Les mots du Grand débat national : quelques résultats et réflexions préliminaires

Sabine Ploux, Michael Genay, Leu Ploux-Chillès

Les mots du Grand débat national : quelques résultats et réflexions préliminaires

Sabine Ploux*[1], Michael Genay[2] et Leu Ploux-Chillès[2]

[1]CAMS, UMR8557 CNRS-EHESS
[2]Les Atlas sémantiques

Résumé. À l'occasion du Grand débat national, lancé le 15 janvier 2019, plusieurs plates-formes comme celles du *Grand débat national*, du *Vrai débat* et d'*Entendre la France* 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 3 plates-formes 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 partagés par les 3 plates-formes contient essentiellement des mots relatifs à l'écologie et à la fiscalité. Les listes propres à chacune d'elles montrent des profils distincts décrits dans l'article. Enfin, deux exemples de réseaux lexicaux sont détaillés : celui du mot *transport* et celui du mot *contre* extraits du corpus de Grand débat national. Pour le mot *contre*, choisi car il révèle les sujets abordés qui font objection pour les contributeurs, nous montrons que la méthode permet d'expliciter et de faire la synthèse de liens sémantiques « dilués » dans les propositions et de révéler leur organisation. Il ressort de cet exemple que le domaine de l'écologie est un domaine pivot, centre organisateur vers lequel convergent les principaux thèmes abordés par les contributeurs.

Contexte

Suite au développement du mouvement des *Gilets jaunes* (nov. 2018-) et à son prolongement, le président de la République française a ouvert le 15 janvier 2019 le Grand débat national (GDN) autour de 4 thèmes : (a) transition écologique, (b) démocratie et citoyenneté, (c) fiscalité et dépenses publiques, (d) organisation de l'état et des services publics. Il s'agissait de consulter

*sabine.ploux@ehess.fr

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 plate-forme de consultation sur le site `granddebat.fr`. Conjointement à ce lancement, d'autres plates-formes de consultation ont été créées, notamment celle du *Vrai débat* (VD dans la suite du texte, `www.le-vrai-debat.fr`) à l'initiative d'un collectif de *Gilets jaunes* et *Entendre la France* (EF dans la suite du texte, `www.entendrelafrance.fr`) principalement dirigée vers une population jeune qui semblait trop absente des autres modes de participation.

Enjeux

Le GDN a produit un corpus de contributions de taille considérable. Son analyse à travers des méthodes automatiques a été et reste donc nécessaire. Plusieurs synthèses ont été proposées. Celle d'Opinionway (en partenariat avec société QWAM pour l'analyse des textes) et commanditée par la mission du GDN. Leurs résultats publiés sur le site du GDN propose pour chacune des questions posées une liste de réponses types avec pourcentage de représentativité. Les méthodes semi-automatiques employées opèrent une reconnaissance des entités nommées (noms de personnes, d'organisations, d'appellations comme l'expression *gilets jaunes*, de lieux, dates, etc.) et un algorithme de catégorisation des réponses. Des cartographies des propositions (`politoscope.org` de l'Institut des sciences complexes, ou `cartolabe.fr` de l'INRIA-CNRS) ont par ailleurs été réalisées et mises en ligne. Elles permettent à la manière d'un moteur de recherche spatial d'entrer 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 et/ou modèles vectoriels 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 plus bas). Enfin, il a été objecté que le fonctionnement d'une plate-forme et la participation en ligne ne permettaient pas de ramener les différentes contributions aux contributeurs et/ou à la population. Ceci 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 postal. L'observatoire des débats (`observdebats.hypotheses.org`) a donc procédé à des enquêtes de terrain en se rendant sur les lieux des réunions locales et en étudiant les profils de la participation.

