Mots. Les langages du politique

108 | 2015
Thèmes et thématiques dans le discours politique

Des mondes lexicaux aux représentations sociales. Une première approche des thématiques dans les débats à l'Assemblée nationale (1998-2014)

From lexical scopes to social representations. A preliminary thematic approach to the National Assembly debates (1998-2014)

Desde los mundos lexicales a las representaciones sociales. Una primera aproximación de las temáticas en los debates en la Asamblea nacional (1998-2014)

Pierre Ratinaud et Pascal Marchand

Édition électronique
URL : https://journals.openedition.org/mots/22006
DOI : 10.4000/mots.22006
ISSN : 1960-6001

Éditeur
ENS Éditions

Édition imprimée
Date de publication : 6 octobre 2015
Pagination : 57-77
ISBN : 978-2-84788-727-3
ISSN : 0243-6450

Référence électronique
Pierre Ratinaud et Pascal Marchand, « Des mondes lexicaux aux représentations sociales. Une première approche des thématiques dans les débats à l'Assemblée nationale (1998-2014) », *Mots. Les langages du politique* [En ligne], 108 | 2015, mis en ligne le 06 octobre 2017, consulté le 23 avril 2022. URL : http://journals.openedition.org/mots/22006 ; DOI : https://doi.org/10.4000/mots.22006

Pierre Ratinaud, Pascal Marchand

Des mondes lexicaux aux représentations sociales. Une première approche des thématiques dans les débats à l'Assemblée nationale (1998-2014)*

Objectif

L'extraction de thématiques constitue un enjeu important lorsqu'on est confronté à un grand corpus textuel. Outre la taille des corpus (considérée aujourd'hui en millions d'occurrences), la question qui se pose est, d'une part, celle de l'homogénéité et, d'autre part, de la recherche d'une description de ce qu'il contient, de l'importance des parties qui le composent et de sa structure, c'est-à-dire de l'organisation des oppositions et des rapprochements entre ces thématiques. Ainsi, lorsque nous avions proposé une analyse de l'intégralité des 251 287 télégrammes du « CableGate » de *Wikileaks* (238 116 128 occurrences), les résultats s'exprimaient sous forme d'une soixantaine de classes, rendant délicate une synthèse interprétative (Ratinaud, Marchand, 2012). Nous notions néanmoins que, si l'on faisait varier le rang et la fréquence des formes lexicales (5 000 premières formes ou 5 000 suivantes dans l'index), les analyses présentaient des similarités certaines et nous semblaient renforcer l'hypothèse des « mondes lexicaux » (Reinert, 1993, 2007). C'est à nouveau cette optique qui nous guide ici, en proposant une méthode à base de classifications non supervisées mais à focales successives[1].

Nous partirons d'un principe très général selon lequel une thématique peut être définie comme un ensemble de formes pleines cotextuelles[2] liées entre

* Ce travail a été réalisé dans le cadre du Labex SMS, portant la référence ANR-11-LABX-0066, a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du programme Investissements d'avenir portant la référence ANR-11-IDEX-0002-02.

1. La démarche proposée ici a déjà été testée dans un autre contexte (Loubère, 2014).
2. Apparaissant dans les mêmes unités textuelles, qu'il s'agisse du texte ou de segments de texte en fonction des orientations de l'analyse.

Université de Toulouse, Laboratoire d'études et de recherches appliquées en sciences sociales (LERASS)
ratinaud@univ-tlse2.fr
Université de Toulouse, Laboratoire d'études et de recherches appliquées en sciences sociales (LERASS)
pascal.marchand@iut-tlse3.fr

elles par leur objet[3] et leur contexte[4]. Notons immédiatement qu'une définition par l'objet seulement nous renverrait à une structure paradigmatique telle qu'elle peut être opérationnalisée par des analyseurs de réseaux sémantiques ou des classifications supervisées. L'introduction du contexte permet d'envisager qu'une classification lexicale puisse définir des sous-classes selon le vocabulaire privilégié par des locuteurs à propos d'une même thématique, ce qui nous amènera à évoquer l'intervention des représentations sociales (Moscovici, 1961).

Le lien entre le texte et ses conditions de production est à l'origine des « mondes lexicaux » tels que définis initialement par Reinert (1993). En effet, le tableau lexical utilisé comme point de départ de l'analyse qu'il propose[5] croise des énoncés, qui renvoient au « point de vue » du sujet énonciateur, et des lexèmes, qui renvoient à l'objet référentiel. Un « monde lexical » était alors défini comme la trace d'un lieu référentiel et l'indice d'une forme de cohérence liée à l'activité spécifique du sujet énonciateur. Dans le cas où le sujet était collectif, Reinert précisait qu'on avait affaire à des « lieux communs » (à un groupe, à une collectivité, à une époque…) : « de ce fait, ils peuvent s'imposer davantage à l'énonciateur qu'ils ne sont choisis par lui, même si celui-ci les reconstruit, leur donne une coloration propre. Un recouvrement avec la notion de représentation sociale apparaît donc ici assez clairement » (*ibid*., p. 12).

Par la suite, Reinert reviendra sur le lien qu'il établissait entre mondes lexicaux et représentations sociales. Son expérience avec des corpus très divers l'amenait, dès 1997, à identifier des similitudes interprétatives dans les résultats des analyses et à poser l'hypothèse d'une stabilisation des mondes lexicaux dont l'origine était à rechercher du côté de la dynamique de fabrication des signes et d'une phénoménologie impliquant une logique du sens. Il distinguait ainsi trois grands types de mondes associés à trois types de discours : « Un monde d'émotions et de perceptions, plus naturel, plus féminin, associé au "discours poétique" ; un monde plus engagé dans le monde réel, politique ou artistique, et associé à un "discours polémique" ; et enfin, un monde plus conceptuel, celui des théories et des projets où Freud et Marx ont une place particulière, associé au "discours philosophique". » Plus récemment, sous l'influence de la sémiotique percienne, Reinert (2007) optera pour des « postures énonciatives », plus génériques : *du Témoin, de l'Acteur, du Patient*.

Il est à noter que les analyses de Reinert concernaient souvent des corpus de tailles relativement réduites (des œuvres littéraires, des récits de cauchemars, six numéros d'une revue sur le surréalisme) et des classifications de trois

3. La référence : ce dont les formes traitent, ce qu'elles décrivent.
4. Le contexte doit s'entendre dans un sens très large. Il peut s'agir d'une unité de temps ou de lieu, de caractéristiques socio-démographiques des locuteurs ou d'une unité situationnelle de la production…
5. Ces analyses ont d'abord été implémentées du logiciel Alceste.

à quatre classes terminales seulement. La perspective présente voudrait s'inspirer de la démarche de Reinert, et notamment de l'idée de « mondes lexicaux stabilisés », mais on devra considérer des corpus d'une taille considérable dont la classification débouche sur plusieurs dizaines de classes lexicales. La stabilisation des mondes lexicaux sera donc liée, pour nous, à l'interaction entre des lexiques, des sujets et des contextes que nous rapprocherons évidemment de la façon dont Doise (1985) définit les représentations sociales.

Rappelons que les représentations sociales sont définies par Doise comme « principes générateurs de prises de position liées à des insertions spécifiques dans un ensemble de rapports sociaux et organisant les processus symboliques intervenant dans ces rapports » (*ibid.*, p. 245).

