HAL
open science

Analyser les questions ouvertes dans les sondages

Cyril Labbé, Dominique Labbé

To cite this version:

Cyril Labbé, Dominique Labbé. Analyser les questions ouvertes dans les sondages. Comment convaincre ? Analyse scientifique de la campagne électorale 2012, Mar 2012, Grenoble, France. halshs-00709115

HAL Id: halshs-00709115

https://shs.hal.science/halshs-00709115

Submitted on 17 Jun 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Journée d'étude

« Comment convaincre ? Analyse scientifique de la campagne électorale 2012 »

Institut d'études politiques de Grenoble
9 Mars 2012

Analyser les questions ouvertes dans les sondages.

Cyril Labbé
Laboratoire d'Informatique de Grenoble - Université Joseph Fourier
(cyril.labbe@imag.fr)

Dominique Labbé
Laboratoire PACTE (CNRS - Institut d'Etudes Politiques de Grenoble)
(dominique.labbe@iep.grenoble.fr)

Résumé :
Présentation d'une méthode d'analyse des réponses aux questions ouvertes dans les enquêtes d'opinion. On décrit d'abord la transcription, la codification et le traitement des réponses. Les caractéristiques particulières de ces réponses (longueur et vocabulaire) se prêtent mal à l'analyse « textuelle » traditionnelle. Grâce aux univers lexicaux et à la classification automatique, on repère les principaux thèmes présents dans les réponses. Cela permet de traiter les questions ouvertes comme les questions fermées.

Beaucoup d'enquêtes d'opinion comportent au moins une question « ouverte », c'est-à-dire une question sans réponses pré-établies, comme dans le reste du questionnaire. L'enquêté est invité à y répondre avec ses propres mots.

A part quelques études pionnières (comme Brugidou 1998), ces réponses sont sous-exploitées. Cette sous-exploitation s'explique par de multiples raisons allant des faiblesses de la transcription par les enquêteurs (Caillot et Moine 2001 ; Marc, 2001 ; Brugidou et Moine 2010) à la difficulté de traiter un matériel essentiellement qualitatif avec les outils statistiques standard (Brugidou et Escofier 2007).

En général, un enquêteur regroupe les réponses sous un nombre limité d'items, ce qui revient à établir *a posteriori* la grille des réponses possibles (Peretti 2005). Autrement dit, les réponses sont *fermées* après l'enquête au lieu de l'être avant comme pour le reste du questionnaire. De ce fait, on n'est jamais sûr que la case, dans laquelle on range l'enquêté, soit bien celle qu'il aurait choisie si on lui avait présenté cette grille, en lui demandant de choisir parmi les items, au lieu de le laisser parler…

Cette communication présente des outils qui pourraient résoudre ces difficultés grâce à une retranscription intégrale, et de qualité, des propos tenus par les enquêtés, propos auxquels sont appliqués divers traitements statistiques ne faisant pas intervenir la subjectivité des opérateurs ni de grille extérieure aux propos tenus.

L'idée de départ est la suivante : l'homme qui parle (ou écrit) créé un monde dont son esprit tient les fils, mais cette création n'est pas arbitraire, elle est enserrée dans une série de règles – de la langue et de la communication -, règles qui, sans avoir la rigidité de la gravitation universelle, n'en sont pas moins bien réelles. Pour comprendre le message, et remonter jusqu'aux idées qui l'ont engendré, il faut "neutraliser" tout ce qui, dans ce message, est le produit des règles dans lesquelles l'esprit doit se couler pour communiquer, et cette "neutralisation" doit se faire de telle manière que la subjectivité de l'observateur ne vienne en rien perturber les observations. Un certain nombre d'outils, statistiques et informatiques, ont été mis au point dans ce but. Ils seront présentés dans cette communication et illustrés à l'aide d'une enquête réalisée lors de la campagne présidentielle de 2007.

I. L'enquête et le traitement des réponses

Le sondage utilisé a été réalisé - auprès d'un échantillon représentatif de la population iséroise en âge de voter - en deux vagues (du 5 au 10 février et du 10 au 14 avril 2007), avec le même questionnaire comportant notamment :

- une question fermée ("Telle qu'elle se déroule actuellement, la campagne électorale pour l'élection présidentielle de 2007 vous donne-t-elle envie d'aller voter ?") avec comme choix : Oui, très envie ; Oui, assez envie ; Non peu envie ; Non pas du tout envie; Ne sais pas.
- suivie d'une question ouverte formulée en fonction de la réponse précédente : « Pour quelles raisons cette campagne vous donne-t-elle (très, assez, peu, pas tu tout) envie d'aller voter ? ».

Les enquêteurs ont relancé systématiquement l'enquêté avec les techniques usuelles (en reprenant, sous forme de question, les derniers mots de la réponse, en demandant si l'enquêté n'avait pas d'autre raisons ou s'il n'avait pas quelque chose à ajouter).

Au total sur les 2 036 enquêtés, 1 467 (soit 72 %) ont accepté que leur réponse à cette question ouverte soit enregistrée. Il semble que ce soit un taux de réponse honorable pour une question ouverte dans un sondage téléphonique.

Pour analyser objectivement ces réponses - et remonter à l'opinion des enquêtés - le texte a été traité de la manière suivante.

Balisage

Le tableau ci-dessous présente les balises ajoutées au texte qui a été enregistré.

Tableau 1. Balisage des réponses

<100545 févr-07>
<Question : Pour quelles raisons cette campagne vous donne-t-elle très envie d'aller voter ?>
<Réponse :>
Ah c'est moi qui dois vous répondre là ?
<Question : oui, pour quelles raisons ?>
<Réponse :>
Pour essayer, euh, d'avoir un changement.

En première ligne, le numéro d'ordre de la réponse. Ce numéro permet d'établir, pour chaque enquêté, un lien entre le contenu de sa réponse à cette question ouverte et celles qu'il a apportées au reste du questionnaire.

Les balises « questions » et « réponse » sont également indispensables pour neutraliser les propos de l'enquêteur et être bien certain que l'analyse ne porte que sur la réponse. Naturellement, il ne faut pas supprimer les propos de l'enquêteur afin de permettre à l'analyste de retrouver exactement le contexte dans lequel la réponse a été fournie (et mesurer éventuellement dans quelle mesure cette réponse aurait été "induite" par l'enquêteur…)

Ensuite, la graphie des mots est contrôlée et standardisée.

Correction et standardisation orthographiques

Cette opération est également indispensable et doit être menée avec soin. En effet, malgré la qualité des transcriptions, les erreurs sont nombreuses. Il s'agit des problèmes classiques de transcription de l'oral dans les grands corpus (Nelson 1997, Labbé 2001). Parmi les difficultés, citons :

- le parti pris phonétique : "y" (pour *il*) "y'a" (pour *il y a*), "pis" (pour *puis*), "parce" voire "parque" (pour *parce que*), "chuis" (pour *je suis*), "j'frais" (pour *je ferais*)… Il s'agit d'éviter les erreurs d'analyse (ainsi "j'frais" : pronom personnel "je" + adjectif "frais"). Naturellement, on respecte intégralement les propos tenus, mais en adoptant la graphie du français. Si l'enquêté a dit "ch'uis pas" on écrit "je suis pas" (ou "j'suis pas") et on ne corrige pas la double négation omise (*je* ***ne*** *suis pas*).

