
Les mots du Grand Débat national : les réseaux lexicaux des contributions déposées sur trois plateformes

The Words of the Grand Débat national (Great National Debate): the Lexical Networks of the Contributions Submitted on Three Platforms

Sabine Ploux, Michael Genay et Leu Ploux-Chillès



Édition électronique

URL : <https://journals.openedition.org/revuehn/2655>

DOI : 10.4000/revuehn.2655

ISSN : 2736-2337

Éditeur

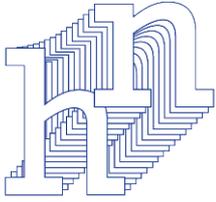
Humanistica

Référence électronique

Sabine Ploux, Michael Genay et Leu Ploux-Chillès, « Les mots du Grand Débat national : les réseaux lexicaux des contributions déposées sur trois plateformes », *Humanités numériques* [En ligne], 4 | 2021, mis en ligne le 01 décembre 2021, consulté le 01 décembre 2022. URL : <http://journals.openedition.org/revuehn/2655> ; DOI : <https://doi.org/10.4000/revuehn.2655>



Creative Commons - Attribution 4.0 International - CC BY 4.0
<https://creativecommons.org/licenses/by/4.0/>



Les mots du Grand Débat national : les réseaux lexicaux des contributions déposées sur trois plateformes

The Words of the Grand Débat national (Great National Debate): the Lexical Networks of the Contributions Submitted on Three Platforms

Sabine Ploux, Michael Genay et Leu Ploux-Chillès

Résumés

À l'occasion du Grand Débat national, lancé le 15 janvier 2019, plusieurs plateformes comme celles du Grand Débat national (GDN), du Vrai Débat (VD) et d'Entendre la France (EF) ont recueilli les contributions de participants sur des sujets de société. Dans cet article, nous présentons une méthode pour extraire et analyser les réseaux lexicaux contenus dans les corpus de textes formés par ces contributions grâce au modèle géométrique des *Atlas sémantiques*. Les résultats qui en découlent donnent, d'une part, les mots et le profil partagés par les trois plateformes, ainsi que les mots et profils propres à chacune d'elles et, d'autre part, pour chacun des mots, son réseau lexical. La liste des mots surreprésentés dans ces corpus et partagés par les trois plateformes contient essentiellement des mots relatifs à l'écologie et à la fiscalité. Les listes propres à chacune d'elles montrent des profils distincts. En particulier, le corpus GDN contient une surreprésentation des mots *incivilité* et *respect* ainsi que des mots relatifs à l'écologie et au collectif, celui du VD une surreprésentation des termes relatifs à des organisations politiques, gouvernementales ou internationales, à des personnalités politiques, à des contingences personnelles, sociales et économiques, à des privilèges, à des modes de participation et de vote. Chez les contributeurs d'EF (moyenne d'âge : 29 ans), on observe une surreprésentation des mots relatifs aux discriminations, à la tolérance et à une écologie des pratiques. Enfin, deux exemples de réseaux lexicaux extraits du corpus GDN sont détaillés : celui du mot *transport* et celui du mot *contre*. Pour le

mot *contre*, choisi car son réseau lexical révèle de nombreux sujets débattus par les contributeurs, nous montrons que la méthode permet d'explicitier et de faire la synthèse de liens sémantiques et de révéler leur organisation. Il ressort de cet exemple que le domaine de l'écologie est un centre organisateur qui articule les principaux thèmes abordés.

During the Grand Débat national, launched on January 15, 2019, several platforms such as the Grand Débat national (GDN), Le Vrai Débat (VD), or Entendre la France (EF) collected contributions from participants on societal issues. In this article, we present a method for extracting and analyzing lexical networks derived from the text corpora formed by these contributions using the Semantic Atlas geometric model. The method permits to obtain (1) words and profiles shared by the three platforms, as well as words and profiles specific to each of them, (2) for each word, its lexical network. The list of words over-represented in these corpora and shared by the three platforms contains mainly words related to environmental issues and taxation. The lists specific to each of the platforms show distinct profiles. In particular, the GDN corpus shows an over-representation of words related to incivility and respect, as well as words related to environmental issues and the collective, the VD corpus shows an over-representation of terms related to political, governmental, or international organizations, political figures, personal, socio-economic contingencies, privileges, or modes of participation and voting. For EF, whose contributors are young (average age 29), one finds an over-representation of words related to discrimination, tolerance, and behaviours based on environmental concern. Finally, two examples of lexical networks extracted from the GDN corpus are detailed: that of the word *transport* (“transport”) and that of the word *contre* (“against”). For the word *contre* – chosen because its lexical network reveals which topics matter to the contributors – we show that the method makes explicit and synthesizes semantic links and reveals their organization. It can be seen from this example that environmental issues are an organizing center around which the main topics addressed by the contributors are articulated.

Entrées d'index

MOTS-CLÉS : analyse sémantique, linguistique et sciences du langage, modèle formel, traitement automatique des langues, visualisation de données

KEYWORDS: semantic analysis, linguistics, formal model, natural language processing, data visualisation

Contexte

¹ Suite au développement du mouvement des Gilets jaunes (à partir de novembre 2018) et à sa poursuite, le président de la République française a ouvert le 15 janvier 2019 le Grand Débat national (GDN), autour de quatre thèmes : « transition écologique », « démocratie et citoyenneté », « fiscalité et dépenses publiques », « organisation de l'État et des services publics ». Il s'agissait de consulter l'ensemble des Français sur ces sujets de société à travers des questions proposées par la mission interministérielle chargée de l'organisation du GDN en lien avec les ministères compétents. Dans ce cadre, différents modes de participation ont été mis en place : réunions, cahiers dans les mairies, mise en ligne d'une plateforme de consultation¹. Conjointement à ce lancement, d'autres plateformes de consultation ont été créées, notamment celle du Vrai Débat (VD dans la suite du texte²), à l'initiative d'un collectif de Gilets jaunes, et Entendre la France (EF dans la suite du texte³), principalement dirigée vers une population jeune qui semblait trop absente des autres modes de participation.

Enjeux

² Le GDN a produit un corpus de contributions de taille considérable. Son analyse par des méthodes automatiques a été et reste donc nécessaire. Plusieurs synthèses ont été proposées, comme celle d'Opinionway (en partenariat avec la société QWAM pour l'analyse des textes), commanditée par la mission du GDN. Leurs résultats publiés sur le site du GDN proposent pour chacune des questions une liste de réponses types avec un pourcentage de représentativité. Les méthodes semi-automatiques employées opèrent une reconnaissance des entités nommées (noms de personnes ou d'organisations, appellations comme l'expression « Gilets jaunes », lieux, dates, etc.) et utilisent un algorithme de catégorisation des réponses. Des cartographies des propositions de l'Institut des systèmes complexes de Paris – Île-de-France⁴, ou de l'INRIA et du CNRS⁵ ont par ailleurs été réalisées et mises en ligne. Elles permettent, à la manière d'un moteur de recherche spatial, de saisir une requête, de localiser sur la carte les propositions les plus pertinentes et de les afficher. La cartographie a de plus l'avantage de dégager de grands thèmes issus d'un balayage des contributions grâce à des méthodes d'analyse des données ou mettant en œuvre des réseaux de neurones, après un même repérage des entités nommées. Ces méthodes cartographiques ont été appliquées sur l'ensemble du corpus constitué des réponses aux questions non fermées (voir ci-dessous). Enfin, il a été objecté que le fonctionnement d'une plateforme et la participation en ligne ne permettaient pas de croiser les différentes contributions et les profils des contributeurs et de la population. Cela est en partie dû au choix initial des organisateurs du GDN de ne demander aux contributeurs aucun autre renseignement personnel qu'une adresse électronique et un code postal⁶. L'observatoire des débats⁷ a donc procédé à des enquêtes de terrain en se rendant sur les lieux des réunions locales et en étudiant les profils des participants.

