Intelligibilité du numérique

N° 5 | 2024
Les plateformes en ligne comme dispositifs de production relationnelle. Produire de la relation, produire par la relation.

Comment les plateformes Reddit et Google différent-elles dans leur médiation sociale du Web ?

Une Analyse Structurale de Réseau de corpus comparés de résultats de recherche

Pascal JOLLIVET-COURTOIS *Maître de conférences*
Technologies et Sciences de l'Homme
COSTECH (EA 2223)
University of Technology of Compiègne

Édition électronique :
URL :
https://intelligibilite-numerique.numerev.com/index.php/articles/revue-5/2184-comment-les-plateformes-reddit-et-google-different-elles-dans-leur-mediation-sociale-du-web
DOI : numerev_1953
Date de publication : 20/03/2024

CertiScience© *Certifié évalué par les pairs*

Cette publication est sous licence **CC BY-NC-ND** (Attribution - No commercial - No derivatives).

Pour **citer cette publication** : JOLLIVET-COURTOIS, P. (2024) Comment les plateformes Reddit et Google différent-elles dans leur médiation sociale du Web ?. *Intelligibilité du numérique*, (5). https://doi.org/10.34745/numerev_1953

La notion de « plateforme », dans sa prétention scientifique, manifeste toujours à la fois une grande polysémie et une non moins grande polychésie. Néanmoins, on s'accorde aisément sur le fait que les « plateformes numériques » se rapportent à une modalité d'instrumentation de la communication - c'est-à-dire la présence *sine qua non* d'un dispositif technique de médiatisation - et que ce dispositif implique un mode d'organisation spécifique (des mises en relations multiples coordonnées par un opérateur). À partir de ces éléments, nous proposons de considérer ces plateformes comme des dispositifs d'intermédiation médiatisée en vue d'une production. Ce faisant, il s'agit d'insister sur le fait qu'un dispositif déborde la seule dimension technique (bien que celle-ci soit prégnante) et se compose également de normes et de conventions, de représentations et de justifications portés par des discours d'escorte, bref, d'une pluralité d'éléments qui, en l'occurrence, participent de la configuration de la relation. Il s'agit également d'insister sur le fait qu'une plateforme en ligne a comme finalité une production et que celle-ci procède de la relation. La question centrale est dès lors : qu'est-ce qui est produit par cette mise en relation ?

Can we systematically analyse and interpret the differences in search results provided by the Reddit and Google platforms, following an identical query? Are these differences structural? Finally, can these dissimilarities be explained overall by the intensity with which a particular engine drives and captures social interactions?
Within the framework of anchored theorization and mixed methods, we will implement a text Structural Network Analysis on the corpus consisting of the respective search results. Three lexical network maps will be generated and compared.
The results of these comparisons show that both quantitative and qualitative differences appear to be significant. The content of these differences suggests an information retrieval based on a greater generalization capability for Reddit than for Google.
The explanation of these differences can only be partial. We hypothesize that they stem from the differentiated intensity with which each platform develops its enactive social mediation function through interaction (high for Reddit, low for Google).

Mots-clés :
Moteurs de recherche, ASR textuelle, Médiation sociale, Méthode mixte, Text SNA, Text SNA, Social mediation, Search engines, Mixed method

1. Introduction

1.1. Problématique

Peut-on qualifier des différences entre les propositions de valeur - sémantique - de deux plateformes dites d'infomédiation ? En l'instance, les plateformes du moteur de recherche *Google* d'une part, et celle de curation sociale *Reddit* d'autre part ? Si nous parvenons à qualifier ces différences de proposition de valeur sémantique (statistiquement via l'analyse structurale de réseaux textuel, puis qualitativement via l'interprétation) serons-nous capables d'en expliquer la genèse ?

Plus spécifiquement, l'hypothèse explicative selon laquelle chaque plateforme mobiliserait l'interactivité de la cognition sociale (Deschryver, 2008) dans une ampleur et dans des modalités différentes (se posant plus ou moins intensément en médiateur des interactions sociales) fournit-elle une voie de réponse ?

Cet article présentant un caractère multidisciplinaire - voire transdisciplinaire - nous ne pourrons pas parcourir chacune des disciplines évoquées. Nous nous focaliserons plutôt sur un chaînon intellectuel qui nous semble clef dans la possibilité même de pratiquer la pluri/transdisciplinarité : la *méthode mixte,* qui articule rigoureusement et finement les approches, longtemps opposées, qualitatives et quantitatives (Creswell et al., 2011; Teddlie & Tashakkori, 2008).

1.2. Approche théorique

Un des concepts qui appelle, pour notre problématique, une approche *multi* voire *transdisciplinaire* est celui de *valeur*, en associant notamment les sciences économiques et de gestion, la sociologie et la linguistique (pour ne pas mentionner les sciences cognitives). Rien qu'au sein des sciences économiques, la question de la valeur connait des lectures plurielles et parfois controversées (Piluso, 2013).

La notion de valeur sémantique, en tant que transverse aux disciplines de gestion et de linguistique, bien qu'abordée par quelques chercheurs (dont Samuel Szoniecky, 2013) peut ainsi d'emblée être considérée comme problématique. Pourtant, dans un capitalisme de régime désormais cognitif, la question de l'économie du sens et de la cognition, tout comme de l'économie de l'attention, s'impose (Citton, 2014; Moulier Boutang, 2007) et nous tenterons donc l'usage de cette locution *valeur sémantique* (au-

delà de sa seule acception en linguistique).

Nous ne l'aborderons cependant ici que de façon pragmatique : suis-je capable, comme sujet chercheur, en associant rigoureusement statistiques de réseau et interprétation de “thèmes” via les verbatim, de repérer des différences entre telle ou telle plateforme, dans leur manière de donner accès et de rendre intelligible le Web en réponse à une requête, présentant donc une valeur ajoutée « sémantique » ?

1.3. Cadre Méthodologique

Dans une approche constructiviste, nous adopterons une méthodologie mixte (*mixed method* en anglais) articulant finement et systématiquement des *moments* quantitatifs et qualitatifs de l’analyse, impulsant un processus de type co-évolutif entre matériau empirique et catégories d’analyse (Johnson & Onwuegbuzie, 2004). Nous tenterons d’opérationnaliser cette approche et sa méthode en mobilisant les sciences des réseaux, appliquées à l’analyse structurale de réseau de corpus textuels, que nous couplerons avec une analyse de contenu (et non pas seulement du *text mining*)(Macia, 2015)

1.4. Analyses et résultats visés

Nous procéderons à la présentation de nos résultats et à leurs analyses en deux temps principaux.

Dans le premier, nous présenterons les cartes lexicales générées avec les résultats de recherche obtenus[1] d’une part via la plateforme Google et d’autre part via la plateforme Reddit. Nous pourrons avoir une première appréhension de l’ampleur des différences entre les corpus de résultats de ces deux plateformes, malgré l’utilisation de la même requête. La méthode utilisée pour effectuer l’analyse des graphes lexicaux sera l’analyse structurale de réseau textuel déjà mentionné et tout particulièrement l’analyse de cluster (statistique) pour faire émerger, par interprétation et prise en compte du contexte (par recours aux verbatim) des thèmes principaux (niveau sémantique).

Dans un deuxième temps, nous approfondirons cette analyse des différences entre corpus de résultats provenant de Google ou de Reddit, dans le cadre de notre requête donnée, en ayant recours à un outil (fonctionnalité du logiciel InfraNodus) particulièrement pertinent pour notre travail : la génération, statistique, d’une carte de *complémentarité*, à savoir une représentation de ce qui est dans un des corpus et de ce qui n’est pas dans l’autre. Par soucis de concision, nous choisirons de générer uniquement la carte de ce qui est présent dans le corpus des résultats via Reddit et qui n’est pas dans ceux via Google (proche du concept de complémentaire du corpus Google dans Reddit). Cet exercice nous permettra, après analyse et interprétation,

d’identifier, pour une même requête, ce que Google ne mentionne pas, alors que Reddit le fait.

Nous tenterons pour conclure d’interpréter ces différences aux prismes des processus socio-techniques à l’œuvre dans la génération des résultats pour chaque plateforme, en prêtant une attention particulière aux dynamiques socio-cognitives en jeu, et plus précisément aux différentes modalités de médiation des interactions sociales (Jeanneret, 2014) proposées par chaque plateforme.

2. Méthode mixte et théorie constructiviste : Analyse Structurelle de Réseau textuel et analyse de contenu

Nos terrains d’étude de traces volumineuses d’interactions sociales médiatées (Jeanneret, 2014) – notamment sur les forums de Reddit (Jollivet, 2018) - nous ont amené à développer une attention particulière aux cadres théoriques permettant de coupler finement analyses quantitatives et qualitatives d’une part, tout en autorisant l’appréhension de processus d’émergence (ici, de sens) d’autre part. Nous avons ainsi progressivement abouti à mobiliser le cadre théorique du *constructivisme* (Avenier, 2011) dans sa déclinaison de la *théorisation ancrée*, et la méthode *mixte*. La théorisation ancrée (*grounded theory*) peu répandue en France, permet au chercheur de prendre part à une coévolution entre constitution du matériau empirique et catégorie d’analyse théorique, qui rompt avec les dichotomies des épistémologies traditionnelles de type *données empiriques/hypothèse théorique*, ou encore *objet/sujet de la recherche* (Guillemette & Luckerhoff, 2009; Méliani, 2013; Paillé, 1994). Ce cadre théorique s’accorde avec l’approche méthodologique connue sous le nom de *méthode mixte* (*mixed method*) qui vise à développer une méthodologie permettant de s’émanciper de la dichotomie entre qualitatif et quantitatif (Creswell, 2013; Creswell & Clark, 2011; Fakis et al., 2013).

2.1. Analyse structurale de réseau textuel et implémentation dans un dispositif logiciel

En tant que déclinaison des sciences des réseaux pour les humanités numériques, nous adoptons la méthode d'Analyse Structurale de Réseaux textuel (*text-SNA* en anglais), avec sa facilité de cartographie visuelle, dans son implémentation par l'outil Infranodus[2] (logiciel SAAS) (Paranyushkin, 2011). Tout d'abord, ce logiciel fournit des moyens efficaces d'explorer ("manuellement", avec la souris) de grands corpus de textes, avec des fonctions de lecture à distance et de zoom avant/arrière, permettant de zoomer jusqu'à la phrase originale elle-même. Ainsi, il est possible d’explorer un corpus en

allant de la lecture distante à l'analyse fine de la textualité du discours. Deuxièmement, parce que les calculs d'analyse de contenu basés sur le réseau (et en particulier l'algorithme de clustering) fournissent un moyen efficace d'identifier de potentiels « sujets principaux », mais aussi les thèmes influents sous-jacents, et les "angles morts" dans un corpus (op. cit.). Enfin, parce que cette implémentation logicielle permet d'éditer "à la volée" le matériel textuel lui-même, avec des étiquettes (*tags*) pour marquer des catégories.