Dans cette étude, nous avons choisi une approche complémentaire. Il s'agit d'analyser les mots employés ainsi que les réseaux lexicaux qui les re-

lient. Cette approche permet, comme d'autres citées précédemment, de faire des synthèses de schéma-types de phrases présentes dans les 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 fondamentales. Pour cela, nous avons analysé par la méthode des *Atlas sémantiques* (AS dans la suite du texte, `www.atlas-semantiques.eu`) décrite ci-après les réponses soumises sur la plate-forme du GDN. Cette même analyse a également été effectuée sur les propositions et les réponses issues des deux autres plates-formes VD et EF. Nous présentons ici des résultats obtenus à partir du corpus du GDN et quelques éléments de comparaison avec les deux autres plates-formes.

Dans la première partie du document, nous donnons un aperçu quantifié des corpus analysés puis décrivons le modèle utilisé. La seconde partie présente des résultats obtenus pour l'ensemble des thèmes et pour chaque thème pris séparément. Ces résultats comprennent le repérage des concepts saillants, leur variabilité de contexte d'emploi, l'organisation de cette variabilité. Le mot *transport* qui recouvre un sujet très présent dans le corpus sert d'exemple-type pour décrire la forme multiple que peut recouvrir un mot et ses emplois. Enfin, à partir du mot *contre* nous montrons comment la méthode permet d'expliciter et de faire la synthèse de liens sémantiques « dilués » dans les propositions et de révéler une organisation des sujets abordés qui font objection pour les contributeurs.

Rapide description des corpus analysés

Le site de la plate-forme de GDN comprenait pour chacun des 4 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 plate-forme le 15 mars 2019. L'ensemble des 82 questions est disponible dans les synthèses du site `grand.debat.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)*.

À titre de comparaison, le Vrai débat (VD) comportait 9 thèmes : (a) démocratie, institutions, (b) transition écologique & solidaire, agriculture & alimentation, transport (c) justice, police, armée (d) Europe, affaires étran-

gères, outre-mer (e) santé, solidarité, handicap (f) économie, finances, travail, comptes publics (g) éducation, jeunesse, enseignement supérieur, recherche et innovation (h) sport, culture (i) expression libre & 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 pour 82 (GDN)) et une possibilité de contribution libre hors thèmes. Comme pour le GDN, pour chacune des plates-formes VD et EF, les réponses et contributions analysées sont celles collectées à leur clôture.

Les variables annexes disponibles pour chaque contribution diffèrent suivant les plates-formes : 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.

La table 1 donne le volume en contributions et en mots repérés des corpus de textes issus des 3 plates-formes. Il apparaît que ce volume est significativement plus important pour le GDN que pour les 2 autres plates-formes.

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

TABLE 1 – Volume des corpus

Les méthodes

Une approche par les mots

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 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), thème qui n'était pas directement questionné apparaît 149178 fois dans l'ensemble des réponses données sur la plate-forme 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 7127 fois (encadré 1). 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'expliciter.

L'analyse sémantique fine

La spécificité du modèle des AS est l'analyse sémantique fine. Cette précision dans l'analyse est rendue possible grâce à la notion de clique détaillée dans plusieurs publications (voir par exemple [7]). 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 l'entrée correspondante du dictionnaire) extraite de l'analyse du GDN, le mot *optimisation* a une valeur négative induite par la présence des autres mots qui marquent le fait que l'optimisation peut être abusive et qu'il faille lutter contre ce phénomène. 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.

La relation lexicale choisie dans cette étude pour le calcul des cliques est la cooccurrence régulière ou contextonymie [1] (notion précisée en annexe). Notons, qu'à la différence des modèles vectoriels (de type Glove [6] ou Word2vec [5], par exemple) qui associent un vecteur à un mot, le modèle géométrique des *Atlas sémantiques* associe un vecteur à une clique et un domaine dans un espace multidimensionnel à un mot. Ce domaine est constitué de l'ensemble des cliques-vecteurs qui contiennent un mot donné. Ainsi, cette approche permet de représenter la variation sémantique et contextuelle d'un mot (ou d'un ensemble de mots) [7] 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) mot(s).