- les difficultés de la langue française qui embarassent les meilleures secrétaires, par exemple : "quelque soit" (pour *quel que soit*), les tirets oubliés (est ce, peut être, c'est à dire, moi même, celui ci, cette fois là…), les accents : l'article "la" confondu avec l'adverbe "là", les accents circonflexes omis (*sur*, *parait*, *plait*, etc), "forcement" (substantif) à la place de *forcément* (adv), etc. Le participe passé confondu avec l'infinitif homophone ("allé voter", "aller voté"…)

- la transcription des noms propres. On trouve « Royale », « Sarkosy », « Bayerou »…

Il faut également reprendre la ponctuation, notamment au sein de la phrase afin d'homogénéiser des pratiques très discordantes entre les opérateurs. La plupart ponctuent en fonction du rythme et non de la syntaxe : une virgule pour un bref silence, un point quand le silence se prolonge (ou trois points de suspension quand l'idée reste en suspens). Mais ce parti-pris non syntaxique rend parfois la réponse incompréhensible…

Ce travail préalable vise en quelque sorte à "neutraliser" le "bruit", engendré par les règles de la graphie du français, qui risque de perturber la compréhension du message. Pour

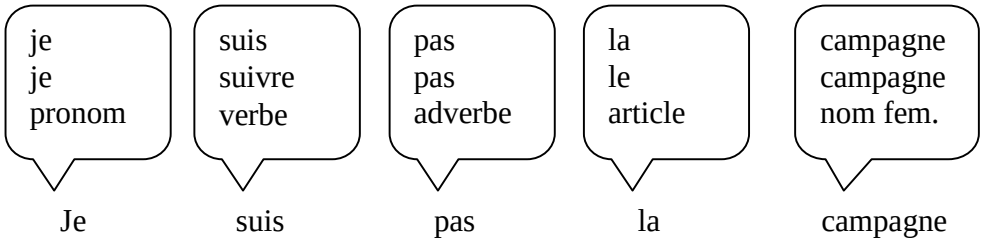
cette compréhension, il faut encore neutraliser les effets considérables du système de la langue. C'est l'objet de la "lemmatisation".

Lemmatisation.

Quand l'ordinateur a commencé à être utilisé pour le dépouillement des textes, il a été proposé d'utiliser les outils de la lexicographie, notamment la nomenclature des dictionnaires. L'implémentation de cette nomenclature dans des algorithmes et des programmes informatiques, et les principales règles de la lemmatisation sont présentées dans un ouvrage en ligne (Labbé 1990). L'application de ces méthodes aux questions ouvertes a déjà été discutée dans Labbé 2001.

L'opération consiste à ajouter une étiquette à chacun des mots du texte (figure ci-dessous).

Tableau 2. Exemple d'étiquettes attachées à chacun des mots du texte



L'étiquette vient s'ajouter au texte sans l'altérer. Elle comporte trois informations :

- la graphie standard : majuscule initiale des mots communs ramenée en minuscule (comme pour "Je"), réduction des formes multiples à une graphie standard (puis et peux, événement et évènement…), etc.

- puis le vocable, c'est-à-dire l'entrée où se trouve la graphie dans le dictionnaire et la catégorie grammaticale, telle qu'elle figure en seconde position dans cette entrée de dictionnaire.

En moyenne, plus du tiers des mots d'un texte en français peuvent renvoyer à plusieurs entrées de dictionnaire. Dans l'exemple ci-dessus :

- *suis* : suivre ou être ?
- *pas* : adverbe ou substantif masculin ?
- *la* : article ou pronom relatif ou note de musique ?

Naturellement, les opérations de standardisation des graphies et de lemmatisation sont confiées à des automates qui résolvent la quasi-totalité des cas et qui, pour les quelques ambiguïtés restantes, proposent à l'opérateur les différentes solutions possibles.

La nomenclature des mots, apprise à l'ordinateur, est systématique (par exemple, si les substantifs se distinguent par le genre, alors tous les substantifs doivent se voir affecter le masculin ou le féminin), elle est exhaustive (tous les mots y trouvent leur place), elle est univoque (une seule entrée par mot), elle exclut tout double compte, elle ne comporte pas de catégorie ad hoc, ou fourre-tout, etc. Enfin l'opération est réversible : on peut retrouver le texte original, sans altération, à partir du fichier étiqueté.

Pour bien comprendre les opérations présentées ci-dessous, il faut se souvenir qu'un texte est une succession de **mots** (en anglais "word tokens", c'est-à-dire "emplacements") – dont le nombre total donne la **longueur** du texte – ces mots étant issus d'un **vocabulaire** nécessairement plus restreint puisque certains **vocables** (en anglais « word types ») sont employés plusieurs fois dans le texte. Par exemple, "le", "les", "la", "l´" – et leurs équivalents avec une majuscule initiale - sont les différentes **graphies**, ou **forme graphiques**, sous

lesquelles l'article ou le pronom "le" apparaissent dans un texte. "le, article" et "le, pronom" sont des vocables (ou "entrées de dictionnaire"). Chacune des **occurrences** de ces deux vocables – sous les formes "le", "la", "les", "l´", "Le", "La", "Les", "L´" – constitue un mot du texte.

Ainsi corrigés, standardisés, balisés et étiquetés, les textes des réponses ouvertes permettent de remonter les fils, dont nous parlions en introduction, qui relient les propos tenus par les enquêtés à leurs opinions et à leurs attitudes politiques.

II. Caractéristiques des réponses à la question ouverte

On examine successivement la longueur de ces questions et leur vocabulaire.

Une longueur très variable

Les principales caractéristiques sont résumées dans le tableau 3 ci-dessous.

Tableau 3. Caractéristiques de l'enquête et des réponses à la question ouverte

Valeurs caractéristiques	Vague 1 (février)	Vague 2 (avril)	Total des 2 vagues
Nombre d'enquêtés	1 026	1 010	2 036
Nombre de réponses à la question ouverte	740	727	1 467
Taux de réponse à la question ouverte	72,1	72,0	72,1
Nombre de mots	63 481	78 813	142 294
Longueur moyenne (mots)	85,8	108,4	97,8
Ecart type de la moyenne	90,9	104,6	98,6
Longueur modale (mots)	46	81	46
Longueur médiane (mots)	59	80	69
Longueur médiale (mots)	120	140	132

Deux constats principaux :

- Si le taux de réponse n'augmente pas entre les deux vagues, en revanche, les valeurs centrales sont toutes en hausse entre février et avril : moyenne et médiane augmentent de 25%. Deux interprétations sont possibles (et non contradictoires). D'une part, une plus forte mobilisation et une plus forte implication des enquêtés au fur et à mesure que le scrutin approche. Toutefois cette interprétation se trouve en partie contredite par le constat que le taux de réponse, lui, n'augmente pas… D'autre part, lors de la 2e vague, les enquêteurs ont nettement plus relancé les enquêtés et une partie de l'allongement des réponses provient de tout ce que les enquêtés ajoutent à leur première réponse sous l'effet de ces relances. Pour cette enquête, le lien entre la longueur et le nombre de relances a déjà été mis en lumière par Brugidou et Moine (2011).