2.2. Approche constructiviste permise par le dispositif logiciel

Ces trois possibilités permettent de déployer et implémenter une posture constructiviste, telle que définie initialement par Berger & Luckmann (1966) et discutée notamment dans Avenier (2011). Tout d'abord, lors de l'exploration des corpus, nous pouvons catégoriser les éléments de texte (avec les étiquettes) au fur et à mesure de leur lecture. Après avoir édité une série d'éléments textuels (*posts* ou phrases), nous pouvons relancer le calcul des clusters de textes, avec la méthode SNA, pour obtenir une nouvelle carte de clusters. Avec cette nouvelle représentation synthétique des corpus, nous pouvons à nouveau explorer (mais différemment) le texte et faire évoluer nos catégories, en fonction de notre nouvelle compréhension/conceptualisation. Comme dans un processus co-évolutif, le matériau empirique évolue en s'enrichissant de *tags* (étiquettes) et les catégories d'analyse évoluent également, puisque le matériel enrichi suggère de nouvelles significations (Hesse-Biber, 2010) par l'intermédiaire des cartes d'agrégats. Comme dans les langages formels de balisage (comme le HTML), le texte et l'analyse du texte (en termes de métadonnées) finissent par s'entremêler. Ce processus contribue à une construction progressive de la signification, s'inscrivant dans une démarche constructiviste (Avenier, 2011).

2.3. Analyse de contenu et Verbatim

Une fois le processus constructiviste considéré par le chercheur comme convergeant vers un état interprétatif stable (*satisficing*) (Dowding, 2013) s'en suit une analyse de contenu agrégat par agrégat, lemmes principaux par lemmes principaux. Nous nous appuyions ainsi sur les agrégats statistiquement construits pour chercher du sens en consultant certains des verbatim associés.

3. Constitution des deux corpus de résultats issus des plateformes Reddit et Google, cartographie,

modalités d’analyse et d’interprétation

En nous inscrivant dans l’approche constructiviste évoquée, nous ne structurerons pas cet article selon l'épistémologie expérimentaliste traditionnelle en trois sections : *matériau empirique/ expérimentation/analyse des résultats*. Nous l'avons structuré autour du processus *design/interprétation*, dans lequel les deux étapes forment un cycle. Nous pouvons toutefois différencier le travail et les moments *d'analyse* - dans lequel nous prêtons principalement attention aux propriétés structurelles statistiques du graphe (clusters, nœuds influents, trous structurels) - et l'étape et le moment *d'interprétation* (où nous travaillons sur les significations)[3].

Nous allons d'abord explorer et analyser les résultats de recherche (notre corpus) sur notre controverse (voir infra, *« Biodiversité, vs. énergies renouvelables (aspects économiques) ? »)* tels qu’ils se présentent via les plates-formes Google Search, puis via Reddit. Ensuite, dans une troisième section, nous générons et analyserons une carte des différences (statistiques) pour tenter d'interpréter l’origine de ces dissimilitudes.

3.1. Nos deux dispositifs de constitution de corpus sur notre problématique

Le volume des données textuelles que nous avons collecté, et donc de nos corpus, a été volontairement limité au niveau compact des 150 premiers résultats par plateforme utilisée. Ce choix correspond notamment à un parti pris de privilégier la similitude avec les usages effectifs d’internautes sur ces plates-formes, et non pas de viser à une représentativité du Web, fut-elle limitée au sujet de la recherche formulée (voir annexe 1).

Nous avons choisi comme questionnement, pour formuler notre requête, une problématique que nous avions déjà travaillée dans une autre recherche (Jollivet, 2022) ce qui facilite notre capacité d’interprétation. Il s’agit d’une controverse scientifique et socio-économique sur les possibles tensions, voire contradictions, entre la généralisation des énergies renouvelables et la protection de la biodiversité, selon un axe problématique économique. Le cas le plus médiatisé en France sur cette controverse (AFP & Libération, 2021) est peut être celui de l’implantation d'éoliennes offshore en baie de Saint-Brieuc et des plaintes et revendications associées : les pêcheurs professionnels, pour motif de perte de revenus occasionnés par la réduction des surfaces de pêches et de chute de la biomasse marine ; les hôteliers professionnels du tourisme en front de mer au motif d’un préjudice de perte de paysage. Un parallèle, dans l’arène scientifique, s’est présentée via la controverse (entre écologues des fonds sous-marins) concernant l’effet négatif ou positif sur les ressources halieutiques à terme d’une artificialisation localisée des sols marins par les infrastructures éoliennes

(Raoux et al., 2017).

Pour chacun des deux dispositifs de médiation socio-informationnelle – qui mettent en relations non seulement des informations, telles que requêtes et résultats, mais également des personnes - que nous avons utilisés (Reddit ou Google) nous avons formulé exactement la même requête, selon la syntaxe booléenne suivante :

```
Biodiversity AND renewable energy AND (economic OR economy)
```

La formalisation sous forme booléenne, plutôt qu'en langage naturel peut être discutée. Nous l'avons choisie principalement dans une perspective d'interopérabilité et de comparabilité entre différentes plateformes (par ex. avec la base de Reddit en ligne requêtable en langage SQL compatible avec le booléen). Cette formulation est à également, ici, volontairement simplifiée[4], reprenant presque mot par mot les termes des attendus formulés par l'Office Français de la Biodiversité dans le cadre du projet de recherche (Jollivet, op. cit.)

3.2. La plate-forme de curation sociale Reddit comme dispositif pour constituer notre corpus d'extraits du Web (1)

Cette plate-forme de curation sociale n'est pas très connue en France, ni en Europe (au temps de cet article). Elle est par contre extrêmement populaire sur le continent américain, du Nord comme du Sud, et se classe parmi les 10 sites les plus fréquentés du Web occidental. De prime abord, cette plateforme s'apparente à un gigantesque enchevêtrement d'une multitude de forums de discussions. En réalité, cette plateforme conserve en partie son projet d'origine, à savoir d'être un portail de curation sociale des ressources Web : la curation (sélection, présentation et commentaire, pointage sur le Web) se réalise de façon contributive par des multitudes en interactions et non pas un (ou plusieurs) expert isolé. Ainsi certains utilisateurs en viennent à s'affranchir des moteurs de recherche dédiés (de type Google, Bing et concurrents) pour directement effectuer leur recherche via la plate-forme Reddit, au titre d'une pertinence des appariements jugée meilleure (Bouvron, 2022)

Nous avions plusieurs techniques à notre disposition pour constituer un corpus d'extraits de fils de discussion issus de Reddit concernant notre requête problématique. Comme nous n'avions pas l'ambition de travailler sur d'importants volumes de corpus textuels, nous n'avons pas recouru à la technique *reine* consistant réaliser des requêtes très structurées en langage SQL directement sur la base de données Reddit accessible sur le Web (ni à un logiciel de Scraping spécifique pour Reddit). Nous avons eu simplement recours à *Google recherche avancé*[5]. Ce recours n'est absolument pas neutre. La première raison en est que, selon les discussions sur certains forums, il semblerait que Google indexe mal, ou incomplètement, Reddit (laofmoonster, 2012).

Ensuite, Google restitue le contenu sélectionné par lui via Reddit sous une forme particulière, différente de celle du moteur de recherche intégré à Reddit lui-même[6].

Malgré ses biais éventuels, il nous est apparu que les avantages du recours à Google, en termes de simplicité, dépassaient ses défauts, et ce d'autant plus que la petitesse des deux corpus nous permettait de réaliser à la main des vérifications et corrections d'ensemble.

La plus importante des caractéristiques de cette plate-forme de curation sociale, pour notre propos, est cependant bien la richesse des interactions sociales qui président à la production de ses contenus textuels. Elle nous semble ainsi faire écho à la thèse de la cognition sociale interactive. Nous y reviendrons.

3.3. La plate-forme Google Recherche comme dispositif pour constituer notre corpus d'extraits du Web (2)

Le site d'infomédiation (qui met en relation requêtes et résultats informationnels) Google Recherche n'est plus à présenter. Leader mondial des moteurs de recherche pour le Web, il ne souffre en Europe d'aucun concurrent sérieux.

Sa non-neutralité ne fait cependant pas de doute - les plus connus des aspects portant sur le filtrage de requêtes et sites considérés comme relatifs au terrorisme (Le Figaro & AFP, 2017) ce que Reddit ne réalise qu'avec beaucoup de précaution (Belanger, 2023). Mais comme le moteur de recherche constitue une référence, ce sont ses rivaux qui semblent devoir justifier de leurs différences. Il peut être intéressant de noter pour notre propos que ce moteur de recherche tend de plus en plus à ne plus pointer vers des ressources Web extérieures à lui-même (positionnement initial du moteur), mais à proposer des synthèses ou des services de son cru, sur sa propre première page de recherche, gardant le client-utilisateur plus longtemps captif de son écosystème (Autorité de la Concurrence, 2021) . En découle une augmentation du temps de captation d'attention, de la valeur de l'exposition de l'utilisateur à la publicité, et donc de la valeur marchande du service proposé par la plate-forme Google.

3.4. Outil, cartographie (design de cartes) analyses, interprétations

Lors de notre production, édition, et analyse des cartes, nous allons procéder selon plusieurs moments, qui pourront s'enchaîner par circularité.

La première étape, et le premier moment, consistera en la production d'une première version de carte, via notre outil InfraNodus, qui implémente une méthode d'analyse structurale de réseau textuel (Paranyushkin, 2019)[7].

S’en suivra une étape et un moment d’édition de cette première version. Nous pourrons ainsi procéder à la suppression de certains lemmes sur la carte ou à la fusion de deux lemmes que nous considérons pour nos propos comme quasi synonymes, ou encore à la concaténation lorsque nous pensons identifier des expressions-formules (par exemple, *changement_climatique*). Cette étape s’apparente à une activité de *design* de cartes (plutôt au sens de conception qu’esthétique) ou plus simplement de cartographie.

La troisième étape, ou moment, consistera à utiliser certains des indicateurs statistiques spécifiques des sciences de réseau pour effectuer une première analyse quantitative de notre corpus.

Le détail de ces indicateurs, sur les 50 premiers lemmes de chaque carte, figure en annexe 2, 3 et 4.