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 mots.

1. Cette notion a été introduite par H. Ji dans sa thèse et les publications résultant (voir [4]). L'appellation qu'il a choisi en anglais était *contexonym*. Ici, en français, nous traduisons par *contextonyme*.

Encadré 1

GDN **s**anté [149178, $0,56 * 10^{-2}$] 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, $0,82 * 10^{-3}$] soin, médecin, dangereux, médical, éducation, hôpital, domaine, secteur, sécurité, enseignement, professionnel, maison, justice, mauvais, environnement, coût, service, accès, manque

VD santé [1773, $0,52 * 10^{-3}$] 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**,

En haut, le mot *santé* suivi, entre crochets, de ses fréquence et fréquence normalisée (ramenée à la taille du corpus) dans les différentes plates-formes, et pour chacune, les premiers contextonymes donnés par ordre d'indice de cooccurrence décroissant. On remarque, par exemple, d'avantage de termes relatifs au système de santé publique (en gras et italique) dans les contextonymes extraits du corpus VD. La fréquence normalisée montre une plus forte présence de ce mot pour le corpus du GDN. Il aurait été intéressant de pouvoir vérifier si la fréquence normalisée du mot *santé* plus de 10 fois supérieure dans le corpus du GDN par rapport à celui d'EF, alors que ces corpus comportent les mêmes thèmes et majoritairement les mêmes questions, n'est pas un effet d'une différence des profils d'âge des deux populations de contributeurs.

En bas, 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 et 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..*).

Premiers résultats

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

Afin de détecter les mots les plus saillants, nous avons comparé leur fréquence dans chacun des 3 corpus à celle d'un large corpus de référence (de 341 millions de mots) comprenant des textes journalistiques et littéraires

et 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.

Voici la liste des mots communs aux corpus des 3 plates-formes parmi les 200 mots les plus surreprésentés dans chacun des corpus par rapport au corpus de référence. Ces mots apparaissent dans chacun des 3 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 mots qui relèvent de l'écologie [2] 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* (proportion normalisée de *pour* par rapport aux *contre* dans chacun des corpus GDN : $1,61$, VD : $1,8$, EF : $1,44$) et une sur-représentation des modalités : *falloir* (GDN : $4,64$, VD : $2,76$, EF : $6,86$), *devoir* (GDN : $3,55$, VD : $3,04$, EF : $3,44$), *pouvoir* (GDN : $1,66$, VD : $1,85$, EF : $1,79$) et une sous-représentation de la modalité *vouloir* (GDN : $0,67$, VD : 0.80, EF : $0,90$). La présence de ces marqueurs semblent indiquer une démarche de proposition.

Les tendances propres pour chacun des 3 corpus sont données dans l'encadré 2. Pour chacun des 3 corpus, la liste contient les mots pour lesquels le rapport de la fréquence normalisée sur 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 mot [3]).

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 et à l'associatif.

Pour le VD, on observe une surreprésentation des termes relatifs à des organisations politiques, gouvernementales, internationales ou à des personnalités politiques (OTAN, Matignon, PS, UE, ONU, Bercy,...), à des contingences personnelles, sociales et économiques (*conjoint*, *invalidité*, *smicard*,

2. Les couleurs sont attribuées automatiquement en fonction de la sélectivité thématique du mot dans les 4 sous corpus du GDN. La couleur vert indique que le mot à une fréquence normalisée au moins 5 fois plus élevée dans le thème de la *transition écologique* que dans chacun des 3 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*.

3. Pour ce calcul ont été retenus, pour chacun des corpus, les 1000 mots les plus surreprésentés par rapport au corpus de référence.

loyer, *indexation*, *sécu*, *frais*, *augmentation*, *facture*, *parental*,..) à des termes décrivant des avantages (*pantouflage*, *pognon*, *lobbyiste*, *dividende*), à des modes de participation et de vote (*révocable*, *uninominal*, *quorum*, *pétition*).

