des informations intéressantes qui éclairent ce texte et facilitent son analyse. En voici quelques exemples.

II. Longueurs et vocabulaires des entretiens

On tire d'abord de ces corpus « étiquetés » des inventaires portant sur le volume des documents traités (longueurs) puis sur leurs vocabulaires.

La longueur

Les réponses des 64 enquêtés comportent au total 722 888 mots et un vocabulaire de 10 475 vocables différents (annexe 1). Ces caractéristiques - que l'on pourrait qualifier de « physiques » - des textes et des corpus devraient figurer dans toutes les recherches afin de permettre les comparaisons.

La longueur moyenne des entretiens est de 11 295 mots et leur longueur médiane 10 269 mots. Le faible écart entre ces deux valeurs indique que les longueurs sont réparties à peu près également des deux côtés de la moyenne.

En une heure, un enquêté prononce en moyenne environ 7 000 mots (sans compter les propos de l'enquêteur). La durée moyenne des entretiens a donc excédé une heure et demie. Le total représente plus d'une centaine d'heures de face-à-face.

Le plus court (n° 23) comporte 3 521 mots et le plus long (124) 36 567. La durée a donc varié de à peine plus d'une demie-heure à plus de 5 heures… L'étendue de la distribution est de 1 à 10, ce qui semble considérable, mais en mettant à part les trois plus longs (124, 109 et 145), l'étendue de la distribution n'est plus que de 1 à 4,8, suggérant une relative homogénéïté des longueurs.

Que signifient ces dimensions ?

D'abord que le corpus FP représente un volume considérable. A titre de comparaison voici la longueur de deux des plus longs romans de la langue française : Hugo (*les Misérables)* : 564 301 mots ; Eugène Sue (*Les mystères de Paris*) ; 578 933 mots.

Le traitement approfondi d'une telle masse de mots par des méthodes purement manuelles est impossible. La statistique lexicale est donc un outil indispensable pour défricher ces entretiens, avant toute interprétation.

Les premiers renseignements fournis au chercheur portent sur le vocabulaire.

Le vocabulaire

Chacun des textes se voit associer un « index » : liste des vocables avec les formes sous lesquelles ils apparaissent et leurs effectifs (ou fréquence absolue). La fusion de ces index donne le tableau lexical complet selon la même architecture. Le tableau 1 en donne un extrait (en négligeant les 64 colonnes correspondant aux textes).

Tableau 1. Extraits de l'index alphabétique complet du corpus FP.

Vocables et formes	Effectifs totaux
est (n m)	30
(…)	
été (n m)	13
(…)	
être (v)	32 334
est	18 433
été	1 613
sommes	92
suis	2 135
(…)	
pouvoir (v)	3 174
pouvoir	195
pouvoir (n m)	363
pouvoir	327
(…)	
somme (n f)	42
somme	31
sommes	11
(…)	
suivre (v)	
suis	26

Par exemple, le verbe *être* est présent 18 433 fois sous la forme « est » et le nom masculin « est » (le point cardinal) apparaît 30 fois. De même, il y a 26 *suis*, verbe *suivre*, noyés parmi les 2 135 *suis* du verbe *être*. Ces 26 *suivre* surviennent dans 18 entretiens, dont 11 – soit près d'un sur six – sont un aveu : « je (ne) suis pas la politique » (ou les nouvelles). L'une (n° 117) le répète cinq fois. Beaucoup d'autres enquêtés laissent transparaître leur ignorance, ou leur désintérêt, mais peu osent l'avouer aussi franchement.

Evidemment, il est impossible de retrouver, à la main, ces cas singuliers dans la mer des formes homographes du verbe, de même pour la belle saison (*été*), les *sommes* ou le *pouvoir* (qui intéresse particulièrement les politologues).

On remarquera que ces exemples ne sont pas anecdotiques : dans tout texte en langue française, *être* est toujours le verbe le plus employé et, d'autre part, les « homographies » concernent plus du tiers des mots. Par conséquent, seul l'étiquetage des mots permet des recherches sur le vocabulaire des textes. Sinon, le chercheur est confronté à un épais brouillard de « formes graphiques ».

Evidemment, ce tableau lexical est très grand et, dans un premier temps, on peut souhaiter limiter l'examen au vocabulaire usuel grâce aux index hiérarchiques qui classent les vocables non plus en fonction de l'ordre alphabétique mais selon leur fréquence. Le tableau 2 ci-dessous donne le début de l'index hiérarchique.

Tableau 2. Index hiérarchique du corpus FP.

Rang	Vocable	Effectifs	Fréquence (‰)
1	le (det)	47 765	66.08
2	de (pré)	34 504	47.73
3	être (v)	32 339	44.74
4	je (pro)	26 041	36.02
5	avoir (v)	25 005	34.59
6	ce (pro)	18 982	26.26
7	il (pro)	15 859	21.94
8	pas (adv)	14 835	20.52
9	à (pré)	14 226	19.68
10	que (cj)	13 901	19.23
(…)			
20	faire (v)	6 728	9.31
21	dire (v)	6 180	8.55
(…)			
53	gens (n m)	2 502	3.46

Cet extrait permet de comprendre deux caractéristiques de tout texte en langue naturelle :

- la surface est essentiellement occupée par des articles, pronoms, adverbes et verbes auxiliaires. En revanche, les verbes (autres que *être* et *avoir*), les substantifs et les adjectifs sont beaucoup plus discrets. Il faut descendre au vingtième rang pour renconter le premier verbe (*faire* puis *dire*) et au 53^{e} rang pour trouver le premier substantif (*gens*) qui apparaît en moyenne une fois tous les 289 mots (3,46 ‰). Il est parfois proposé de négliger les « mots outils ». A cela deux objections principales. Ces mots couvrent plus de la moitié de la surface des textes, or une analyse scientifique doit commencer considérer l'ensemble du matériel sans hypothèse *a-priori* sur ce qui est significatif et ne l'est pas. De plus, pourquoi éliminer les

pronoms personnels alors qu'il est admis qu'ils jouent un rôle central notamment dans les situations d'interlocution (pour un exemple : Arnold & Labbé 2015), spécialement le « je »…

- les densités sont faibles : dès le 20e mot (« faire ») la fréquence est inférieure à 1%. C'est la raison pour laquelle on raisonne en « pour mille » (‰). La statistique lexicale étudie un grand nombre d'événements rares survenant dans de vastes populations. Pour de telles analyses, la plupart des calculs usuels – notamment en probabilité - sont peu adéquats.