Le premier des indicateurs statistiques que nous mobiliserons sera le classique *cluster* ou agrégat, via une variante de l’algorithme de clustering de Louvain (Blondel et al. 2008) disponible dans InfraNodus, avec tous les calculs et indicateurs usuels disponibles. Il peut être noté que, rompant avec la séquentialité des étapes, le chercheur se sert aussi de ces structurations et visualisation (sous forme d’agrégats) pour reprendre son travail d’édition (à la main) en choisissant à nouveau de supprimer ou de fusionner tel ou tel élément. Un autre aspect du réseau textuel, topologique, est généralement exploré également : il s’agit de l’analyse de distances entre différents *clusters* (qui peuvent être calculées, mais que nous réaliserons visuellement). Enfin, l’indicateur statistique *d’intermédiarité*[8] pourra être extrêmement utile pour l’interprétation de l’étape suivante.

Cette dernière étape consistera à réaliser une interprétation de notre carte, amorçant l’étape sémantique de l’analyse. Il faudra différencier ici une première activité interprétative qui se voudra la plus proche du texte (à savoir les verbatim) d’une seconde activité interprétative, plus proche de l’analyse de contenu et qui tentera d’accéder à une partie d’implicite, ou de sous-entendus, véhiculé par les textes.

Nous avons mentionné comment nous considérons l’activité d’édition ou de design de carte comme partie intégrante d’une approche constructiviste, dans le contexte d’une méthode mixte.

Il est intéressant de noter qu’à chaque opération réalisée sur la carte par l’utilisateur, le logiciel recalcule l’intégralité de la carte et en fournit une nouvelle version, de manière interactive. Ce dispositif permet ainsi, par un processus d’essai-erreur, d’explorer un corpus textuel (structuré selon l’analyse de réseaux) en appréciant, étape après étape, l’effet de nos actions d'édition sur l’intelligibilité du réseau textuel (selon notre propre entendement). Loin d’être anecdotique, ce dispositif logiciel nous semble fournir aux chercheurs une vraie *encapacitation*, via ce que certains dénomment, au sein du paradigme de l’enaction en sciences cognitives, une *interface enactive* (Lenay, 2006,

2008)

Pour conclure ce préambule sur les modalités d'interprétation de ces cartes d'analyse structurale de réseau textuel, il convient de désigner un absent sur ces visualisations : les lemmes constituant la requête initiale. Cette pratique (retirer les termes d'une requête de la carte) est assez courante (Center for History and New Media, 2017). Elle permet de mieux visualiser les résultats eux-mêmes qui risqueraient d'être *écrasés* visuellement par l'importance statistique de ces termes de la recherche. Mais cette pratique nécessite un petit exercice intellectuel lors de l'analyse, puis l'interprétation : il faut s'imaginer les termes de la requête comme omniprésent sur la carte, à la manière d'un grand titre qui la surplomberait.

Nous ne ferons qu'aborder l'analyse et l'interprétation des deux premières cartes correspondant respectivement à l'analyse de corpus via Reddit et via Google, en nous limitant aux caractéristiques topologiques globales et à quelques clusters. En effet, nous avons jugé plus judicieux de nous concentrer sur l'analyse et l'interprétation de la carte ultime : la carte qui croise ses deux cartes, mettant en exergue, ce qui existe dans le corpus Reddit mais est *absent* du corpus Google.

4. Elaboration de la carte issue de la recherche via Google, analyse et interprétation

Le premier indicateur topographique global qui nous est proposé est dénommé dans notre logiciel *variabilité sémantique*[9] (Paranyushkin, 2019) . Il apparaît ici (en bas à droite du visuel) clairement dans le rouge. Cet indicateur vise à fournir une évaluation du caractère *équilibré* du réseau textuel (en multidimensionnel) dans l'esprit de ce qui est dénommé par certains *chercheurs en Traitement Automatique des Langues (TAL) un graphe bien balancé*.

4.1. Biodiversité et énergies renouvelables (aspects éco) : carte des 150 résultats de Google



Figure 1 : Carte-réseau des résultats de recherche via Google

4.2. Analyse de la carte

Il apparaît ainsi que le terme composé *changement climatique* (*Climate change*) est doté d'un poids statistique qui apparait visuellement relativement disproportionné, dans le contexte d'un discours, par rapport aux autres lemmes.

Nous avons cependant décidé après réflexion de ne pas le supprimer. Nous voulions pouvoir explorer en cela les performances respectives de Google et de Reddit dans une certaine capacité *d'abstraction*, ou plus précisément dans une capacité qui pourrait s'apparenter à une faculté de *généralisation*. En effet, dans notre requête, beaucoup d'implicites sont présents ou, autrement dit, beaucoup de connaissances du monde – non explicitées - sont requises pour comprendre l'énoncé.

Comment le dispositif Google Search a-t-il fait pour "comprendre" qu'une relation essentielle *entre « énergie renouvelable » et « biodiversité »* demandait de passer par le concept - plus général - de *réchauffement climatique*? Car pour la majorité des lecteurs, l'implicite selon lequel la motivation du développement des énergies renouvelables est l'excès d'émissions de gaz à effet de serre, cause première du réchauffement climatique anthropique, est évident.

Nous avons donc conservé cette première carte avec ce lemme dans l'optique d'apprécier s'il existait, à travers la comparaison des deux cartes sur ce lemme, des différences perceptibles dans cette capacité de généralisation de ces deux plates-formes, selon l'exemple susmentionné[10].

4.3. Analyse et Interprétation du cluster 1 (17% du corpus) : climate change/loss/circular

Le premier cluster semble de prime abord, à titre d'hypothèse interprétative, plutôt homogène et cohérent sémantiquement (à savoir interprétable pour nous). Nous avons vu le lien entre le lemme locution *changement climatique* et les termes de notre requête (généralisation). Le deuxième terme, *perte* (*loss* en anglais) s'associe directement à un des termes de notre requête : la biodiversité. Le troisième terme semble d'un registre relativement différent. Il s'avère (via la lecture de verbatim) s'insérer dans l'expression *économie circulaire*. La juxtaposition sous forme d'agrégats de ces trois lemmes fait-il sens. ?

Nous avons bien ici une cooccurrence, dans cet extrait procuré par Google :

> *« For example in areas like renewable energy, biodiversity or circular economy. The goal is to reach a climate-neutral economy in the EU by 2050, ...4 days ago »*

Complété par cet extrait :

> *« 23 juin 2022 — circular economy cloud behind renewable energy … Mining threats to biodiversity will increase as more mines target materials for renewable … »*

Mais il n'est pas étonnant que nous ne trouvions pas une phrase avec tous les termes souhaités (ceux du cluster et ceux de la requête) ne serait-ce que du fait que les corrélations retenues bénéficient d'une certaine transitivité.

Notre interprétation globale de ce cluster, aidé par les indicateurs statistiques et par les verbatim, est que *la principale réponse économique que l'on peut apporter à la crise écologique constituée par le réchauffement climatique et la chute de la biodiversité est l'économie circulaire, couplée au développement des énergies renouvelables.*

Mais ce résultat interprétatif ne nous intéresse pas tant *dans l'absolu.* C'est sa similitude ou sa dissemblance avec les clusters de la carte sourcée sur Reddit qui nous importe ici.

4.4. Cluster 2 : Offshore/Water/Land (16% du corpus)

Ce verbatim avec la co-occurrence des trois termes de l'agrégat semble assez éclairant sur le sens de ce cluster :

> *« 27 janv. 2021 — Renewable Energy on Public Lands and in Offshore Waters. … protection for our lands, waters, and biodiversity and creating good jobs. 4 days ago »*

Les énergies renouvelables sur des espaces publics, et tout particulièrement en mer (offshore), sont ici présentées comme constituant la solution à presque tous les problèmes écologiques (le réchauffement climatique pourrait être considéré comme implicitement présent à travers la décarbonisation permises par les EnR) y compris sur le plan économique.

4.5. Cluster 3 : Wind/power/plant (13% du corpus)

L'interprétation suggérée semble directe : la "solution" à la triple crise serait constituée par le développement de centrales énergétiques mues par l'énergie éolienne offshore.

> *"17 janv. 2023 — Renewable energy is an opportunity to unite action on biodiversity and climate. Ocean-based renewable energy, such as offshore wind power..."*

4.6. Cluster 4 : Carbon/Policy/Emission (13% du corpus)

Cette triade parait la plus simple à interpréter. Le verbatim suivant semble le confirmer :

> *« The economics of climate change has encouraged that thought. Development of cheap renewable energy sources would help to reduce carbon emissions to …. 4 days ago »*

Cet extrait semble provenir d'un rapport d'expert-économiste - dont la posture est, typiquement, de conseiller (voire de prescrire) des *politiques* (*policy* en anglais) - en prenant grand soin de formuler la prescription de manière impersonnelle, comme si seule la science s'exprimait.

L'interprétation globale de cet agrégat, au vu d'autres verbatim, peut être résumée ainsi : *Le changement climatique présente un tel coût économique qu'il est urgent de réduire les émissions carbonées en développant les énergies renouvelables.*

Avec l'examen de notre deuxième carte, celle portant sur les résultats de requête sur Reddit, nous allons pouvoir commencer à avoir un aperçu des similitudes ou différences entre les deux plateformes dans leurs modalités d'offrir de *faire sens* du web, dans le contexte d'une réponse à une requête.

5. Elaboration de la carte issue de la recherche via Reddit, analyse et interprétation

L'indicateur global d'équilibre « sémantique » (*semantic variability*) est très différent de celui de la carte issue de Google : il se positionne à un gradient plutôt équilibré, bien que focalisé. Il est possible cependant que cette différence entre les deux graphes soit le fait d'un artefact: nous n'avons pas, en effet, sur ce graphe, effectué la fusion entre *climate* et *change* ce qui, vu le poids de ces deux lemmes, tend à mieux répartir l'inertie dans le graphe. Cependant, l'importance relative des lemmes *work* et *collapse* (qui n'existaient pas dans notre corpus précédent de manière significative) peut contribuer à expliquer cet indicateur plus équilibré.

Cet indicateur est censé souligner le caractère plus diversifié des champs lexicaux (et supposément des champs sémantiques) de notre corpus issu de Reddit. C'est un premier résultat de comparaison, que nous reprendrons.

5.1. Biodiversité et énergies renouvelables (aspects éco) : carte des 150 résultats de Reddit



Figure 2 : Carte-réseau des résultats de recherche via Reddit

5.2. Cluster 1 Work/capitalisme/future[11] (16 % du corpus)

Dans une quête de cohérence du sens, de prime abord, la présence du lemme *future* parmi les deux précédents étonne. L'examen des verbatim nous fournit une explication.