Enfin, pour EF dont les contributeurs sont jeunes avec une moyenne d'âge de 29 ans (donnée issue du rapport `reponses.entendrelafrance.fr/rapport`), on observe une surreprésentation des mots relatifs aux discriminations et à la tolérance (en gras) : *sexisme(-iste)*, *racisme*, *discrimination*, aux migrations (*migration, migrant, migratoire, migrer*, *réfugié*), à l'écologie. On remarque une différence des 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*,...), pour les seconds une référence plus soutenue à des questions et matériels énergétiques (*chaudière*, *isolation*, *solaire*, *photovoltaïque*, *fossile*,. . .).

Encadré 2 – Tendances propres à chacun des corpus [2]

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]	chaine [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 mots et leur réseau lexical

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 informatifs). Pour le GDN, les résultats sont en ligne sur le site des AS. Il y a 5 listes de mots interrogeables, une par thème et une pour l'ensemble du corpus tous thèmes confondus.

Chacun de ces mots a son propre réseau de contextonymes organisé selon la topologie de la carte calculée par la méthode donnée plus haut. En plus de la carte, il est possible de consulter les regroupements (appelés *constellations* dans le modèle des AS) et donnés par classification hiérarchique des proximités. La figure 1, 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 3 constellations, on observe :

1. une constellation de contextonymes (en jaune) qui est la plus importante et aussi la plus centrale si on se réfère à la carte. Les mots renvoient aux transports de proximité, au maillage, à la fréquence, aux abonnements, au covoiturage, etc.
2. une autre constellation (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 (rouge) aborde la question de la taxation et du kérosène.

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. Seront alors distingués les thèmes (1) du vélo et des pistes cyclables, (2) de l'accès aux métropoles et les relations centre et périphérie ou rural, et enfin (3) les questions des abonnements, des fréquences, du maillage, de la régularité et des dessertes pour le métro, RATP, TER, SNCF, TRAM.

Ainsi le réseau lexical du mot *transport* révèle la diversité des problématiques et propositions à travers (et c'est la plus grande part 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).

Contre : ce que révèle la formulation des objections des contributeurs

Ici, nous détaillons le cas du mot *contre*, choisi car il concentre l'expression des objections formulées dans cette plate-forme. La figure 2 donne la carte construite à partir de ce mot dans l'ensemble des 4 thèmes.