Enfin, si l'on postule que la fréquence est un indice de l'importance accordée à une notion, il faut prendre garde à ce que certaines choses peuvent être exprimées de différentes manières. Prenons un exemple simple : il y a 282 « Mitterrand » et 238 « Giscard ». Cependant, on aurait tort d'en conclure qu'il est plus question du premier que du second car il y a également 54 « Giscard d'Estaing »…

On trouvera en annexe 2 la liste des vocables les plus utilisés dans ce corpus classés par catégories grammaticales. Chacune de ces listes comporte de nombreux enseignements… Par exemple, celle des verbes illustre des phénomènes lexicaux peu étudiés. Dans tout texte en français, *être, avoir* et *faire* occupent, dans cet ordre, les trois premières places dans la liste des verbes. Ensuite, le rang de « dire » et de « aller » est typique de la communication orale. Le verbe *aller* est un « pseudo-auxiliaire » marquant le futur sans avoir à utiliser ce temps (ou pour s'épargner la concordance des temps) : « je vais dire ». Dans le discours politique, la modalité de la nécessité (*falloir*) et de la volonté (*vouloir*) l'emporte habituellement sur celles du possible (*pouvoir*), de la connaissance (*savoir*) et de l'obligation (*devoir*) (Labbé & Labbé 2013). Ainsi se marque l'une des principales singularités du discours politique par rapport au langage de la vie courante…

Un tel classement repose sur un postulat implicite : l'importance d'un vocable se mesure au volume de ses emplois. Or les 64 entretiens commençaient toujours sur le thème suivant : « Qu'est-ce que la gauche et la droite selon vous ? » et les deux interviewers relançaient assez systématiquement à propos de la coupure droite/gauche. Certes ces deux mots figurent bien dans la liste de tête des substantifs mais ils sont largement précédés par «gens », « chose » et « an ».

Cela n'a rien d'anormal car, dans tout discours oral improvisé sur quelque sujet que ce soit, *gens* puis *chose* figurent en tête des substantifs les plus utilisés. Le premier a la même fonction que *people* en anglais, désignant à la fois le peuple, les concitoyens, les autres, voire un groupe de personnes que l'on a du mal à nommer… Quant à *chose*, il peut se substituer à tout objet dont le nom n'est pas immédiatement présent à l'esprit du locuteur. Chez les hommes politiques, ces deux mots sont très peu employés, même dans les entretiens « impromptus », preuve que les professionnels de la politique n'improvisent pas et utilisent des « éléments de langage » soigneusement préparés.

Cette discussion permet également de comprendre qu'il faut prendre en considération non pas les mots isolément mais en contexte.

Les combinaisons de mots

Un mot tire son sens de sa place dans le lexique et de ses combinaisons avec les autres mots dans les propos tenus par chacun des enquêtés. Pour retrouver ces combinaisons, on utilise d'abord les « concordances ».

Par exemple, dans le corpus FP, l'index indique qu'il y a 1034 occurrences de la forme « politique(s) » et que :
- 634 sont des substantifs féminins : *la, une, les politique(s)* ;

- 389 sont des adjectifs (par exemple «parti politique »)
- 11 des substantifs masculins (« le, un politique »).

Quel emploi font les enquêtés de ce substantif masculin ? Le « contexte » de ces 11 emplois est connu grâce au « concordancier » (annexe 3) qui montre qu'il s'agit autant de(s) homme(s) politique(s) que de l'abstraction au cœur de la science politique (*le politique*). L'index montre également que les emplois péjorants pour désigner les politiques sont rares, notamment 3 *politicard* ; 13 *politicien* (n m) et 14 *politicien* (adj.).

Ces concordances sont donc utiles pour dénouer la polysémie de certains mots. Par exemple, le « pouvoir » : s'agit-il toujours du pouvoir politique ? Sur les 363 occurrences de ce substantif, le tableau ci-dessous donne les principales combinaisons dans lesquelles il apparaît. Il suffit de demander au concordancier les substantifs, adjectifs et verbes utilisés devant ou derrière « pouvoir » en négligeant les prépositions, déterminants, adverbes situé entre eux (qui sont placés dans le tableau 3 entre parenthèses), pour les verbes, toutes les flexions sont regroupées sous l'infinitif.

Tableau 3. Principales combinaisons avec le substantif « pouvoir ».

syntagmes	Effectifs absolus	% de « pouvoir »
être (au) pouvoir	80	22,0
avoir (le) pouvoir	49	13,5
pouvoir (d') achat	49	13,5
gauche (au) pouvoir	27	7,4
prendre (le) pouvoir	16	4,4
gens (au) pouvoir	13	3,6
total	234	64,5

Les six premiers syntagmes couvrent à eux seuls pratiquement les deux tiers de tous les emplois du substantif « pouvoir ». En dehors de « pouvoir d'achat » (49 occurrences), il s'agit bien du *pouvoir politique*.

En généralisant cette recherche à tous les substantifs et verbes usuels, on obtient les « syntagmes répétés » qui permettent d'accéder simplement aux principaux thèmes présents dans un corpus (Pibarot & Labbé 1998). Les principaux syntagmes du corpus FP sont reproduits en annexe 2.

Ces outils simples mais puissants apporteraient sans doute une aide utile aux chercheurs en sciences sociales qui sont, pour la plupart, confrontés à des textes dont la masse est trop importante pour une exploitation manuelle.

Des exploitations plus sophistiquées sont possibles, notamment des classifications.

III. Classifications

En lisant les articles d'E. Schweisguth cités au début de cette communication, on se rend compte qu'il a eu un certain mal à classer ses entretiens sur l'axe droite/gauche et qu'il existe un résidu important rangé dans la rubrique « divers ». Dès lors, pourquoi ne pas recourir à des outils statistiques pour classer ces entretiens en fonction de leur seul contenu et sans faire intervenir d'autre élément (comme le genre de l'enquêté, sa profession ou son vote) ?

La classification automatique répond à cet impératif. Elle opère des regroupements dans de vastes populations sans aucune intervention de l'observateur. La meilleure classification possible sera celle qui minimise les distances entre les individus classés ensemble et qui maximise les distances entre les différents groupes ainsi créés (Sneath & Sokal 1973).

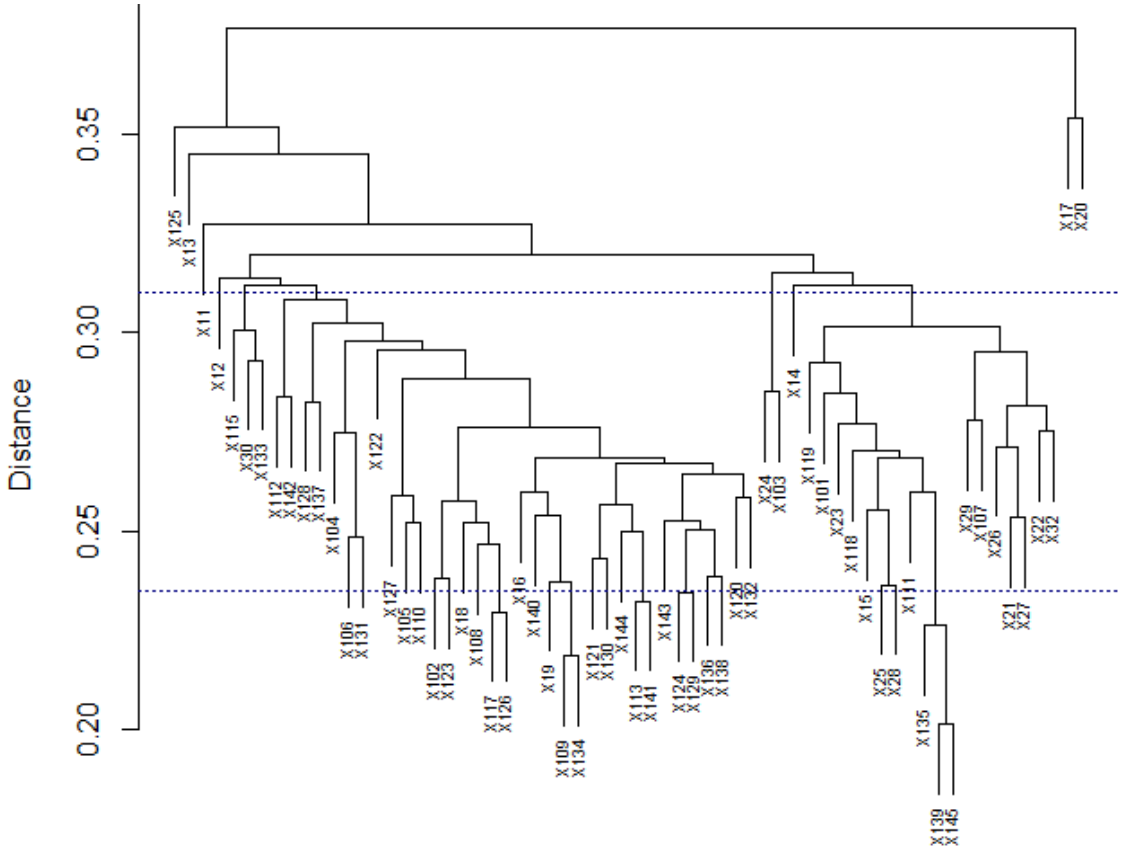
En effet, toute classification repose sur le calcul d'une distance séparant les objets à classer. Le calcul de la distance entre deux textes (distance « intertextuelle ») consiste à superposer leurs vocabulaires et à compter les différences (Labbé & Labbé 2003). Sous certaines conditions (remplies dans le cas présent), la distance intertextuelle présente les caractéristiques d'une distance dans un espace euclidien, ce qui permet de réaliser des classifications représentant, sans déformation, la population étudiée.