Ces « générateurs de prises de position », de nature sociétale, organisent donc la manière dont les individus perçoivent les antagonismes sociaux au sein d'un groupe social et se modulent aussi en fonction de leurs rapports aux valeurs communes qui y servent de repères (Doise, 2011). La « stabilisation » du lexique serait donc ici renvoyée à celle de la structure sociale : il y aurait une organisation normative des mots, qui permettrait d'investir des lexiques pour montrer notre appartenance groupale dans un contexte qui rend saillante et importante la catégorisation sociale. Pour pouvoir être correctement interprétés, encore faut-il que ces « marqueurs socio-langagiers » (Scherer, Giles, 1977) soient socialement organisés en structures normatives. Leur fonction est alors de maintenir le système social en identifiant les catégories sociales, en reconnaissant les membres qui y occupent des rôles et des positions hiérarchiques variés et en adaptant nos conduites à leur égard (Smith, Giles, Hewstone, 1979). Nous avons ainsi pu montrer, par l'analyse d'entretiens ou de mots associés, que les objets sociaux, dans leur diversité, étaient susceptibles d'être investis par les locuteurs pour marquer, au-delà des thématiques, des appartenances et des relations entre groupes (Marchand, 2008).

C'est cette double influence, d'une part, de l'existence de groupes d'appartenances, avec les exigences identitaires que cela implique, et d'autre part, de la sensibilité à un contexte en perpétuelle évolution, qui rend particulièrement pertinente l'analyse diachronique des discours politiques. Le positionnement droite/gauche reste effectivement un régulateur important des formes que revêtent les représentations sociales en politique française, tout en étant susceptible d'évoluer dans le temps. L'histoire des discours politiques peut ainsi révéler la base lexicale, syntaxique et sémantique sur laquelle s'organisent les interventions, et mettre en évidence les effets normatifs (Marchand, 2007).

Nous poserons donc qu'un corpus de comptes rendus des séances de débats à l'Assemblée nationale de 1998 à aujourd'hui devrait permettre de mettre en évidence des thématiques récurrentes qui révéleront la stabilité des « mondes lexicaux » structurant les débats, mais également des évolutions chronologiques et des prises de position politiques qui donneront à voir

la façon dont les groupes investissent lexicalement ces thématiques selon les contextes, les époques, les représentations et l'orientation des politiques des majorités en place.

Méthodologie

Récupération et formatage du corpus

Le site web de l'Assemblée nationale[6] propose, dans l'une de ses rubriques, la possibilité d'accéder aux comptes rendus des débats dans leur intégralité[7]. Ces comptes rendus sont disponibles à partir de l'année 1998, soit la seconde année de la XI^e^ législature. Nos premiers essais d'analyse de ce corpus ont été réalisés à partir d'une extraction du site, effectuée par Pierre Molette dans le cadre du développement du logiciel *Owledge*[8]. La collection des comptes rendus de séances étant incomplète, nous avons réaspiré ce site en limitant les requêtes aux répertoires qui recensent l'ensemble des documents relatifs aux différentes législatures, dont les comptes rendus intégraux font partie[9]. Dans l'arborescence de ce site web, les comptes rendus sont regroupés dans un répertoire particulier (répertoire « cri ») pour chaque législature. Ils se présentent sous la forme de fichiers HTML relativement simples dans leur structure, à l'exception des sessions de 2001-2002 dont le format a rendu impossible une extraction automatique[10]. Tous ces fichiers ont ensuite été transformés en format « texte »[11]. La dernière étape du formatage a consisté à réunir ces fichiers dans un corpus unique et à repérer à l'intérieur de ces textes les dates des séances de façon à les utiliser comme métadonnées. Finalement, ce corpus regroupe les comptes rendus de toutes les séances de l'Assemblée nationale qui ont eu lieu entre le 5 octobre 1998 et le 18 septembre 2014, à l'exception des séances de la session 2001-2002[12].

Description du corpus

Entre ces deux dates, l'Assemblée a accueilli 4 009 séances qui composent donc 4 009 textes distincts dans le corpus. Ces sessions sont réparties sur

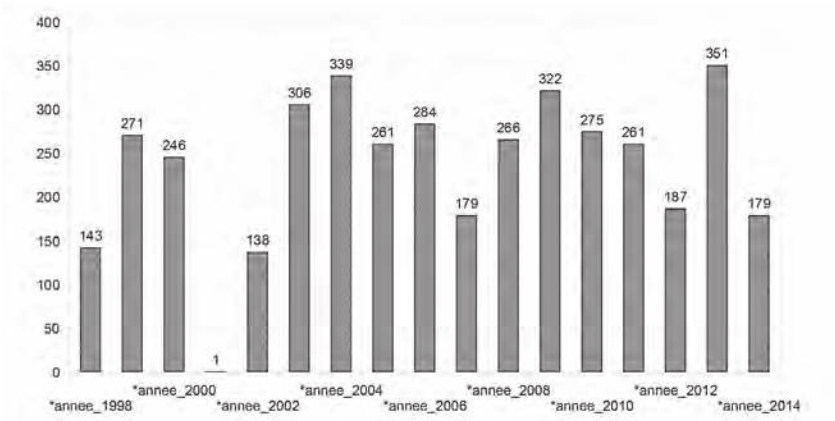
6. http://www.assemblee-nationale.fr/.
7. http://www.assemblee-nationale.fr/14/debats/index.asp.
8. http://www.owledge.org/.
9. Nous avons utilisé le logiciel libre wget pour cette tâche (http://www.gnu.org/software/wget/).
10. Nous disposons effectivement de ces données mais leur formatage en fichier texte nécessitera un travail supplémentaire.
11. Nous avons utilisé le logiciel libre w3m pour cette tâche (http://w3m.sourceforge.net/).
12. Depuis 2011, les comptes rendus des débats de l'Assemblée sont disponibles sous forme de fichiers XML très facilement exploitables (http://data.journal-officiel.gouv.fr/index.php?dir=Debats%2FAN%2F). Cette collection ne commence malheureusement qu'avec les débats de 2011.

1 779 jours (une journée peut donner lieu à plusieurs séances, souvent deux ou trois, rarement davantage : voir tableau 1 ci-après). Le graphique 1, *infra*, présente la répartition des textes par année et permet de visualiser la quasi-absence de l'année 2001, ainsi que la diminution du nombre de séances les années d'élections législatives (2002, 2007 et 2012).

Tableau 1 : Nombre de séances par jour

Nombre de séances	Effectif
1	217
2	908
3	646
4	5
6	3
Total	1779

Graphique 1. Fréquence des sessions parlementaires par année



Le corpus représente 137 988 880 occurrences pour 185 565 formes différentes (dont 55 721 hapax, soit 30 % des formes et 0,04 % des occurrences). La forme la plus fréquente, le pronom « le », comptabilise 6 988 580 occurrences.

Le parti pris thématique amène à opter pour quelques interventions lexicales. Tout d'abord, la reconnaissance morphosyntaxique des formes lexicales permet d'effectuer un tri pour distinguer les « mots pleins », porteurs de significations et considérés comme actifs dans l'analyse, et les « mots-outils », constructeurs des séquences textuelles et considérés comme illustratifs (ou

supplémentaires). Dès lors, la lemmatisation appliquée aux verbes, substantifs et adjectifs permet de concentrer la signification sur les formes réduites sans considérer leurs flexions. Encore une fois, ces choix, pertinents dans une optique d'extraction thématique, seraient discutables dans une autre, si l'on devait considérer l'argumentation, par exemple. Notons qu'ils ne sont pas définitifs dans l'outil utilisé et qu'une fois les thématiques extraites, elles peuvent être envisagées dans leur intégralité lexicale.