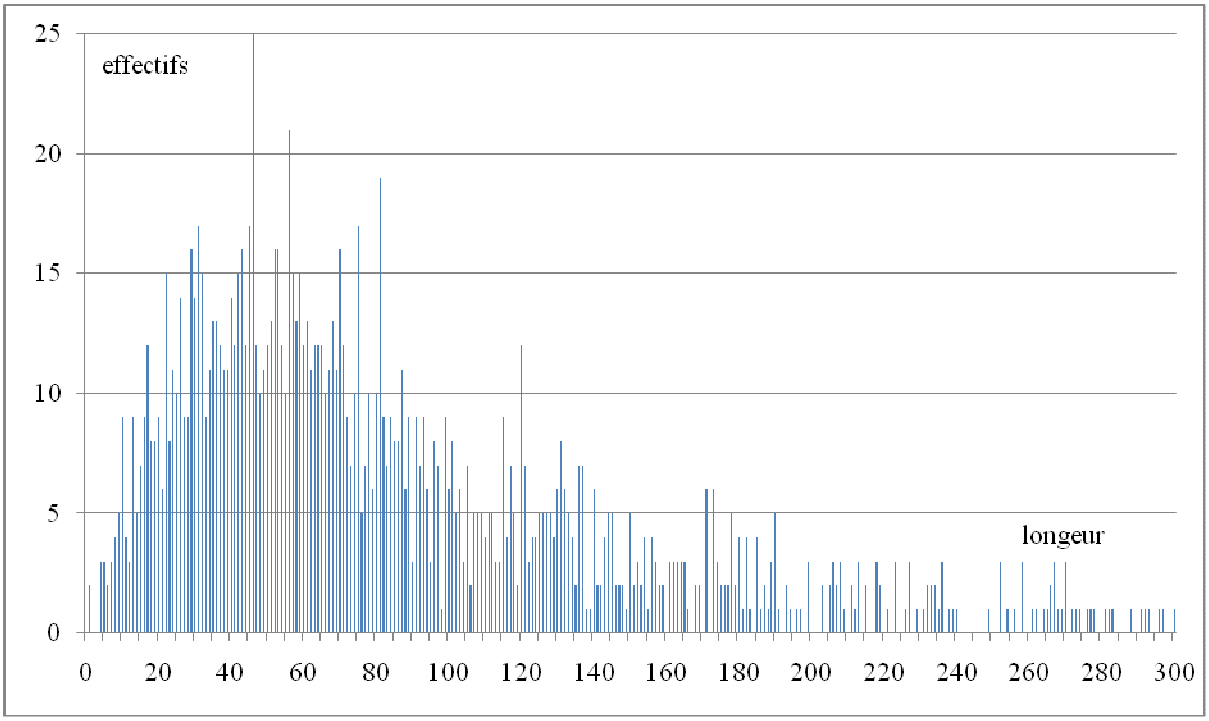
- une forte asymétrie : médiane et mode sont nettement inférieurs à la moyenne. Un grand nombre de réponses assez brèves et quelques réponses très longues tirent la moyenne vers le haut. Par exemple, la moitié des réponses comptent moins de 69 mots (médiane) mais la moitié des mots sont prononcés dans des réponses de 132 mots et plus (médiale). Le graphique ci-dessous résume cette asymétrie

La grande majorité des réponses sont comprises entre 1 et 120 mots et leur distribution

épouse grossièrement le profil d'une courbe en cloche. En revanche, au-delà, on n'observe aucune régularité.

Ces caractéristiques soulèvent évidemment une question : comment neutraliser ces différences considérables de longueur dans la comparaison entre les enquêtés ?

Tableau 4. Histogramme des longueurs des réponses classées par longueur croissante (effectifs absolus)



Vocabulaire des réponses

Au total les 142 294 mots contenus dans les réponses correspondent à 5 873 formes graphiques différentes mais seulement 3 712 vocables différents. C'est l'un des principaux intérêts de la lemmatisation : réduction du vocabulaire (- 37 %) et augmentation des unités de forte fréquence, sur lesquelles porte le calcul. Cependant, cet avantage reste limité par deux caractéristiques qu'illustre le tableau en annexe 1 (les vocables les plus fréquents).

Premièrement, pour comparer les deux colonnes de ce tableau en annexe, les **effectifs** – chiffres absolus – ont été transformés en **fréquences** (chiffres relatifs). Ces chiffres relatifs sont exprimés en "pour mille mots". Même ainsi, les proportions deviennent rapidement assez petites quand on descend dans la hiérarchie. On passe au-dessous de 1% après le 21e vocable. De plus, la surface du texte est très inégalement répartie. La dernière ligne du tableau indique que les 100 vocables les plus fréquents couvrent, à eux seuls, plus des trois quarts de cette surface et donc que les 3 612 vocables restants se partagent moins d'un quart de celle-ci. Pourtant ce sont eux qui véhiculent la plupart des informations !

Ces déséquilibres se retrouvent toujours dans les corpus en langue naturelle. Du point de vue statistique, tout texte peut se définir comme une *population constituée d'un grand nombre d'individus rares et d'effectifs très inégaux*.

La deuxième remarque concerne la présence massive, en tête de ce tableau en annexe, des "mots outils" (articles, prépositions, adverbes, pronoms). Parmi les mots les plus employés, aucun substantif ni adjectif. En caractères gras, les cinq pronoms (*je*, *il*, *nous*, *on* et *vous*) et les quatre verbes (*être*, *avoir*, *faire* et *dire*). Seule la présence de *dire* est significative

puisque, dans tout corpus de textes français, les trois verbes les plus employés sont toujours *être*, *avoir* et *faire*, dans cet ordre (Labbé & Labbé 2010).

On note enfin que *aller* (19e), *voter* (25e), *envie* (46e), *donner* (52e), *campagne* (67e), *raison* (76e) figurent dans les mots les plus utilisés parce que beaucoup d'enquêtés répètent plus ou moins la question de l'enquêteur (« pour quelles raisons la campagne électorale vous donne-t-elle (très, assez, peu, pas du tout) envie d'aller voter ? »). Si on retire ces vocables "soufflés" à l'enquêté, il ne reste alors, dans la liste des 100 premiers vocables, que 13 verbes, 9 substantifs et un adjectif…

Un tableau lexical quasiment vide

Conséquence logique de ces caractéristiques, le tableau lexical entier présente une structure particulière (tableau 5).

Tableau 5. Caractéristiques du tableau lexical entier

Vocables (*j*) Réponses (*i*)	1 à (prép.)	2 abaisser (v.)	(…)	3712 zone (n. fem)	Total
1	1	0	(…)	0	27
2	0	0	(…)	0	31
(…)	(…)	(…)	(…)	(…)	(…)
1467	1	0	(…)	0	66
Total	2 259	1	(…)	1	142 294

En ligne les *i* réponses non vides (*i* variant de 1 à 1467) ; en colonne les *j* vocables employés dans ces réponses (*j* variant de 1 à 3 712, classement alphabétique).

A l'intersection de la *ième* ligne et de la *jième* colonne, le nombre de fois que le répondant n°*i* a utilisé le vocable n° *j*. Par exemple le premier répondant a utilisé une fois la préposition *à* mais zéro fois le verbe *abaisser* ou le substantif féminin *zone*. Au total, sa réponse compte 27 mots dont 20 vocables différents, ce qui signifie encore que dans la première ligne du tableau lexical entier, 20 cases sont supérieures à zéro et…3 682 sont égales à zéro (nulles).

Il y a 1 467 * 3 712 = *5 445 504* intersections ou cases dans ce tableau, sans compter la dernière ligne et la dernière colonne (les totaux). Parmi ces cases,

- 5 392 865 sont nulles (le répondant *i* n'a pas utilisé le vocable *j*), soit 99,03% des cases
- 52 639 contiennent un chiffre positif (le répondant *i* a utilisé au moins une fois le vocable *j*), soit 0,97 % des cases.

Autrement dit, dans ce tableau, plus de 99 % de ces cases sont vides. Les manuels de statistique indiquent que ce genre de tableau "creux" se prête mal au calcul et recommandent *de regrouper certaines lignes et/ou certaines colonnes pour supprimer le maximum de cases vides*. Par exemple, Lebart et Salem (1994) limitent l'analyse aux "formes graphiques" les plus fréquentes. Le tableau 6 donne le "bénéfice" obtenu par cette limitation (réduction de la proportion de cases non vides) et le "coût" : perte d'information par élimination de certains mots.

On remarque que :

- même en restreignant l'analyse aux 170 vocables d'effectifs au moins égal à 100, 84% des cases du tableau lexical restent vides. Les classifications ou les calculs ont-ils encore un sens quand ils sont effectués sur de tels tableaux ?