Les 64 entretiens comparés deux à deux génèrent 2016 couples différents. Pour représenter ces 2016 comparaisons, deux méthodes principales sont disponibles (pour les algorithmes graphiques : Paradis 2006).

La classification hiérarchique ascendante

L'algorithme commence par regrouper les deux textes les plus proches (ici les n° 139 et 145) puis il calcule les distances entre ce groupe et tous les autres textes en effectuant la moyenne arithmétique simple des distances originelles et ainsi de suite jusqu'à la formation d'un ensemble unique. Ces groupements successifs sont représentés dans un "dendrogramme" avec en ordonnées les distances relatives correspondants aux différents niveaux d'agrégation (Figure 1) (pour la méthode : Roux 1985).

Figure 1. Classification hiérarchique ascendante par la méthode de la moyenne des distances.



Graphique réalisé avec R (version 3.5.1 – librairie APE) août 2018

L'ordre dans lequel sont présentés les textes sur l'axe horizontal n'a pas d'importance. Seule compte la hauteur du trait horizontal qui marque la distance entre les différents textes groupés sous ce trait. Il suffit donc de couper le graphique à différentes hauteurs significatives pour identifier les textes les plus proches au seuil de confiance choisi. Sur la figure ci-dessus, sont portés les deux traits horizontaux correspondant à la moyenne des distances (0,313) et à cette moyenne diminuée de deux écarts-types (0,239).

Ce second seuil permet d'isoler quelques couples très proches : en utilisant la totalité du corpus comme norme, on a moins de 5% de chance de se tromper en affirmant que ces entretiens sont « anormalement » proches (entre parenthèse, leur classification par E. S.) :
139 – 145 – 135 (GaucheHuma58, GaucheHuma64 et DroiteLibéral54)
109 – 134 (DroiteRigo30 et DroiteRigo53)
117 – 126 (Divers36 et Divers 45)
113 – 141 (Divers34 et DroiteRigo60)
124 – 129 (Divers43 et Petits48)

Le deuxième seuil (la moyenne) isole deux groupes mais laisse à part une dizaine d'entretiens.

Dans les deux cas, on constate que les classes constituées ne correspondent pas toujours à celles qu'avait effectuées E. S. (cette question est discutée plus bas).

Cette première méthode de classification présente une difficulté qu'illustre l'importance du résidu non classé et notamment les textes n°17 et 20 qui sont groupés ensemble en haut à droite à l'écart de tous les autres. Ces textes sont-ils réellement atypiques et n'ont-ils aucun lien avec les autres comme semble le suggérer le graphique ? Avec la classification hiérarchique ascendante, il est impossible de répondre car la méthode produit des "effets de chaîne" qui effacent les liens individuels entre les textes groupés et ceux qui restent en dehors de ce groupe. Ces inconvénients sont moins sensibles dans la classification arborée.