³ Dans cette étude, nous avons choisi une approche complémentaire. Il s'agit d'analyser les mots employés ainsi que les réseaux lexicaux qui les relient. Cette approche permet, comme d'autres citées précédemment, de synthétiser les schémas de phrases typiques des contributions. Elle offre de plus la possibilité de repérer la variabilité du sens lexical en contexte. En effet, des personnes peuvent employer un même mot sans pour autant lui associer le même contenu. Dans le cadre d'un débat, le repérage et l'analyse de cette polysémie sont fondamentaux. Pour cela, nous avons analysé par la méthode des *Atlas sémantiques* (voir le site Web⁸ et la description ci-dessous) les réponses soumises sur la plateforme du GDN. Cette même analyse a également été effectuée sur les propositions et les réponses issues des deux autres plateformes, VD et EF. Nous présentons ici des résultats obtenus à partir du corpus du GDN et quelques éléments de comparaison avec les deux autres plateformes.

⁴ Dans la première partie de cet article, nous donnons un aperçu quantifié des corpus analysés, puis décrivons le modèle utilisé. La seconde partie présente des résultats obtenus pour l'ensemble des plateformes et pour chaque plateforme prise séparément. Ces résultats comprennent le repérage des concepts saillants, leur variabilité de contexte d'emploi et l'organisation de cette variabilité. Le mot *transport* qui recouvre un sujet très présent dans le corpus sert d'exemple-type pour décrire les formes multiples que peuvent revêtir un mot et ses emplois. Enfin, à partir du mot *contre*, nous montrons comment la méthode permet d'explicitier et de faire la synthèse de liens sémantiques « dilués » dans les propositions et de révéler une organisation des sujets abordés.

Description sommaire des corpus analysés

⁵ Le site de la plateforme de GDN comprenait, pour chacun des quatre thèmes rappelés ci-dessus, deux séries de questions : la première constituée de questions fermées – les participants devaient faire un choix dans une liste de réponses proposées –, la seconde constituée de questions ouvertes, en ce sens que les participants devaient taper leur réponse sous forme de texte libre. Dans cette première étude, nous nous sommes intéressés aux réponses aux questions de la seconde série. Ces réponses sont celles collectées à la clôture de la plateforme le 15 mars 2019. L'ensemble des 82 questions est disponible dans les synthèses du site granddebat.fr. Notons que certaines semblent guider le type de réponses possibles ou au moins attendues : *Quels sont selon vous les impôts qu'il faut baisser en priorité ?* (Thème « fiscalité et dépenses publiques ».) D'autres sont plus véritablement ouvertes : *Y a-t-il d'autres points sur les impôts et les dépenses sur lesquels vous souhaiteriez vous exprimer ?* (Thème « fiscalité et dépenses publiques ».)

⁶ À titre de comparaison, le Vrai Débat (VD) comportait 9 thèmes : « démocratie, institutions », « transition écologique et solidaire, agriculture et alimentation, transport », « justice, police, armée », « Europe, affaires étrangères, outre-mer », « santé, solidarité, handicap », « économie, finances, travail, comptes publics », « éducation, jeunesse, enseignement supérieur, recherche et innovation », « sport, culture » et « expression libre et sujets de société ». Entendre la France (EF), dirigé vers un public jeune, reprenait les mêmes thèmes que le GDN, avec une simplification

des questions (49 questions, contre 82 dans le GDN) et la possibilité de contribuer librement hors des thèmes soumis à consultation. Comme pour le GDN, pour chacune des plateformes VD et EF les réponses et contributions analysées sont celles collectées à leur clôture⁹.

7 Le tableau 1 donne le volume en contributions et en mots repérés des corpus de textes issus des trois plateformes. Il apparaît que ce volume est significativement plus important pour la plateforme GDN mise en place par la mission interministérielle que dans le cas des deux autres plateformes. Ces variations pourraient en partie dériver de l'étendue de la communication mise en œuvre pour chacune d'elles, ainsi que de la sensibilité politique et du profil des contributeurs. Notons aussi que l'organisation du VD comme un débat interactif, dans lequel il était possible de soumettre une contribution mais aussi de voter ou de répondre aux contributions soumises, pourrait avoir favorisé un nombre plus faible de contributions et un nombre de mots par contribution plus important (GDN : $4,6 \times 10^4$; VD : $8,5 \times 10^4$; EF : 3×10^4).

TABLEAU 1. VOLUME DES CORPUS

| | Nombre de contributions (en milliers) | Nombre de mots comptabilisés (en millions) |
|-----|--|---|
| GDN | 569,02 | 26,4 |
| VD | 25,41 | 2,16 |
| EF | 54,73 | 1,65 |

Les méthodes

Une approche par les mots

8 De façon complémentaire à l'approche précédemment citée et retenue par la mission du GDN, nous avons fait le choix d'étudier les réponses dans leur ensemble et non question par question. Ce choix est fondé sur plusieurs observations. Tout d'abord, les sujets abordés par les participants dépassent les questions. Ainsi, le mot *impôt* apparaît dans tous les thèmes. Il est alors intéressant de l'analyser en fonction de la diversité de ses contextes d'emploi, qu'il s'agisse d'écologie, de société, etc. Autres exemples, le mot *santé* (encadré 1), qui renvoie au thème éponyme et n'ayant pas été soumis au débat, apparaît 149 178 fois dans l'ensemble des réponses données sur la plateforme du GDN. Et le mot *loisir*, qui semblerait encore plus éloigné des préoccupations entourant le mouvement des Gilets jaunes et le débat qui a suivi, est présent 7 127 fois (encadré 1).

9 Étudier les mots et leur contexte d'emploi au-delà des réponses directes aux questions posées permet donc de mettre en évidence des thématiques latentes, qu'une étude uniquement fondée sur les couples question-réponse n'aurait pas permis d'explicitier.

GDN santé [149178, $0,56 \times 10^{-2}$] : mental, prévention, pharmaceutique, honoraire, complémentaire, vieillesse, désert, nocif, généraliste, patient, dentaire, soignant, médecine, soin, médicament, tabac, infirmier, médecin, lunettes, couverture, dépendance, alcool, hôpital, optique, médical, éducation...

EF santé [864, $0,82 \times 10^{-3}$] : soin, médecin, dangereux, médical, éducation, hôpital, domaine, secteur, sécurité, enseignement, professionnel, maison, justice, mauvais, environnement, coût, service, accès, manque...

VD santé [1773, $0,52 \times 10^{-3}$] : complémentaire, prévention, mutuel(le), éducation, environnement, alimentation, hôpital, professionnel, soin, sécurité, problème, assurance, établissement, maison, **ministère**, domaine, humain, bon, centre, maladie, **système**, protection, médecin, matière, meilleur, frais, **public**, dépense...

GDN loisir [7127] : ski, sport, vacances, croisière, cahier, tourisme, sportif, voyage, chasse, cantine, bateau, centre, avion, luxe, crèche, course, vêtement, culturel, alimentation, parc, nourriture, culture, déplacement, aérien, activité, **accéder**, commercial, équipement, commerce, mobilité, courant, espèce, adulte, soin, **garantir**, carburant, **promouvoir**, scolaire, accès, consommation, **taxer**...

¹⁰ En haut de l'encadré, le mot *santé* suivi, entre crochets, de sa fréquence et de sa fréquence normalisée (ramenée à la taille du corpus) dans les différentes plateformes, et pour chacune des plateformes, les premiers contextonymes donnés par ordre d'indice de cooccurrence décroissant. On remarque, par exemple, davantage de termes relatifs au système de santé publique (en gras et italique) dans les contextonymes extraits du corpus VD. La fréquence normalisée du mot *santé* dans le corpus du GDN est près de 10 fois supérieure à celle d'EF alors que ces corpus comportent les mêmes thèmes et majoritairement les mêmes questions. Il aurait été intéressant de pouvoir vérifier si cet écart n'est pas l'effet d'une différence des profils d'âge entre les deux populations de contributeurs.

¹¹ En bas de l'encadré, le mot *loisir* suivi de sa fréquence entre parenthèses et des premiers contextonymes dans le corpus GDN. Les verbes sont en gras. On note que les contextonymes renvoient (1) aux types de loisirs, en particulier à la sphère du luxe (*luxe, croisière...*) ; (2) aux centres de loisirs et à ce qui touche au périscolaire (*centre, crèche, cantine, scolaire...*).

