La cartographie des traces textuelles comme méthodologie d'enquête en sciences sociales

Jean-Philippe Cointet

Jean-Philippe Cointet

La cartographie des traces textuelles comme méthodologie d'enquête en sciences sociales

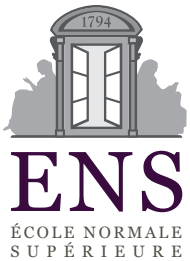
Mémoire de synthèse en vue de l'obtention de l'Habilitation à Diriger des Recherches

Rapporteurs:

- Katherine Stovel (Prof. University of Washington)
- Eric Fleury (Prof. ENS Lyon)
- Dominique Boullier (Prof. EPFL)

Jury:

- Thierry Poibeau (DR CNRS, ENS Paris), Garant
- Dominique Boullier (Prof. EPFL)
- Florence Maillochon (DR CNRS - ENS Paris)
- Dominique Cardon (Prof. Sciences Po)
- Alberto Cambrosio (Prof. Mc Gill)
- Jean-Daniel Fekete (DR INRIA - Saclay)



Remerciements

J'adresse en premier lieu mes remerciements à Thierry Poibeau, qui a accepté de se porter garant de ce mémoire. Des rades de la Gare du Nord au Comptoir du Relais, son soutien aura été constant. Merci Thierry pour tes suggestions au plus près du texte, sans pitié pour ma syntaxe chancelante mais toujours pleines d'esprit. Merci de m'avoir pointé du doigt l'autre Wittgenstein et merci pour ta bienveillance.

J'aimerais également remercier et dans un même temps présenter mes excuses auprès de l'ensemble des membres du jury pour les avoir fait patienter si longtemps alors que je repoussais régulièrement la date de soutenance puis de rendu de ce manuscrit alors même que je l'avais moi-même fixée. Merci en particulier aux rapporteurs Katherine Stovel, Eric Fleury et Dominique Boullier d'avoir accepté cette responsabilité sans hésiter. J'adresse également toute ma gratitude aux autres membres du jury, à mes collègues au long cours Dominique Cardon et Alberto Cambrosio qui me supportent et soutiennent depuis bientôt 10 ans (je tiens à remercier spécialement et publiquement Alberto qui a accepté de braver son anxiété des transports aériens pour assister à la soutenance), merci à Florence Maillochon d'avoir fait ce pas de côté, merci à Jean-Daniel Fekete de se rendre à nouveau disponible pour moi alors que je sais son emploi du temps très chargé.

Bientôt huit ans depuis la soutenance de ma thèse et la rédaction de mes derniers remerciements officiels, et que de rencontres et de chemin parcouru ! Comme pour tout voyage, certains lieux s'impriment dans la mémoire, même si je serais bien en peine d'en placer certains sur une carte. Saint-Pierre-des-Corps, Torun, Bois l'Etang, Trélou, San Diego, et à chaque lieu ses visages, les familiers que l'on reverra avec le plaisir de l'habitude, les plus fugaces que l'on retrouvera tôt ou tard.

On aimerait ne pas avoir à ordonner, faire le tri, mais il serait trop simple de se soustraire à l'exercice. Je commencerai donc à adresser toute ma gratitude à Marc Barbier, la personne avec qui j'aurai probablement passé le plus de temps en réunion (et ce en dépit de l'adoption à l'unanimité des méthodes radicales du *lean manangement*), merci Marc pour ta confiance et la liberté que tu m'as offerte, merci aussi de m'avoir appris le souci du collectif. Je salue aussi la patience de tous mes collègues de la plateforme durant ces sept années de travail en commun et leur abnégation alors que la salle serveur coulait et que météor prenait feu. Voilà pour le cercle resserré de l'équipe de CorText, j'espère que ce mémoire sera aussi l'occasion de poser un regard rétrospectif sur le chemin accompli ensemble pour envisager le futur sous les meilleurs auspices alors que je m'apprête à poursuivre l'aventure à Sciences Po. Je remercie aussi collectivement les membres du LISIS (feu SenS (feu TSV)) et en profite pour saluer Pierre-Benoît Joly qui a su garder le cap et maintenir la même ambition malgré

les re-configurations institutionnelles. J'en profite également pour remercier sincèrement François Houllier dont le soutien sans équivoque à l'INRA a été précieux durant toutes ces années.