Tableau 6. Caractéristiques du tableau lexical en fonction du nombre de vocables retenus.

Effectifs des vocables compris dans l'analyse	Nombre de vocables compris dans l'analyse	% du vocabulaire compris dans l'analyse	Nombre de cases du tableau	Nombre de cases non nulles	Proportion des cases non nulles (%)
Tous	3712	100,0	5 445 504	52 639	0.94
> 4	1138	30.7	1 669 446	50 158	3,00
> 9	718	19,3	1 053 306	48 580	4,61
> 49	261	7,0	382 887	43 295	11,31
> 99	170	4,6	249 390	39 906	16,00

- comme l'indique l'annexe 1, les vocables les plus fréquents sont en majorité des "mots outils" considérés comme peu intéressants du point de vue de la théorie de la communication. Autrement dit, en limitant l'analyse aux vocables les plus fréquents, l'essentiel des substantifs, des adjectifs et des verbes disparaissent alors qu'ils sont considérés comme les vecteurs essentiels de la communication ;

- quand on leur soumet un tableau de ce genre, les logiciels de statistique ne refusent pas de les traiter mais leurs réponses sont-elles pertinentes ?

Deux solutions sont habituellement proposées :

- regrouper les colonnes en rattachant les mots à des thèmes (voir par exemple le logiciel Tropes : Piolat & Bannour 2009). Mais c'est postuler que le sens des mots est univoque et que ce sens provient du système de la langue. Ou, pour le dire plus simplement, c'est apposer sur le corpus la grille de lecture a priori de celui qui aura défini les thèmes. Est-ce acceptable ? Ne vaudrait-il pas mieux tirer cette nomenclature des réponses elles-mêmes ?

- regrouper les lignes par caractéristiques démographiques (sexe, tranches d'âge), sociales (CSP) ou politiques (intentions de vote). C'est postuler que ce ne sont pas des individus qui parlent mais des tranches d'âge, des classes sociales, des courants politiques... Est-ce acceptable ?

Nous proposons d'explorer une autre voie : effectuer ces regroupements, sans intervention manuelle et en utilisant exclusivement le discours tenu par les enquêtés. Les procédures sont les suivantes :

- afin de réduire le nombre des colonnes, un calcul puis une classification identifient, dans les réponses, les principaux thèmes abordés qui se caractérisent chacun par un certain vocabulaire ;

- afin de réduire le nombre de lignes, un tri regroupe les enquêtés en fonction des thèmes qu'ils privilégient ou qu'ils omettent…

III. A la recherche des thèmes

Pour reconstituer les univers de pensée évoqués en introduction, plusieurs méthodes sont concevables. Ainsi, l'utilisation des "syntagmes répétés" se révèle particulièrement féconde (Pibarot & Labbé 1998). On propose ici une voie nouvelle : les univers lexicaux (Hubert & Labbé 1995 ; Labbé & Labbé 1994 et 2005 ; Labbé 2010).

Les univers lexicaux

L'univers lexical est constitué comme les galaxies que nous évoquions en introduction. C'est l'ensemble des relations d'attirance ou de répulsion qu'un vocable entretient avec tous les autres, comme les interactions entre étoiles et les planètes dans une galaxie (ici l'ensemble

des réponses des enquêtés qui est supposé restituer les champs de force constituant l'opinion au début de 2007).

Pour repérer ces relations, on relève toutes les réponses contenant le vocable considéré. Leur vocabulaire est comparé à celui de l'ensemble du corpus. Lorsque la fréquence (relative) dans ce sous-ensemble dépasse significativement celle observée dans le corpus entier, l'association est "positive" : les deux vocables sont associés dans l'esprit du (ou des) locuteurs, ou encore l'utilisation d'un de ces mots est fortement prédictif de la survenue de l'autre. A l'inverse, si la fréquence relative est significativement inférieure à celle observée dans le corpus entier, l'association est négative (antonymie). Les deux mots se repoussent.

En annexe, un exemple : l'univers lexical de "devoir, substantif masculin".

Beaucoup d'associations sont logiques comme *aller voter est un devoir*, *le devoir électoral*, *le devoir civique*, *important*, *essentiel*, etc. On note l'association avec des substantifs comme *démocratie*, *citoyen*, *citoyenne*, *droit* (et *obligation*). On remarque que les gens qui utilisent cette notion de *devoir électoral* parlent plus que les autres de la *France* (ou de *Le Pen*) mais mentionnent moins que les autres *Ségolène Royal*…

Dans les associations négatives (les vocables sous-employés par les réponses qui mentionnent le *devoir électoral*) se trouvent : le *pouvoir*, les *programmes*, les *hommes*, les *promesses*, les *débats*, les *idées*. Certains adjectifs sont également évités comme *intéressant*, *concret*, *principal*, *politique*…

On propose de considérer que cet univers est présent dans une réponse si, même en l'absence du mot *devoir*, cette réponse contient un grand nombre des mots positivement associés à ce vocable – par exemple *droit*, *citoyen* ou *démocratie* – et que, à l'inverse, on y rencontre peu de mots négativement associés à ce vocable (par exemple, *pouvoir*, *proposition*, *candidat*, *média*, *politique*…)

Le logiciel relit toutes les réponses en affectant à chacune (*i* variant de 1 à 1444) et pour chaque univers (*u*) un « score » qui est la somme de toutes les associations positives entre *i* et *u* diminuée du nombre d'associations négatives entre les deux. Afin de pouvoir comparer des réponses de longueurs différentes, le score absolu est divisé par le nombre de mots contenus dans la réponse (longueur). Le score de la réponse *i* pour l'univers *u*, s'écrit :

$$\text{score}_{iu} = \frac{\text{Nombre d'associations positives}_{iu} - \text{Nombre d'associations négatives}_{iu}}{\text{Longueur de la réponse}_{i}}$$

- un score supérieur à 0 indique que le répondant partage plutôt l'univers intellectuel étudié (ici *devoir électoral*), même s'il n'a pas employé ces deux mots-là ;
- un score inférieur à 0 indique que le répondant se situe en opposition par rapport à l'univers intellectuel du *devoir électoral*.

77 univers ont été retenus. Il s'agit de tous les substantifs, adjectifs, verbes et pronoms personnels apparaissant au moins 100 fois dans les réponses – à l'exclusion de : *raison campagne*, *donner*, *envie*, *aller*, *voter* (comme expliqué ci-dessus, c'est la question que beaucoup d'enquêtés ont repris dans leur réponse). Ces 77 univers contiennent au total 1096 vocables différents soit 96% des vocables utilisés 5 fois ou plus (qui sont les seuls à pouvoir entrer dans le calcul des univers).

On obtient un tableau de 1444 lignes (les *i* réponses) et de 77 colonnes (les *u* univers) dont l'intersection contient le score de la réponse *i* pour l'univers *u*. Au total ce tableau comporte 111 188 cases dont 105 637 (95%) sont non nulles, à comparer avec les proportions du tableau 6 ci-dessus.

Des univers aux thèmes

On applique à ce tableau une classification (Roux 1985 & 1994), en commençant par les colonnes. Cette classification a deux objectifs :

- rechercher les meilleurs groupements possibles sans intervention humaine. Deux critères sont utilisés : d'une part, les distances entre les colonnes composant un même groupe doivent être aussi petites que possible ; d'autre part, les distances séparant les différents groupes ainsi constitués, doivent être les plus grandes possibles ;
- offrir la meilleure représentation possible, en deux dimensions de ces groupements alors que le phénomène représenté comporte un très grand nombre de dimensions, comme on représente le "plan" d'une galaxie…

En pratique, on calcule le profil de chacune des colonnes – en fonction des 1444 lignes – puis les distances entre ces profils considérés deux à deux.