L'analyse sémantique fine

- ¹² La spécificité du modèle des *Atlas sémantiques* est l'analyse sémantique fine. Cette précision dans l'analyse est rendue possible par la notion de clique, détaillée dans plusieurs publications (voir par exemple Ploux, Boussidan et Ji 2010). Une clique est un ensemble de mots fortement liés les uns aux autres. Ces liens croisés contraignent le sens de chacun d'eux. Ainsi, dans la clique « *lutter, contre, optimisation, abusif* » – comme nous l'expliquons ci-dessous, les mots sont ramenés à leur lemme, c'est-à-dire à l'entrée correspondante du dictionnaire –, extraite de l'analyse du GDN, le mot *optimisation* a une valeur négative¹⁰ induite par la présence de l'adjectif *abusif* qui le qualifie. Dans une autre clique, « *mécanisme, dispositif, PME, optimisation* », le même mot *optimisation* fera référence aux mécanismes et dispositifs pour l'optimisation des PME, sans nécessairement comporter un caractère négatif.
- ¹³ La relation lexicale choisie dans cette étude pour le calcul des cliques est la cooccurrence régulière ou contextonymie¹¹ (notion précisée en annexe). Notons qu'à la différence des modèles vectoriels, de type Glove (Pennington, Socher et Manning 2014) ou Word2vec (Mikolov *et al.* 2013), par exemple, qui associent un vecteur à un mot, le modèle géométrique des *Atlas sémantiques* associe un vecteur à une clique et un domaine d'un espace multidimensionnel à un mot. Ce domaine est constitué de l'ensemble des cliques-vecteurs qui contiennent un mot donné. Ainsi, cette approche (Ploux, Boussidan et Ji 2010) permet de représenter la variation sémantique et contextuelle d'un mot (ou d'un ensemble de mots) par une distinction en différentes régions du domaine construit. Ces régions correspondent à des ensembles de cliques distincts et donc à différentes façons d'employer ce ou ces mots.
- ¹⁴ Les annexes détaillent le calcul des fréquences des mots et des contextonymes, le calcul des cliques et l'obtention des cartes de cliques et de contextonymes.

Premiers résultats

Les mots les plus saillants par comparaison avec un corpus de référence

- ¹⁵ Afin de détecter les mots les plus saillants, nous avons comparé leur fréquence dans chacun des trois corpus à celle d'un large corpus de référence¹² (de 341 millions de mots) compilé dans notre laboratoire. La détection consiste à repérer les mots pour lesquels le rapport des fréquences normalisées entre le corpus choisi et le corpus de référence est élevé.
- ¹⁶ La figure 1 donne la liste des mots communs aux corpus des trois plateformes parmi les 200 mots les plus surreprésentés dans chacun des corpus par rapport au corpus de référence. Ces mots apparaissent dans chacun des trois corpus GDN, VD et EF avec un rapport des fréquences normalisées supérieur à 26.

anti, plateforme, cyclable, isolation, défavorisé, pollueur, malus, assisté, drastiquement, optimisation, doublon, lambda, média, responsabiliser, migrant, TVA, taxer, réfugié, écologique, régalien, progressivité, taux, pesticide, polluant, obsolescence, raisonné, éolien, recyclage, stop, gaspillage, biodiversité, participatif, kérosène, taxe, impôt, citoyen, sécu, démuni, contraint, lobby, éolienne, niche, fraudeur, emballage, polluer, écologie.

Les couleurs sont attribuées automatiquement en fonction de la sélectivité thématique du mot dans les quatre sous-corpus du GDN associé à chacun des thèmes. La couleur verte indique que le mot a une fréquence normalisée au moins 5 fois plus élevée dans le thème de la transition écologique que dans chacun des trois autres thèmes. Même chose pour la couleur violette et le thème « fiscalité et dépenses publiques », le rouge et « démocratie et citoyenneté », et le bleu et « organisation de l'État et des services publics ».

- 17 Les mots qui relèvent de l'écologie sont très présents dans cette liste, ainsi que ceux qui relèvent de la fiscalité. Cette comparaison avec le corpus de référence révèle aussi d'autres contrastes. Ainsi, nous avons noté plus de *pour* que de *contre* (comme le montre la proportion normalisée des *pour* par rapport aux *contre* dans chacun des corpus, GDN : 1,61, VD : 1,8, EF : 1,44), une surreprésentation des modalités *falloir* (GDN : 4,64, VD : 2,76, EF : 6,86), *devoir* (GDN : 3,55, VD : 3,04, EF : 3,44) et *pouvoir* (GDN : 1,66, VD : 1,85, EF : 1,79), ainsi qu'une sous-représentation de la modalité *vouloir* (GDN : 0,67, VD : 0,80, EF : 0,90). L'analyse fine de ces marqueurs demanderait un retour aux textes des contributions afin d'examiner la possible prévalence d'une démarche de proposition et d'un discours qui positionne son propos moins en fonction de demandes ou de souhaits qu'en fonction de nécessité, d'obligation, de capacité, etc.
- 18 Les tendances propres pour chacun des trois corpus sont données dans la figure 2. Pour chaque corpus, la liste contient les mots pour lesquels le rapport entre la fréquence normalisée et la plus haute des fréquences normalisées des deux autres corpus est le plus élevé (le rapport des fréquences normalisées est indiqué entre crochets à la suite de chaque mot¹³).
- 19 Pour le GDN, les tendances propres mettent en évidence une surreprésentation des mots *incivilité* et *respect* par rapport aux deux autres corpus, la présence de nombreux mots relatifs à l'écologie et de mots relatifs au collectif ou à l'associatif.
- 20 Pour le VD, on observe une surreprésentation des termes relatifs à des organisations politiques, gouvernementales et internationales ou à des personnalités politiques (OTAN, Matignon, PS, UE, ONU, Bercy, etc.), à des contingences personnelles, sociales et économiques (*conjoint*, *invalidité*, *smicard*, *loyer*, *indexation*, *sécu*, *frais*, *augmentation*, *facture*, *parental*, etc.), à des termes décrivant des avantages (*pantouflage*, *pognon*, *lobbyiste*, *dividende*) ou à des modes de participation et de vote (*révocable*, *uninominal*, *quorum*, *pétition*).
- 21 Enfin, pour EF, dont les contributeurs sont jeunes, avec une moyenne d'âge de 29 ans (donnée issue du rapport disponible en ligne¹⁴), on observe une surreprésentation des mots relatifs aux discriminations et à la tolérance, composés en gras dans la figure 2 (*sexisme/-iste*, *racisme*, *discrimination*), aux migrations (*migration*, *migrant*, *migratoire*, *migrer*, *réfugié*) et à l'écologie. On remarque une différence entre les mots employés par les contributeurs d'EF et ceux du GDN pour parler de l'écologie. Pour

les premiers, les mots reflètent plus une écologie des pratiques (*trier, réutiliser, recycler, etc.*), tandis que les seconds font davantage référence à des questions et matériels énergétiques (*chaudière, isolation, solaire, photovoltaïque, fossile, etc.*).