J'ai aussi énormément appris au contact des nombreux étudiants avec qui j'ai eu la chance de travailler à commencer par Elisa Omodei et Antoine Mazières dont j'ai eu le privilège de co-diriger le travail de thèse. Mais il y a aussi Benjamin, Ian, Nicolas, Gabriel, etc. Ce sont ces expériences qui donnent tout son sens à la rédaction d'une habilitation.

Il y a enfin les collègues les plus proches, on ose d'ailleurs à peine les qualifier de collègues sauf avec une pointe d'accent marseillais. Mentionner leur nom sonne comme une évidence, ça n'enlève rien à leur mérite. C'est sans doute grâce à Sylvain Parasie que l'idée de ce mémoire a germé. Depuis il n'a eu de cesse d'entretenir la fragile pousse. Mais c'est surtout lui qui, à force de patience et de pédagogie, m'a donné goût au raisonnement sociologique. Andrëi Mogoutov tient évidemment une place particulière de par son intelligence incandescente et sa différence à tous égards - merci de partager encore tes découvertes avec moi. Il y a déjà plus de trois ans, je rencontrai Alix Rule et Peter Bearman à New York. C'est peu de le dire, le nouveau monde a eu des vertus très revigorante pour ma curiosité scientifique. Mais c'est avant tout leur générosité qui a été déterminante. Je peux maintenant réciter les 45 présidents américains sur le bout des doigts mais parle toujours aussi mal anglais. Merci à eux de continuer à ne pas m'en tenir rigueur.

Résumé

Ce mémoire se situe au cœur de la zone d'échange où se rencontrent big data et sciences sociales. Par « big data » on fait référence à la double transition que constituent d'une part la profusion de traces numériques, souvent produites en ligne, qui permettent de tracer les comportements individuels à des résolutions et des échelles inédites, et d'autre part le développement de nouvelles formes d'analyse de données inspirées des algorithmes d'apprentissage automatique. Nous nous concentrerons sur une pratique de l'analyse de données très particulière en sciences sociales : l'analyse automatique de contenu. Ce mémoire débute par un regard rétrospectif sur l'histoire de ces méthodes. Nous nous efforçons de décrire les opérations matricielles en jeu dans les méthodes factorielles, de revenir sur les hypothèses sociologiques de l'analyse par mots associés, de restituer le travail d'enquête que permet Prospero, etc. Une typologie commune est proposée pour distinguer ces approches en fonction des théories sociologiques qu'elles embarquent, des stratégies de modélisation de l'énonciation qu'elles adoptent et des modes de calculs et d'intelligibilité du social qu'elles permettent. À travers cette même grille, des approches plus récentes nées dans les mondes de l'informatique et de l'intelligence artificielle sont analysées : notamment topic models, et plongements de mots. Nous défendons enfin la cartographie de réseau comme une méthode à part entière qui est systématiquement comparée aux autres approches. Le dernier chapitre est l'occasion d'examiner la façon dont le web modifie la pratique de l'enquête empirique de corpus textuels. Comment les notions de locuteurs, d'énonciation et plus généralement l'épistémologie même de l'enquête en sciences sociales est-elle déplacée avec les traces numériques ? Entre analyse critique historique et description méthodologique, ce mémoire original est également traversé de nombreuses références à des projets empiriques menés durant ces huit dernières années qui illustrent la diversité de la pratique de l'analyse de corpus.

Abstract

This manuscript is situated in the trading zone where big data and social sciences meet. In this instance, "big data" refers to two inter-twined transformations. One transformation is the wealth of digital traces, often produced online, that allow us to trace individual behaviors at scales and resolutions never seen before. The other transformation is the blossoming of news analytical tools inspired by machine learning. We will focus on a very particular kind of data analysis in social sciences, namely automatic content analysis. This research report starts with a retrospective look at the history of content analysis methods for social sciences. The matrix operations of factorial methods are detailed, sociological assumptions underlying co-word analysis are discussed, the practice of sociological investigation with Prospero is described, and so on. A general typology is introduced to distinguish these approaches in terms of the sociological theories that they build on such as their strategies for modeling the enunciation, and modes of calculation and intelligibility of the social they open. Using the same grid, the more recent approaches of artificial intelligence and computer sciences are analyzed : in particular, topic modeling and word embedding. In the second chapter, network mapping is defended as a method in its own right, and systematically compared with the other approaches. The last chapter is an opportunity to examine how digital traces produced online is likely to change the way empirical investigation of textual corpora is lead by social scientists. How the notions of speakers, enunciation and more broadly the very epistemology of social science practice is shifted with the advent digital traces ? Mixing historical critical analysis and methodological description, this original manuscript is also populated with numerous references to empirical projects carried out during the last eight years, that illustrate the diversity of the practice of corpus analysis.