Tout d'abord, on peut noter l'importance des discussions dans Reddit sur l'ouvrage et la problématique du *futur du travail* (dans lequel l'auteur diagnostique, une nouvelle ère de substitution capital /travail avec l'avènement des robots dans l'économie ... capitaliste) : que devient le travail (et les travailleurs) dans un futur (assez proche) ou la logique d'accumulation renouvelée du capital engendrera une nouvelle génération de robots qui se substitueront dans bien de nouveaux secteurs à l'homme ?

Mais quel rapport cette discussion entretient-elle avec la biodiversité, et les énergies renouvelables ?

Un lien qui se dégage est celle de la triple crise (écologique, sociale, économique) annonçant un futur inquiétant menaçant de déclencher à un effondrement :

> "*I'am Federico Pistono, author of "Robots Will Steal Your Job ...https://www.reddit.com* ***Futurology***  *comments* *i_a..... automation, existential risks, and the* ***Future*** *of* ***Work****. My book ["Robots Will Steal Your Job, But That's OK: How to Survive the Economic* ***Collapse*** *and Be... a year ago*"

Un lien qui s'exprime ici de manière plus proactive :

> "*A more natural* ***future*** *for the Netherlands in 2120 - Reddithttps://www.reddit.com* *comments*  *eonro0* *nl_212...Jan 14, 2020* *It unfolds around the themes of water management,* ***energy****, agriculture, circular* ***economy****, urbanization and* ***biodiversity***"

5.3. Cluster 2 : collapse/futurology/political

De prime abord, le terme *futurology* peut paraître déplacé ici (connotation quelque peu ésotérique en français). Pourtant, comme déjà évoqué, la traduction en français en est souvent, selon le contexte (scientifique), le terme *prospective*, ce qui nous rapproche du champ sémantique des sciences économiques.

La thématique de *l'écroulement* est la deuxième (après le lemme *work* du cluster précédent) à sembler être *spécifique* au corpus des résultats par Reddit. En français ce terme peut évoquer une connotation catastrophiste (voire complotiste). Le verbatim qui suit en propose une analyse bien plus analytique, soulignant les interdépendances complexes entre économique, social et *politique* :

*“collapse - ﷺ wiki >ﷺBy **collapse**, I mean a drastic decrease in human population size and/or **political**/economic/social complexity over a considerable area...”*

5.4. Cluster 3 : growth/farm/carbon

L’examen des verbatim souligne la polysémie du terme *growth (*du moins pour notre étude) : la croissance *économique*, la croissance de *forêts* et la croissance des *émissions carbones*. L’extrait suivant en donne un aperçu :

*“Should we encourage economic **collapse** to prevent ... - collapse ﷺ comments ﷺ should_...Feb 11, 2020 ﷺ We aren't seeing decoupling between **economic growth** and carbon ... full substitution with **renewable energy** on the temperature rise, too”*

Le terme *farm* apparaît, lui, au sein de la problématique de l’impératif de changement de pratiques agricoles et d’élevage, pour limiter les émissions de CO2, en sortant d’une logique de croissance productiviste en monoculture.

5.5. Cluster 4 : climate/change/crisis

Nous avions mentionné plus haut notre motivation pour ne pas supprimer du graphe - malgré son poids déséquilibré et son caractère de contexte - la locution *réchauffement climatique* : essayer d’apprécier les capacités comparées de chaque plateforme à se comporter “comme si” elles réalisaient une opération cognitive de type *généralisation*. Sans que ce cas suffise pour conclure, on ne constate qu’aucune des plateformes n’a “échoué” dans cet exercice particulier, dans notre contexte[12].[13]

6. Elaboration de la carte des « différences » entre les recherches issues de Reddit et de Google : analyse et interprétation

Nous abordons maintenant le cœur de notre propos, à travers cette partie : comment peut-on qualifier les différences entre nos deux plateformes dans l'accès qu’elles fournissent au Web en réponse à une requête ?

Nous allons tout d’abord, synthétiser rapidement les différences, qui peuvent d’emblée apparaître à partir de la comparaison, agrégat par agrégat, de nos deux cartographies précédentes. Nous passerons ensuite à l’examen de la génération de la carte des différences entre les réseaux textuels de nos deux plates-formes.

6.1. Comparaison visuelle rapide des deux cartes existantes selon leurs agrégats respectifs (aperçu)

Nous constatons d'emblée en termes de comparaison d'agrégats qu'aucun des agrégats présents sur une carte n'est présent sur l'autre. Ceci constitue déjà un certain résultat. Ensuite, bien sûr, lorsque l'on ne s'attache pas exclusivement aux agrégats, en tant que tel, mais aux lemmes principaux qui les caractérisent, il est possible de trouver des similitudes (telles que *climate change*). Mais ce qui nous semble le plus intéressant et que s'offrent à nous des dissimilitudes nettes en termes de lemmes présents ou absents.

Ainsi, les deux termes d'agrégats, qui, d'un point de vue statistique, apparaissent comme très spécifiques à Reddit sont : *work* et *collapse.* Nous développerons en infra l'interprétation de cette différence qui apparait plus explicitement encore dans la nouvelle carte des « différences » suivante.

Contentons-nous pour l'instant de noter que sur la carte de Reddit semblent figurer de nombreux éléments abordant les aspects socio-politiques et d'économie politique, alors qu'ils sont presque absents (en écartant donc l'économie du changement climatique et des énergies renouvelables) dans la carte Google.

6.2. Biodiversité et énergies renouvelables (aspects éco) : carte des « différences » (ce qui est dans Reddit, mais pas dans Google) : analyse et interprétation

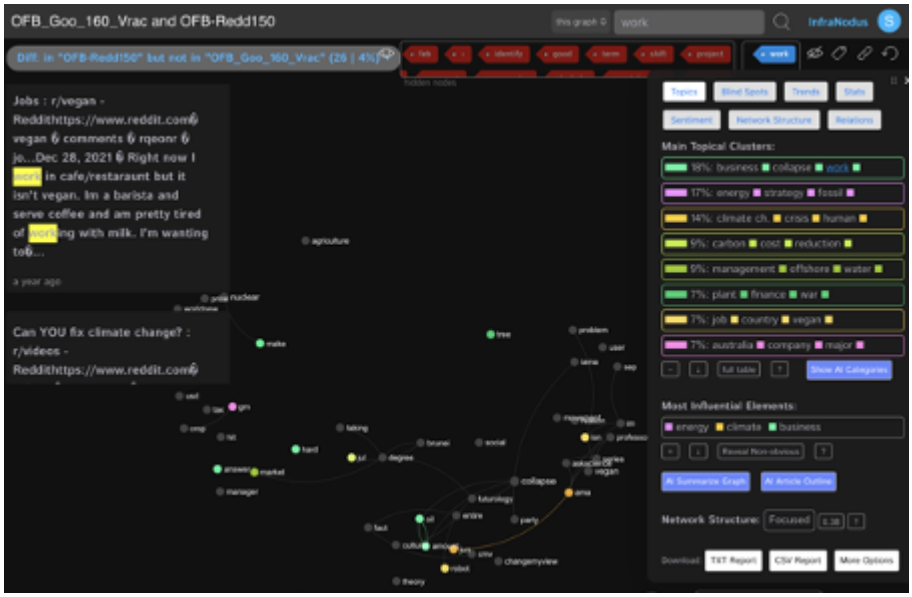


Figure 3 : Carte-réseau des "différences" de résultats de recherche Reddit/Google

La partie du lexique des corpus qui est inclue dans les résultats de la recherche sur Reddit mais qui n'est pas dans ceux de Google peut paraître, visuellement, marginale (cf. fig. 3). Mais bien que le logiciel ne fasse pas figurer l'intégralité des lemmes sur sa carte lexicale (uniquement ceux avec un certain seuil de co-fréquence) il propose des statistiques qui sont, elles, exhaustives. Ainsi, statistiquement, ce lexique représente 25% du total du lexique des corpus fusionnés, ce qui est loin d'être anecdotique. Au-delà de la simple quantification, il convient d'apprécier également si, *qualitativement* (notamment au niveau du sens), ce lexique diffère bien de celui de son complémentaire.

Pour procéder à cette analyse interprétative, nous allons successivement parcourir chaque cluster (statistique) et, en faisant des allers-retours pour consulter les verbatim correspondants (leur contexte), tenter une interprétation permettant de dégager du sens. Nous obtiendrons alors des « sujets » (« topics ») dont « l'originalité » par rapport au lexique complémentaire pourra alors être discutée.

7. Interprétation des « différences » via le cluster 1 « Business/Collapse/Work » : comment renouveler le travail et les affaires face à l'écroulement à venir?

7.1. Interprétation lemme par lemme du Cluster 1 :

7.1.1. Le lemme « work » et son interprétation en contexte

Le premier lemme qui se distingue, dans le premier cluster est « work » (travail). Ce terme est bien sûr polysémique et il convient de lever les ambiguïtés (en explorant les verbatim) avant de conclure à une interprétation. Il se révèle ainsi que le mot « work » présente au moins trois sens différents dans le corpus des verbatim :

- Le travail *labeur* comme dans « my work consists in … », « work for money » ou … « work and skills » (qui constitue environ 25% des occurrences du lemme), tous présents dans les verbatim;
- La *coordination avec*, comme dans « Work closely with the administration »,
- *S'occuper de* ou *traiter* comme dans « We will work on that », toujours présents dans les verbatim.

C'est le sens 1 qui est intéressant ici pour notre problématique. Une des phrases les plus représentatives en est :

> *"Jobs : r/vegan - ⁄ vegan ⁄ comments ⁄ rqeonr ⁄ jo…Dec 28, 2021 ⁄ Right now*

I work in cafe/restaurant but it isn't vegan. Im a barista and serve coffee and am pretty tired of working with milk. I'm wanting to quit but .... a year ago"

Cet extrait se situe dans le contexte d'un subreddit nommé : « what can YOU do for climate?». Il fait écho à une préoccupation assez présente sur la plateforme Reddit : la demande de conseils dans le cadre d'une recherche d'emploi. Ici, le lien avec notre problématique initiale (biodiversité/renouvelables/économie) transite par l'enjeu de l'alimentation (non carnée et même végétalienne) via le travail/emploi (démissionner pour cause d'incompatibilités « alimentaires »).