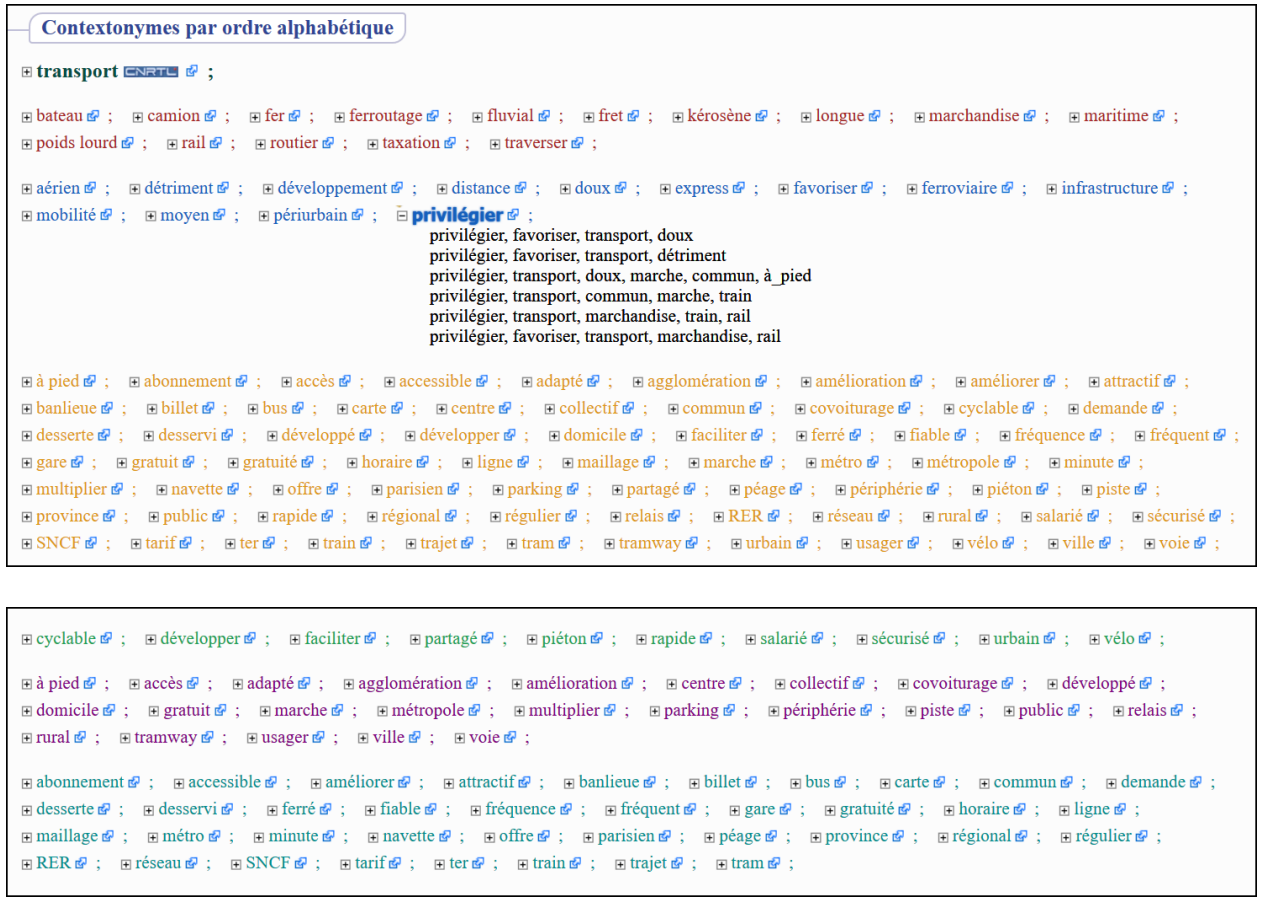

FIGURE 1 – En haut, les 3 premières constellations du mot *transport* du thème *Transition écologique* du GDN. Les cliques du contextonyme *privilégier* sont affichées. En bas, la plus grande constellation a été subdivisée en 3.

Les thèmes principaux La carte générée, en forme de triangle, fait apparaître 3 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.

Dans la constellation relative à l'écologie (numérotée 5 sur la carte 2) 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 les contributeurs s'y opposent. Cela révèle que chacun des mots est régulièrement présent à l'intérieur de phrases contenant *contre* et dans lesquelles peuvent apparaître des objections ici en matière d'écologie et de 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 cliques et convergence des enveloppes) soulignent donc que l'écologie est un thème important, organisateur des principaux thèmes représentés sur la carte.

Dans la branche relative aux questions sociales se trouvent 3 constellations : *combattre pauvreté, forme, dérive, extrême* (numérotée 9 sur la carte), *racisme, discrimination, religieux, communautarisme, exclusion, haine, harcèlement, musulman, femme, blanc* (numérotée 10) et *violence, anti, délinquance, drogue, délit, terrorisme, agression, incivilité, précarité, courant, manifestation, voter* (numérotée 11).