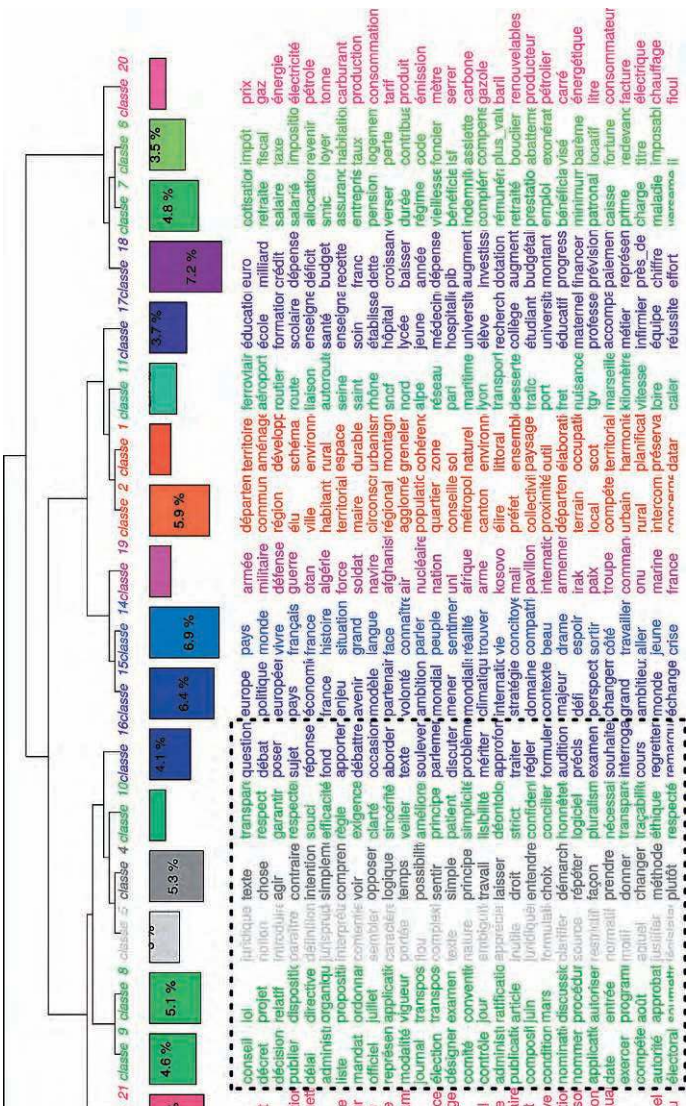
Afin de soumettre ce corpus à une analyse selon la méthode Reinert (Reinert, 1983, 1990 ; Ratinaud, Marchand, 2012), le logiciel Iramuteq (Ratinaud, 2009) l'a découpé en 3 909 220 segments de textes de 35,3[13] occurrences en moyenne. Rappelons que cette méthode d'analyse part d'un tableau de présence/absence qui croise les segments de texte avec les formes pleines du corpus. L'objectif de l'analyse, qui opère par bipartitions successives du tableau sur la base d'une analyse factorielle des correspondances, est de réunir les segments de texte qui ont tendance à contenir les mêmes formes dans des ensembles que l'on nomme « classe ».

À la hache

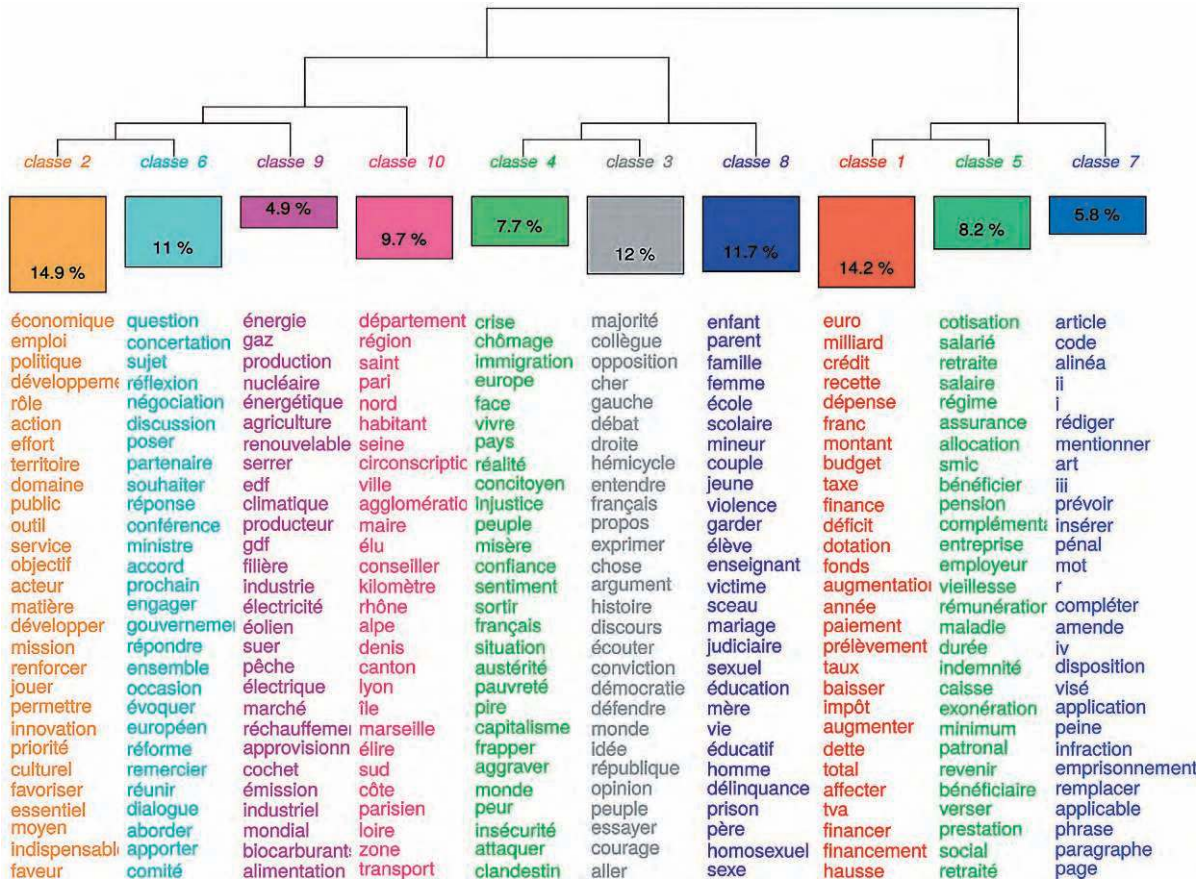
La première analyse que nous proposons est donc une classification de ces segments de texte. Nous avons conservé ici les 9 957 formes pleines les plus fréquentes. Avec ce paramétrage, la forme la moins fréquente représentée dans l'analyse apparaît 198 fois dans le corpus. Nous avons utilisé le mode « patate » proposé par IRaMuTeQ. Ce paramètre supprime la seconde phase de chaque étape de la classification (reclassement de chacune des lignes de la matrice : voir Ratinaud, Marchand, 2012). La perte de précision engendrée est très modeste au regard du gain de temps pour le traitement d'une matrice de presque 39 milliards de cases. Enfin, nous avons demandé au logiciel de construire 80 classes et de conserver celles contenant au moins 48 865 segments de texte (3 909 220/80). On notera que dans les résultats de cette classification, dont l'unité est le segment de texte, une même forme peut être présente dans plusieurs classes, d'une part selon les éventuelles ambiguïtés homographiques, d'autre part et de façon plus fondamentale, selon les thématiques dans lesquelles elle apparaît.

Le dendrogramme 1 (*infra*) rend compte des 22 classes terminales obtenues (91,37 % des segments de texte du corpus ont été classés) et d'une partie du lexique qui caractérise chacune d'entre elles (par ordre décroissant du chi2 de liaison aux classes).