7.1.2. Premier Interprétation du différentiel

La question du *travail* peut de prime abord sembler bien éloignée de la problématique initiale. Sa présence, dans les résultats fournis par Reddit, ne parait ainsi pas nécessairement favoriser l'intelligibilité du web vis à vis de la requête formulée. Pourtant, quand on se réfère au terme de recherche - bien explicite - *économie/sciences économiques*, la présence de l'enjeu du travail (comme celui du *monde des affaires*) apparaît plutôt « naturelle ». L'étonnement face à la différence de lexique peut même se renverser : comment se fait-il que Google « choisisse » de ne pas inclure la question du travail, vis à vis de la transition écologique, lorsqu'il a pour ambition de nous favoriser l'intelligibilité du web « pertinent » suite à notre question ?

7.1.3. Le lemme « collapse » et son interprétation en contexte

Le terme de « collapse » (effondrement) peut de prime abord sembler également presque « hors sujet », étant donnée notre requête initiale. Google serait ici plus précis, pertinent, que Reddit, doté donc d'une valeur ajoutée sémantique supérieure dans sa capacité à *pointer* des ressources pertinentes sur le Web.

Pourtant, ne serait-ce que dans le champ académique (écologues, archeo-anthropologues ...) cet enjeu existe bien, et est lié à notre questionnement initial, notamment à travers la question de la biodiversité et de l'effondrement des populations de pollinisateurs. Dans l'arène de la « société civile », la question de l'effondrement (biologique, sociétal, économique ...) est également bien présente, comme le post qui suit s'en fait écho :

"*Should we encourage economic **collapse** to prevent [ecological collapse] ? ... - Redditht*[*tps://www.reddit.com*](https://www.reddit.com)*collapse comments should_...Feb 11, 2020 We aren't seeing decoupling between economic growth and carbon ... full substitution with **renewable energy** on the temperature rise, too.82 answers 37 votes:Problem is that an economic crisis at the current time will likely lead to regionalisation ...*"

Nous sommes amenés ici aussi à renverser l'étonnement premier : comment se fait-il que Google exclue l'enjeu de l'effondrement de ce qu'il *nous donne à voir* du web ?

7.1.4. Le lemme « business » et son interprétation en contexte

Ici encore l'on peut s'étonner sur une telle présence du *monde des affaires* dans un questionnement sur la biodiversité et les Energies Renouvelables (EnR).

La réponse semble claire. Ce qui distingue la plateforme Reddit de Google dans leurs résultats à notre requête consiste d'abord en un questionnement sur *comment continuer à aller travailler et à faire des affaires en entreprise lors d'un effondrement socio-économique et bioclimatique jugé inéluctable.*

> "*It's far too late to do anything about climate change. : r/collapsehttps://www.reddit.com*[https://www.reddit.com]*�* ***collapse*** *� comments � Dec 27, 2021 � These* ***businesses*** *and our governments have merged a long time ago. ... I am legitimately worried that economic/societal collapse will occur�... a year ago*"

8. Interprétation des « différences » via le cluster 2 « Energy/Strategy/Fossil » : quelles stratégies énergétiques pour sortir des ressources fossiles (sans trop affecter la biodiversité)?

8.1. Interprétation

L'extrait suivant campe bien le contexte de cette triade dans notre problématique initiale, questionnant les tensions existantes entre développement des énergies renouvelables et sauvegarde de la biodiversité lors de l'arrêt des énergies fossiles.

> "*18 juin 2019 — Shift to renewable energy could have biodiversity cost, ... the long-awaited move to a fossil fuel-free economy will not come without its ... a day ago*"

Le fil de discussion suivant met bien en relation les trois termes à travers la question de la stratégie énergétique à suivre (arrêt immédiat ou progressif des énergies fossiles, substitution via nucléaire ou renouvelable seul)

> "*No country on Earth is taking the 2 degree climate target ... � Futurology �*

> *comments · no_c...Oct 6, 2016 · Edit: about the serious economic decline from immediately halting fossil ... and that nuclear power and renewable energy are our only hope.*"

8.2. Analyse des différences

D'un point de vue sémantique, l'association de ce « thème » n° 2 à la requête initiale semble faire tout à fait sens. Il nous reste à analyser comment Google Search a pu ne pas accéder à cette association (et pourquoi Reddit y est parvenu).

9. Interprétation des « différences » via le cluster 3 « Climate_change/Crisis/Human »: la responsabilité humaine dans la crise du changement climatique

En descendant dans les clusters et leur « représentativité » statistique nous avons de moins en moins de publications de fil de discussion correspondants aux mots clefs des triades. Ainsi, la recherche du mot clé « human » fournit seulement 3 publications, dont la suivante :

> "*It's too late. We can't save Earth. : r/ecology - ecology · comments · its_too...Oct 23, 2019 ·* ***Biodiversity*** *loss, pollution (water, air, and light), massive habitat loss. And all of that is due to* ***humans****, even without* ***climate change****...*"

Le mot « crisis » apparaît en d'autres occurrences associé à « climate change » dans les verbatim, faisant que le sens global approché de la triade peut être approché par: *la responsabilité humaines dans la crise du changement climatique*.

Ces trois clusters achèvent la liste de ceux qui représentent plus de 10% du corpus, et en deçà duquel les occurrences sont si faibles qu'elles rendent l'interprétation, via les verbatim, hasardeuse.

L'interprétation de ces trois clusters en des thèmes principaux constitue, sur le plan de la méthode suivie, les principaux résultats de de cette étape de notre étude.

Nous allons maintenant passer à la section de discussion de ces résultats, en tentant de trouver quelques hypothèses interprétatives pouvant venir expliquer ces résultats

9.1. Hypothèses explicatives des « différences » de résultats et des interprétations trouvées et discussion

9.1.1. Hypothèses explicatives

La section précédente a montré qu'il y a bien des distinctions importantes entre les résultats restitués par la plateforme Reddit comparés à ceux de Google Search pour la même requête de notre questionnement initial. Et ce, tant au niveau quantitatif (25% des lemmes sont spécifiques au corpus de résultat Reddit) qu'au niveau qualitatif (ces lemmes spécifiques forment bien des clusters et des thèmes principaux qui font sens et qui paraissent a priori pertinent pour le questionnement).

Nous devons maintenant tenter de comprendre pourquoi de telles dissemblances entre ces deux plateformes « informationnelles » de médiation sociale existent, quelles causes sont-elles susceptibles d'avoir, et s'il est possible de détecter un sens transversal à ces dissemblances ?

De prime abord, l'on serait tenté d'expliquer ces dissemblances par l'intensité de la contribution sociale activée dans telle ou telle plateforme. En effet, Reddit, comme plateforme de curation sociale, vise précisément à *encapaciter* (*enable* en anglais) les utilisateurs dans leurs activités contributives de curation, consistant à identifier, commenter et pointer des ressources sur le Web. Cette plateforme joue son rôle de médiation sociale à un niveau élevé. Google Search, à l'opposé, ne solliciterait pas la moindre *activité* (consciente) de la part de ses utilisateurs.

Pourtant, cette dernière plateforme de recherche réalise également, de *facto,* des opérations de crowdsourcing (captation de traces laissées par les internautes) auprès de ses utilisateurs et des internautes et ce au moins de deux manières. La première, désignée souvent par l'appellation *travailleurs du clic* (Jones, 2021) se réalise lorsqu'un utilisateur ayant entré une requête dans le moteur, clique sur tel ou tel lien hypertexte de la page de résultats qui s'offre à lui. L'algorithme de ranking de Google capitalise sur ces successions de clic sélectifs pour améliorer ses performances en termes d'adéquation entre une requête et les réponses.

Pourrait être opposé cependant l'argument selon lequel l'activité *d'indexation* du Web par contre, est contrairement à Reddit, l'œuvre exclusive d'un robot logiciel, et non pas de contributions sociales. Cette vision serait pourtant superficielle car l'on sait aussi que l'automate *crawler* de Google parcourt le Web pour l'indexer en sautant de lien en lien hypertextuel. Or, ces liens sont bien des ouvrages réalisés par des êtres humains, qui y expriment des associations entre pages web et, se faisant entre lexiques voire entre concepts. L'utilisation centrale de ces liens hypertextuels pour alimenter l'algorithme de la plateforme Google constitue une forme de crowdsourcing, puisant

donc également dans les contributions des internautes. En mettant en relation internautes cherchant des ressources et internautes en fournissant, via l'architecture socio-cognitive du Web des réseaux de liens hypertextuels (Ghitalla, 2021), Google se pose donc également comme plateforme de médiation sociale.

Les deux plateformes ont donc un recours *central*, bien que différent, au crowdsourcing - ou contributivité des multitudes - pour délivrer leurs services informationnels. Les processus à l'œuvre, explicites ou sous-jacents, dans la production de services informationnels par les deux plateformes sont donc de nature *socio-cognitive*, dans le sens ici ou la production d'information - ou plutôt de connaissances - procède indissociablement de processus sociaux *et* cognitifs. Ainsi, dans les forums de Reddit, les connaissances qui se développent sont le fruit direct d'interactions sociales actives et conscientes (discussions entre humains). Dans le moteur de recherche, nous l'avons vu, les contributions effectuées par les *travailleurs du clic* sont non seulement plus passives et peu conscientes mais surtout il n'y a pas d'interaction sociale directe explicite (deux *click workers* n'ont pas d'interactions sociales *directes* et conscientes entre eux)

9.1.2. L'hypothèse interprétative des schèmes interactionnistes de la cognition sociale

Nous formulons l'hypothèse explicative des dissemblances de résultats entre Google et Reddit en mobilisant la théorie de la cognition sociale et, plus précisément, l'hypothèse des schèmes interactionnistes de la cognition sociale. Certains spécialistes de la didactique des langues ont en effet depuis longtemps souligné les liens - parfois orientés - entre interactions sociales et apprentissages (Mondada & Pekarek Doehler, 2000). Des recherches plus récentes en sciences cognitives ont montré qu'il existait de véritables schèmes interactionnistes de la cognition sociale et que ceux-ci pouvaient donner lieu à expérimentations, dans le cadre du paradigme de l'enaction (Lenay, 2008)

Notre hypothèse d'explication des dissimilitudes entres les résultats fournis par Google et par Reddit est la suivante : l'intensité des interactions sociales en jeu dans les processus de cognition sociale mobilisés pour parvenir à une adéquation entre *questions* et *réponses* varie fortement d'une plateforme à l'autre, Google présentant une intensité faible et Reddit, forte.

Peut-on maintenant repérer, au niveau des contenus respectifs présentés par les deux plateformes, des régularités dans les dissemblances ? L'exercice est ici hautement spéculatif mais il semblerait, à partir de l'échantillon (minimal et non représentatif) que nous avons collecté, que les champs sémantiques renvoyés par Google soient systématiquement plus restreints dans leur extension que ceux renvoyés par Reddit.