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

L'écologie comme élément pivot et domaine de convergence de l'ensemble des thèmes La constellation 5, relative à l'écologie, est reliée à celles relatives au social à travers une autre 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*, *protéger* dont l'enveloppe s'étend de la constellation 5 pour atteindre les constellations $4, 9, 10$ à thème

1	**chômeur**
2	**évasion, optimisation, fraude, taxation, fonds, groupe, efficacement, 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**

FIGURE 2 – La carte du mot **contre** (GDN) et 11 constellations

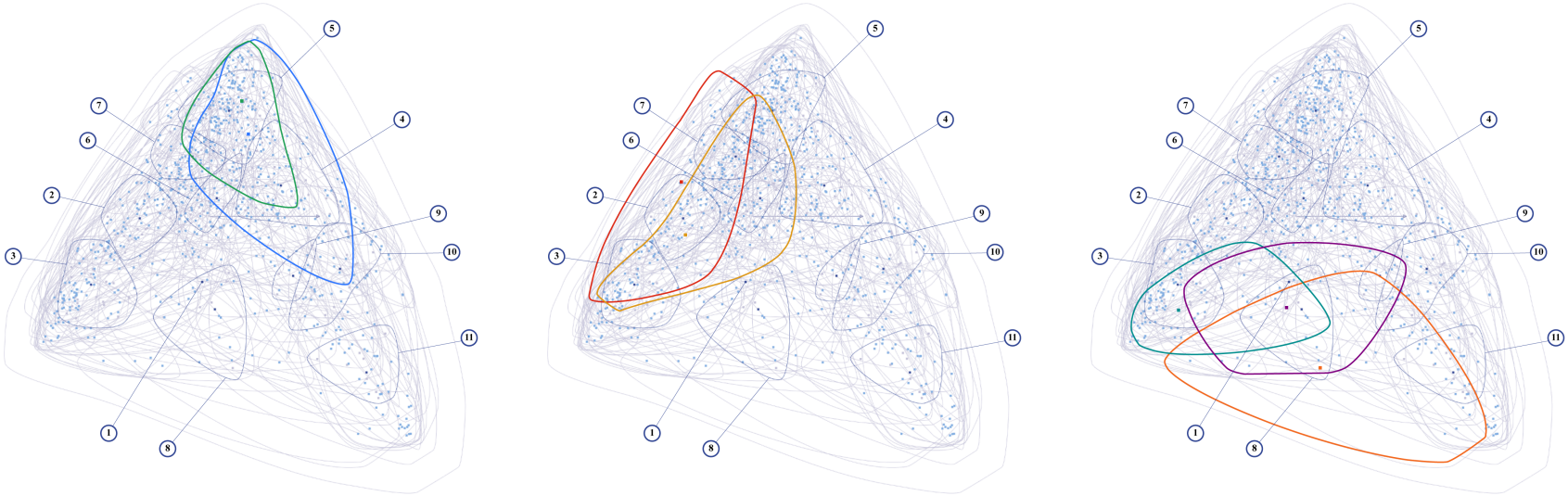

FIGURE 3 – La carte du mot **contre** (GDN) : liens entre les thèmes majeurs.

social (voir la figure 3), est aussi un élément charnière de ce lien. Protection de la nature et de la diversité des espèces représentée, par exemple, par la clique : *lutter, **protéger**, contre, pesticide, biodiversité, lobby, chasse*, et des personnes à travers la clique *lutter, **protéger**, contre, précarité, violence*.

La constellation 5 est reliée à celles relatives aux questions économiques et fiscales précédemment citées par celle intermédiaire (numérotée 7) évoquant 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 3, sont mises en évidence les enveloppes des mots *efficacement* et *mesure*.

Notons aussi la présence plus ténue, mais liant directement les thèmes sociétaux et 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 3, 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.