13. Cette longueur est le résultat du mode de découpage des segments de texte qui utilise à la fois la ponctuation (l'algorithme essaie de respecter le découpage naturel de la langue en appliquant un coefficient aux différents signes de ponctuation) et la longueur demandée (ici, 40 occurrences).



Dendrogramme 1. Classification sur le corpus complet, taille des classes et extrait des lexiques caractéristiques des classes
(par chi2 décroissant de liaison aux classes)



Dendrogramme 2. Classification sur le sous-corpus thématique, taille des classes et extrait des lexiques caractéristiques des classes (par chi2 décroissant de liaison aux classes)

Le dendrogramme de cette analyse met en évidence les grandes forces de structuration qui organisent ce corpus. Les classes 22, 3, 12 et 13 (entourées en trait plein) sont les premières à se différencier. Elles regroupent les noms propres des intervenants (notamment énumérés dans les sommaires des séances) et le vocabulaire lié à l'organisation des tours de parole (classes 13 et 12), ainsi que les éléments hors débats qui apparaissent dans les comptes rendus, comme les applaudissements ou les interpellations provenant des bancs de l'Assemblée (classe 22) et le lexique lié aux votes et au comptage des voix (classe 3).

Un second groupe de classes (entouré en pointillés) présente un ensemble de champs lexicaux relatifs à l'élaboration de textes juridiques (classes 16, 10, 4, 5, 6 et 9). Ces classes renvoient à la façon de discuter les lois beaucoup plus qu'aux thématiques dont elles traitent.

De façon à centrer notre analyse sur les thèmes abordés (et non pas sur la façon de les aborder), nous avons extrait de cette analyse toutes les autres classes. Cela revient à générer automatiquement un nouveau corpus en éliminant l'ensemble des classes que nous venons de décrire, soit 47,7 % de segments classés. Cette opération de génération automatique de sous-corpus représente une avancée déterminante dans l'approche de très gros corpus, comme c'est ici le cas.

À la machette

Un tel sous-corpus, composé de toutes les classes non encadrées sur le dendrogramme 1, a été soumis à une classification, à nouveau selon la méthode de Reinert. L'intérêt est ici de travailler uniquement sur les segments de texte qui semblent effectivement rendre compte des thématiques qui suscitent les débats.

Description du corpus et analyse

4 006 des 4 009 séances originales sont représentées dans cette partie. Ce chiffre semble confirmer que la première analyse a effectivement permis de distinguer, séance par séance, les propos abordant le fond des sujets de ceux relatifs aux débats ou à la rédaction des textes. Les 1 867 369 de segments qui composent ce corpus (47,7 % du corpus total) contiennent 67 989 913 occurrences (dont 50 126 hapax qui représentent 33,99 % des formes et 0,07 % des occurrences).

Pour la classification, nous avons conservé les 9 932 formes pleines les plus fréquentes. La fréquence de la forme la moins représentée tombe ici à 103. Nous avons demandé 35 classes et retenu toutes celles de plus de 53 353 segments de texte (1 867 369/35).

Résultats

La classification produit 10 classes terminales qui regroupent 85,53 % du corpus (voir dendrogramme 2).

La lecture des profils de ces classes permet de constater qu'elles semblent effectivement composées de segments dont le lexique dessine les contours cohérents de questions sociétales régulièrement débattues. Nous verrons plus loin que la description que nous proposons ici ne peut dévoiler la complexité de chacune d'entre elles. Pour autant, cette classification nous semble rendre compte de l'importance relative de chacune de ces grandes thématiques dans les discussions des députés pour la période concernée. Nous proposons pour chaque classe une courte description ainsi que les trois segments de texte les plus représentatifs (les formes en italique dans ces segments correspondent aux formes pleines présentes dans le profil de ces classes).

Classe 2 : emploi, territoires et développement économique (14,9 %)

Cette classe regroupe les vocabulaires du champ économique au sens large.

rep_14 *file_2013 *annee_2012 *am_201211 *amj_20121129
dont la *compétence* en *matière* de *développement économique* fait les *acteurs principaux* de cette première *décentralisation* de la *politique industrielle* la *bpi constitue* un *outil pratique nécessaire* au *développement économique local* et de ce fait au *service* de l *emploi*

rep_11 *file_2001 *annee_2000 *am_200010 *amj_20001003
alors que les *nouvelles donnes* en *matière* de *développement local* et de *politique publique* de l *emploi conduisent* de plus en plus vers le *territoire* le *pays* le *bassin* d *emploi* c_est_à_dire vers l *initiative* d *acteurs locaux* qui se réunissent autour de *projets partagés* de *développement économique* et *social*

rep_11 *file_2001 *annee_2000 *am_200011 *amj_20001102
il *traduit* la *volonté* du *gouvernement* de *valoriser* ce *secteur important* du *développement économique* et *social* de notre *pays créateur* de *richesses* et d *emplois* l *action* pugnace et *efficace* que vous *menez* est *largement reconnue* et les *moyens* vous sont *donnés* d *impulser* une *véritable politique touristique* de gauche

Classe 6 : négociation et concertation (11 %)

Cette classe regroupe les segments contenant le lexique lié à la négociation et à ces modes de structuration à différents niveaux institutionnels.

rep_13 *file_2008 *annee_2008 *am_200806 *amj_20080618
eh bien l *union européenne* si elle veut être mieux *entendue* par les peuples doit *prendre* l habitude avant de *répondre* non aux *questions* qui sont *posées* d *étudier* les *sujets* en *concertation* avec l *ensemble* des *états* pour *dégager* des *solutions concrètes*

rep_14 *file_2013 *annee_2013 *am_201302 *amj_20130228
madame la *ministre* dans le *cadre* de notre *discussion* je *souhaiterais* vous *poser* plusieurs *questions* dès le lendemain de votre *nomination* vous avez *engagé* un tour des capitales *européennes* pour *convaincre* nos *partenaires* des autres *états membres* de la *pertinence* de l instrument de *réciprocité*

rep_11 *file_1999 *annee_1998 *am_199812 *amj_19981201
monsieur le *député* vous avez vous_même *répondu* à la *question* que vous *venez* de *poser* dans la mesure où vous avez *indiqué* que la *concertation* était *engagée* dans le cadre du *groupe* de *travail mis* en *place* par ma *collègue ministre* de l *aménagement* du *territoire* et de l *environnement*

Classe 9 : énergie, agriculture et écologie (4,9 %)

Les thématiques de l'énergie et de l'agriculture sont regroupées dans cette classe. Il y est aussi question d'écologie, notamment au travers de la question du réchauffement climatique.

rep_12 *file_2004 *annee_2004 *am_200406 *amj_20040616
l *impératif environnemental* d une réduction des *gaz* à *effet* de *serre* la *diversification* de notre *production énergétique* qui passe par le *développement* des *énergies renouvelables* la *maîtrise* des *consommations* et la rénovation du pôle *nucléaire*

rep_12 *file_2006 *annee_2005 *am_200511 *amj_20051104
cependant ce n est pas uniquement en *matière* de *sûreté* ou de *production* d *énergie nucléaire* que l état doit jouer son rôle mais sur toute la *chaîne énergétique électricité pétrole gaz hydroélectricité énergies renouvelables* politique *industrielle* politique des *transports*

rep_12 *file_2004 *annee_2004 *am_200404 *amj_20040415
assurément le *développement* des *énergies renouvelables* la *diversification énergétique* la réduction de l *émission* de *gaz* à *effet* de *serre* la création d emplois la *compétitivité économique* des *filières renouvelables* la décentralisation de la *production*