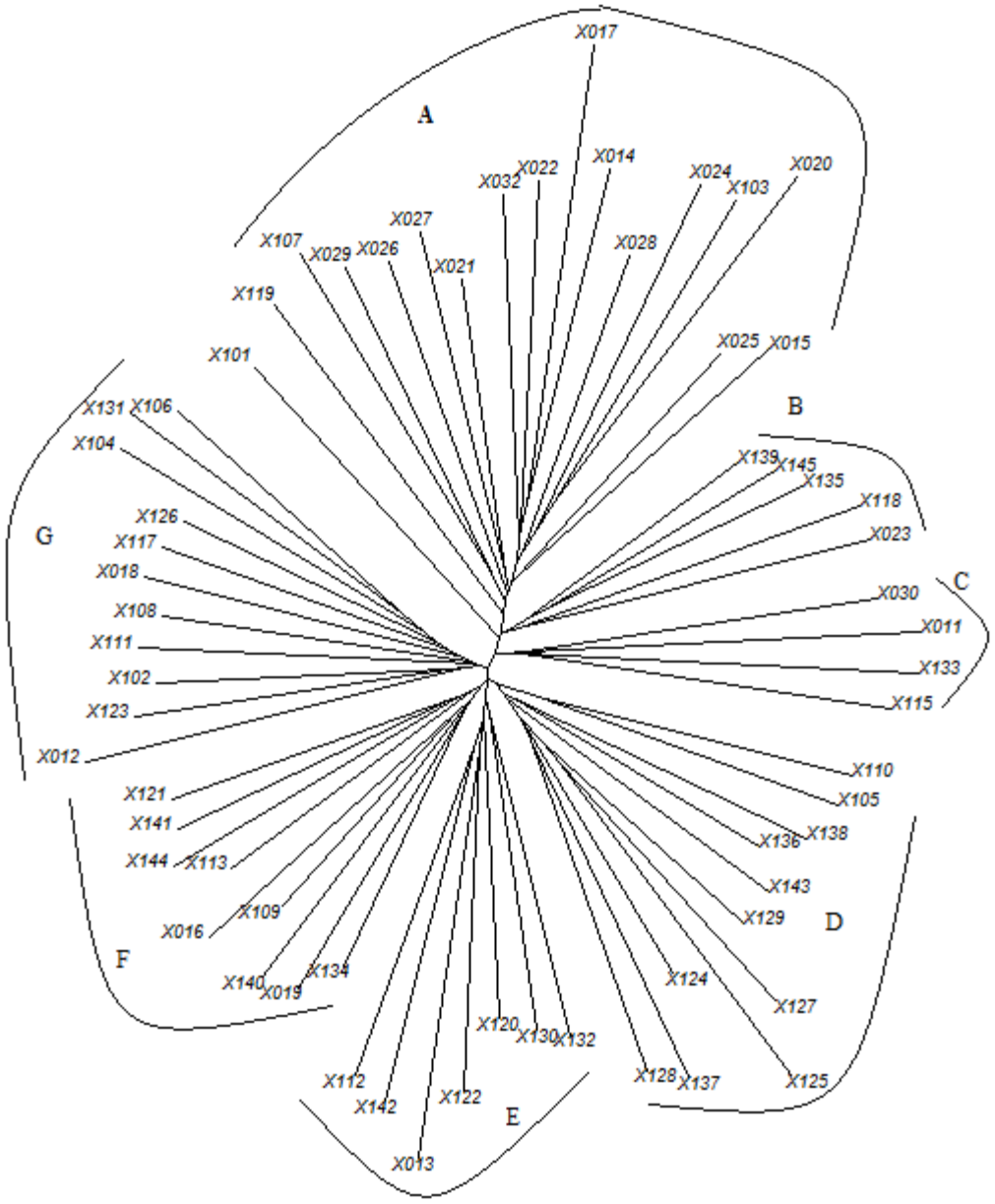
Classification arborée

La "classification arborée" est classique en génétique (Felsenstein 2004) ou en linguistique historique (Embleton 1986, Holm 2007). Elle repose sur la propriété suivante : si les distances séparant les individus étudiés présentent les propriétés requises d'une distance dans un espace euclidien, il existe un "arbre" qui représente, le mieux possible sur un plan, les positions respectives de ces individus les uns par rapport aux autres et les meilleurs groupements possibles entre eux. L'algorithme est présenté dans X. Luong (1988 et 1994). Voir également, Barthélémy & Guénoche 1988 ; Labbé & Labbé 2006.

La figure 2 présente l'arbre du corpus FP.

Dans cette figure, les feuilles terminales figurent les textes ; les nœuds intermédiaires donnent les meilleurs groupements possibles, c'est-à-dire ceux pour lesquels les distances entre les éléments qui composent le groupe sont les plus faibles et les distances les séparant des autres les plus grandes possibles. La distance entre deux points quelconques est figurée par le chemin unissant ces points et la longueur de ce chemin est proportionnelle à la distance originelle correspondante. Il ne faut pas attacher d'importance au placement des branches rattachées à un même nœud, spécialement pour les branches terminales.

Figure 2. Classification arborée du corpus FP



Graphique réalisé avec R (version 3.5.1 – librairie APE) août 2018
Dans le sens des aiguilles d'une montre :
A { 107, 29} 26 { 27, 21 } 32, 22 { 17, 14 } { 28, 24, 103, 20 } { 25, 15 }
B { 139, 145, 135}, { 118, 23 }
C { 30, 11, 133, 115 }
D { 110, 105 } { 138, 136, 143 } { 129, 127, 125, 124, 137, 128 }
E { 132, 130 }, 120, {122, 13 } {142, 112 }
F { 134, 19, 140, 109, 16 } { 113, 144, 141, 121 }
G { 12, 123, 102} 111, 108, 18, 117, 126 { 104, 131, 106 }
Non classés : 101, 119

Autre intérêt : la qualité de la classification (Labbé et Labbé 2006). Ici tous les groupes sont agrégés à partir de nœuds dont l'indice de qualité est supérieur à 95%. Autrement dit, il y a moins de 5% de chances de se tromper en affirmant que les 7 groupes isolés par la

classification correspondent au « meilleur » classement possible d'ordre 7 et que, à ce seuil de confiance, il n'est pas possible d'aller plus loin dans les regroupements.

Cette analyse ne contredit pas la précédente mais ne souffre pas des mêmes limites. Par exemple, l'arbre rétablit les proximités (relatives) des textes n° 17 et 20 avec les 13 autres formant le groupe A et notamment les n° 103, 24 et 14 qui étaient également « mal classés » dans le dendrogramme. Pour le reste, tous les couples isolés par la classification hiérarchique se retrouvent dans un même groupe sur l'arbre.

Une fois les 7 groupes constitués, on recherche, pour chacun, son vocabulaire caractéristique par rapport à tous les autres (en suivant la méthode présentée par Lafon 1984). Mais avant de faire ce travail, il faut s'interroger sur la portée de ces classifications. La distance intertextuelle entre deux textes dépend de quatre facteurs principaux. Dans l'ordre d'importance : le genre, l'auteur, l'époque, les thèmes. Dans le corpus FP, deux de ces facteurs sont neutralisés. Tous les textes appartiennent au même genre (oral spontané en situation de conversation) et ils ont été enregistrés à la même époque. Par conséquent, la distance entre deux entretiens ne devrait provenir que des différences de style individuel (auteur) et des thèmes traités par chacun.

En ce qui concerne l'influence de l'auteur, le corpus FP offre une illustration amusante. Les deux textes les plus proches sont les n°139 et 145. C'est un couple dont E. S. a interrogé séparément l'homme (n° 139) puis la femme (n° 145) mais au cours du premier entretien, la femme est intervenue pour environ 10% du texte. En plus des influences mutuelles, il y a donc présence partielle du même auteur dans les deux entretiens – même si cette personne ne tient pas les mêmes propos les deux fois -, ce qui se traduit par une distance plus faible que dans le cas de deux auteurs différents et indépendants l'un de l'autre…

Nous ne pousserons pas davantage cette analyse car un autre facteur intervient et brouille la classification.

IV. L'influence de l'enquêteur sur le contenu de l'entretien.

Les entretiens 11 à 32 ont été réalisés par un enquêteur professionnel (X). Les numéros 101 à 145 ont été menés par Etienne Schweisguth (E. S.). Abrial et Al. (2017) ont noté des différences significatives entre ces deux séries. Nous avons relevé deux caractéristiques, au moins, qui confirment cette conclusion :

Premièrement, la classification sépare nettement les deux séries. En effet, en l'absence d'influence de l'enquêteur, on attend à peu près la proportion « un X pour deux E.S. » (21/43) dans chaque groupe. La légende de la Figure 2 indique que la composition de tous les groupes diffère significativement de cette proportion ; le groupe A contient 13 des 21 entretiens menés par X. A l'inverse, le groupe D compte le tiers des entretiens de E. S. (11) et aucun de X…

Deuxièmement, la longueur des entretiens diffère significativement. Les entretiens menés par X ont une longueur moyenne de 8 848 mots contre 12 490 mots pour ceux de E. S. (soit 40 % de plus) ;

Deux indices confirment l'influence de l'enquêteur :

- les distances moyennes entre les entretiens menés par E. S. sont un peu plus faibles que celles séparant les entretiens de X. Autrement dit, E. S. a un peu plus « recadré » ses interlocuteurs sur le thème de l'entretien que ne l'a fait X ;

- enfin, comme l'ont remarqué Abrial et Al. (2017), ES relance plus systématiquement ses interlocuteurs (ce qui peut expliquer que ses entretiens sont plus longs).

Certes, avec des échantillons de cette taille, il est impossible d'écarter *a priori* l'hypothèse selon laquelle ES aurait rencontré des gens différents de ceux interrogés par X (plus bavards et plus proches entre eux) mais le dernier indice (le nombre de relances) permet d'écarter cette hypothèse et de conclure à l'influence des enquêteurs. Il est même probable que cette influence soit le principal facteur explicatif des classifications présentées ci-dessus (avant les thèmes et les personnalités). Cette conclusion explique pourquoi nous n'avons pas creusé davantage l'analyse des « clusters » découpés par les deux classifications.