FIGURE 2. TENDANCES PROPRES AUX TROIS CORPUS GDN, VD ET EF

| GDN | VD | EF |
|-----------------------------|--------------------------|-------------------------------|
| incivilité [9,29] | OTAN [21,1] | sexisme [9,64] |
| chaudière [5,12] | Matignon [6,31] | lavable [9,37] |
| bénévolat [3,98] | Giscard [4,95] | végétarien [8,84] |
| respect [2,52] | rémunération [4,68] | discrimination [6,78] |
| habitation [2,38] | conjoint [4,19] | réutilisable [6,22] |
| isolation [2,37] | indexation [4,16] | racisme [5,87] |
| dérèglement [2,22] | révocable [4,09] | migration [5,44] |
| régalien [2,13] | invalidité [3,70] | vrac [5,00] |
| incitatif [2,08] | dividende [3,68] | faciès [4,47] |
| associatif [2,07] | instauration [3,44] | réutiliser [4,38] |
| bénévole [2,04] | pantouflage [3,32] | sexiste [4,30] |
| incitation [1,990] | détaché [3,32] | harcèlement [4,07] |
| piéton [1,95] | autoroute [3,30] | trier [4,02] |
| solaire [1,93] | chaîne [3,28] | alternative [3,84] |
| sanctionner [1,90] | rétablissement [3,19] | migrant [3,72] |
| entraide [1,90] | additif [3,15] | migratoire [3,67] |
| collectivité [1,86] | employeur [3,00] | biodégradable [3,66] |
| strate [1,86] | Chirac [2,87] | plastique [3,53] |
| comptabiliser [1,82] | PS [2,79] | migrer [3,39] |
| commune [1,79] | député [2,78] | réfugié [3,30] |
| cyclable [1,79] | rembourser [2,71] | recycler [3,25] |
| fiable [1,77] | loyer [2,67] | supermarché [3,24] |
| consultatif [1,76] | uninominal [2,65] | sensibilisation [3,14] |
| allocation [1,75] | PMA [2,56] | sensibiliser [3,08] |
| civique [1,75] | frais [2,54] | emballage [3,06] |
| foncier [1,74] | reformer [2,44] | paperasse [2,90] |
| pôle [1,74] | prestation [2,41] | jetable [2,86] |
| pénaliser [1,72] | quorum [2,41] | poubelle [2,82] |
| territorial [1,70] | suppression [2,39] | ostentatoire [2,79] |
| panneau [1,68] | UE [2,38] | voiture [2,78] |
| départemental [1,68] | augmentation [2,37] | récapitulatif [2,76] |
| chauffage [1,68] | ONU [2,36] | compost [2,75] |
| photovoltaïque [1,67] | lobbyiste [2,35] | laïcité [2,74] |
| laxisme [1,66] | parental [2,33] | tri [2,71] |
| vertueux [1,66] | pognon [2,31] | sélectif [2,68] |
| valoriser [1,65] | dette [2,27] | légume [2,67] |
| concitoyen [1,61] | facture [2,27] | bénéfique [2,64] |
| responsabiliser [1,61] | Bercy [2,22] | consommer [2,61] |
| usager [1,5] | cotisation [2,21] | accessibilité [2,51] |
| pollution [1,58] | plafonnement [2,20] | déchet [2,48] |
| limitation [1,54] | salarial [2,20] | politicien [2,39] |
| communal [1,54] | anormal [2,19] | biologique [2,31] |
| regroupement [1,49] | entreprendre [2,18] | efficace [2,30] |
| financièrement [1,48] | indirect [2,15] | écolo [2,30] |
| biodiversité [1,47] | revaloriser [2,12] | améliorer [2,29] |
| administré [1,47] | pétition [2,09] | primordial [2,28] |
| fossile [1,46] | aidant [2,09] | surconsommation [2,23] |
| contrepartie [1,45] | stop [2,082] | impact [2,22] |
| véhicule [1,44] | majoré [2,07] | crèche [2,17] |
| performant [1,44] | smicard [2,06] | SDF [2,12] |

Les couleurs sont attribuées de la même manière que dans la figure 1 ; les termes en gras sont relatifs à des organisations politiques, gouvernementales et internationales ou à des personnalités politiques.

Les mots et leur réseau lexical

22 Nous avons calculé les réseaux lexicaux de contextonymie pour chacun des 300 mots les plus fréquents des différents corpus (indépendamment du corpus de référence et hors mots de fonction non caractéristiques d'un débat ¹⁵). Pour le GDN, les résultats sont en ligne sur le site des *Atlas sémantiques* ¹⁶. Il y a cinq listes de mots interrogeables, une par thème et une pour l'ensemble du corpus, tous thèmes confondus.

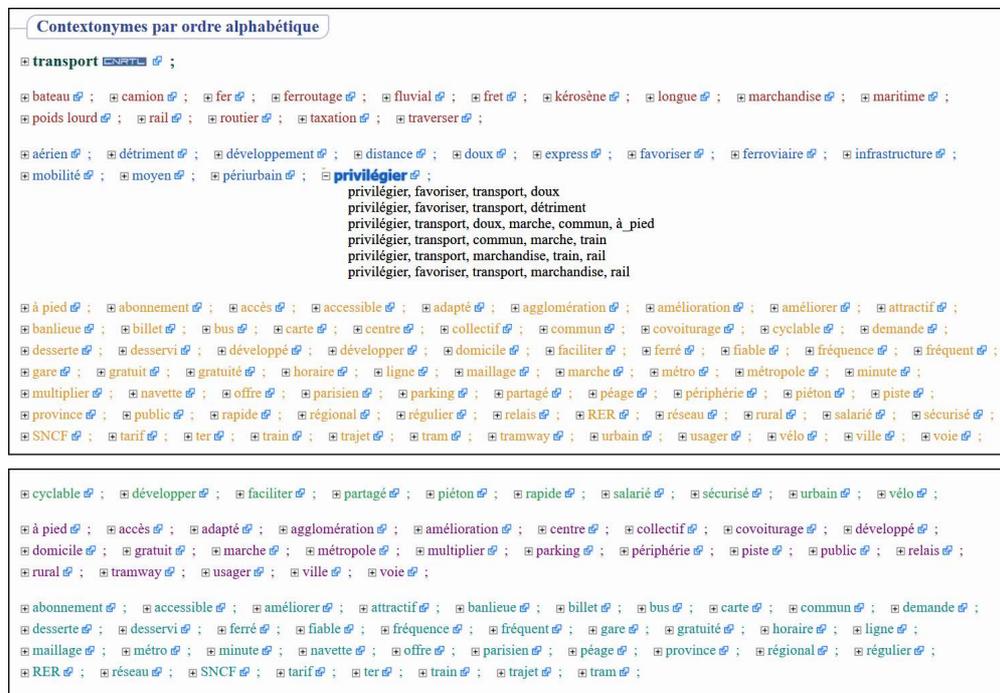
23 Chacun de ces mots a son propre réseau de contextonymes organisé selon la topologie de la carte calculée par la méthode détaillée en annexe. En plus de la carte, il est possible de consulter les regroupements – appelés *constellations* dans le modèle des *Atlas sémantiques* –, qui sont calculés par classification hiérarchique des contextonymes. La figure 3 donne le résultat obtenu pour le mot *transport*, qui est un des mots les plus fréquents de l'ensemble du corpus du GDN et le mot de contenu le plus fréquent du thème « transition écologique ». Pour une classification en trois constellations, on fait les observations suivantes :

1. Dans la constellation de contextonymes (en jaune), la plus importante, les mots renvoient aux transports de proximité, au maillage, à la fréquence, aux abonnements, au covoiturage, etc.
2. Une autre constellation (en bleu) est relative à l'aérien et à l'idée de privilégier le ferroviaire. Les cliques du mot *privilégier* sont affichées. Elles enjoignent l'utilisation des transports doux, du rail.
3. Enfin, la dernière constellation (en rouge) aborde la question de la taxation et du kérosène.

24 Notons qu'il est possible de modifier le nombre de constellations afin d'obtenir des regroupements de plus en plus précis. En particulier, la constellation la plus importante peut être divisée (seconde partie de la figure 3). Seront alors distingués plusieurs thèmes : le vélo et les pistes cyclables ; l'accès aux métropoles et les relations entre centre et périphérie ou monde rural ; enfin, les abonnements, les fréquences, le maillage, la régularité et les dessertes pour le métro, la RATP, les TER, la SNCF, les trams.

25 Ainsi le réseau lexical du mot *transport* révèle la diversité des problématiques et propositions à travers – et c'est la part la plus foisonnante en nombre de contextonymes et de cliques de ce réseau lexical – des demandes d'amélioration de transport du quotidien, mais aussi une approche plus vertueuse des transports à longue portée, qu'il s'agisse de l'aérien ou du transport commercial (bateau, camion).

FIGURE 3. LES PREMIÈRES CONSTELLATIONS DU MOT *TRANSPORT*



En haut, les trois premières constellations du mot *transport* du thème « transition écologique » du GDN. Les cliques du contextonyme *privilegier* sont affichées. En bas, la plus grande constellation a été subdivisée en trois.