Classe 10 : Aménagement du territoire (9,7 %)

La politique locale et l'aménagement du territoire sont les thèmes majeurs de cette classe.

rep_11 *file_1999 *annee_1998 *am_199811 *amj_19981105
si jusqu_à présent le rhône l *isère* la *loire* les trois *départements* de la *région picardie* et la *seine saint denis* ont été *équipés* cette couverture s *étendra* à la fin de l année prochaine à *paris* à la *petite couronne* et à la *corse*

rep_13 *file_2010 *annee_2010 *am_201005 *amj_20100527
peuplé de seulement 130 752 *habitants répartis* sur 177 *communes* dont 31 comptent moins de cent *habitants* mon *département* compte aujourd_hui trente *conseillers généraux* et quatre *conseillers régionaux* qui sont des *élus* du *nord* de la *région provence alpes côte* d *azur*

rep_14 *file_2014 *annee_2014 *am_201407 *amj_20140718
des *communes* et des *communautés* d *agglomération essentiellement limitrophes* du *grand paris* ont pu assez logiquement *souhaiter rejoindre* cette *métropole composée* des quatre *départements* de *paris* du *val* de *marne* des *hauts* de *seine* et de la *seine saint denis*

Classe 4 : Un monde en crise (7,7 %)

Cette classe aborde le rapport au monde, à l'Europe et à la crise économique.

rep_13 *file_2010 *annee_2010 *am_201009 *amj_20100928
qui doivent gérer des *situations impossibles* et faire *face* à une *dérive* de leurs finances une *pression* à la hausse sur le *chômage* qui n en a pourtant pas besoin avec la *crise* et la perte de substance *économique* que *vit* le *pays* depuis trente ans

rep_14 *file_2013 *annee_2013 *am_201306 *amj_20130617
bernard accoyer alors que dans un *monde* en pleine reprise l *europe* stagne dans le *marasme* et *demeure* en *crise* et que notre *pays* est *confronté* à d *immenses difficultés récession* niveau *record* du *chômage*

rep_14 *file_2013 *annee_2013 *am_201304 *amj_20130410
chômage déficits publics activité *économique* à l arrêt perte de notre *influence* en *europe* et bien entendu *crise morale* qui nous *frappe* aujourd_hui mais le plus *incroyable* c est que vous vous *acharnez* à *casser* tout ce qui *marche* dans ce *pays*

Classe 3 : Des débats entre majorité et opposition (12 %)

Il s'agit ici de nouveau d'une classe regroupant des segments de texte qui semblent davantage porter sur la mise en scène des débats entre majorité et opposition que sur les thématiques de fond. Ces relations d'alliances et d'oppositions entre les groupes parlementaires intéressent sans doute davantage les politologues que l'analyse des contenus telle que nous la développons ici.

rep_14 *file_2014 *annee_2014 *am_201406 *amj_20140630
ce *débat oppose* d_abord *majorité* et *opposition droite* et *gauche* mais tout le *monde* le sait il traverse aussi la *majorité* comme l *opposition* nous l avons *vu* lors des *votes* sur le *discours* du premier *ministre*

rep_13 *file_2010 *annee_2010 *am_201001 *amj_20100122
l *œil* sur le chronomètre et le *doigt* sur la *bouche* pour *empêcher* ses *collègues* de *parler* l *opposition essayant* pour sa part *modestement* de *susciter* quelques *débats* dans cet *hémicycle* débats *refusés* par le *gouvernement* et la *majorité*

rep_14 *file_2013 *annee_2012 *am_201211 *amj_20121112
nous n avons pas été *convoqués* chez le premier *ministre cher collègue* quand on *entend* les *députés* de l *opposition* à cette *tribune* on *comprend* la *différence* entre la *politique* menée par cette *majorité* et l *ancienne*

Classe 8 : Famille et éducation (11,7 %)

Nous utiliserons cette classe pour montrer l'intérêt d'une approche en abîme pour ce genre de corpus. Elle sera donc à son tour soumise à une classification. À ce stade, elle semble regrouper les thématiques de la famille et de l'éducation.

rep_14 *file_2013 *annee_2013 *am_201303 *amj_20130312
la *réussite scolaire* des *enfants* est *liée* à l *existence* d une continuité *éducative* entre l *école* et la *famille* laquelle se matérialise par une participation *accrue* des *parents* dans la *vie scolaire* de leurs *enfants* et une *meilleure connaissance* des *situations familiales* par les *professeurs*

rep_11 *file_1999 *annee_1999 *am_199903 *amj_19990330
mais dans les *familles recomposées* c_est_à_dire lorsque les *parents* se sont remariés ou *vivent* en *couple* mais pas avec le *père* ou la *mère* de l *enfant* ce sont 8 de ces *adolescents* qui ont tenté de se suicider et 5 5 d entre eux qui ont *subi* des *violences sexuelles*

rep_13 *file_2010 *annee_2010 *am_201006 *amj_20100629
le premier est que nos collègues sénateurs *abandonnant* le concept de *violences familiales* nous *cantonnent* aux *violences* de *couple* des *jeunes femmes* sont en *effet* parfois *victimes* de leur *frère* et des *parents victimes* de leurs *enfants*

Classe 1 : Budget, recettes et dépenses (14,2 %)

Cette classe est la plus importante de l'analyse. Elle regroupe des segments de texte caractérisés par le vocabulaire de la gestion budgétaire de l'État.

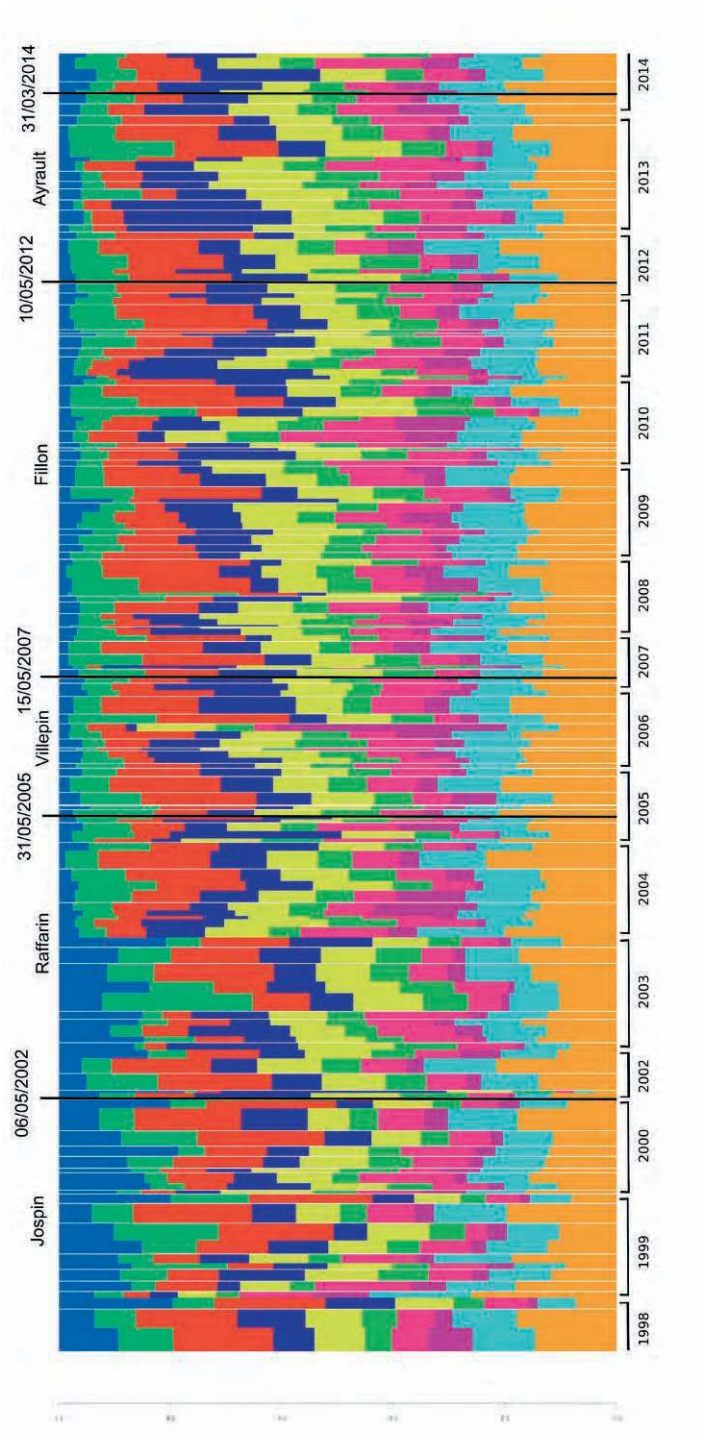
rep_12 *file_2002 *annee_2002 *am_200207 *amj_20020718
en *recettes* comme en *dépenses* du côté des *dépenses* ce *projet* ouvre des *crédits* pour un *montant* d environ 5 *milliards* d *euros* a défaut l *etat* n aurait pas été en *mesure* d honorer ses *engagements* au *titre* de l *année* 2002 m