Table des Matières

Introduction

Ce mémoire se situe au cœur de la « zone d'échange »[1] qui fait se rencontrer big data et sciences sociales. Par « big data » on souhaite ici faire référence à deux transitions contemporaines majeures. D'une part la profusion de données numériques, souvent produites en ligne, permet de tracer les comportements individuels à des résolutions inédites (par la taille des populations concernées, la précision des inscriptions, et la variété des types d'action qu'elles indexent). D'autre part le développement de nouvelles formes d'analyse de données comme les méthodes d'apprentissage automatique, fortement influencées par les mondes de l'ingénierie, modifient en profondeur la nature des connaissances produites.

Alors que certains prophétisaient il y a huit ans déjà l'avénement de « sciences sociales computationnelles » (Lazer et al., 2009), quelle est la situation actuelle ? Il semble de prime abord qu'à l'enthousiasme premier a rapidement succédé un certain scepticisme voire une position de repli par rapport à des méthodes et des données qui sont mal ajustées aux questions traditionnelles des sciences sociales (Boyd et Crawford, 2011). Pour autant les expérimentations sur et avec ces données numériques qu'elles soient le fait de sociologues ou de modélisateurs[2], renvoient à un paysage loin d'être manichéen où s'hybrident de nouvelles façons d'enquêter sur le social.

Certains voient même dans les traces numériques l'occasion de renouveler l'édifice théorique des sciences sociales. Pour Latour et al. (2012), les possibilités d'enquête sur le web permettent enfin de circuler d'entités individuelles à collectives, sans faire appel à des transformations ontologiques complexes comme le changement d'échelle mais simplement en suivant des liens hypertextes ou en listant les attributs des profils de chaque entité. Sur un tout autre mode, mais toujours en mobilisant Tarde, Lee et Martin (2015) soulignent également la façon dont les big data permettent de tenir compte de toutes les singularités plutôt que d'avoir à « liquéfier chaque individualité dans une soupe homogène »[3].

1. Le concept de trading zone est évidemment emprunté par McFarland, Lewis, et Goldberg (2015) à Gallison.

2. Il faudra préciser ce que ces sociologues et modélisateurs représentent, en termes d'épistémologie et de domaines de recherche, mais mieux vaut entretenir le flou pour ne pas tout de suite réduire la discussion à des querelles de chapelles.

3. dans la version originale : « we liquefied everyone into a homogenous soup despite how much they varied individually ».

De façon paradoxale et en contradiction partielle avec le point de vue précédent, il est souvent difficile de caractériser les acteurs sur le web. Les actions des individus (quand ce ne sont pas des robots !) sont enregistrées avec toujours plus de précision. Mais ces acteurs semblent comme flotter dans un espace indéfini, manquant cruellement d'épaisseur sociale (Bowker, 2014; Boellstorff, 2013). A ces critiques, l'apprentissage automatique est prompt à répondre en montrant comment l'état civil des utilisateurs de médias sociaux peut être prédit depuis leurs seules traces d'activité (Kosinski et al., 2013; Sloan et al., 2015), comme si les catégories classiques des sciences sociales n'étaient plus qu'une propriété émergente parmi d'autres. A ceux qui s'étonnent du manque de profondeur historique de données délivrées sous forme de flux ininterrompus, et sitôt menacées de péremption, on leur objecte qu'il faut adapter les concepts de relations et de structures sociales à une logique d'événements (de Nooy, 2015).

C'est que ce monde de traces, tout entier fait de corrélations[4], interroge directement l'épistémologie de la méthode sociologique. A nouveau, le débat est extrêmement ouvert opposant entre autres lignes les promoteurs d'une reconfiguration des sciences sociales qui épouse le modèle de la résolution de problème cher aux ingénieurs (Watts, 2017b) et des chercheurs qui voient dans les big data l'occasion pour les sciences sociales de se dégager du carcan du raisonnement hypothético-déductif à condition que celles-ci soient utilisées pour générer de nouvelles théories (Goldberg, 2015). Pour autant les sciences sociales sont elles solubles dans une épistémologie purement inductive que les progrès récents réalisés en apprentissage automatique ont propulsé sur le devant de la scène (Manning, 2016) ? Est-on entré dans une ère où les théories n'importent plus ou les nouvelles capacités exploratoires des big data sont elles l'occasion de développer une pratique des sciences sociales plus abductive (Timmermans et Tavory, 2012; Kitchin, 2014) ?