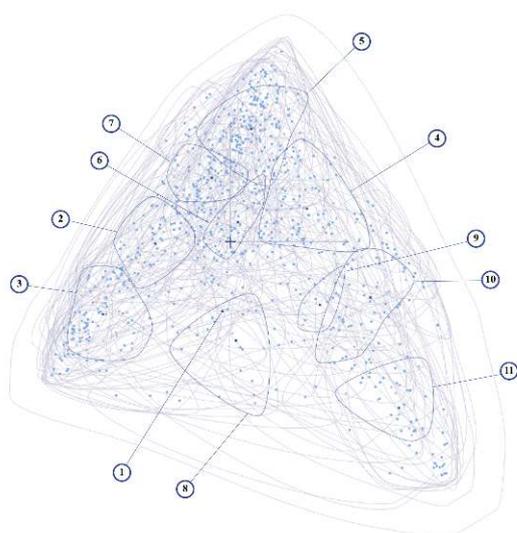
Ce que révèle la structure du réseau lexical du mot *contre*

26 Ici, nous détaillons le cas du mot *contre*, choisi sous l'hypothèse qu'il concentre davantage l'expression d'éléments de débat formulés dans cette plateforme. La figure 4 donne la carte construite à partir de ce mot dans l'ensemble des quatre thèmes.

Les thèmes principaux

27 La carte générée, en forme de triangle, fait apparaître trois zones principales. Au sommet supérieur du triangle se trouve un réseau lexical relatif à l'écologie, à droite un réseau relatif à des questions sociales et sociétales, à gauche un réseau relatif à la fiscalité, à l'économie.

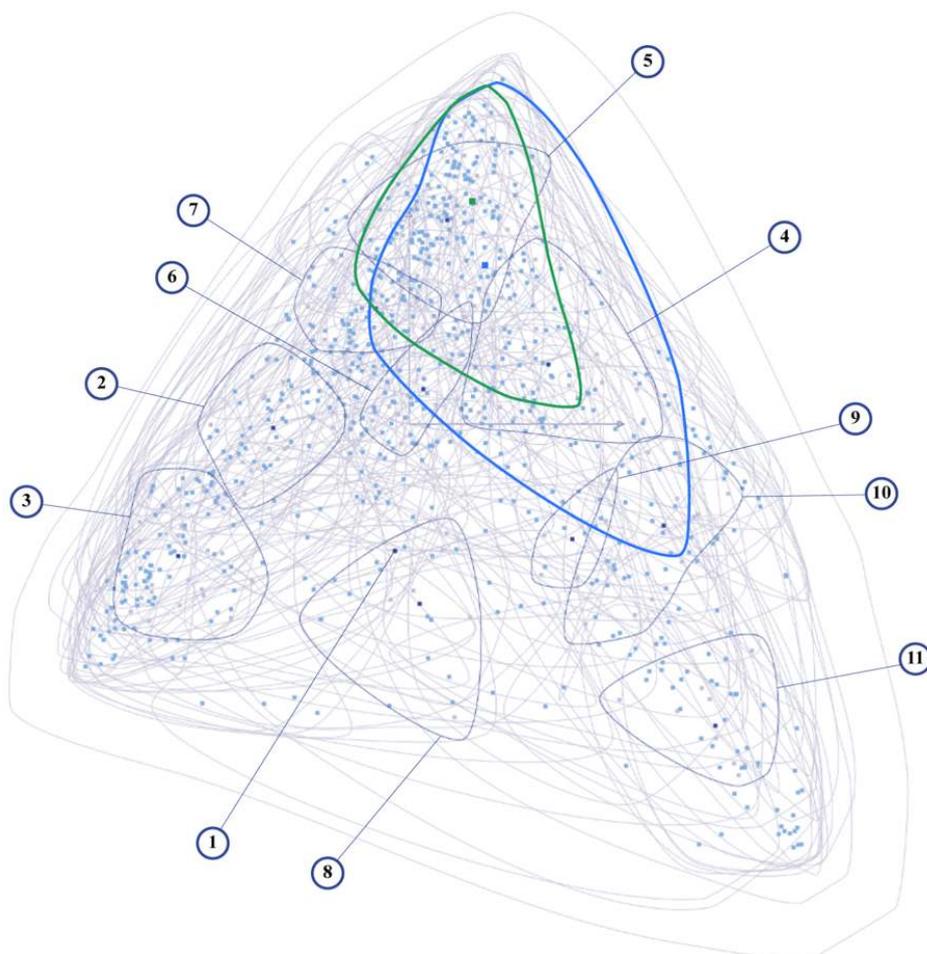
FIGURE 4. LA CARTE DU MOT *CONTRE* (GDN)



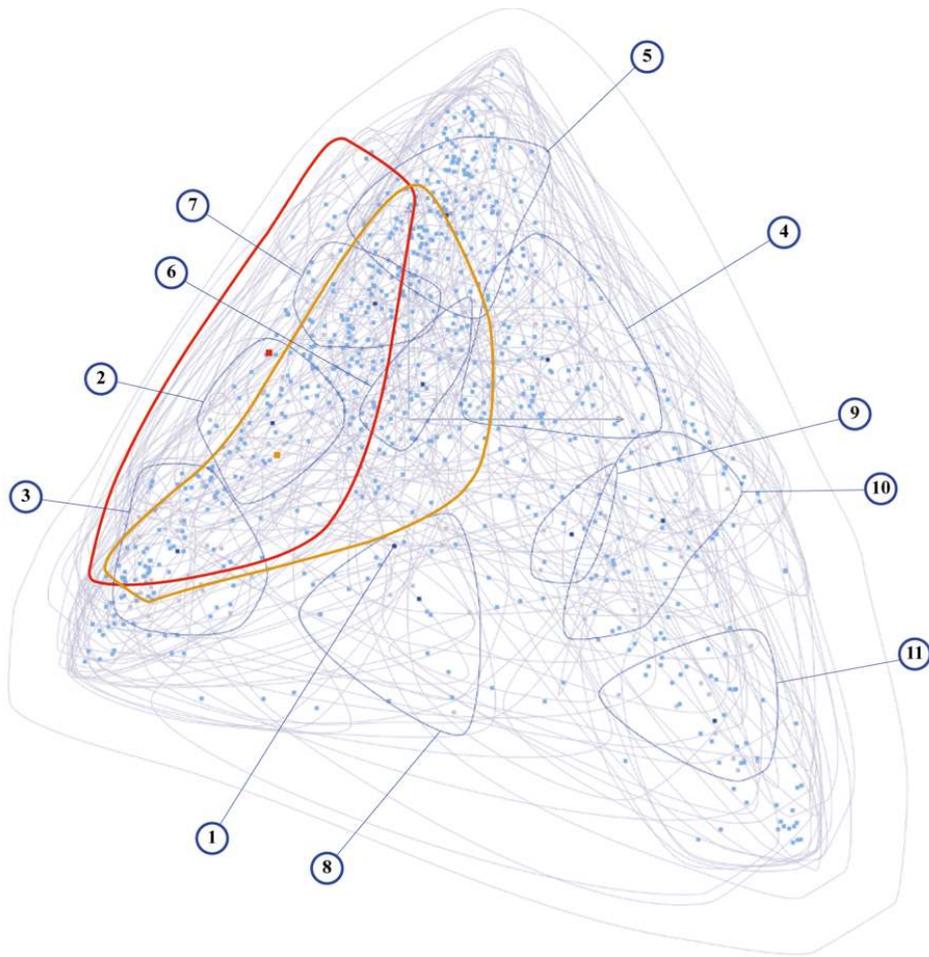
- | | |
|----|--|
| 1 | chômeur |
| 2 | évasion, optimisation, fraude, taxation, fonds, groupe, effacement, fortement, mesure, concurrence, fraudeur... |
| 3 | GAFa, paradis, ISF, fiscal, bénéfice, multinational, fortune, niche, banque, capital, transaction, milliard, taxer, PME, rétablir, riche, français, CICE, partie, luxe, productif, imposé, résultat, actif, probant, concertation, biaisé |
| 4 | corruption, trafic, forces, guerre, ordre, arme, illégal, sanction, garantir, conflit, défense, prévention, pollution, crime, clandestin, passeur |
| 5 | chasse, biodiversité, espèce, interdiction, protection, protéger, soutenir, agir, soutien, climatique, priorité, plan, réchauffement, animal, pesticide, mener, sauvage, neuf, défendre, CO2, promouvoir, dérèglement, recours, urbain, humanité |
| 6 | lutter, lutte, abus, au noir, faux, gaspillage, se battre, légal, médical, désert |
| 7 | international, renforcer, européen, renforcement, contrôle, lobby, aérien, vente, commercial, renforcé, obsolescence, programmé... |
| 8 | inégalité, million, pauvre, injustice, chômage, voté, monter |
| 9 | combattre, pauvreté, forme, dérive, extrême |
| 10 | racisme, discrimination, religieux, communautarisme, exclusion, haine, harcèlement, musulman, femme, blanc |
| 11 | violence, anti, délinquance, drogue, délit, terrorisme, agression, incivilité, précarité, courant, manifestation, voter |