Certes, Google parvient (comme Reddit) à faire le lien entre notre problématique *Biodiversité & Energies Renouvelables (aspects économiques)* et la question à l'origine

de cette problématique (et qui la *subsume*) : le *réchauffement climatique*.

Mais Google ne fait jamais le lien entre les trois dimensions de la crise contemporaine systémique (dont on peut pourtant constater la présence sur le web): écologique (dérèglement, climatique et chute de la biodiversité) socio-politique (montée des populismes & nationalismes) et économique (crise du modèle de la croissance sans fin) (Laurent, 2015). La plate-forme Reddit, elle, fait explicitement ce lien. Ou peut-être plutôt, elle note les liens qui sont faits en ce sens par ses contributeurs (*redditors)*. Cette différence est particulièrement perceptible avec le lemme et la problématique de *l'effondrement*. En effet, la terminologie même de l'effondrement suggère une interdépendance entre les crises, selon les trois dimensions précitées. Nous avons d'ailleurs constaté dans les verbatim (de Reddit) un usage du mot *effondrement* plus souvent axé sur les phénomènes humains (effondrement de la taille de la population par ex.) que sur les phénomènes biologiques, les deux étant souvent associés.

Tout se passe comme si Reddit disposait, de fait, d'une forme *d'état du monde*, lorsqu'il cherche à apparier une question (requête) avec des réponses (fournies par lui), alors que Google n'en disposerait pas, ou alors bien plus restreinte. Cette hypothèse interprétative peut paraître paradoxale quand on sait que Google investit effectivement pour construire (via le web3) un tel *état du monde* structuré et explicite (évoquant la quête d'une ontologie générale), alors que Reddit ne le fait pas.

Cet *état du monde* de facto, présent chez Reddit, serait en fait une forme de propriété émergente, pragmatique, de processus interactionnistes animant la cognition sociale, processus nécessaire à la mise en intelligibilité du Web.

Il pourrait cependant être objecté que lorsque Google ne mentionne pas les deux autres dimensions de la crise contemporaine, il ne fait que faire preuve de *précision* (au sens de la recherche documentaire)(Hudon, 2013) s'en tenant - quasi littéralement - à la question posée. Reddit, elle, ferait preuve *d'absence de précision* en générant donc du *bruit* (informations non pertinentes) (op.cit) en donnant une extension excessive à la question.

Il est en fait difficile de trancher dans l'absolu si tels résultats de recherche sont plus pertinents (sémantiquement) que tels autres. Restons-en pour l'instant à constater que des différences significatives existent entre les résultats des deux plateformes, et qu'elles semblent être structurelles (non aléatoire).

10. Discussion

La taille de notre échantillon peut bien sûr être discutée. Surtout si l'on considère qu'il aurait fallu explorer de nombreux questionnements différents (requêtes) sur plusieurs domaines de connaissances pour pouvoir établir des régularités dans les différences de

restitution des deux plateformes.

Une critique fréquemment rencontrée concernant ces travaux en *méthode mixte* (mêlant travaux qualitatif et quantitatif) porte sur la scientificité de l'approche (Morgan, 2007). Une première réticence peut s'exprimer ainsi vis à vis du travail *d'édition* de la carte lexicale que nous réalisons, en supprimant des lemmes, ou en en fusionnant certains, et en en gardant d'autres. Également des réserves peuvent être formulées en matière de scientificité concernant la subjectivité des interprétations de clusters statistiques (op. cit.).

En effet, dans une épistémologie classique des sciences, les sujets et objets de la recherche sont radicalement séparés (Firode, 2018). Il ne peut y avoir de subjectivité du chercheur mobilisée dans la modification de l'objet de la recherche (comme nous le faisons pourtant).

Une porte de sortie nous est cependant proposée vers une autre épistémologie, applicable ici: le constructivisme et plus spécifiquement la *théorisation ancrée*. Ici, les « données » et les « catégories » d'analyse se construisent de concert et co-évoluent, jusqu'à ce qu'elles fassent sens pour le chercheur (et qu'il soumette ses interprétations à des pairs)(Creswell et al., 2011; Dowding, 2013).

Concernant nos recherches à venir, nous souhaiterions approfondir, dans cette approche en méthode mixte, les hypothèses interprétatives de dissimilitudes de résultats entre les deux plateformes, formulées dans cet article, en tentant de développer des indicateurs quali-quanti de ces dissimilitudes.

11. Conclusion

Comment des plateformes « informationnelles » peuvent-elles créer ou capter de la valeur (marchande) en créant du sens, en rendant plus intelligible le Web pour des internautes effectuant des recherches ? Car le concept de *fournisseurs de services informationnels* semble réducteur pour appréhender toute la complexité (socio-cognitive) des étapes de la chaîne de valeur des plateformes telles que Reddit ou Google, avec toutes les médiations sociales qu'elles mobilisent.

Nous avons montré comment ces deux plateformes, face à une requête donnée, pouvaient restituer des résultats sensiblement différents, alors même qu'elles procèdent toutes les deux à du *crowdsourcing* (contributivité des multitudes, ou captation de l'intelligence collective).

Peut-on envisager d'attribuer une valeur comparée - d'abord d'usage, puis marchande - de ces appariements réalisés par les plateformes entre requête et résultats restitués ? Nous en somme resté, pour l'instant, à des hypothèses interprétatives qualitatives :

Reddit restituerait des résultats « supérieurs » (en qualité, à savoir en pertinence sémantique pour l'utilisateur) que Google (au moins dans le cas de notre requête) grâce au fait qu'elle stimulerait et capterait davantage des processus de cognition sociale à l'œuvre dans la mise en intelligibilité du Web pour l'utilisateur.

Mais cette hypothèse demanderait bien d'autres travaux pour pouvoir être validée (ou invalidée) dans la lignée des quelques travaux initiés sur l'économie sémantique sur des Forums de discussion (Szoniecky, 2013).

Qu'en est-il alors de la compétitivité comparée des deux plateformes et de leurs captations de valeur marchande associée à leur offre (de nature socio-cognitive) ? Les deux entreprises ont recours à au moins un mécanisme commun : capter et monétiser l'attention (Citton, 2014) par l'exposition à de la publicité (fut-elle très ciblée pour Google, peu pour Reddit)[14]. Mais si Reddit était si supérieur à Google (au moins qualitativement selon notre hypothèse) pour effectuer des recherches pertinentes, pourquoi le premier serait-il estimé à 15 Milliards de Dollars de capitalisation, alors que le second a pu atteindre les 1.500 Milliards, soit 100 fois plus (La Tribune, 2022) ? La diversification - modérée – opérée par le moteur de recherche ne suffit pas à l'expliquer. Par contre, la firme vient effectivement d'être condamnée par le « gendarme » français de la concurrence à une lourde amende pour abus de position dominante sur le marché de la publicité (Autorité de la Concurrence, 2021) ce qui témoigne des obstacles artificiels que la concurrence (dont Reddit) rencontre pour se développer.

Mais ces considérations n'épuisent pas la question et il reste bien des recherches à mener sur la construction de valeur par les plateformes de médiation sociale à l'ère du capitalisme de régime cognitif.

12. Bibliographie

AFP, & Libération. (2021). Eoliennesâ¯: Après les manifestations, le bras de fer judiciaire en baie de Saint-Brieuc. *Libération*. [En ligne] https://www.liberation.fr/environnement/eoliennes-apres-les-manifestations-le-bras-de-fer-judiciaire-en-baie-de-saint-brieuc-20211002_7SGHNRV6V5FFXG4GLZIK3LA5YU/

Autorité de la Concurrence. (2021, juin 7). *L'Autorité de la concurrence sanctionne Google à hauteur de 220 millions d'euros pour avoir favorisé ses propres services dans le secteur de la publicité en ligne*. Autorité de la concurrence. [En ligne] https://www.autoritedelaconcurrence.fr/fr/article/lautorite-de-la-concurrence-sanctionne-google-hauteur-de-220-millions-deuros-pour-avoir

Avenier, M. (2011). Les paradigmes épistémologiques constructivistesâ¯: Post-modernisme ou pragmatismeâ¯? *Management & Avenir*, *43*, 372-391.

Belanger, A. (2023, janvier 20). *Supreme Court allows Reddit mods to anonymously defend Section 230*. Ars Technica. [En ligne] https://arstechnica.com/tech-policy/2023/01/supreme-court-allows-reddit-mods-to-anony

mously-defend-section-230/
Bouvron, M. (2022). *Reddit, ou la chronique programmée de la mort de Google Search*. [En ligne] https://fr.linkedin.com/pulse/reddit-o%C3%B9-la-chronique-programm%C3%A9e-de-mort-google-martin-bouvron-
Center for History and New Media. (2017). *Zotero Quick Start Guide*. [En ligne] http://zotero.org/support/quick_start_guide
Citton, Y. (Éd.). (2014). *L'économie de l'attentionâ¯: Nouvel horizon du capitalismeâ¯?* La Découverte.
Creswell, J. W. (2013). *Research Designâ¯: Qualitative, Quantitative, and Mixed Methods Approaches*. SAGE.
Creswell, J. W., & Clark, V. L. P. (2011). *Designing and Conducting Mixed Methods Research*. SAGE.
Creswell, J. W., Klassen, A. C., Plano Clark, V. L., & Smith, K. C. (2011). Best practices for mixed methods research in the health sciences. *Bethesda (Maryland): National Institutes of Health*, *2013*, 541-545.
Deschryver, N. (2008). *Interaction sociale et expérience d'apprentissage en formation hybride*. 501.
Dowding, D. (2013). *Best Practices for Mixed Methods Research in the Health Sciences John W. Creswell, Ann Carroll Klassen, Vicki L. Plano Clark, Katherine Clegg Smith for the Office of Behavioral and Social Sciences Research; Qualitative Methods Overview Jo Moriarty*. Sage Publications Sage UK: London, England.
Fakis, A., Hilliam, R., Stoneley, H., & Townend, M. (2013). Quantitative Analysis of Qualitative Information From Interviews A Systematic Literature Review. *Journal of Mixed Methods Research.* [En ligne] https://doi.org/10.1177/1558689813495111
Firode, A. (2018). La critique de l'épistémologie classique et ses implications pédagogiques chez John Dewey et Karl Popper. *Recherches en éducation*, *34*.[En ligne] https://doi.org/10.4000/ree.1943
Ghitalla, F. (2021). Qu'est-ce que la cartographie du webâ¯?â¯: Expéditions scientifiques dans l'univers des données numériques et des réseaux. In D. Boullier & M. Jacomy (Éds.), *Qu'est-ce que la cartographie du webâ¯? â¯: Expéditions scientifiques dans l'univers des données numériques et des réseaux*. OpenEdition Press. [En ligne] http://books.openedition.org/oep/15358
Guillemette, F., & Luckerhoff, J. (2009). L'induction en méthodologie de la théorisation enracinée (MTE). *Recherches qualitatives*, *28*(2), 3-20.
Hesse-Biber, S. (2010). Emerging Methodologies and Methods Practices in the Field of Mixed Methods Research. *Qualitative Inquiry*, *16*(6), 415-418. [En ligne] https://doi.org/10.1177/1077800410364607
Hudon, M. (2013). *Analyse et représentation documentairesâ¯: Introduction à l'indexation, à la classification et à la condensation des documents*. Presses de l'Université du Québec.
Jeanneret, Y. (2014). *Critique de la trivialité*. [En ligne] https://editions-non-standard.com/books/critiquedelatrivialite
Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed Methods Researchâ¯: A Research Paradigm Whose Time Has Come. *Educational Researcher*, *33*(7), 14-26. [En ligne]