L'algorithme commence par regrouper en une seule colonne les deux dont les profils sont les plus proches puis il calcule le profil de cette nouvelle colonne et ses distances avec toutes les autres restantes et regroupe la paire la plus proche, etc. jusqu'à ce qu'il n'en reste plus qu'une.

Une figure (dendrogramme) résume les étapes de cette classification (tableau 7). Elle illustre l'une des difficultés de l'analyse : les 77 étiquettes sont le maximum que l'on puisse représenter, de manière lisible, sur l'axe horizontal…

En ordonnées, les distances correspondantes aux différents niveaux d'agrégation.

En coupant le graphe, horizontalement et au plus près de l'un des seuils considérés comme significatifs, on isole les univers très proches, relativement proches, etc. Ces groupes étant isolés, on peut étudier en quoi leurs vocabulaires diffèrent grâce à l'étude de leurs spécificités.

Quelques remarques préalables :

- les branches les plus basses correspondent aux univers agrégés en premier : ils forment les couples les plus proches et les plus sûrs. Normalement ce sont aussi les univers les plus "centraux", c'est-à-dire ceux qui sont séparés de tous les autres par les distances les plus courtes. Plus on s'élève, plus les individus sont décalés, plus les classes constituées sont hétérogènes et plus l'interprétation devient complexe. C'est le cas notamment du groupe à l'extrême-gauche du graphe ;
- quelle que soit leur position sur l'axe horizontal, la proximité entre deux individus ou groupes d'individus, est mesurée par la hauteur à laquelle se rejoignent les traits les unissant. Par exemple, le couple le plus proche est "Ségolène + Royal", ce qui est logique, puis "droite + gauche"… ;
- le calcul neutralise les différences de taille entre les univers, cependant il faut éviter de comparer des individus trop différents ;
- la technique produit parfois des "effets de chaîne" (un graphe en "escalier" signale habituellement ce genre d'effet). Certaines proximités entre individus ne sont plus discernables car les sommets qui les relient sont effacés par des agrégations effectuées à un niveau inférieur. Le graphe doit donc être utilisé avec prudence. L'appartenance de chacun des individus à une classe donnée sera éventuellement contrôlée sur la matrice des distances. Il n'est pas mauvais non plus de recalculer la distance moyenne de chacun des individus à l'ensemble de ses "voisins" supposés.
- ces calculs n'ont de signification que si les textes ont été dépouillés en utilisant la même norme afin de ne pas interpréter les fluctuations dans les graphies comme des distances réelles car elles peuvent entraîner des différences quantitativement significatives mais sans contenu lexical...

Tableau 7. Classification automatique des univers

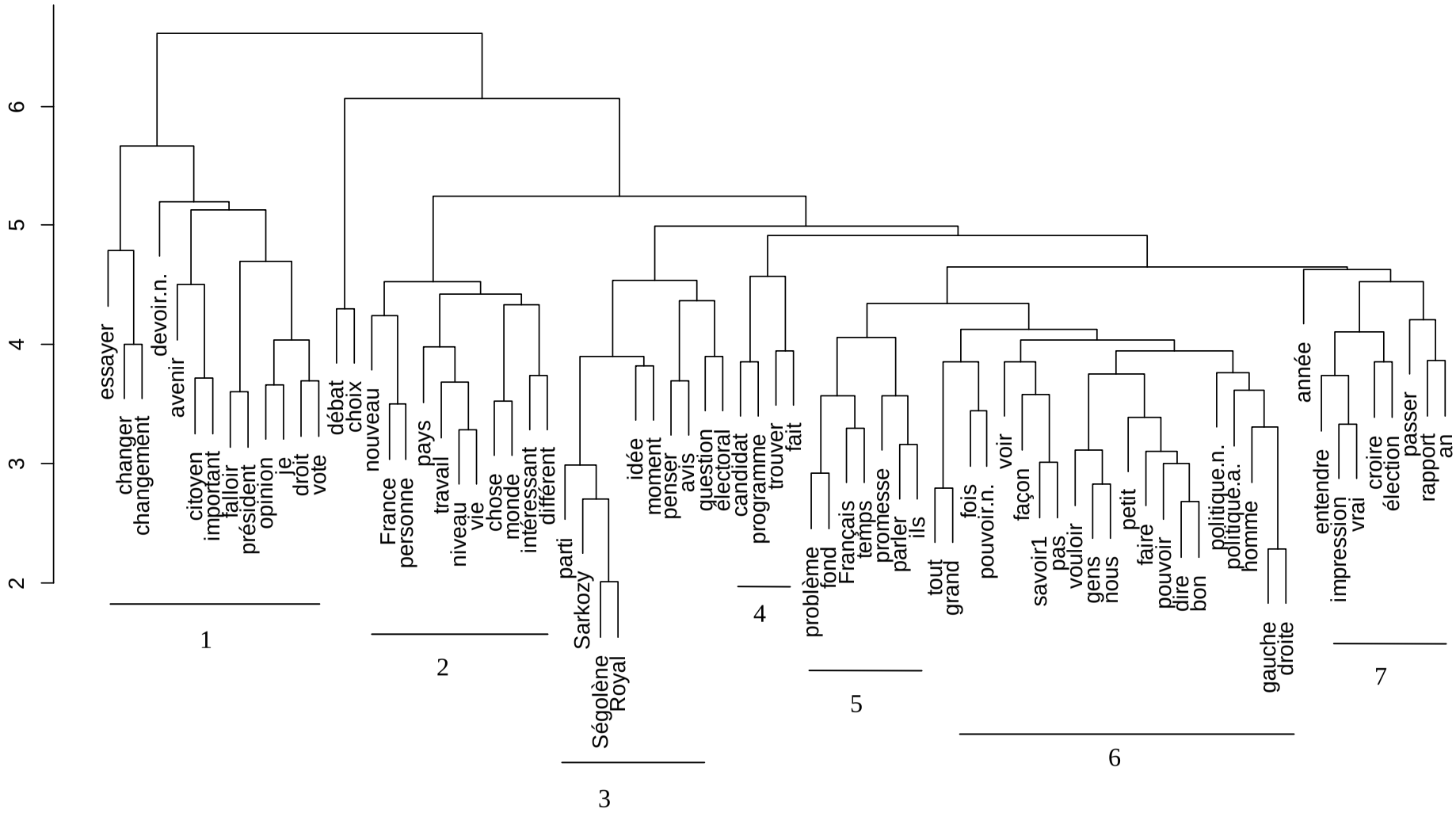


Figure réalisée avec le logiciel R (Meyer et Al. 2008)

Dans le cas présent, la classification oppose deux groupes principaux. L'un à gauche autour du « je » contient une douzaine d'univers et se révèle hétérogène (niveaux d'agrégation élevés). L'autre à droite regroupe une soixantaine d'univers que l'on peut subdiviser en 6 sous-groupes. On aboutit aux classes suivantes (de gauche à droite) :

1. Le groupe autour du « je » : *droit* (de) *vote*, *opinion*, *falloir* (un) *président*, *devoir* (du) *citoyen*, *avenir*. A ce groupe est attaché un trio : *essayer* (de) *changer*, (le) *changement*.

2. La *France* : *pays*, *travail*, *niveau* (de) *vie*, *chose intéressant*(e), *différent*(e), *nouveau*(elle), *monde*.

3. *Sarkozy*, *Ségolène Royal* : *parti*, *idée*(s) (du) *moment*, groupe auquel se rattachent : *penser*, *avis*, *question électoral*(es)

4. *Programmes* (des) *candidat*(s), *trouver*, *fait*(s).