La carte comporte 145 contextonymes, 700 cliques, 11 constellations. La croix représente l'origine des axes (inertie cumulée : 5,1 %). Les mots de chacune des constellations sont ordonnés par nombre décroissant de cliques auxquelles ils appartiennent.

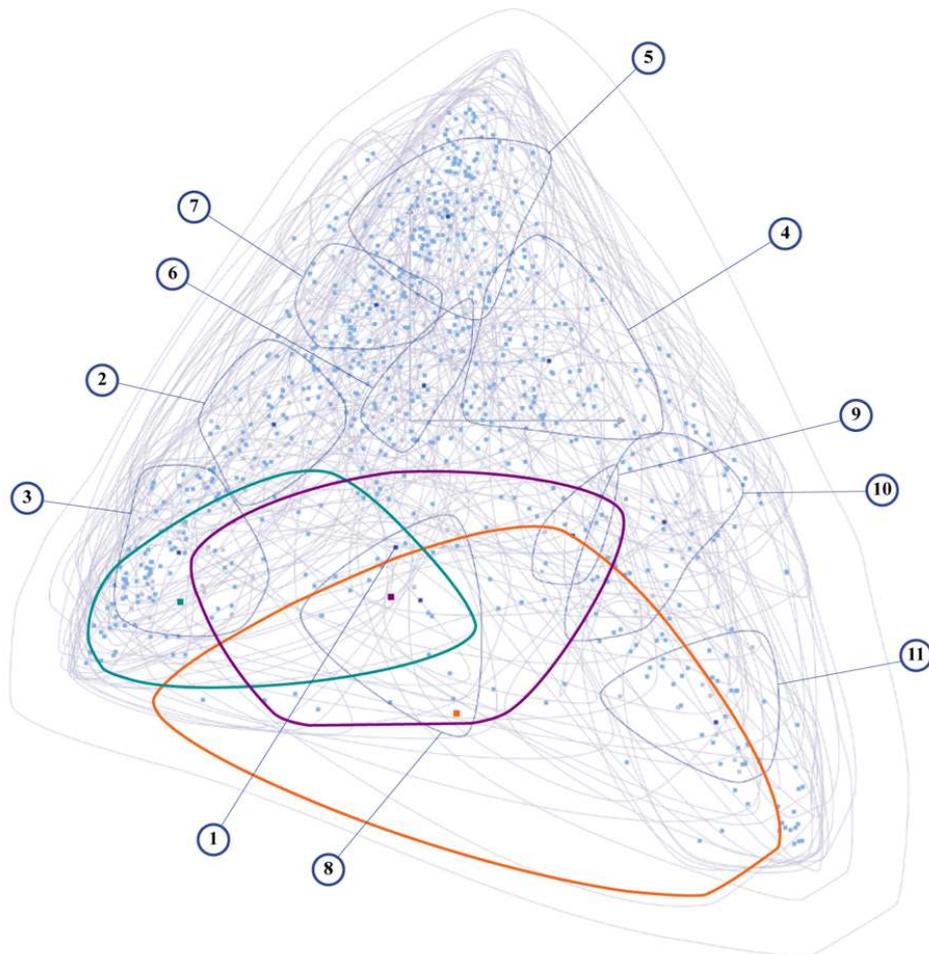
FIGURE 5. LA CARTE DU MOT *CONTRE* (GDN) : MISE EN ÉVIDENCE DE LIENS ENTRE LES THÈMES MAJEURS



protection, protéger



mesure, efficacment



injustice, pauvre, luxe

28 Dans la constellation relative à l'écologie (numérotée 5 sur les figures 4 et 5) se trouvent les mots : *chasse, biodiversité, espèce, interdiction, protection, protéger, soutenir, agir, soutien, climatique, priorité...* Précisons que la présence de mots comme *biodiversité* dans le réseau lexical de *contre* n'induit pas que le mot *contre* porte sur la biodiversité. Cela révèle que chacun des mots est régulièrement présent à l'intérieur de phrases contenant *contre* et dans lesquelles peuvent apparaître des objections dans le cadre de débats sur l'écologie et la biodiversité. Par exemple, des cliques de *contre* mettent en évidence un lien entre biodiversité et pesticides, lobbies ou chasse. La constellation 5 est la plus dense en cliques (les cliques sont figurées par des points bleu clair sur la carte). C'est aussi une constellation où convergent la plupart des enveloppes de mots (contours grisés sur la carte). Le fait que l'enveloppe d'un mot chevauche, au moins en partie, une constellation indique que ce mot, même s'il se rapporte à d'autres thèmes, appartient néanmoins à des cliques liées à la thématique de la constellation (ici l'écologie). Ces éléments (densité des cliques¹⁷ et convergence des enveloppes) soulignent donc que l'écologie est un thème important, organisateur d'une part significative des thèmes représentés sur la carte.

29 Dans la branche relative aux questions sociales se trouvent trois constellations : *combattre, pauvreté, forme, dérive, extrême* (numérotée 9 sur la carte) ; *racisme, discrimination, religieux, communautarisme, exclusion, haine, harcèlement, musulman, femme, blanc* (numérotée 10) ; *violence, anti, délinquance, drogue, délit, terrorisme, agression, incivilité, précarité, courant, manifestation, voter* (numérotée 11).

30 Dans la branche relative aux questions économiques et fiscales se trouvent principalement les constellations suivantes : *GAFa, paradis, ISF, fiscal, bénéfice, multinational, fortune, niche, banque, capital, transaction, milliard, taxer, PME, rétablir, riche, français, CICE...* (numérotée 3) ; *évasion, optimisation, fraude, taxation, fonds, groupe, efficacement, fortement, mesure, concurrence, fraudeur...* (numérotée 2).

L'écologie comme élément pivot et domaine de convergence

31 La constellation 5, relative à l'écologie, est reliée à celles relatives au social à travers une constellation intermédiaire (numérotée 4) évoquant les conflits, la pollution et les déplacements de personnes qui en résultent (*corruption, trafic, forces, guerre, ordre, arme, illégal, sanction, garantir, conflit, défense, prévention, pollution, crime, clandestin, passeur*). La protection, avec la présence des mots *protection* et *protéger* – dont l'enveloppe part de la constellation 5 pour atteindre les constellations 4, 9 et 10, à thème social (voir la figure 5) – est aussi un élément charnière de ce lien. La protection de la nature et de la diversité des espèces est représentée, par exemple, par la clique *lutter, protéger, contre, pesticide, biodiversité, lobby, chasse*, et la protection des personnes, par la clique *lutter, protéger, contre, précarité, violence*.

32 La constellation 5 est reliée à celles relatives aux questions économiques et fiscales, précédemment citées, par celle (numérotée 7) qui évoque entre autres l'international, les contrôles et l'obsolescence (programmée). En outre, les mots du thème économique et fiscal dont l'enveloppe recouvre en partie la constellation 5 renvoient aux mesures et à

leur modalité : *efficacement, mesure, taxation, taxer, rétablir*. Sur la figure 5 sont mises en évidence les enveloppes des mots *efficacement* et *mesure*.