4. Certains n'hésitent d'ailleurs pas à les qualifier de douteuses (Calude et Longo, 2016).

Ce mémoire vise donc à discuter ce moment particulier que certains analysent comme l'entrée des sciences sociales dans une nouvelle ère (Boullier, 2015). Néanmoins afin d'interroger la façon dont les données du web chamboulent la pratique et l'épistémologie des sciences sociales, nous nous concentrons sur un certain type d'analyse : l'analyse automatique des données textuelles. En dépit de cette réduction, le paysage est encore immense. Les sciences humaines et sociales qu'elles s'appuient sur des archives, des entretiens ou des sondages ont toujours entretenu un rapport étroit au texte pour établir des inférences sur le monde social à des niveaux de profondeurs très variés : des normes tacites qui gouvernent les situations de communication aux cadres collectifs qui structurent un débat public (Evans et Aceves, 2016). Mais dans un contexte où les sciences sociales sont sommées de se plier (Venturini et al., 2014b) à l'urgence du flot de traces générées sur Internet, il est utile de prendre un peu de recul et de se souvenir du chemin parcouru. Et

l'histoire de l'analyse de corpus textuels est riche d'enseignements de ce point de vue. La sociologie n'est en effet pas entièrement naïve quant à l'analyse de corpus textuels pour enquêter sur le social. C'est pourquoi nous passerons un certain temps à décrire des méthodes d'analyse textuelle passées dont on interrogera la pertinence dans ce nouveau monde de traces. Ce regard rétrospectif est aussi utile pour apprécier les différentes configurations qui se sont mises en place en leur temps alors que des outils et des théories venues d'autres domaines (la linguistique, les statistiques) interpellaient déjà la sociologie.

Mais ce manuscrit doit avant tout être lu comme une exploration de ces méthodes d'analyse en situation : confrontées à un matériau empirique particulier et au service d'une question de recherche précise. Il est conçu comme une pérégrination le long d'une série d'expérimentations que nous avons menées ces dernières années et qui ont soulevé des questions de méthodologie de tout ordre : comment coder des commentaires publics en ligne ? comment concilier un modèle géométrique du sens sans adopter un point de vue atomiste sur les individus ? Ce mémoire fait donc référence à de nombreux travaux, menés depuis la fin de ma thèse et réalisés alors que j'étais chercheur au LISIS [5] avec des collègues à Marne la Vallée et ailleurs. Ce manuscrit, rendant compte de projets finalisés, avortés, ou encore en cours, témoigne ouvertement d'une recherche en marche ouverte aux expérimentations. Un exposé honnête de mon expérience personnelle de la pratique de l'analyse de données en sciences sociales offre sans doute la meilleure illustration possible des tensions qui ont brièvement été soulevées précédemment. Inscrite au sein de projets de recherche inter-disciplinaires, elle s'appuie toujours sur un contexte empirique et théorique précis, achoppe sur des problèmes techniques, mais échoue aussi parfois à convaincre pour des raisons théoriques. C'est à travers ces exemples que l'on peut mieux saisir la variété de ses formes et des horizons qu'elle ouvre. C'est aussi grâce et à travers l'expérimentation collective et le design d'une infrastructure de recherche comme CorText [6] qu'émergent et se stabilisent progressivement des modalités de travail de la sociologie avec les traces textuelles.

5. Pour être précis, au gré des fusions, et autres déménagements, j'ai successivement été membre des laboratoires TSV (Transformations Sociales liées au Vivant, SenS (Sciences en Société) et depuis le 1er janvier 2015 LISIS (Laboratoire Interdisciplinaire Sciences Innovations Sociétés).

6. CorText est une application en ligne dédiée à l'analyse de corpus textuels développée au sein de l'IFRIS à laquelle j'ai largement contribué ces sept dernières années.

Ce document se décompose en trois courts chapitres : (i) dans un premier temps, on adopte une distance historique et critique pour décrire la façon dont les sciences sociales ont mobilisé l'analyse textuelle dans le passé et ce que les méthodes actuelles, souvent issues du monde de l'intelligence artificielle ou de l'informatique, peuvent leur offrir, (ii) partant de cette typologie, le second chapitre vise à présenter et mettre en relief les avantages et limites de l'approche cartographique que nous défendons (iii), enfin, on interroge dans le dernier chapitre la façon dont les données du web et le type de traitement qu'elles appellent reconfigurent la nature de l'enquête sociologique voire du social lui même.