5. *Ils* (ne) *parlent* (pas des) *problèmes* (de) *fond* (des) *Français*

6. *Nous* (les) *gens* : (ne) *pas savoir*, (ne) *pas vouloir*, *pouvoir dire*, *façon* (de) *voir*, *tout grand*, *le pouvoir*, (les) *petits*, la *politique*, (les) *hommes politiques*, (de) *droite (*comme de) *gauche.*

7. *Impression* (avoir l') : (c'est) *vrai*, *croire* (que l') *élection* (ne se) *passer*(a) (pas comme l') *an* (2002).

Ne sont pas classés : *débat*, *choix* et *année* (qui rejoignent les autres trop haut dans la classification pour pouvoir être rattachés à l'un des groupes).

La cohérence de la plupart des regroupements est évidente. Par exemple, dans le groupe 1, on retrouve l'expression « droit de vote », le verbe *changer* comme quasi-synonyme du substantif *changement*, etc. D'autres paraissent évidents *a posteriori* mais ne l'étaient pas *a priori*. Ainsi l'association de la première personne avec *opinion*, *avenir*, *falloir* (exercer) le *droit de vote* et *essayer* le *changement*… Ou encore : *nous les petits*, *nous les gens*. De même, le fait que la *politique*, les *politiques*, les *hommes politiques* soient opposés aux *problèmes* des *Français*, etc.

Enfin, certains regroupements sont contre-intuitifs : *droite* et *gauche* ou *Royal* et *Sarkozy* sont des univers très fortement associés. Ce sont, après *Ségolène + Royal*, les couples de colonnes dont les profils sont les plus proches et qui, pour cette raison, sont les premiers à être agrégés. Cela signifie que, dans l'esprit du plus grand nombre des enquêtés, ces mots vont ensemble – quand on parle de l'un, l'autre vient à l'esprit et on leur associe le même vocabulaire - alors qu'on attendrait plutôt des univers opposés (quand on choisit l'un, on repousse l'autre…)

En définitive, on peut réduire l'infinie diversité des réponses à 7 classes principales que l'on peut résumer ainsi.

L'enquêté a (ou n'a pas) envie de voter :

1. *Parce que c'est son devoir de citoyen, c'est important d'élire un président et d'essayer de changer* ;

2. *Pour qu'il se passe des choses intéressantes, différentes ou nouvelles en France, qu'il y ait du travail et un meilleur niveau de vie* ;

3. *Pour le parti (ou Sarkozy, Ségolène Royal) qui est le plus proche de mes idées en ce moment* ;

4. *A cause du programme du candidat qu'il a choisi* ;

5. *Parce qu'ils font des promesses et ne parlent pas des problèmes de fond des Français* ;

6. *Parce que les hommes politiques, de droite comme de gauche, veulent le pouvoir et ne savent pas comment les gens vivent* ;

7. *Parce qu'il ne faut pas que l'élection se passe comme en 2002*.

Il est donc possible de compresser le nombre des colonnes (de 77 à 7) et de classer chacun des enquêtés en fonction de sa proximité (ou de son éloignement) par rapport à ces 7 opinions. Puis un intervalle est délimité autour de la moyenne des scores, de ± un écart type. On considère que :

- un score supérieur à cet intervalle indique une forte adhésion de l'enquêté à l'opinion. On propose de considérer qu'il partage "tout à fait" cette opinion (pour utiliser le jargon des enquêtes d'opinion…)

- un score compris dans cet intervalle et supérieur à la moyenne indique que l'enquêté partage plutôt cette opinion mais ne la privilégie pas ;

- un score compris dans cet intervalle et inférieur à la moyenne indique que l'enquêté se distancie par rapport à cette opinion sans la rejeter tout à fait ;

- un score inférieur à cet intervalle signifie que l'enquêté n'adhère pas du tout à cette opinion.

Lorsque le vocabulaire utilisé ne permet pas de classer l'enquêté, on considère qu'il ne l'a pas abordé.

Le tableau 8 résume les résultats.

Tableau 8. Synthèse des 1444 réponses ouvertes selon que l'enquêté adhère ou rejette les 7 classes principales d'opinion

Opinion	Partage tout à fait		Partage plutôt		Plutôt pas d'accord		Pas du tout d'accord		Ne l'aborde pas	
	N	%	N	%	N	%	N	%	N	%
1	337	23,3	546	37,8	144	10,0	7	0,5	410	28,4
2	208	14,4	485	33,6	379	26,2	127	8,8	245	17,0
3	155	10,7	542	37,5	361	25,0	101	7,0	285	19,7
4	228	15,8	423	29,3	353	24,4	237	16,4	203	14,1
5	161	11,1	408	28,3	404	28,0	289	20,0	182	12,6
6	174	12,0	507	35,1	371	25,7	149	10,3	243	16,8
7	34	2,4	450	31,2	579	40,1	284	19,7	97	6,7

La première opinion (*le devoir du citoyen, l'importance de l'élection, le changement*) recueille un large assentiment : 23% la privilégient et 38% lui sont plutôt favorables (soit 61% des enquêtés). En revanche, si pratiquement personne ne s'affiche résolument contre, 28% des enquêtés ne l'ont pas mentionnée. Il s'agit d'une sorte d'opinion commune, ou de convention, dans laquelle se sont réfugiés un grand nombre d'enquêtés. Certains ont refusé d'en dire plus, d'autres l'ont complétée à la suite des relances de l'enquêteur. Le fait que la majorité des répondants partagent cette opinion rend évidemment la classification plus compliquée. Dans un tel cas, il est possible de scinder cette opinion en deux ou trois classes plus homogènes. En contrepartie, les effectifs en dernière colonne augmenteront sensiblement…

Les deux opinions suivantes (*je vais voter pour qu'il se passe des choses neuves* et *pour le parti le plus proche de mes idées*) sont également reprises à son compte par une majorité relative importante (48%) contre 35% qui la rejettent. A l'opposé, la 5e (*ils font des promesses et ne s'intéressent pas aux problèmes des Français*) est partagée seulement par 39% des enquêtés (et rejetée par 48 %). Quant à la dernière (*éviter la situation de 2002*), elle n'est partagée que par 34% des enquêtés (et rejetée par 6 enquêtés sur 10).

Enfin, la dernière colonne indique que le précédent de 2002 était présent dans les propos de pratiquement tous les enquêtés et que c'était l'opinion la plus clivante, avec le rejet de la classe politique (exprimé en 5 et 6).

Conclusions

Du point de vue technique, l'opération consiste à "fermer" les réponses ouvertes qui peuvent maintenant être analysées à l'aide des mêmes outils que les autres. En particulier, on peut la croiser avec les variables socio-démographiques, les attitudes et les intentions de vote.

Naturellement, cette analyse ne fait qu'estimer la réponse probable de l'enquêté si on lui avait soumis l'opinion reconstituée a posteriori. Rien n'assure que tel aurait bien été le cas.

On pourrait imaginer une procédure en deux temps. Premièrement, des entretiens semi-directifs sur un panel limité, analyse lexicométrique de ces verbatim, débouchant sur les opinions les plus fréquemment exprimées avec une formulation canonique issue des mots les plus utilisés dans les entretiens. Deuxièmement, dans un sondage classique, présentation de ces courtes phrases aux enquêtés sous forme de questions fermées.

Il faudrait également aborder la question de la compétence politique. En effet, dans les années 1980, une controverse a eu lieu en Amérique du nord autour de cette notion appliquée aux questions ouvertes (Geer 1988 & 1991). Une majorité de chercheurs estimaient que cette technique n'avait guère d'intérêt car les réponses mesuraient, non pas des opinions, mais une capacité plus ou moins grande à produire un discours (selon le niveau culturel de l'enquêté et son intérêt pour la politique). Certains ont même affirmé que la plupart des enquêtés ne faisaient que répéter des formules toutes faites entendues dans les médias, sans livrer leur opinion réelle. Plus récemment, Brugidou et Moine ont repris cette question - de l'influence de la compétence politique sur les réponses aux questions ouvertes - dans une optique nouvelle (2011). Pour avancer sur cette question, il faudrait croiser les opinions exprimées dans la question ouverte avec les réponses aux questions portant sur l'intérêt pour la politique ou le niveau culturel des enquêtés.