33 Notons aussi la présence plus ténue, mais liant directement les thèmes sociétaux, économiques et fiscaux sans passer par la constellation de l'écologie, de deux constellations numérotées 1 (*chômeur*) et 8 (*inégalité, million, pauvre, injustice, chômage, voté, monter*). Notamment, sur la figure 5 est indiquée l'enveloppe du mot *injustice*, centrale entre les thèmes sociétaux et économiques, ainsi que celles des mots contrastés *pauvre* et *luxe*, qui la recouvrent partiellement.

34 En somme, cette carte offre une explicitation des thèmes qui font débat pour les contributeurs en dégageant des articulations fortes, principalement organisées autour de la question écologique, mais aussi reliées par le thème des inégalités et de l'injustice.

Perspectives

35 Ce corpus est d'une grande richesse. Nous avons cherché à montrer comment la méthode d'analyse automatique des *Atlas sémantiques* pouvait être utile à son exploitation. Des travaux en cours permettront, à partir de l'exploration des réseaux lexicaux et des cliques, le retour aux extraits pertinents des contributions.

36 Pour aller plus loin, certains points demanderaient de nouvelles investigations. En particulier, les contributions pouvant être individuelles ou collectives, il serait intéressant de les analyser séparément afin d'étudier d'éventuels marqueurs d'une posture plus personnelle dans les premières ou un effacement de ces marqueurs dans des synthèses collectives.

37 En outre, le mouvement des Gilets jaunes a pu être interprété comme la conséquence d'une fracture entre les métropoles et les autres territoires (Tendil 2019 ; Lainé 2017). Il semble donc essentiel de mieux comprendre ce phénomène en analysant, si elles existent, les variations du contenu des concepts sociétaux en fonction de l'origine géographique des contributeurs. Pour cela, nous utiliserons le code postal associé à chaque contribution afin d'extraire la densité de population correspondante. Autre point : la teneur des contributions ayant pu évoluer en fonction des événements et de l'actualité, nous projetons des analyses qui étudient l'évolution des champs lexicaux en fonction de la date de soumission de la contribution. Enfin, nous souhaiterions pouvoir exploiter le contenu numérisé des cahiers déposés dans les mairies. Il est raisonnable de faire l'hypothèse que les contributeurs de ces cahiers et ceux des plateformes numériques ont des profils distincts. Les exploiter apporterait une image plus exhaustive de l'ensemble des réponses au Grand Débat national.

REMERCIEMENTS. Nous remercions vivement le collectif CodeforFrance, l'équipe du Vrai Débat et celle d'Entendre la France qui nous ont permis d'analyser les différents corpus.

Bibliographie

Benzécri, Jean-Paul. 1980. *L'Analyse des données : l'analyse des correspondances*, Paris : Dunod.

Boussidan, Armelle, Anne-Lyse Renon, Charlotte Franco, Sylvain Lupone et Sabine Ploux. 2012. « Repérage automatique de la néologie sémantique en corpus à travers des représentations cartographiques évolutives : vers une méthode de visualisation graphique dynamique de la diachronie de la néologie ». *Cahiers de lexicologie* 100 : 117-136.

Ji, Hyungsuk, Sabine Ploux et Éric Wehrli. 2003. « Lexical Knowledge Representation with Contexonyms ». Dans *Proceedings of the 9th Machine Translation Summit*, 194-201. Association for Computational Linguistics. <https://aclanthology.org/2003.mtsummit-papers.26/>.

Lainé, Frédéric. 2017. « Dynamique de l'emploi et des métiers : quelle fracture territoriale ? ». Note d'analyse 53. France Stratégie. <https://www.strategie.gouv.fr/sites/strategie.gouv.fr/files/atoms/files/na53-fractures-territoriales-ok.pdf>.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado et Jeff Dean. 2013. « Distributed Representations of Words and Phrases and Their Compositionality ». Dans *Proceedings of the 26th International Conference on Neural Information Processing Systems*, 3111-3119. New York : Curran Associates Inc. <https://dl.acm.org/doi/10.5555/2999792.2999959>.

Pennington, Jeffrey, Richard Socher et Christopher Manning. 2014. « Glove : Global Vectors for Word Representation ». Dans *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1532-1543. Association for Computational Linguistics. <http://www.aclweb.org/anthology/D14-1162>.

Ploux, Sabine, Armelle Boussidan et Hyungsuk Ji. 2010. « The Semantic Atlas : an Interactive Model of Lexical Representation ». Dans *Proceedings of the 7th International Conference on Language Resources and Evaluation*, Malte. European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2010/pdf/592_Paper.pdf.

Saporta, Gilbert. 2006. *Probabilités, analyse des données et statistique*. Paris : Éditions Technip.

Schmid, Helmut. 1994. « Probabilistic Part-of-Speech Tagging Using Decision Trees ». Dans *International Conference on New Methods in Language Processing*. Manchester, Royaume-Uni.

Tendil, Michel. 2019. « Le premier “baromètre des territoires” montre la fracture du pays ». Banque des territoires. 12 février. <https://www.banquedesterritoires.fr/le-premier-barometre-des-territoires-montre-la-fracture-du-pays>.

Annexe

Fréquences, cooccurrence et contextonymie

Un traitement du corpus des réponses au GDN a été effectué sur les quatre sous-corpus obtenus pour chacun des thèmes, ainsi que sur l'ensemble (tous thèmes confondus). Parce qu'ils sont peu volumineux¹⁸, les deux corpus VD et EF ont été traités séparément à partir de l'ensemble de leurs contributions, sans tenir compte des thèmes qu'ils incluent. Le traitement des textes comporte plusieurs étapes et choix. Tout d'abord, les corpus ont été lemmatisés¹⁹. La lemmatisation consiste à remplacer chaque mot du corpus par l'entrée correspondante du dictionnaire (*garantit* devient *garantir*, *travaux* devient *travail*). Ce choix est plus spécifiquement utile pour le traitement des verbes dont le grand nombre de flexions empêcherait la juste prise en compte s'ils n'étaient pas regroupés sous un même lemme. Ensuite, la fréquence de chaque mot, la fréquence de cooccurrence avec l'ensemble des autres mots et la fréquence de cooccurrence ordonnée (le nombre de fois où le mot apparaît avant un autre dans une phrase) sont calculées. La fenêtre de calcul de la cooccurrence est la phrase. Les cooccurrents dont la fréquence de cooccurrence avec le mot étudié est faible ou au contraire dépasse significativement celle du mot étudié sont retirés²⁰. Pour chaque mot i , la fréquence de cooccurrence $f_{i,j}$ avec un autre mot j est ensuite divisée par la fréquence de j , (f_j). Cette normalisation permet de tenir compte du lien réciproque qui lie le cooccurrent au mot étudié²¹. La liste des cooccurrents de chaque mot est ensuite triée par ordre décroissant des indices de cooccurrence $f_{i,j}/f_j$. Finalement, un paramètre α fixe la portion des premiers cooccurrents qui seront analysés. Ce paramètre tient compte de la distribution des indices de cooccurrence²². On appelle *contextonymes* les cooccurrents réguliers ainsi calculés. À chacun des contextonymes sont associées deux valeurs : l'indice de cooccurrence avec le mot étudié et la proportion de phrases dans lesquelles le mot étudié i apparaît avant le contextonyme j (notée $p_{i,j}$ dans la suite). Les premiers contextonymes du mot *santé* pour les trois plateformes et du mot *loisir* pour celle du GDN sont donnés dans l'encadré 1. Une fois les liens de contextonymie calculés, le système construit les cliques.