rep_13 *file_2008 *annee_2007 *am_200711 *amj_20071107
bien au_delà des *crédits* de la *mission* dont le *montant* s *élève* à 12 3 *milliards* d *euros* il *inclut* en *effet* les *dépenses fiscales* 9 6 *milliards* d *euros* ainsi que les *recettes fiscales compensant* les *allégements* de *charges près_de* 27 *milliards* d *euros*

rep_12 *file_2003 *annee_2002 *am_200211 *amj_20021115
monsieur le ministre les *crédits proposés* pour 2003 au *titre* du *ministère* de l *économie* des *finances* et de l industrie s *élèvent* à *près_de* 15 *milliards* d euros soit 5 4 des *dépenses totales nettes* du *budget général*

Classe 5 : Système de protection sociale (8,2 %)

Toutes les facettes du système français de protection sociale sont présentes dans cette classe (retraite, prestations sociales, droit du travail, sécurité sociale, etc.).



Graphique 2. Proportion de chacune des classes par mois, de novembre 1998 (à gauche) à septembre 2014 (à droite)

rep_11 *file_2000 *annee_1999 *am_199912 *amj_19991202
les *employeurs* de 10 *salariés* ou plus sont *redevables* sur le *salaire versé* aux *apprentis* des *cotisations supplémentaires* d *accidents* du *travail* et de *retraite complémentaire* pour les *parts patronales* et *salariales* les *entreprises* doivent donc tout de même *acquitter* certaines *cotisations patronales*

rep_12 *file_2003 *annee_2003 *am_200306 *amj_20030604
un *chèque emploi entreprises* peut être utilisé pour *rémunérer* les *salariés* et pour *simplifier* les *déclarations* et *paiements afférents* aux *cotisations* et *contributions* dues au *régime* de *sécurité sociale* au *régime* d *assurance chômage* et aux institutions de *retraites complémentaires* et de *prévoyance* au *titre* de ces *salariés*

rep_12 *file_2004 *annee_2004 *am_200401 *amj_20040122
ce *seuil* d *activité minimale* de 50 *correspond* déjà à l *affiliation* des *salariés* au *régime* de la *mutualité sociale agricole* le *secteur* des *entreprises* paysagistes représente 10 de la masse *salariale agricole* 15 des *cotisations msa* et 33 des *versements* à la *caisse* de *retraite complémentaire* des *cadres* de l agriculture

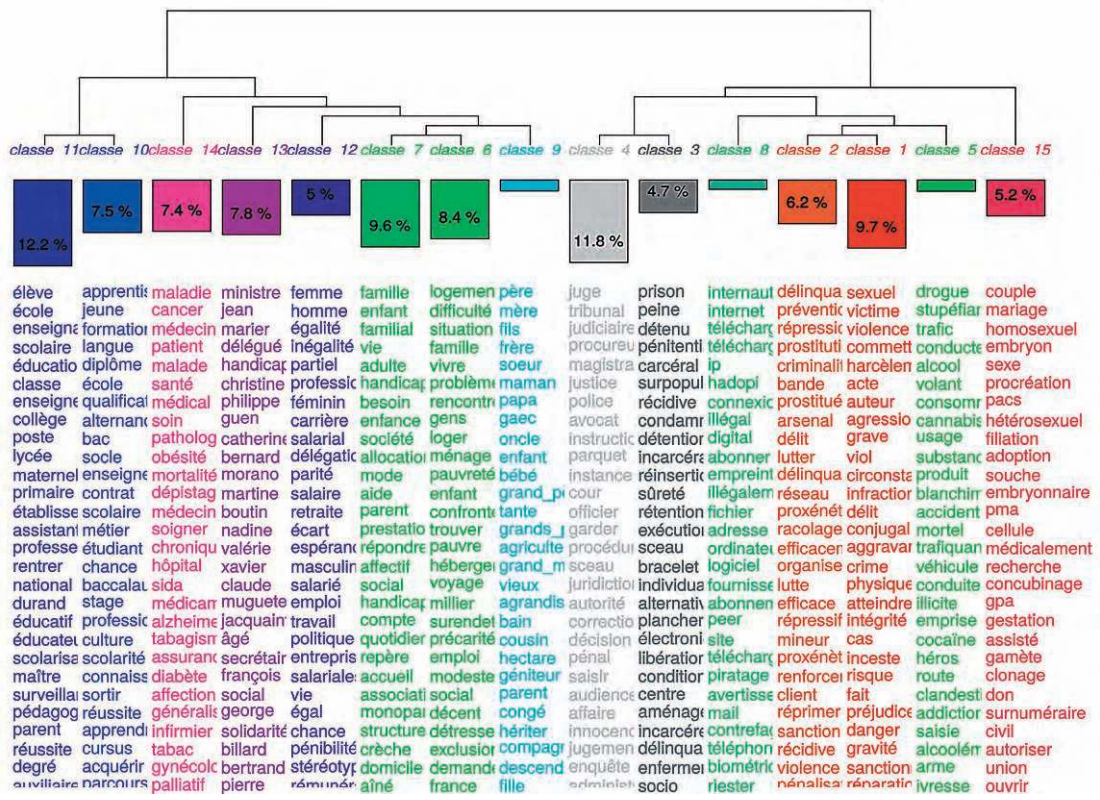
Classe 7 : Dans la structure des textes (5,8 %)

Cette classe regroupe un vocabulaire technique (*article*, *code*, *alinéa*…) spécifique de l'élaboration des textes juridiques. Nous verrons que l'apparition de ce vocabulaire a peut-être été régulée institutionnellement.