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

Perspectives

Ce corpus est d'une grande richesse. Nous avons cherché à montrer comment la méthode d'analyse automatique des AS pouvait être utile à son exploitation. Pour aller plus loin, certains points demanderaient de nouvelles investigations. En particulier, il est apparu que des différences pouvaient exister entre les contributions individuelles et collectives. Il serait intéressant de relancer les analyses en séparant ces deux types de contributions. En outre, le mouvement des *Gilets jaunes* a pu être interprété comme une conséquence d'une fracture entre les métropoles et les autres territoires ([1, 3]). 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 attaché à chaque contribution afin d'extraire la densité de population correspondante. Autre point : la teneur des contributions a pu évoluer en fonction des événements et de l'actualité, nous projetons des analyses qui étudient l'évolution des champ 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 plates-formes

numériques ont des profils distincts. Exploiter ces cahiers, apporterait une image plus exhaustive de l'ensemble des réponses au grand débat.

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.

Références

[1] Le premier "baromètre des territoires" montre la fracture du pays. `www.banquedesterritoires.fr/le-premier-barometre-des-territoires-montre-la-fracture-du-pays`, 2019.

[2] Jean-Paul Benzécri. *L'analyse des données : l'analyse des correspondances.* Bordas, Paris, 1980.

[3] Frédéric Lainé. Dynamique de l'emploi et des métiers : quelle fracture territoriale ? `https://www.strategie.gouv.fr/sites/strategie.gouv.fr/files/atoms/files/na53-fractures-territoriales-ok.pdf`, 2017.

[4] H. Ji, S. Ploux, and E. Wehrli. Lexical knowledge representation with contexonyms. *Proceedings of the 9th Machine Translation Summit*, pages 194–201, 2003.

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[6] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove : Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[7] Sabine Ploux, Armelle Boussidan, and Hyungsuk Ji. The Semantic Atlas : an Interactive Model of Lexical Representation. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 2010. LREC.

[8] Gilbert Saporta. *Probabilités, analyse des données et statistique.* Editions Technip, 2006.

Annexes

Fréquences, cooccurrence et contextonymie

Un traitement du corpus des réponses au GDN a été effectué sur les 4 sous-corpus obtenus pour chacun des thèmes ainsi que sur l'ensemble (tous thèmes confondus). Parce qu'ils sont peu volumineux [4], les deux corpus du VD et d'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 lemmatisés. 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 variantes 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 ainsi que la fréquence de cooccurrence ordonnée (le nombre

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

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 retirés [5]. Ainsi par exemple, sont retirés les déterminants (*le*, *la*, *les*, *une*, *un*, etc.) avec lesquels tout mot peut avoir une fréquence de cooccurrence très importante sans que cela soit informatif du point de vue sémantique. 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 étudié [6]. 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 cooccurrence [7]. On appelle *contextonymes* les cooccurrents réguliers ainsi calculés. A 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 3 plates-formes 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 cliques [8] 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 autres [9]. Une fois les cliques obtenues, elle 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 (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.

5. Ici, les paramètres sont fixés à $f_{i,j} > 3$ et $f_{i,j} < 1,5 * f_i$ où f_i est la fréquence du mot étudié i et $f_{i,j}$ la fréquence de cooccurrence du terme j avec i.

6. 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*.

7. Il s'agit des cooccurrents du mot étudié dont l'indice de cooccurrence est supérieur à $m_i + \alpha * \sigma_i$ 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 AS, α est fixé à 4%.

8. 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 arrêtes les liens de contextonymie.

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

Construction des cartes

Pour chaque mot étudié (ou pour un ensemble de mots), une analyse factorielle des correspondances [2] 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 associée un point. Chaque mot est figuré par l'enveloppe englobant les cliques qui le contiennent. Une classification hiérarchique par la méthode de Ward [8] est appliquée sur le nuage des points afin de dégager les regroupements cohérents de sens. Les cartes données ci-dessous illustrent le résultat de cette construction, l'ensemble des résultats est disponible sur le site des AS (`www.atlas-semantiques.eu/GDN_as.html?l=FR`). On pourra se reporter à la référence [7] pour plus de détails sur les méthodes de calcul des cartes.