https://doi.org/10.3102/0013189X033007014
Jollivet, P. (2022). *Biodiversité vs Energies Renouvelablesâ¯? Structures des débats économiques in Les leviers d'intégration de la biodiversité dans les projets d'énergies renouvelables, Office Français de la Biodiversité (OFB)(commanditaire) /PriceWaterhouseCooper (PWC) (coordinateur)/Institut de la Transition Environnementale (ITE)(partenaire). Rapport intermédiaire. Dans le cadre du projet LIFE BTP « Biodiversité intégrée dans les territoires et les politiques », soutenu par le programme LIFE de l'Union européenne.*
Jones, P. (2021). *Work Without the Workerâ¯: Labour in the Age of Platform Capitalism*. Verso Books.
La Tribune. (2022, novembre 16). *GAFAMâ¯: 1.500 milliards de dollars de valorisation partis en fumée en un mois*. La Tribune. [En ligne] https://www.latribune.fr/technos-medias/internet/gafam-1-500-milliards-de-dollars-de-valorisation-partis-en-fumee-en-un-mois-940861.html
laofmoonster. (2012, décembre 1). *Has anyone else noticed how poorly Google indexes Reddit? How does this affect Reddit?* [Reddit Post]. [En ligne]. www.reddit.com/r/TheoryOfReddit/comments/143m3k/has_anyone_else_noticed_how_poorly_google_indexes/
Laurent, É. (2015). La social-écologieâ¯: Une perspective théorique et empirique. *Revue française des affaires sociales*, *1-2*, 125-143. [En ligne] https://doi.org/10.3917/rfas.151.0125
Le Figaro, & AFP. (2017). Google détaille ses actions contre les contenus « terroristes ». *LEFIGARO*. [En ligne] https://www.lefigaro.fr/flash-eco/2017/06/19/97002-20170619FILWWW00327-google-detaille-ses-actions-contre-les-contenus-terroristes.php
Lenay, C. (2006). Enaction, externalisme et suppléance perceptive. *Intellectica. Revue de l'Association pour la Recherche Cognitive*, *43*(1), 27-52. [En ligne] https://doi.org/10.3406/intel.2006.1326
Lenay, C. (2008). Enaction et techniqueâ¯: Approche expérimentale minimaliste. In B. Stiegler (Éd.), *Le design de nos existences à l'époque de l'innovation ascendante* (p. 211-237). Mille et une nuits.
Macia, L. (2015). Using Clustering as a Toolâ¯: Mixed Methods in Qualitative Data Analysis. *The Qualitative Report*. [En ligne] https://doi.org/10.46743/2160-3715/2015.2201
Méliani, V. (2013). Choisir l'analyse par théorisation ancréeâ¯: Illustration des apports et des limites de la méthode. *Recherches qualitatives*, *Hors Sérieâ¯: Du SIngulier à l'universel*(15), 435-452.
Mondada, L., & Pekarek Doehler, S. (2000). Interaction sociale et cognition situéeâ¯: Quels modèles pour la recherche sur l'acquisition des languesâ¯? *Acquisition et interaction en langue étrangère*, *12*, Article 12. [En ligne] https://doi.org/10.4000/aile.947
Morgan, D. L. (2007). Paradigms Lost and Pragmatism Regained Methodological Implications of Combining Qualitative and Quantitative Methods. *Journal of Mixed Methods Research*, *1*(1), 48-76. [En ligne] https://doi.org/10.1177/2345678906292462
Moulier Boutang, Y. (2007). Le capitalisme cognitif. *Comprendre la Nouvelle Grande*

Transformation et ses Enjeux.
P. Jollivet. (2018). Has “Collective Bargaining” evolved through timeâ¯? An experiment with text-network analysis and topic modeling with big data from Reddit. *Proceedings of ECOMOD2018 - International Conference on Economic Modeling (Fondazione Università Ca’ Foscari Venezia)*, 30.
Paillé, P. (1994). L’analyse par théorisation ancrée. *Cahiers de recherche sociologique*, *23*, 147. [En ligne] https://doi.org/10.7202/1002253ar
Paranyushkin, D. (2019). InfraNodusâ¯: Generating insight using text network analysis. *The world wide web conference*, 3584-3589.
Piluso, N. (2013). Keynes et les synthèses néoclassico-keynésiennes :. Les raisons d’un divorce analytique. *Idées économiques et sociales*, *174*(4), 51-60. [En ligne] https://doi.org/10.3917/idee.174.0051
Raoux, A., Tecchio, S., Pezy, J.-P., Lassalle, G., Degraer, S., Wilhelmsson, D., Cachera, M., Ernande, B., Le Guen, C., Haraldsson, M., Grangeré, K., Le Loc’h, F., Dauvin, J.-C., & Niquil, N. (2017). Benthic and fish aggregation inside an offshore wind farmâ¯: Which effects on the trophic web functioning? *Ecological Indicators*, *72*, 33-46. [En ligne] https://doi.org/10.1016/j.ecolind.2016.07.037
Szoniecky, S. (2013). *Dictionnaires de catégories pour la génération automatique de proverbesâ¯: Vers une économie sémantique de l’interprétation*.
Teddlie, C., & Tashakkori, A. (2008). *Foundations of Mixed Methods Researchâ¯: Integrating Quantitative And Qualitative Approaches In The Social And Behavioral Sciences*. SAGE Publications Inc.

14. Annexes

14.1. Annexe 1 - Échantillon, représentativité statistique et édition du corpus/de la carte

Ce travail ne prétend pas à la représentativité statistique *stricto sensu* (ne serait-ce que du fait que nous n’avons extrait que les 150 premiers résultats respectifs de nos deux plateformes « informationnelles », pour un total de plus de 30 millions de résultats pour Google et 175 pages « affichables »). Ce corpus est susceptible par contre ce se faire l’écho de pratiques des internautes. L’effectif de 25 résultats par pages étant le format par défaut de Google sur des navigateurs, 150 résultats représentent alors 6 pages, alors que l’immense majorité des internautes ne dépassent pas la première page de leurs résultats affichés.

Le travail d’édition effectué sur les réseaux textuels (et donc sur les corpus), dans une approche constructiviste, peut être regroupé en trois catégorie :

- Édition (suppression) de scories directement liées aux imperfections des techniques de scraping utilisées (récolte automatique de données). Une bonne

partie est constituée de métadonnées. Ainsi, le mois « jun. » sera retiré car il s'agit d'une mention de la date du post et non pas d'une date dans le post. Idem pour « cité » (tant de fois…)

- Des termes trop généraux dans le langage courant. Ex : important.
- Des termes trop généraux pour notre recherche spécifique (ex : sustainability).

14.2. Annexe 2 – Carte-réseau des résultats de Google : principales statistiques d'analyse structurale de réseau (40 premiers lemmes)

node name	degree	frequency	betweenness	topic	conductivity	locality	diversivity
sum total	**NaN**	**834**	**3.109078**	**n/a**	**3652.50**	**445**	**5285.10**
sum / 221 nodes	**NaN**	**3.77**	**0.014068**	**n/a**	**16.53**	**2.01**	**23.91**
climate ch.	58	0	0.519719	4	89.6	1	0
circular	25	19	0.174917	4	70	0	92.1
loss	17	15	0.171412	4	100.8	0	114.3
offshore	22	11	0.147772	1	67.2	0	134.3
carbon	16	8	0.107297	11	67.1	0	134.1
wind	14	10	0.102032	2	72.9	0	102
water	25	11	0.096245	1	38.5	1	87.5
land	15	6	0.071792	1	47.9	0	119.7
emission	10	7	0.065136	11	65.1	0	93.1
policy	15	9	0.055102	8	36.7	0	61.2
service	12	5	0.051373	3	42.8	0	102.7
finance	9	6	0.049274	10	54.7	0	82.1
eu	12	12	0.047466	8	39.6	0	39.6
essential	12	7	0.045469	1	37.9	0	65
strategy	15	9	0.044228	8	29.5	1	49.1
ecosystem	10	13	0.043144	3	43.1	0	33.2
power	7	7	0.034568	2	49.4	0	49.4
natural	12	7	0.031674	3	26.4	0	45.2
juin	10	9	0.030663	3	30.7	0	34.1
protection	11	5	0.030176	1	27.4	0	60.4
building	10	7	0.029247	1	29.2	0	41.8
gas	4	3	0.027379	11	68.4	0	91.3
rbc	3	2	0.027248	11	90.8	0	136.2
life	8	4	0.025721	2	32.2	0	64.3
crisis	5	4	0.025273	4	50.5	0	63.2
risk	10	6	0.024951	1	25	0	41.6
support	8	5	0.024101	3	30.1	0	48.2
decoupling	3	2	0.022655	2	75.5	0	113.3
framework	8	3	0.022144	8	27.7	0	73.8
plant	6	6	0.020710	2	34.5	0	34.5
net	8	5	0.019262	11	24.1	0	38.5
solar	7	7	0.019262	2	27.5	0	27.5
job	7	6	0.019035	5	27.2	0	31.7
enhance	3	2	0.018999	4	63.3	0	95
national	5	3	0.018637	1	37.3	0	62.1
home	5	5	0.018341	4	36.7	0	36.7
goal	11	5	0.018297	4	16.6	1	36.6
corporate	4	2	0.018253	11	45.6	0	91.3
investment	8	4	0.017639	6	22	0	44.1