Calcul et ordonnancement des cliques de contextonymie

Les relations de contextonymie sont utilisées pour le calcul des cliques²³ qui contiennent un mot étudié. Il s'agit des ensembles maximaux de mots contenant le mot étudié et des contextonymes de ce mot, tous contextonymes eux-mêmes les uns des autres²⁴. Une fois les cliques obtenues, elles sont réordonnées afin de tenir compte de l'ordre le plus fréquent des mots dans les phrases du corpus. Pour cela, pour chaque terme i de la clique, la somme $s_i = \sum_{j \neq i} p_{i,j}$ est calculée. Les termes de la clique sont alors triés suivant les valeurs s_i croissantes. Voici quelques exemples de cliques triées (pour GDN) :

- *lutter, contre, évasion, montage, optimisation, fiscal, illégal, légal*
- *lutter, contre, optimisation, abusif*
- *mesure, arbitraire, contre, million, signature, mobiliser, manifestation, combattre*
- *instaurer, contrôle, évaluation, indépendant*

On peut reconnaître dans ces exemples des schémas-types de phrases ou de parties de phrases sous leur forme lemmatisée et qui représentent des motifs récurrents du corpus analysé.

Construction des cartes et constellations

Pour chaque mot étudié (ou pour un ensemble de mots), une analyse factorielle des correspondances (Benzécri 1980) est appliquée sur la matrice qui contient en ligne ses cliques et en colonnes ses contextonymes. Cette matrice est composée de 0 ou de 1 suivant qu'un contextonyme-colonne appartient (1) ou pas (0) à une clique-ligne. Le résultat est un espace dans lequel à chaque clique est associé un point. Chaque contextonyme est figuré par l'enveloppe (par défaut en gris pâle) englobant les cliques qui le contiennent. Afin de dégager les regroupements cohérents de sens appelés constellations, une classification hiérarchique par la méthode de Ward (Saporta 2006) est appliquée sur les centres de gravité du nuage de points-cliques²⁵ contenant un contextonyme donné et à partir de leurs premières coordonnées, dont le nombre est ici fixé à 2. Le nombre de constellations est un paramètre de la classification choisi ici par les auteurs et modifiable sur l'interface des *Atlas sémantiques* dont l'adresse est donnée ci-dessous. Ces constellations peuvent, elles aussi, être représentées par leur enveloppe (par défaut tracée en gris plus soutenu). Les cartes données dans le présent article illustrent le résultat de cette construction ; l'ensemble des résultats est disponible sur le site de *Atlas sémantiques* (http://www.atlas-semantiques.eu/GDN_as.html?l=FR). On pourra se reporter à la référence Ploux, Boussidan et Ji 2010 pour plus de détails sur les méthodes de calcul des cartes.

Notes

1 <https://granddebat.fr>.

2 https://fr.wikipedia.org/wiki/Le_Vrai_Débat/.

3 <https://www.entendrelafrance.fr>.

4 <https://politoscope.org>.

5 <https://cartolabe.fr>.

6 Les variables annexes disponibles pour chaque contribution diffèrent suivant les plateformes : code postal du contributeur et date de soumission pour le GDN ; score, nombre de votes, pourcentage de votes favorables, date de soumission pour le VD. En effet, le VD a été organisé comme un débat interactif entre, d'une part, des contributions et, d'autre part, des votes et réponses à ces contributions. Enfin, pour EF, les variables annexes sont plus nombreuses : code postal, commune, type de commune, département, sexe, âge, formation, profession, taille de l'organisation, position vis-à-vis du mouvement des Gilets Jaunes, date de soumission.

7 <https://observdebats.hypotheses.org>.

8 <http://www.atlas-semanticues.eu>.

9 La plateforme du VD a clos le recueil des contributions le 3 mars 2019, celle d'EF, le 17 mars 2019.

10 Notons que les cliques ne donnent pas accès au point de vue des contributeurs. En effet des phrases du corpus, les plus proches de la clique, pourraient contenir des négations... Elles représentent des régularités, des motifs qui trament les textes étudiés.

11 Cette notion a été introduite par H. Ji dans sa thèse et les publications en résultant (voir Ji, Ploux et Wehrli 2003). L'appellation qu'il a choisie en anglais était *contextonym*. Ici, en français, nous traduisons par *contextonyme*.

12 Ce corpus de référence comprend l'ensemble des articles du journal *Le Monde* sur la période 1997-2007 et l'ensemble des textes littéraires de la base *Frantext* (<https://www.frantext.fr>) pour les XIX^e et XX^e siècles. Les composantes journalistique et littéraire permettent de couvrir à la fois un vocabulaire politique, économique et social et un vocabulaire ancré dans le langage courant (*pognon, décent, dignement...*).

13 Pour ce calcul ont été retenus, pour chacun des corpus, les 1 000 mots les plus surreprésentés par rapport au corpus de référence.

14 <http://reponses.entendrelafrance.fr/rapport/>.

15 Ont été retirés les déterminants, les pronoms et des conjonctions.

16 Voir http://www.atlas-semanticues.eu/GDN_as.html?l=FR/.

17 La densité des cliques découle de la multiplication des emplois du mot, ici *contre* dans le paradigme de l'écologie (voir Boussidan *et al.* 2012 pour une étude plus détaillée de ce phénomène).

18 Le calcul des cooccurrences sur un corpus de taille trop faible ne permettrait pas de repérer des régularités.

19 Nous avons utilisé le logiciel TreeTagger (<https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger>, Schmid 1994) complété par des routines d'ajustement aux corpus étudiés, que nous avons développées.

20 Ici, les paramètres pour la sélection d'un cooccurent sont fixés à $f_{i,j} > 3$ et $f_{i,j} < 1,5 \times f_i$, où f_i est la fréquence du mot étudié i et $f_{i,j}$ la fréquence de cooccurrence du terme j avec i .

21 Par exemple, l'adjectif *petit(e)* pourrait avoir une fréquence de cooccurrence plus élevée que celle du mot *isolation* avec le mot étudié *maison*. *Petit* ayant un profil d'emploi très étendu et une fréquence élevée, la division de sa fréquence de cooccurrence avec *maison* par sa propre fréquence diminue la portée sémantique de *petit* quand il s'agit de *maison*. Au contraire, *isolation* ayant un profil d'emploi plus spécifique et ciblé sur quelques mots, dont *maison*, le rapport de sa fréquence de cooccurrence avec *maison* sur sa fréquence totale sera plus important et reflétera ainsi un lien sémantique plus important entre *maison* et *isolation*.

22 Il s'agit des cooccurrents du mot étudié dont l'indice de cooccurrence est supérieur à $m_i + \alpha \times \sigma_i$, où m_i est la moyenne des indices de cooccurrence du mot i étudié et σ_i l'écart-type des indices de cooccurrence. Dans les résultats consultables sur le site des *Atlas sémantiques*, α est fixé à 4 %.

23 Du point de vue mathématique, une clique est un sous-graphe maximal complet connexe. Le graphe ici étudié a pour sommets les mots et pour arêtes les liens de contextonymie.

24 Seules les cliques d'au moins 3 mots sont conservées.

25 Il est également possible d'appliquer cette classification directement sur les points-cliques. Dans ce cas, la classification repose directement sur les cliques, unités de sens du modèle géométrique. Cependant, certains contextonymes dont l'enveloppe recouvre deux ou plusieurs constellations pourraient alors ne figurer dans aucune des constellations.

Auteurs

Sabine Ploux

UMR 8557 CAMS, CNRS-EHESS, Paris, France

Sabine Ploux est chercheure au CNRS. Ses travaux portent sur la modélisation de la sémantique lexicale.

ORCID [0000-0002-4876-657X](https://orcid.org/0000-0002-4876-657X)

sabine.ploux@ehess.fr

Michael Genay

Les Atlas sémantiques, Nanterre, France

Michael Genay est ingénieur informaticien. Il a participé au projet *Les Atlas sémantiques* depuis ses débuts et a contribué à la production des résultats de l'étude.

Leu Ploux-Chillès

Les Atlas sémantiques, Nanterre, France

Leu Ploux-Chillès a contribué à la dernière version du site *Les Atlas sémantiques* et à la mise en ligne des résultats obtenus dans cette étude. Il a proposé l'algorithme d'ordonnement des cliques de contexte utilisé dans l'étude.

Droits d'auteur



Les contenus de la revue *Humanités numériques* sont mis à disposition selon les termes de la [Licence Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/).