En fonction de ce constat, il faudrait séparer le corpus en deux sous-ensembles (les entretiens de X. et ceux de E. S.) et refaire les analyses ci-dessus. Au passage, cela permettrait aussi de savoir précisément sur quoi a porté cette fameuse « influence » de l'enquêteur[2].

Conclusions

Beaucoup d'autres questions n'ont pu être abordées dans le cadre restreint de cette communication. Citons notamment :

- le sens des mots. Par exemple, quelle(s) signification(s) donnent les enquêtés – ou certains d'entre eux – à des vocables polysémiques comme « droite », « gauche » ou « politique (n f) » ? Au-delà des concordances présentées dans cette communication, la méthode des univers lexicaux apporte une réponse plus approfondie (présentation dans Labbé & Labbé 2005).

- existe-t-il des différences selon le genre des enquêtés et comment mesurer ces différences ? (il y a 21 femmes et 43 hommes).

- l'influence du niveau d'étude, et plus largement du « capital culturel » (longueur des entretiens, richesse du vocabulaire, plus ou moins grande réticence envers la politique, etc) ;

- l'identification des thèmes et de leurs poids respectifs… Par exemple, dans les 1 500 réponses à une question ouverte dans un sondage, cette méthode isole 8 thèmes principaux et reclasse chaque réponse en fonction de l'adhésion ou du rejet de l'enquêté face à chacun de ces thèmes (Labbé & Labbé 2012).

Cette liste n'est pas limitative. Chacun pourra ajouter de nouvelles recherches : il suffit de formuler la question d'une manière qui soit « programmable » afin de la traduire en modèles, tests statistiques, algorithmes et programmes informatiques. L'objectif étant toujours le même : fournir le maximum de données au chercheur et retarder le moment où l'on libère son imagination…

Néanmoins, nous espérons avoir montré combien la lexicométrie peut être un outil utile pour la recherche en sciences sociales, spécialement pour l'analyse du matériel verbal recueilli lors des entretiens ou dans les questions ouvertes des sondages.

Le principal coût réside dans une transcription soignée de l'intégralité des entretiens et la correction des erreurs d'orthographe. En l'absence de logiciels efficaces de reconnaissance de la parole, ce travail est à la charge du chercheur…

[2] E. Schweisguth a bien voulu nous communiquer à ce sujet des éléments inédits qu'on lira en annexe de cette communication.

Enfin, cette présentation plaide pour un archivage des entretiens sociologiques dans une véritable base de données comme celle offerte par le Réseau Quételet. On pourrait aussi communiquer une version standardisée, balisée et étiquetée…

Quelle utilité aurait une telle archive ?

— avant de lancer une nouvelle enquête, elle permettrait un débroussaillage du problème, une réflexion sur les questionnaires (ou les guides d'entretien), et sur la sélection des enquêtés ;

— elle fournirait une base de comparaison pour l'analyse des résultats de la nouvelle enquête. Qu'apporte-t-elle ? Que confirme-t-elle ? Quelles "nouveautés" annonce-t-elle ? Avec la possibilité d'un regard rétrospectif, cette base serait un moyen précieux pour identifier les tendances lourdes au sein de la population, ce qu'une enquête isolée ne peut faire ;

- en l'absence d'enquête d'usage récente - prolongeant l'étude pionnière de Gougenheim (1956 et 1958) — cette base pourrait offrir un outil de connaissance de la langue française telle qu'on la parle effectivement. C'est dans cette optique que nous nous sommes intéressés à l'enquête d'Etienne Schweisguth que nous remercions.

Reconnaissance

Etienne Schweisguth a piloté l'enquête « les Français et la politique », réalisé la majeure partie des entretiens et les a mis dans le domaine public. Il a bien voulu relire cette communication et nous communiquer ses réactions qu'on lira ci-dessous.

Le Réseau Quételet et le Centre de Données Socio-Politiques de SciencesPo (Paris) nous ont communiqué ces documents.

Depuis 25 ans, Cyril Labbé (Laboratoire d'Informatique de Grenoble) a aidé à la mise au point des logiciels.

Edward Arnold, Guy Bensimon, Jean-Guy Bergeron, Mathieu Brugidou, Pierre Hubert, Cyril Labbé, Nelly & Jean Leselbaum, Xuan Luong, Thomas Merriam, Denis Monière, Gaétan Paéquin, Jacques Picard, André Pibarot, Mathieu Ruhlman et Jacques Savoy ont collaboré à la mise au point des outils de lexicométrie.

Certains de nos corpus étiquetés sont disponibles en libre accès auprès du Centre de Linguistique de Corpus (Université de Neuchatel).

Les logiciels de lexicométrie peuvent être obtenus sur demande auprès de Dominique Labbé.

Références

Toutes nos communications citées dans ce texte sont disponibles en ligne dans les archives ouvertes du CNRS (HAL) et ResearchGate.

Abrial Stéphanie, Brugidou Mathieu & Salomon Annie-Claude (2017). Types idéologiques et classe résiduelle dans l'enquête d'Etienne Schweisguth; les Français et la politique, 1982-1988. Réanalyse à partir de deux familles de logiciels, CAQDAS et ADT. ARQ

(Association pour la Recherche Qualitative). *La réanalyse des enquêtes qualitatives à l'épreuve de l'expérimentation*, Hors-Série "Les Actes" (21), pp.29-53.

Arnold Edward & Labbé Dominique (2015). Vote for me. Don't vote for the other one. *Journal of World Languages*. Routledge, 2015, p. 1-18.

Barthélémy Jean-Pierre et Guénoche Alain (1988). *Les arbres et les représentations de proximité*. Paris, Dunod.

Brugidou, Mathieu (2008). *L'opinion et ses publics. Une approche pragmatiste de l'opinion publique*. Paris : Presses de Sciences Po.

Embleton Sheila M. (1986). *Statistics in Historical Linguistics*. Bochum : Brokmeyer.

Felsenstein Joseph (2004). *Inferring Phylogenies*. Sunderland : Sinauer Ass.

Garcia Guillaume (2013). *L'inventaire de l'enquête : Les Français et la politique (Enquête par entretiens individuels approfondis réalisée par Etienne Schweisguth avec des citoyens anonymes sur le thème du clivage gauche droite)*. Paris : beQuali (CDSP UMS 828 CNRS-Sciences Po) : 2013.

Gougenheim Georges, en collaboration avec Michea René, Rivenc Paul, Sauvageot Aurélien (1956). *L'élaboration du français élémentaire : étude sur l'établissement d'un vocabulaire et d'une grammaire de base*. Paris : Didier. Réédition augmentée en 1964 sous le titre : *L'élaboration du français fondamental. Etude sur l'établissement d'un vocabulaire et d'une grammaire de base*, Paris : Didier.

Gougenheim Georges (1958). *Dictionnaire fondamental de la langue française*. Paris : Didier. Nouvelle édition revue et augmentée, Didier, Paris, 1977.

Grimmer Justin & Stewart Brandon M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*. 21(3): July 2013, p. 267-297.