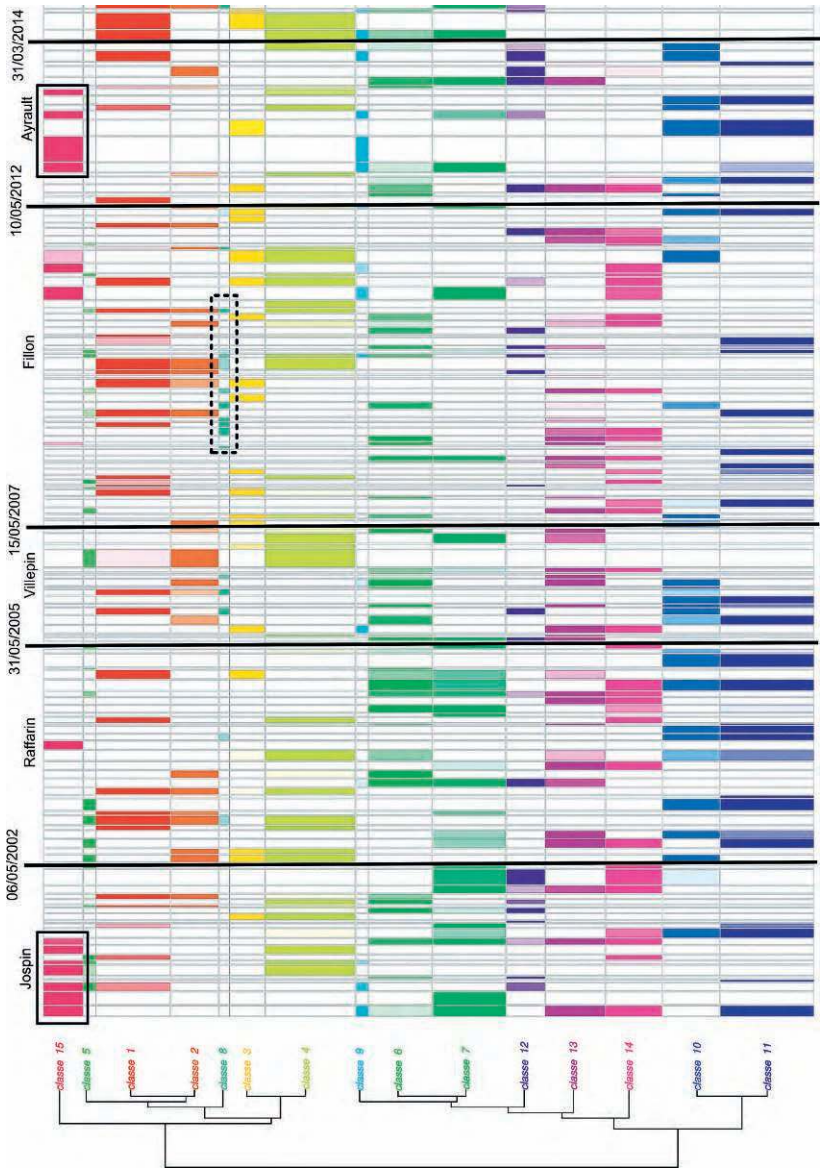
Ce dont rend compte cette première approche des thématiques abordées dans les débats, ce sont des objets de représentation, plus que des représentations elles-mêmes, qui sont principalement mobilisés par les députés et dont on peut alors penser qu'ils sont ceux qui préoccupent le plus le législateur. Le graphique 2 permet de voir que toutes ces thématiques pourraient, à ce niveau de généralité, être trivialement qualifiées de « serpents de mer », traversant toutes les législatures avec une rythmicité et une constance étonnantes.

Seule la classe 7 (en bleu tout en haut du graphique) présente une forte variation dans le temps. Elle disparaît presque totalement après novembre 2003. Nous pensons que ce phénomène est lié à une modification dans la façon de saisir les comptes rendus.

L'étape suivante de l'analyse consisterait donc à préciser le contenu de chacune de ces « méta-thématiques » en les soumettant chacune à une nouvelle classification. Il n'est bien sûr pas possible de rendre compte de toutes ces analyses ici. Nous avons donc choisi de nous centrer sur l'une d'entre elles : la classe 8 qui porte sur la famille au sens large.



Dendrogramme 3. Classification du sous-sous-corpus « famille », taille des classes et extrait des lexiques caractéristiques des classes (par chi2 décroissant de liaisons aux classes)



Graphique 3. De gauche à droite – dendrogramme de la classification et graphique de la sur-représentation des classes par mois, de novembre 1998 à septembre 2014

Au scalpel

Description du corpus et analyse

La classe 8, considérée comme un sous-sous-corpus (c'est-à-dire comme un sous-corpus du sous-corpus précédent), compte 6 766 785 occurrences séparées en 186 000 segments de texte. Elle a, à son tour, été soumise à une classification sur les 6 000 formes pleines les plus fréquentes, ce qui permet ici de retenir toutes les formes apparaissant au moins 26 fois dans le texte. Nous avons demandé 20 classes et retenu celles qui regroupaient au moins 1 000 segments de textes. Cette valeur a été choisie pour permettre une plus grande précision dans la différenciation des sous-thématiques qui composent ce nouveau corpus. À ce stade, la taille du corpus et son homogénéité relative justifie l'utilisation de tels paramètres.

Résultats

L'analyse présente 15 classes terminales qui regroupent 99,7 % des segments de texte. Le dendrogramme 3 rend compte de cette classification.

Comme on pouvait s'y attendre, les « mondes lexicaux » exprimés dans les classes de cette analyse apparaissent beaucoup plus homogènes. De gauche à droite dans ce graphique, nous voyons les discours sur l'enseignement scolaire (classe 11) se distinguer de ceux relatifs à l'apprentissage et à l'insertion professionnelle (classe 10). Nous trouvons ensuite une classe traitant du champ de la santé (classe 14) et de nouveau une classe principalement caractérisée par des noms propres (classe 13). La classe 12, qui traite de l'égalité homme/femme, semble inscrire cette thématique prioritairement dans le champ professionnel. Apparaissent ensuite une classe sur les politiques familiales et le handicap (classe 7), une classe sur le logement et l'exclusion (classe 9) et une classe sur la filiation (classe 9). L'autre partie de l'arbre présente d'abord une grande classe sur la justice (classe 4), distincte de celle traitant de la question des peines (classe 3), puis la thématique du piratage sur Internet (classe 8), le traitement de la délinquance (classe 2), la question spécifique des violences sexuelles (classe 1), les stupéfiants et la sécurité routière (classe 5) et, enfin, la question du couple, de l'homosexualité et de la procréation (classe 15). Si ces ensembles lexicaux nous paraissent familiers, c'est d'une part, parce qu'ils sont moins hétérogènes que les classes des analyses précédentes et, d'autre part, parce qu'ils renvoient à des thématiques récurrentes dans les communications sociales quotidiennes (communications médiatiques notamment).

La lecture chronologique de cette analyse est également plus précise que la précédente. Ce niveau de précision permet de différencier des périodes dans le corpus qu'il est facile d'associer à des débats marquants de ces législatures.

Le graphique 3 rend compte de la surapparition de certaines thématiques à certaines dates (en utilisant le mois comme unité). Il reprend la stratégie que nous proposions dans un autre contexte (Ratinaud, 2014). Ici, la largeur des cellules est proportionnelle à l'intensité des débats (mesurée par le nombre de segments de textes), la hauteur des lignes est proportionnelle à la taille des classes et les cases sont coloriées lorsque la surreprésentation d'une date est significative (au sens du chi2) pour la classe concernée. Ce graphique permet donc de suivre précisément l'évolution chronologique des thématiques. Nous pouvons repérer facilement, par exemple, la surreprésentation de la classe sur le couple (en haut du graphique, zones encadrées en traits pleins) lors du débat sur le PACS pour la période où Lionel Jospin était Premier ministre, et lors du débat sur le « mariage pour tous » pour la période où Jean-Marc Ayrault occupait Matignon. L'activation de cette même classe dans les périodes des gouvernements de Jean-Pierre Raffarin et de François Fillon correspond aux révisions de loi sur la bioéthique, qui aborde notamment les questions de procréation médicalement assistée (PMA). Ces phases sont également des moments de surreprésentation de la classe sur la filiation (petite classe bleue au centre du graphique). C'est un excellent exemple de ce que nous souhaitons mettre en évidence dans le cadre d'une interprétation fondée sur la théorie des représentations sociales : ici, un champ lexical restreint permet l'expression de deux interprétations différentes d'une réalité sociale.