Enfin il faut souligner que ces traitements sont entièrement automatiques. A aucun moment n'interviennent des éléments extérieurs aux propos tenus par les enquêtés. En supposant que la question était pertinente et qu'elle a été administrée dans les règles de l'art à un échantillon représentatif de l'électorat, ces techniques permettent de reconstituer l'état de l'opinion, non pas à partir d'une grille pré-établie, mais en utilisant les propos, réellement tenus par les enquêtés, pour remonter jusqu'aux univers intellectuels qu'ils expriment.

Remerciements

Les responsables de l'enquête de 2007 qui ont bien voulu nous communiquer les données utilisées dans cette présentation.

Mathieu Brugidou et Michèle Moine qui ont opéré une première remise en forme des réponses et qui nous ont communiqué leur propre analyse.

Bibliographie

Brugidou M. (1998). Epitaphes, l'image de François Mitterrand à travers l'analyse d'une question ouverte posée à sa mort. *Revue française de science politique*, 48(1), p. 97-120.

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

Brugidou Mathieu et Escoffier Caroline. Questions ouvertes et opinion publiques discursives. In Marc Xavier et Tchernia Jean-François (dir). *Etudier l'opinion*. PUG : Grenoble, 2007, p. 91-111.

Brugidou Mathieu & Moine Michèle (2010). Normes émergentes et stigmatisation. Bolasco Sergio, Chiari Isabella, Giuliano Luca (Eds). *Proceedings of 10th International Conference Statistical Analysis of Textual Data*. Rome : Edizioni Universitarie di Lettere Economia Diritto, 2010, Vol 1.

Brugidou Mathieu (2011). Stigmatisation et dénonciation : entre adhésion doxique et distance critique. Congrès de l'Association française de sociologie. Grenoble.

Caillot Philippe et Moine Michèle (2001). Mais quelle est la réponse ? *Journal de la Société Française de Statistique*. 142-4, décembre 2001, p. 73-90.

Geer John G. (1988). What Do Open-Ended Questions Measure? *Public Opinion Quarterly* 52(3):365-371.

Geer John G. (1991). "Do Open-Ended Questions Measure 'Salient' Issues? *Public Opinion Quarterly*, 55(3), p 358-368.

Hubert Pierre & Labbé Dominique (1995). "La structure du vocabulaire du général de Gaulle". Communication aux 3e journées internationales d'analyse des données textuelles. In Bolasco Sergio et al. *IIIe Giornate internazionali di Analisi Statistica dei Dati Testuali*. Rome : Centro d'Informazione e stampa Universitaria, 1995, tome II, p. 165-176.

Labbé Cyril & Labbé Dominique (1994). Que mesure la spécificité du vocabulaire ? Grenoble : CERAT, décembre 1994 et juin 1997. Reproduit dans *Lexicometrica*, 3, 2001.

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 (2010). La modalité verbale en français contemporain. Les hommes politiques et les autres. *Communication aux XIe Journées de l'ERLA*. Brest : 19 novembre 2010.

Labbé Cyril & Labbé Dominique (2011). La classification des textes. *Images des mathématiques*. 28 mars 2011. (http://images.math.cnrs.fr/La-classification-des-textes.html).

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

Labbé Dominique (2001). Normalisation et lemmatisation d'une question ouverte. Les femmes face au changement familial. *Journal de la Société Française de Statistique*. 142-4, décembre 2001, p. 37-58.

Labbé Dominique (2010). *Le calcul du sens des mots. La lexicologie assistée par ordinateur*. Université de Neuchâtel, Institut de Mathématiques. Séminaire mathématique et société. 3 décembre 2010.

Lebart Ludovic et Salem André (1994). *Statistique textuelle*. Paris : Dunod.

Marc Xavier (2001). Les modalités de recueil des réponses libres en institut de sondage. . *Journal de la Société Française de Statistique*. 142-4, décembre 2001, p. 21-28.

Meyer D., Hornik K & Feinerer I. (2008). *Text mining infrastructure in r*. 25(5):569–576.

Nelson Gerald (1997). "Standardizing Wordforms in a Spoken Corpus". *Literary and Linguistic Computing*, 12, 2 , p 79-85.

Peretti Gaël de (2005). La "mise en variables" des textes : mythe ou réalité ? *Bulletin de méthodologie sociologique*. Octobre 2005, p. 5-30.

Pibarot André et 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.

Piolat Annie et Bannour Rachid (2009). EMOTAIX : un Scénario de Tropes pour l'identification automatisée du lexique émotionnel et affectif. *L'Année Psychologique*. 109-4, pp. 655-698.

Roux Maurice (1985). *Algorithmes de classification*. Paris : Masson (ouvrage disponible : http://www.imep-cnrs.com/docu/mroux/algoclas.pdf)

Roux Maurice (1994). *Classification des données d'enquête*. Dunod.

Annexe 1.

les 100 vocables les plus employés dans les réponses à la question ouverte

Rang	Lemme et catégorie grammaticale	Fréquence absolue	Fréquence relative (‰)
1	le (det)	9893	69.5
2	de (pré)	7500	52.7
3	être (v)	5802	40.8
4	je (pro)	5462	38.4
5	ce (pro)	4143	29.1
6	que (cj)	3664	25.8
7	avoir (v)	3641	25.6
8	pas (adv)	3366	23.7
9	et (cj)	3032	21.3
10	il (pro)	2515	17.7
11	à (pré)	2259	15.9
12	un (det)	2184	15.4
13	on (pro)	2170	15.3
14	pour (pré)	2047	14.4
15	qui (pro)	1808	12.7
16	parce que (cj)	1733	12.2
17	ça (pro)	1643	11.6
18	euh (loc)	1636	11.5
19	***aller (v)***	1573	11.1
20	ben (loc)	1521	10.7
21	y (pro)	1511	10.6
22	ne (adv)	1390	9.8
23	ils (pro)	1331	9.4
24	que (pro)	1198	8.4
25	***voter (v)***	1138	8.0
26	***faire (v)***	1066	7.5
27	***dire (v)***	1043	7.3
28	en (pré)	1040	7.3
29	se (pro)	895	6.3
30	voilà (pré)	830	5.8
31	mais (cj)	794	5.6
32	donc (cj)	793	5.6
33	non (adv)	742	5.2
34	tout (pro)	719	5.1
35	heu (loc)	703	4.9
36	quoi (pro)	699	4.9
37	plus (adv)	693	4.9
38	peu (adv)	683	4.8
39	dans (pré)	656	4.6
40	***chose (n f)***	655	4.6
41	***candidat (n m)***	643	4.5
42	tout (det)	628	4.4
43	***falloir (v)***	627	4.4
44	***savoir (v)***	607	4.3
45	moi (pro)	590	4.2
46	***envie (n f)***	555	3.9
47	bien (adv)	517	3.6
48	vous (pro)	515	3.6
49	***penser (v)***	503	3.5
50	***pouvoir (v)***	501	3.5
51	le (pro)	496	3.5
52	***donner (v)***	487	3.4
53	oui (adv)	485	3.4
54	***voir (v)***	466	3.3
55	***trouver (v)***	464	3.3
56	sur (pré)	462	3.3
57	en (pro)	460	3.2
58	par (pré)	458	3.2
59	ce (det)	451	3.2
60	mon (det)	451	3.2
61	puis (adv)	442	3.1
62	là (adv)	436	3.1
63	***vouloir (v)***	432	3.0
64	très (adv)	423	3.0
65	quand (cj)	408	2.9
66	si (cj)	396	2.8
67	***campagne (n f)***	388	2.7
68	enfin (adv)	385	2.7
69	même (adv)	375	2.6
70	trop (adv)	373	2.6
71	bon (adv)	372	2.6
72	comme (cj)	363	2.6
73	autre (pro)	357	2.5
74	beaucoup (adv)	346	2.4
75	toujours (adv)	338	2.4
76	***raison (n f)***	337	2.4
77	***France***	331	2.3
78	***fait (n m)***	330	2.3
79	***gens (n m)***	320	2.3
80	nous (pro)	316	2.2
81	***changer (v)***	314	2.2
82	vraiment (adv)	314	2.2
83	ou (cj)	285	2.0
84	bah (loc)	284	2.0
85	hein (loc)	278	2.0
86	leur (det)	269	1.9
87	assez (adv)	267	1.9
88	***débat (n m)***	266	1.9
89	***parler (v)***	266	1.9
90	alors (adv)	265	1.9
91	***idée (n f)***	254	1.8
92	***programme (n m)***	251	1.8
93	***important (adj)***	248	1.7
94	autre (det)	244	1.7
95	***problème (n m)***	240	1.7
96	rien (pro)	239	1.7
97	comment (adv)	238	1.7
98	passer (v)	223	1.6
99	déjà (adv)	217	1.5
100	deux (num)	213	1.5
		109 180	767,3