Labbé Cyril & Labbé Dominique (2001). Discrimination et classement au sein d'un groupe d'entretiens. Le cas du confort électrique. . *Communication aux journées d'études du CIDSP*. Grenoble : 9 mars 2001.

Labbé Cyril & Labbé Dominique (2003). La distance intertextuelle. *Corpus*, 2-2003, p 95-118.

Labbé Cyril & Labbé Dominique (2005). How to measure the meanings of words ? Amour in Corneille's work. *Language Resources Evaluation*, 39, p. 335-351.

Labbé Cyril & Labbé Dominique (2006). A Tool for Literary Studies: Intertextual Distance and Tree Classification. *Literary and Linguistic Computing*. 21-3, p 311-326.

Labbé Cyril & Labbé Dominique (2012). Analyser les questions ouvertes dans les sondages. *Journée d'étude : Comment convaincre ? Analyse scientifique de la campagne électorale 2012*. Grenoble : Institut d'études politiques de Grenoble, 9 Mars 2012.

Labbé Cyril & Labbé Dominique (2013). La modalité verbale en français contemporain. Les hommes politiques et les autres. In Banks David. *La modalité, le mode et le texte spécialisé*. Paris : L'Harmattan, p. 33-61.

Labbé Dominique (1990). *Normes de saisie et de dépouillement des textes politiques*. Grenoble : Cahiers du CERAT.

Labbé Dominique (2002). *Analyse des représentations du confort électrique à partir d'un corpus d'entretiens*. Rapport pour le GREST-EDF. Grenoble : CERAT, juin 2002.

Lafon Pierre (1984). *Dépouillements et statistiques en lexicométrie*. Genève-Paris: Slatkine-Champion.

Luong Xuan (1988). *Méthodes d'analyse arborée. Algorithmes, applications*. Thèse pour le doctorat ès sciences. Paris : Université de Paris V.

Luong Xuan (1994). *L'analyse arborée des données textuelles : mode d'emploi*. Travaux du cercle linguistique de Nice. 16, p. 25-42.

Muller Charles (1963). Le mot, unité de texte et unité de lexique en statistique lexicologique. Reproduit dans : *Langue française et linguistique quantitative*. Genève-Paris: Slatkine-Champion, 1979, p 125-143.

Muller Charles (1977). *Principes et méthodes de statistique lexicale*. Paris : Hachette.

Paradis Emmanuel (2006). *Analysis of Phylogenetics and Evolution with R*. New York : Springer.

Pibarot André, Picard Jacques & Labbé Dominique (1995). Un outil de statistique textuelle : le lemmatiseur. *Travaux scientifique du Service de Santé des Armées*. XVI, p. 305-307.

Pibarot André & Labbé Dominique (1998). Les syntagmes répétés dans l'analyse des commentaires libres. in Mellet Sylvie (ed). *4e Journées d'analyse des données textuelles*. Nice, 1998, p 507-516.

Pionchon S. (2001). *Les Françaises et la politique*. Thèse pour le doctorat de science politique, Institut d'Etude Politique, Grenoble.

Roux Maurice (1985). *Algorithmes de classification*. Paris : Masson.

Schweisguth Etienne (1985). Fenêtre sur… L'égalité d'aujourd'hui. *Intervention*. Vol 11, janv-fév-mars 1985, p. 15-25.

Schweisguth Etienne (1986). Les avatars de la dimension gauche/droite. In Dupoirier Elisabeth & Grunberg Gérard. *La drôle de défaite de la gauche*. Paris : PUF, p. 51-70.

Schweisguth Etienne (1987). Discours des candidats, discours des électeurs : l'exemple français. In Beltran Miguel (Dir.) *Política y sociedad : estudios en homenaje a Francisco Murillo Ferrol*. Madrid : Centro de Investigaciones Sociológicas. Volume 1, p. 401-416.

Schweisguth Etienne (1988). La dimension gauche-droite en France. *Communication au XIVe congrès mondial de l'American Political Science Association*. Washington : août 1988.

Sneath Peter & Sokal Robert (1973). *Numerical Taxonomy*. San Francisco: Freeman, 1973.

Annexe 1
Le corpus des entretiens « Les Français et la politique »

N°	Classification E. Schweisguth	Longueur (mots)	Vocabulaire
011	GaucheHuma1	5 943	662
012	Divers2	5 350	620
013	DroiteRigo3	11 614	790
014	GaucheAnti4	7 311	825
015	GaucheAnti5	10 308	1 284
016	DroiteRigo6	10 837	1 051
017	GaucheAnti7	8 892	1 082
018	Divers8	4 056	512
019	Divers9	11 577	1 204
020	GaucheAnti10	4 743	917
021	DroiteLibe11	12 390	1 124
022	DroiteLibe12	13 668	1 311
023	Divers13	3 521	526
024	DroiteLibe14	6 593	949
025	GaucheHuma15	9 323	1 051
026	DroiteLibe16	7 845	854
027	GaucheHuma17	13 008	1 233
028	GaucheHuma18	9 083	1 051
029	Divers19	7 192	863
030	GaucheHuma20	9 038	1 006
032	Divers21	13 520	1 310
101	GaucheAnti22	13 818	1 279
102	Divers23	16 913	1 184
103	Divers24	12 523	1 413
104	Divers25	9 849	751
105	Petits26	11 150	953
106	Divers27	10 114	770
107	GaucheHuma28	16 817	1 464
108	DroiteRigo29	13 140	1 090
109	DroiteRigo30	22 283	1 872
110	Divers31	11 832	981
111	Divers32	11 183	1 272
112	Divers33	4 019	465
113	Divers34	10 269	971
115	DroiteRigo35	8 542	979
117	Divers36	10 144	857
118	DroiteLibe37	7 834	949
119	DroiteLibe38	9 989	1 113
120	DroiteRigo39	13 312	1 315
121	GaucheHuma40	17 087	1 084
122	DroiteRigo41	4 082	644
123	DroiteLibe42	14 054	1 288
124	Divers43	36 567	1 840
125	Divers44	9 421	924
126	Divers45	13 642	1 077
127	Petits46.	6 200	607
128	Petits47	8 186	583
129	Petits48	15 915	1 076
130	GaucheHuma49	11 800	1 007
131	Divers50	17 006	1 190
132	Divers51	9 224	921
133	Divers52	6 045	863
134	DroiteRigo53	14 347	1 387
135	DroiteLibe54	13 932	1 506
136	Divers55	10 108	937
137	DroiteRigo56	15 031	1 202
138	Divers57	13 411	1 087
139	GaucheHuma58	16 885	1 465
140	Divers59	8 596	922
141	DroiteRigo60	10 102	854
142	DroiteLibe61	11 516	859
143	DroiteLibe62	10 971	1 052
144	GaucheHuma63	9 974	850
145	GaucheHuma64	19 243	1 561
		722 888	10 475

Annexe 2 Le vocabulaire usuel de l'enquête FP (classé par catégories grammaticales)

Les vingt premiers verbes

Rang	Vocable	Effectif	Fréquence (‰)
1	être	32339	44.74
2	avoir	25004	34.59
3	faire	6728	9.31
4	dire	6180	8.55
5	aller	3561	4.93
6	pouvoir	3174	4.39
7	savoir	2929	4.05
8	voir	2665	3.69
9	falloir	2564	3.55
10	vouloir	2553	3.53
11	croire	1881	2.60
12	penser	1774	2.45
13	trouver	1380	1.91
14	travailler	1147	1.59
15	passer	1051	1.45
16	prendre	998	1.38
17	devoir	898	1.24
18	arriver	893	1.24
19	parler	835	1.16
20	donner	814	1.13