1

Panorama des méthodes d'analyse textuelle en sciences sociales

Dans ce chapitre, on se propose d'établir un panorama des méthodes quantitatives développées et/ou employées par les sciences humaines et sociales depuis un demi-siècle pour l'analyse de corpus de textes. Pour chacune d'entre elles on s'efforcera d'apporter une description technique de leurs fondements mathématiques mais aussi de décrire les modèles de langage auxquels elles font appel. Enfin et surtout on interrogera le type de connaissance qu'elles sont susceptibles d'apporter en sciences sociales. Cette prise de distance historique et critique vise à mettre en perspective l'analyse des traces textuelles par la cartographie de réseau que nous décrirons et dont nous défendrons les avantages plus particulièrement dans la deuxième partie de ce mémoire. Mais ce panorama vise également à comparer des approches qui, de par les mondes et les périodes où elles ont circulé, ne se sont que rarement confrontées les unes aux autres. Cette comparaison qui traverse disciplines et décennies sera aussi l'occasion de constater la variété des hypothèses linguistiques posées, et d'interroger les ressemblances entre formalismes mathématiques qu'elles mobilisent, tout en précisant la variété des champs de recherche en sciences humaines et sociales dans lesquelles elles s'inscrivent. On conclura en proposant une typologie générale des méthodes que nous avons couvertes.

Avant d'énumérer les méthodes que l'on examinera individuellement, il semble opportun de distinguer d'ores et déjà deux grandes familles de stratégies. D'un côté, les méthodes nées avant les années 2000 telles que l'analyse des correspondances, la socio-informatique des controverses, ou l'analyse de mots-associés qui sont souvent indissociables d'une théorie du social bien définie et qui ont d'ailleurs toutes été créées ou co-construites avec des sociologues. De l'autre côté, on dénombre des méthodes plus récentes, contemporaines des

« big data », provenant des mondes de la physique ou de l'informatique et souvent moins bien connues en sciences sociales telles que les « topic models » ou les modèles encore plus récents de plongement de mots.

1.1 *De Benzécri à Prospero, les écoles françaises de l'analyse de corpus textuels*

Nous tâcherons dans cette première section de décrire le panorama des méthodes automatiques d'analyse de corpus nées en France avant les années 2000. Nous mettons volontairement de côté le paysage anglo-saxon, sans doute par méconnaissance, mais aussi car il nous semble mieux cadré théoriquement. Mentionnons simplement le champ de l'analyse de contenu dont le nom, « content analysis », remonte à 1941 (Berelson et Lazarsfeld, 1948) et qui se définit à ses origines comme :

> « *a research technique for the objective, systematic, and quantitative description of the manifest content of communication.* » (Berelson, 1952, p. 18) [1]

1. « une technique de recherche pour la description objective, systématique et quantitative du contenu explicite de la communication. »

Fortement marquée par l'analyse de la presse, elle s'est progressivement enrichie conceptuellement (notamment avec l'émergence d'une analyse de de contenu « qualitative ») et diversifiée dans ses sources jusqu'à intégrer des contenus non verbaux (Krippendorff, 2004). La plupart des logiciels d'analyse de contenu textuel qui se sont développés outre-atlantique (NVivo et ATLAS.ti en sont les deux principaux représentants) appartiennent à la famille des CAQDAS [2] dont l'objectif premier est d'accompagner la recherche qualitative (et portant donc souvent sur des données d'entretien) dans des opérations de codage relevant de la *Grounded Theory* (Glaser et Strauss, 1967) [3].

2. Analyse Qualitative de Données Assistée par Ordinateur - « Computer Assisted Qualitative Data AnalysiS »

3. La mise en garde initiale était justifiée : Grounded theory et analyse de contenu qualitatif sont loin d'être entièrement équivalents (Cho et Lee, 2014).

Même en nous restreignant au seul paysage français, la profusion d'options et surtout de postures épistémologiques et d'orientations théoriques est telle qu'elle fait à certains « l'effet d'un maquis dans lequel il est difficile de se repérer et de s'orienter » (Demazière et al., 2006). Néanmoins, nous pourrons constater dans les sections suivantes que ces approches partagent parfois des formalismes mathématiques ou des principes d'analyse en réalité très proches.