Nous pouvons également remarquer l'importance relative des classes sur la délinquance et la justice dans la période gouvernementale de Dominique de Villepin (classes orange et jaune). Cette période est également marquée par des débats sur l'école et l'insertion professionnelle, liés à la tentative de mise en place du Contrat première embauche (en bleu en bas du graphique). La récurrence des réformes liées à l'Éducation nationale est d'ailleurs évidente, ainsi que les liens de cette thématique avec celle de l'insertion professionnelle (à l'exception de la période Fillon durant laquelle ces thématiques semblent décorrélées). Enfin, de façon plus anecdotique, la zone encadrée en pointillés signale les débats sur le piratage sur Internet qui ont abouti à la mise en place de l'Hadopi[14].

L'analyse lexicale proposée par Reinert permet d'explorer les thématiques qui traversent les corpus textuels. Mais lorsque la taille de ces corpus augmente, dans les dimensions devenues courantes avec les bases de données textuelles actuelles, les premières analyses livrent soit un petit nombre de classes qui peuvent renvoyer à des banalités, soit un très grand nombre de classes qui rendent difficile l'interprétation. Nous préconisons alors de procéder par étapes successives, en ciblant à chaque étape les parties les plus pertinentes au regard des hypothèses du chercheur. Des logiciels comme IRaMuTeQ

14. Haute Autorité pour la diffusion des œuvres et la protection des droits sur Internet.

ou TXM proposent différents moyens de partitionner simplement les corpus, à partir de métadonnées, de requêtes de concordance ou de classification. Cette démarche en abîme nous affranchit partiellement de la taille du type de corpus que nous venons de présenter, sans pour autant perdre la temporalité des textes.

Au-delà de l'aspect purement méthodologique, et sans en amoindrir l'intérêt, on passe, d'une opération à l'autre, à la mise en évidence d'ensembles qui, pour nous, relèvent de deux niveaux théoriques distincts. La première partition nous a permis de distinguer les thématiques, objet de notre recherche présente, des formalismes parlementaires et des séquences interactionnelles, qui pourraient, chacun, faire l'objet d'une étude propre. Les thématiques ainsi dégagées peuvent être extraites et analysées séparément, comme nous l'avons proposé dans cet article pour « la famille », pour dégager des prises de position liées à des appartenances, ici opérationnalisées selon le gouvernement (et son Premier ministre) en place. C'est donc bien à l'intérieur de thématiques génériques que se construisent des représentations sociales destinées à promouvoir des dynamiques identitaires, c'est-à-dire des rapports d'alliance et d'opposition entre groupes sociaux.

Évidemment, il faudrait reproduire cette procédure pour toutes les classes thématiques et il ne peut s'agir ici que d'une première approche, d'un travail qui devra s'inscrire non seulement dans la durée, mais dans le croisement des regards en fonction d'intérêts scientifiques divers. Cette ressource textuelle préformatée en accès libre[15] devrait permettre à quiconque s'intéresse à une thématique particulière et à son évolution de construire des corpus d'analyse. Ces corpus seront aussi des outils de test et/ou de développement dans la communauté des logiciels libres de lexicométrie.

On notera néanmoins quelques limites qui restent à lever : l'année 2011 est absente du corpus et devra sans doute faire l'objet d'une saisie particulière. Les noms propres peuvent par ailleurs donner lieu à un traitement spécifique : il devrait être possible de les repérer plus précisément afin de les extraire des formes analysées et de les coder comme métadonnées en les associant éventuellement au parti politique du locuteur. Moyennant quoi, on peut envisager que ce corpus permette d'imaginer des explorations variées intéressant nombre de chercheurs.

Références

Doise Willem, 1985, « Les représentations sociales. Définition d'un concept », *Connexions*, n° 45, p. 243-253.

— 2011, « Sistema e metassistema », *Teoria das Representaçoes Sociais. 50 Anos*,

15. Le corpus des débats à l'Assemblée nationale est libre de droits et sera mis en ligne.

A. M. de Oliveira Almeida, M. F. de Souza Santos, Z. Arauja Trindade éd., Brasilia, Technopolitik, p. 123-156.

Ghiglione Rodolphe, Landré Agnès, Bromberg Marcel, Molette Pierre, 1998, *L'analyse automatique des contenus*, Paris, Dunod.

Loubère Lucie, 2014, « Le traitement des TICE dans les discours politiques et dans la presse », *Actes des 12e Journées internationales d'analyse statistique des données textuelles*, Paris, Inalco / Sorbonne nouvelle, p. 433-445.

Marchand Pascal, 2007, *Le grand oral. Les discours de politique générale de la Ve République*, Bruxelles, De Boeck.

— 2008, « Quelques déterminants du discours politique », *Cognition, santé et vie quotidienne*, vol. I, G. Chasseigne éd., Paris, Publibook Université (Sciences humaines et sociales - Psychologie), p. 162-189.

Martin Éveline, 1993, *Reconnaissance de contextes thématiques dans un corpus textuel. Éléments de lexico-sémantique*, Paris, Didier Érudition.

Moscovici Serge, 1961, *La psychanalyse, son image, son public*, Paris, Presses universitaires de France.

Ratinaud Pierre, 2009, *Iramuteq : interface de R pour les analyses multidimensionnelles de textes et de questionnaires*, http://www.iramuteq.org (consulté le 3 mars 2015).

— 2014, « Visualisation chronologique des analyses Alceste : application à Twitter avec l'exemple du hashtag #mariagepourtous », *Actes des 12e Journées internationales d'analyse statistique des données textuelles*, Paris, Inalco / Sorbonne nouvelle, p. 553-565.

Ratinaud Pierre, Marchand Pascal, 2012, « Application de la méthode Alceste à de "gros" corpus et stabilité des "mondes lexicaux" : analyse du "CableGate" avec Iramuteq », *Actes des 11e Journées internationales d'analyse statistique des données textuelles*, Université de Liège, p. 835-844.

Reinert Max, 1983, « Une méthode de classification descendante hiérarchique. Application à l'analyse lexicale par contexte », *Les Cahiers de l'analyse des données*, vol. VIII, n° 2, p. 187-198.

— 1990, « Alceste, une méthodologie d'analyse des données textuelles et une application : *Aurélia* de Gérard de Nerval », *Bulletin de méthodologie sociologique*, n° 26, p. 24-54.

— 1993, « Les "mondes lexicaux" et leur "logique" à travers l'analyse statistique d'un corpus de récits de cauchemars », *Langage et société*, n° 66, p. 5-39.

— 1997, « Les "mondes lexicaux" des six numéros de la revue *Le surréalisme au service de la révolution* », *Cahiers du Centre de recherche sur le surréalisme (Mélusine)*, vol. XVI, p. 270-302.

— 2007, « Postures énonciatives et mondes lexicaux stabilisés en analyse statistique de discours », *Langage et société*, n° 121/122, p. 189-202.

Scherer Klaus R., Giles Howard éd., 1977, *Social Markers in Speech*, Cambridge University Press / MMSH.

Smith Philip H., Giles Howard, Hewstone Miles, 1979, « Sociolinguistic. A social psychological perspective », *The Social and Psychological Contexts of Language*, R. N. St Clair, H. Giles éd., Hillsdale (NJ), L. Erlbaum Associates.