Les vingt premiers substantifs

Rang	Vocable	Effectif	Fréquence (‰)
1	gens	2502	3.46
2	chose	2418	3.34
3	an	1479	2.05
4	gauche	1355	1.87
5	droite	1230	1.70
6	heure	1034	1.43
7	fait	1002	1.39
8	problème	855	1.18
9	temps	832	1.15
10	façon	807	1.12
11	travail	742	1.03
12	niveau	719	0.99
13	monde	678	0.94
14	moment	661	0.91
15	fois	641	0.89
16	politique (nf)	634	0.88
17	côté	631	0.87
17	exemple	631	0.87
19	jour	622	0.86
20	vie	578	0.80

Les vingt premiers mots à majuscule initiale

Rang	Vocable	Effectif	Fréquence (‰)
1	France	660	0.91
2	Français	300	0.42
3	Mitterrand	282	0.39
4	Giscard	238	0.33
5	Peugeot	166	0.23
6	Chirac	130	0.18
7	Gaulle	120	0.17
8	Paris	109	0.15
9	Américain	79	0.11
9	CGT	79	0.11
11	Allemagne	75	0.10
12	Barre	71	0.10
13	Etats-Unis	70	0.10
14	Eglise	66	0.09
15	Europe	65	0.09
16	PS	64	0.09
17	Angleterre	61	0.08
18	Giscard d'Estaing	54	0.07
19	Brest	53	0.07
20	Marchais	52	0.07

Les vingt premiers adjectifs

Rang	Vocable	Effectif	Fréquence (‰)
1	petit	1245	1.72
2	sûr	755	1.04
3	bon	656	0.91
4	vrai	515	0.71
5	grand	406	0.56
6	actuel	389	0.54
6	politique	389	0.54
8	social	342	0.47
9	pareil	338	0.47
10	difficile	312	0.43
10	important	312	0.43
12	seul	282	0.39
13	différent	266	0.37
14	possible	249	0.34
15	normal	245	0.34
16	gros	228	0.32
17	économique	227	0.31
18	mauvais	223	0.31
19	français	216	0.30
20	certain	206	0.28

Les vingt premiers pronoms

Rang	Vocable	Effectif	Fréquence (‰)
1	je	26041	36.02
2	ce	18982	26.26
3	il	15859	21.94
4	on	11864	16.41
5	ça	11733	16.23
6	qui	8873	12.27
7	ils	7124	9.85
8	y	7122	9.85
9	que	4750	6.57
10	vous	4438	6.14
11	se	4248	5.88
12	moi	3975	5.50
13	le	3897	5.39
14	en	3013	4.17
15	tout	2114	2.92
16	nous	1825	2.52
17	quoi	1823	2.52
18	rien	1304	1.80
19	lui	1050	1.45
20	autre	1034	1.43

Les vingt premiers adverbes

Rang	Vocable	Effectif	Fréquence (‰)
1	pas	14835	20.52
2	ne	13288	18.38
3	bon	5453	7.54
4	bien	4690	6.49
5	plus	4459	6.17
6	là	3829	5.30
7	enfin	3706	5.13
8	alors	3662	5.07
9	oui	3654	5.05
10	même	2952	4.08
11	puis	2485	3.44
12	non	2281	3.16
13	peu	2084	2.88
14	peut-être	1832	2.53
15	très	1822	2.52
16	aussi	1701	2.35
17	beaucoup	1520	2.10
18	maintenant	1357	1.88
19	toujours	1235	1.71
20	où	995	1.38

Annexe 3 Concordances des emplois de « politique » (substantif masculin) dans le corpus FP.

015		
ique, finalement, je crois que c'est tellement récupéré par les	politiques	que finalement il dessert même le christianisme, maintenant, et j
yser, quoi, ça… les événements se succédaient, bon, sans que le	politique	ne m'interpelle, quoi, alors que maintenant, bon, eh bien, je ne s
022		
équitable de… équitable et plus juste de… des revenus ! Mais le	politique	et l'économique sont tellement imbriqués ! Ah oui ! Mille neuf cen
025		
litique. Hein ? C'est-à-dire qu'à ce moment-là l'économie et le	politique	, enfin il y avait … vraiment tout ça : l'argent, le politique … e
le politique, enfin il y avait… vraiment tout ça : l'argent, le	politique	… et on sentait bien que ça fonctionnait comme ça ! Donc si on a d
blé à vendre coûte que coûte ! Et bien on va influencer sur le	politique	pour que si on a des actions à mener contre … je veux dire par exe
ose ! Donc à ce moment là et bien… et bien… on voit bien que le	politique	lui ben … il va bien s'incliner devant ça ! Alors c'est les réalités
101		
parti politique c'est pour ça que j'ai voté ! Parce que, bon, un	politique	où j'ai aucun pouvoir personnellement en tant qu'individu … je sui
119		
cette personne, pour moi, c'est l'opinion que j'en ai, plus qu'un	politique	à droite, enfin il était à droite par opposition mais ce n'est pas
c'était plutôt un financier, c'était un technocrate plutôt qu'un	politique	, en tant que tel, j'entends bien sûr, il est obligé d'en faire p
138		
vous voyez, donc, je n'ai pas d'idées vraiment préconçues sur le	politique	. J'aime bien ce qui est vrai, et ce qui est vrai, actuellement,

Note d'Etienne Schweisguth (16 septembre 2018)

Cher collègue

D'une certaine manière la méthode lexicométrique que vous utilisez fait ses preuves. Il est frappant et très intéressant de voir qu'elle permet de constituer des groupes d'enquêtés sur la base du critère de l'enquêteur qui a réalisé les entretiens.

Ceci dit, il semble que les groupes obtenus par la méthode de la classification arborée correspondent aussi à un autre critère : celui de l'appartenance sociale. Je joins ci-dessous un tableau que j'ai constitué, indiquant l'appartenance sociale des enquêtés de chacun des groupes. On voit que le groupe D est essentiellement constitué d'enquêtés de milieu populaire, et le groupe A de salariés des classes moyennes.

Il y a manifestement un « effet enquêteur » dans le choix des enquêtés. Ce qui, bien sûr, n'exclut pas que « l'effet enquêteur » soit plus large. Le mode de recrutement des enquêtés en fait sans doute partie. J'ai essayé le plus possible de recruter les enquêtés en sonnant aux portes au hasard (en variant bien sûr les localisations géographiques et les habitats), alors que X procédait exclusivement en contactant des gens qui lui étaient indiqués par relations.

Cette méthode pose aussi le problème de l'interprétation des résultats obtenus. Il serait intéressant de savoir quels sont les mots qui contribuent le plus à la constitution d'un groupe.

Par ailleurs la méthode de la classification arborée paraît échouer à constituer des groupes pourvus d'un minimum d'identité idéologique. Dans le groupe C, par exemple, l'enquêté 105 adhère à un système de valeurs humanistes et soutient vigoureusement Mitterrand, alors que le 115 est un parfait raciste de droite.