14.3. Annexe 3 – Carte-réseau des résultats de Reddit : principales statistiques d'analyse structurale de réseau (40 premiers lemmes)

node name	degree	frequency	betweenness	topic	conductivity	locality	diversivity
sum total	**NaN**	**892**	**3.194135**	**n/a**	**3361.30**	**285**	**5234.70**
sum / 219 nodes	**NaN**	**4.07**	**0.014585**	**n/a**	**15.35**	**1.30**	**23.90**
climate	47	45	0.437420	1	93.1	1	97.2
change	34	30	0.284475	1	83.7	0	94.8
work	43	28	0.240766	9	56	1	86
collapse	31	38	0.140840	5	45.4	1	37.1
company	21	13	0.122941	6	58.5	0	94.6
futurology	23	19	0.120626	7	52.4	0	63.5
crisis	12	8	0.087412	1	72.8	0	109.3
political	16	8	0.080356	5	50.2	0	100.4
power	12	11	0.069804	10	58.2	0	63.5
current	9	4	0.064318	1	71.5	0	160.8
loss	13	8	0.056670	8	43.6	0	70.8
growth	16	8	0.050367	4	31.5	1	63
oil	9	6	0.045910	5	51	0	76.5
carbon	11	9	0.045507	4	41.4	0	50.6
farm	4	3	0.040471	3	101.2	0	134.9
cultural	5	3	0.037983	0	76	0	126.6
capitalism	8	4	0.037843	9	47.3	0	94.6
nuclear	7	9	0.036837	10	52.6	0	40.9
forest	10	7	0.036677	10	36.7	0	52.4
manage	9	6	0.035687	6	39.7	0	59.5
production	11	5	0.035044	4	31.9	0	70.1
robot	8	3	0.034709	7	43.4	0	115.7
future	7	5	0.033477	9	47.8	0	67
water	9	4	0.027658	8	30.7	0	69.1
cost	8	5	0.026297	1	32.9	0	52.6
increase	6	5	0.025453	9	42.4	0	50.9
environment	7	3	0.024903	1	35.6	0	83
job	9	4	0.023923	7	26.6	0	59.8
people	11	9	0.023489	16	21.4	1	26.1
isn	8	3	0.021202	7	26.5	0	70.7
positive	9	3	0.019455	4	21.6	0	64.8
fuel	6	3	0.019068	4	31.8	0	63.6
electricity	3	2	0.018901	3	63	0	94.5
avoid	7	3	0.018841	7	26.9	0	62.8
askscience	5	8	0.018785	4	37.6	0	23.5
profit	3	2	0.017860	3	59.5	0	89.3
investment	15	6	0.017307	4	11.5	1	28.8
anxiety	4	2	0.016642	1	41.6	0	83.2
human	15	8	0.016310	5	10.9	1	20.4

14.4. Annexe 4 – Différence entre cartes Reddit/Google :

principales statistiques d'analyse structurale de réseau (40 premiers lemmes)

node name	degree	frequency	betweenness	topic	conductivity	locality	diversivity
sum total	NaN	2081	2.146790	n/a	883.00	4666	1013.20
sum / 250 nodes	NaN	8.32	0.008587	n/a	3.53	18.66	4.05
energy	190	290	0.798645	2	42	5	27.5
climate ch.	95	0	0.390541	6	41.1	2	0
business	37	26	0.065028	5	17.6	2	25
collapse	31	38	0.055576	5	17.9	2	14.6
work	52	31	0.054803	5	10.5	5	17.7
management	36	28	0.048289	9	13.4	3	17.2
power	20	18	0.046213	3	23.1	1	25.7
offshore	20	11	0.035066	9	17.5	1	31.9
carbon	21	17	0.030332	1	14.4	1	17.8
crisis	19	12	0.029793	6	15.7	1	24.8
plant	10	7	0.029254	7	29.3	0	41.8
job	12	10	0.026566	8	22.1	1	26.6
human	16	10	0.023802	6	14.9	1	23.8
loss	21	23	0.022986	6	10.9	2	10
water	30	15	0.022333	9	7.4	4	14.9
australia	9	9	0.020145	0	22.4	0	22.4
strategy	13	11	0.016031	2	12.3	1	14.6
country	16	12	0.015182	8	9.5	2	12.7
wind	15	12	0.014929	3	10	2	12.4
fossil	10	8	0.014888	2	14.9	1	18.6
nuclear	10	9	0.013158	3	13.2	1	14.6
investment	23	10	0.012635	2	5.5	4	12.6
vegan	3	4	0.011417	8	38.1	0	28.5
production	15	9	0.011172	2	7.4	2	12.4
credit	12	4	0.011027	2	9.2	1	27.6
science	11	20	0.009721	4	8.8	1	4.9
company	27	18	0.009221	0	3.4	8	5.1
political	13	9	0.008538	5	6.6	2	9.5
mar	8	10	0.008473	10	10.6	1	8.5
model	10	6	0.008103	5	8.1	1	13.5
developing	6	6	0.008097	8	13.5	0	13.5
market	6	5	0.008065	9	13.4	0	16.1
infrastructure	4	5	0.008065	2	20.2	0	16.1
finance	9	7	0.008032	7	8.9	1	11.5
part	10	9	0.008032	6	8	1	8.9
plan	9	8	0.007967	6	8.9	1	10
target	8	6	0.007950	6	9.9	1	13.3
war	3	3	0.007901	7	26.3	0	26.3
cost	9	8	0.007836	1	8.7	1	9.8

13. Notes

[1] Les corpus sont transformés en graphes de lemmes selon la méthode double suivante (Paranyushkin, 2019): selon la co-occurrence de lemmes dans une phrase donnée (ou autre entité textuelle choisie), à savoir à l'intérieur d'une ligne de la table textuelle ; et selon la co-occurrence de lemmes entre deux phrases quelconques appartenant au corpus, à savoir entre deux lignes différentes de la table textuelle.

[2] Nous avons utilisé, dans des publications antérieures, le logiciel open source Iramuteq, qui possède des fonctionnalités d'analyse linguistique plus riches qu'InfraNodus, avec un positionnement plus proche de l'analyse du contenu, davantage de capacités de traitement informatique, mais avec moins de facilités exploratoires (comme le zoom avant et arrière, de la macro visualisation des corpus à la lecture in extenso des verbatim).

[3] C'est au cours d'un tel processus que nous avons pu identifier, grâce au contexte des verbatim, trois sens différents du lemme Work présentant une importance interprétative pour notre travail (et non pas dans l'absolu). Nous pourrions alors dans une démarche plus systématique - pertinente pour de plus gros corpus textuel - étiqueter de manière semi-automatique chaque occurrence du lemme Work par son trait sémantique, pour enfin pouvoir générer une nouvelle version de notre carte (enrichie par nous) dans laquelle les différents sens du lemme apparaîtraient alors différenciés. Cette nouvelle carte pourrait faire apparaître non seulement de nouveaux clusters statistiques (agrégats) mais également, potentiellement, de nouveau thèmes principaux porteurs d'un sens nouveau.

[4] Dans nos premiers travaux de recherches portant sur le contenu du Web, via des moteurs, y compris privés, nous introduisions pléthores de mots-clés pour être certain de ne rien rater dans la collecte. L'expérience nous a montré que les différents dispositifs utilisés, le moteur de recherche lui-même et la structure sous-jacente du Web, réalisaient eux-mêmes, la plupart du temps, ces liens implicites (que dans un premier temps nous voulions expliciter).

[5] Ce module de Google recherche permet d'effectuer des recherches via le moteur exclusivement sur un site cible donné, qui est ici, bien sûr Reddit.

[6] Google procède ainsi à un déroutant agencement d'extraits de résultats parmi les différents fils de discussion lui semblant pertinents, trouvés sur Reddit. Son dispositif de recherche/restitution ressemble à une forme de synthèse avec citation.

[7] Le choix de la méthode de clustering utilisé ici par le logiciel (Paranyushkin, 2011, 2019) mériterait d'être discuté. Il l'est en partie dans l'article synthétique de Baray (Baray et al., 2022). Il mentionne que l'algorithme ForceAtlas s'apparente à l'algorithme de Louvain, en étalant cependant généralement davantage les résultats de la spatialisation. Nous ne pouvons faire état, à notre niveau, d'une seule petite expérimentation de comparaison de résultats en fonction de différents logiciels. Sur une base de données textuelles issues de discussions sur Reddit constituées par nos soins (P. Jollivet, 2018), nous avions eu l'heureuse surprise de constater que le nombre et la « qualité » des clusters générés par InfraNodus et par Gephi étaient quasiment identiques (algorithme de ForceAtlas 2) (Bastian et al., 2009). Ceci ne constitue aucunement une

preuve bien sûr.

[8] L'indicateur statistique d'intermédiarité structurale, dans un réseau, identifie un nœud qui fait office de « pont » entre 2 ou plusieurs clusters. En sociologie, dans un réseau social, il peut être interprété comme passeur.

[9] Nous ne validons pas intégralement cette expression (« sémantique »), mais nous nous en servons pour apprécier une forme de répartition ou de concentration du réseau lexical, incluant les poids de fréquence et de co fréquence des lemmes.

[10] Nous avons été cependant surpris par la sensibilité du dispositif Google Search à effectuer, (apparemment) cette opération de généralisation, selon le nombre de résultats que nous intégrions dans notre corpus de réponse : en ne sélectionnant que les 60 premiers résultats de notre recherche, via Google , le lemme « changement climatique » n'apparaissait presque plus.

[11] La traduction technique du terme future en français n'est pas tout à fait triviale. En effet, en économie, le champ de la prospective se traduit souvent en future studies (et non pas forcément en prospective) . Ainsi, autant, en français, la coupure entre prospective et futurologie est assez nette, elle semble l'être moins en anglais. D'où, peut être, la co-présence, en anglais, des lemmes future et futurology dans nos résultats.

[12] Ce résultat n'est pas tout à fait trivial. Lors de notre précédente comparaison, certes déséquilibrée dans les effectifs respectifs (50/150), le corpus de résultats via Google ne présentait pas l'expression réchauffement climatique , au contraire de Reddit, pour la même requête.

[13] La robustesse de notre interprétation aurait pu être améliorée en étudiant non seulement les résultats spécifiques à Reddit (i.e. qui sont dans Reddit en n'étant pas dans Google) mais aussi les résultats spécifiques à Google . Les contraintes d'espaces ne permettaient pas d'en faire état ici.

[14] Reddit fait partie de ses rares plateformes commerciales qui présentent un design socio-technique de privacy by design au moins sur les adresses IP (qu'elle ne collecte pas).