Annexe II.
les principaux vocables

Les premiers verbes

Rang	Vocable	Effectif	Fréquence (‰)
1	être	5802	40.77
2	avoir	3641	25.59
3	aller	1573	11.05
4	voter	1138	8.00
5	faire	1066	7.49
6	dire	1043	7.33
7	falloir	627	4.41
8	savoir	607	4.27
9	penser	503	3.53
10	pouvoir	501	3.52

Les premiers substantifs

Rang	Vocable	Effectif	Fréquence (‰)
1	chose	655	4.60
2	candidat	643	4.52
3	envie	555	3.90
4	campagne	388	2.73
5	raison	337	2.37
6	France	331	2.33
7	fait	330	2.32
8	gens	320	2.25
9	débat	266	1.87
10	idée	254	1.79
11	programme	251	1.76
12	problème	240	1.69
13	devoir	206	1.45
14	changement	192	1.35
15	Sarkozy	189	1.33
16	personne	187	1.31
17	question	186	1.31
18	rapport	167	1.17
19	niveau	164	1.15
20	impression	162	1.14

Les premiers adjectifs

Rang	Vocable	Effectif	Fréquence (‰)
1	important	248	1.74
2	politique	192	1.35
3	petit	181	1.27
4	bon	164	1.15
5	vrai	144	1.01
6	intéressant	113	0.79
7	électoral	102	0.72
8	nouveau	81	0.57
9	différent	76	0.53
10	grand	73	0.51

Rang	Vocable	Effectif	Fréquence (‰)
1	je	5462	38.39
2	ce	4143	29.12
3	il	2515	17.67
4	on	2170	15.25
5	qui	1808	12.71
6	ça	1643	11.55
7	y	1511	10.62
8	ils	1331	9.35
9	que	1198	8.42
10	se	895	6.29

Annexe 3

Univers lexical de "devoir" (substantif masculin)
Vocables significativement sur-employés dans l'univers
(Seuil : 5, classement par catégories grammaticales et spécificité décroissante)

Noms propres : Le Pen, France
Verbes : aller, voter, falloir, aider, assumer, soutenir, estimer, vouloir, évoluer, donner, participer, croire, travailler, penser, attendre
Substantifs : envie, campagne, citoyen, droit, démocratie, chance, président, citoyenne, pays, obligation, vote, enfant, valeur, peur, avis, chose, façon, république, point, avenir, raison
Adjectifs : électoral, citoyen, civique, spécial, essentiel, normal, important, français
Pronoms : je, ce, le, nous, leur, vous, il, moi
Adverbes : non, bien, même, d'abord, autant, spécialement, où, bon, d'accord, là, enfin, toujours, puis
Déterminants : un, mon, notre, son, tout
Conjonctions et prépositions : de, mais, si, quand, pour, depuis

Vocables significativement sous-employés dans l'univers
(Seuil : 5, classement par catégories grammaticales et spécificité décroissante)

Noms propres : Ségolène, Royal,
Verbes : aborder, tenir, trouver, dire, avoir, voir, parler, importer, passer, sembler, tirer, savoir
Substantifs : rapport, pouvoir, exemple, proposition, fait, parti, impression, homme, discours, instant, confiance, promesse, candidat, média, fond, problème, débat, sujet, personne, question, idée, droite
Adjectifs : intéressant, concret, principal, politique
Pronoms : se, que, un, autre, qui, ils, ça, même, quoi, tout, dont
Adverbes : vraiment, très, trop, assez, peu, plus, justement, plutôt, maintenant, peut-être, comment
Déterminants : leur, deux, le, aucun, ce
Conjonctions et prépositions : que, à, et, par, sur, entre, avec, donc

Phrases les plus significatives de l'univers lexical de *devoir* avec leurs scores

Ben, parce qu'il faut aller voter, hein ben, c'est un devoir, pour répondre à un devoir à un devoir de citoyen. (0.508)

Bah, c'est un droit et un devoir d'aller voter comme tout citoyen. (0.500)

Il s'agit de l'avenir de notre pays, de faire avancer les choses, voilà ben, je vais faire mon devoir de citoyen, c'est quelque chose de normal, on me demande mon avis, c'est pas tous les jours. (0.450)

Parce que voter c'est un devoir, tout simplement ben qu'il faut, il nous faut un président, il faut quelqu'un qui dirige la France donc, voilà, mais j'ai pas de raison spéciale d'aller voter quoi, il faut le faire. (0.400)

Parce que je pense que c'est important d'aller voter, enfin une obligation en tant que citoyen, j'en ai pas spécialement, c'est pas une obligation, c'est un droit et un devoir, je dirais : rien de spécial. (0.380)

Parce que j'estime que c'est important de… d'exercer son droit de vote, parce que c'est un devoir, parce qu'on est peut-être arrivé à un tournant, où cette campagne-là va peut-être changer la politique de la France, pour une envie de changement. (0.300)

Parce que toutes les campagnes présidentielles et électorales me donnent très envie de voter, parce qu'en démocratie, c'est le moment où l'on doit donner son avis et voilà, je me sens investi de ce devoir pour l'avenir de mes enfants. (0288)

C'est un devoir citoyen c'est-à-dire qu'on a la chance de pouvoir voter, il faut y aller s'il peut y avoir un changement, enfin bon entre guillemets, on va dire. (0.281)

Et bien parce que, disons que l'enjeu est important et, si l'on veut que notre pays soit gouverné de façon positive, c'est ce qui me motive d'aller voter pour Nicolas Sarkozy, ben, pour faire mon devoir de citoyen, j'ai jamais manqué un vote, donc c'est pas aujourd'hui que je vais commencer. (0.270)

Parce que j'estime que c'est notre devoir d'aller voter, nos grands-mères se sont battues pour le droit de vote aux femmes, donc je pense que c'est un devoir et un privilège. (0.257)

Euh ben, moi, la première raison, c'est-à-dire que je vais tout le temps voter parce que je considère ça comme un devoir, donc c'est pas spécialement cette campagne, c'est pour ça que je dis assez envie, parce que j'y vais systématiquement, ben, je vous dis c'est pour cette raison là principalement, parce que je fais mon devoir de citoyen. (0.250)