Sans doute la méthode utilisée, en prenant en compte l'ensemble des mots utilisés, donne-t-elle plus de poids à la variable sociale qu'aux variables idéologiques et politiques.

Quelques caractéristiques des enquêtés regroupés par la classification arborée

Groupe A

N+	Enquêté	Enquêteur	Mode de recrutement
107	Homme, 34 ans, enseignant, Provence	ES	Relation
29	Homme, 34 ans, inspecteur PTT	X	Relation
26	Homme, 54 ans, gérant de société, Bretagne	X	Relation
27	Femme, 29 ans, infirmière,	X	Relation
21	Homme, 41 ans, assureur, Bretagne	X	Relation
32	Homme, 61 ans, fonctionnaire	X	Relation
22	Homme, 35 ans, cadre de banque, Bretagne	X	Relation
17	Homme, 44 ans, agriculteur, Bretagne	X	Relation
14	Homme, 32 ans, secrétaire-comptable, Bretagne	X	Relation
28	Homme, 21 ans, contrôleur aérien en formation, Brest	X	Relation
103	Homme, 43 ans, ingénieur, Provence	ES	Relation
20	Homme, 41 ans, enseignant, Brest	X	Relation
25	Femme, 40 ans, enseignante	X	Relation
15	Homme, 44 ans, enseignant, Brest	X	Relation

Groupe B

	Enquêté	Enquêteur	Mode de recrutement
139	Homme (+ femme), 26 ans, technicien, région parisienne	ES	Porte à porte
145	Femme (compagne du n° 139), 21 ans, actrice au chômage	ES	Porte à porte
135	Femme, 31 ans, documentaliste, région parisienne	ES	Relation
118	Homme, 69 ans, prêtre, région parisienne	ES	Porte à porte
23	Femme, 47 ans, institutrice (privé), Brest	X	Relation

Groupe C

	Enquêté	Enquêteur	Mode de recrutement
30	Homme, 43 ans, coiffeur, région parisienne	X	Relation
11	Homme, 38 ans, ouvrier, Bretagne	X	Relation
133	Homme, 55 ans, commerçant, région parisienne	Y	Porte à porte
115	Homme, 51 ans, opticien, région parisienne	ES	Relation

Groupe D

	Enquêté	Enquêteur	Mode de recrutement
105	Homme, 34 ans, ouvrier, Provence	ES	Relation
110	Homme, 62 ans, ouvrier, retraité, région parisienne	ES	Porte à porte
138	Homme, 36 ans, ouvrier, région parisienne	ES	Porte à porte
136	Homme, 44 ans, ouvrier, région parisienne	ES	Porte à porte
143	Femme, 66 ans, retraitée, ex-serveuse, région parisienne	ES	Porte à porte
129	Homme (+ femme), 35 ans, ouvrier OS, Montbéliard	ES	Porte à porte
127	Homme, 35 ans, pompier, Montbéliard	ES	Porte à porte
125	Femme, 79 ans, veuve d'ouvrier, Montbéliard	ES	Porte à porte
124	Homme, 59 ans, préretraité, Montbéliard	ES	Porte à porte
137	Homme, 36 ans, artisan-taxi, région parisienne	ES	Porte à porte
128	Homme, 35 ans, ouvrier OS, Montbéliard	ES	Porte à porte

Groupe E

	Enquêté	Enquêteur	Mode de recrutement
132	Homme, 37 ans, employé de commerce (produits de luxe) région parisienne	ES	
130	Femme, 49 ans, cadre moyen, région parisienne	ES	Relation
120	Homme, 55 ans, VRP, région parisienne	ES	Porte à porte
122	Femme de cadre, 63 ans, retraitée, région parisienne	ES	Porte à porte
13	Femme, 22 ans, vendeuse, Brest	X	Relation
142	Femme de cadre, 26 ans, région parisienne	ES	Porte à porte
112	Femme, 26 ans, dactylo, région parisienne	ES	Porte à porte

Groupe F

	Enquêté	Enquêteur	Mode de recrutement
134	Homme, 46 ans, technicien, région parisienne	ES	Porte à porte
19	Homme, 54 ans, cadre assurance	X	Relation
140	Homme, 29 ans, maître-nageur, région parisienne	ES	Porte à porte
109	Homme, 32 ans, artisan, région parisienne	ES	Relation
16	Homme (+ femme), 47 ans, agriculteur, Bretagne	X	Relation
113	Homme, 38 ans, employé ministère, région parisienne	ES	Porte à porte
144	Femme, 21 ans, VRP, région parisienne	ES	Porte à porte
141	Homme, 36 ans, technicien, région parisienne	ES	Porte à porte
121	Femme d'électronicien, 40 ans, région parisienne	ES	Porte à porte

Groupe G

	Enquêté	Enquêteur	Mode de recrutement
17	Homme, 44 ans, agriculteur, Bretagne	X	Relation
123	Femme, 27 ans, kiné, Montbéliard	ES	Porte à porte
102	Femme, 25 ans, infirmière, région parisienne	ES	Porte à porte
111	Femme, 38 ans, secrétaire, région parisienne	ES	Porte à porte
108	Femme de médecin, 35 ans, Provence	ES	Relation
18	Femme, 47 ans, aide-soignante au chômage, Bretagne	X	Porte à porte
117	Femme, 30 ans, secrétaire, région parisienne	ES	Porte à porte
126	Femme d'ouvrier, Montbéliard	ES	Porte à porte
104	Homme, 45 ans, agriculteur, Provence	ES	Relation
131	Homme, 31 ans, vendeur fruits et légumes, région parisienne	Y	Porte à porte
106	Homme, contremaître, Provence	ES	Relation

Réponse de D. Labbé à E. Schweisguth

Nous remercions E. S. pour sa note qu'il a accepté de voir reproduire à la fin de notre communication.

Cette note nous semble très pertinente (notamment à propos de l'influence du milieu social de l'enquêté sur la classification automatique). Nous y ajouterons deux précisions.

En ce qui concerne ces classifications. En laissant de côté l'"effet enquêteur", la distance entre les textes provient essentiellement de l'auteur et des thèmes abordés. Le premier facteur est sans doute le plus lourd : personnalité, culture, intérêt pour le sujet et, aussi, aptitude à l'expression orale sont très inégalement répartis et expliquent probablement les regroupements sur une base sociologique que E. S. a signalés fort justement. La classification automatique n'invalide donc pas la sienne qui est fondée sur des critères objectifs (principalement le comportement politique). Comme indiqué en conclusion de notre communication, il existe des moyens pour mesurer les proximités thématiques entre enquêtés.

Deuxièmement, comme suggéré dans notre communication, le terme "effet enquêteur" désigne la capacité à obtenir de l'enquêté qu'il aille au bout de son propos et quitte progressivement la réserve normale en face d'un inconnu. Cette inégale aptitude chez les enquêteurs se traduirait notamment par des entretiens plus ou moins riches et plus ou moins longs, mais aussi plus ou moins centrés sur le sujet. Naturellement, cela ne signifie pas, comme le prétendent certains, que l'enquêteur aurait la capacité de changer les idées de l'enquêté (du moins si l'enquêteur respecte les règles du genre !). De ce point de vue, la lecture des entretiens d'E. S. sera une bonne formation pour les jeunes chercheurs débutants.