La sociologie française a énormément expérimenté avec les méthodes d'analyse de contenu textuel entre les années 60 et 90 bien avant la mode du text-mining (Beaudoin, 2016). D'abord avec Jean-Paul Benzécri dont l'analyse des correspondances offre un espace géométrique dans lequel les sociologues peuvent librement projeter des individus, des variables socio-démographiques classiques telles que la profession, l'âge, ou le sexe, mais aussi des mots ou des modalités d'expression lorsque du contenu textuel soumis à analyse (section 1.1.3). Un peu plus tard, dans les années 90, Max Reinert associe à l'analyse des correspondances de Benzécri une théorie de la langue bien

spécifique dans le logiciel Alceste qui ambitionne de donner aux sociologues un moyen de révéler les « mondes lexicaux » qui structurent les corpus textuels. La méthode Alceste, qui relève de l'approche lexicométrique, est encore employée dans un certain nombre de travaux en sciences sociales, sa diffusion étant maintenant assurée par le logiciel Iramuteq (section 1.1.4). Dans les années 80, Michel Callon posait les bases théoriques de l'analyse par mots-associés[4] pour saisir les dynamiques scientifiques suivant la tradition de la sociologie de la traduction (section 1.1.1). Enfin, nous nous intéresserons au logiciel Prospero conçu par Francis Chateauraynaud avec l'informaticien Jean-Pierre Charriau, à partir des années 90, qui s'inscrit dans le programme de socio-informatique des controverses et embarque des hypothèses très fortes héritées de la sociologie pragmatique (section 1.1.5). Enfin, la section 1.1.2 retrace brièvement l'histoire de l'école de la lexicométrie politique française. Sans égaler l'exhaustivité encyclopédique de Jenny (1997) qui recense 27 logiciels différents (un véritable inventaire à la Prévert : de *Pat-Miroir* à *Coconet* en passant par *Modalis*, *Patate* et autre *Saint-Chef*), nous en retenons néanmoins les grandes catégories.

4. *co-word analysis*

1.1.1 *Analyse par mots-associés*

Si on se réfère à l'article séminal de Callon, Courtial, Turner, et Bauin (1983), les mots-associés offrent une méthode analytique pour répondre à la question suivante : comment caractériser et visualiser les « réseaux de problèmes » que génère l'activité scientifique ? S'appuyant largement sur l'anthropologie de laboratoire (Latour et Woolgar, 2013, 1979) qui décrit l'activité scientifique comme produisant des « inscriptions littéraires » du spectographe de masse à la publication scientifique, l'analyse par mots-associés ne se limite pas à la seule production académique mais considère aussi la littérature grise ou les textes politiques comme matériaux de départ. Elle promet ainsi de « libérer » les dynamiques scientifiques menacées d'être artificiellement clusterisées dans des paradigmes kuhniens purement cognitifs auxquels l'analyse de co-citations (Small, 1973) les condamne.

En s'intéressant aux mots au sein des textes, la méthode des mots-associés vise à restituer la stratégie des auteurs qui pour capturer l'intérêt des ses lecteurs (le fameux *funelling of interest* de John Law) font interagir un certain nombre de notions qui définissent alors un problème. L'analyse par mots-associés est donc par définition dynamique. Elle ambitionne de capturer des processus d'attachement et de détachement entre termes. Elle est également ouverte, au sens où elle permet de lier la recherche fondamentale au paysage socio-politique plus large, les dimensions sociales aux dimensions cognitives dans une même mouvement. Elle propose finalement une méthode pour

mettre en pratique la théorie de l'acteur-réseau (Callon et al., 1986).

En dépit de la circulation incessante des entités décrites par la théorie de l'acteur-réseau, la méthode des mots-associés vise bien à identifier des structures stables qui émergent de la répétition des occurrences d'un mot dans une série de contextes différents : quelles configurations d'équilibre émergent de la façon dont les acteurs associent les mots ou posent des problèmes (Callon et al., 1983) ? Ni internaliste, ni externaliste, la méthode des mots-associés suit l'injonction caractéristique de la théorie de l'acteur-réseau (elle même importée de l'ethnométhodologie) : « suivre les acteurs » (Latour, 2014), ou au moins leur production textuelle, témoin privilégié du travail de négociation permanente des problèmes traités par la science (Callon et Latour, 2006).

L'approche par mots-associés, après avoir bénéficié d'une large diffusion en France mais aussi à l'étranger sous l'impulsion du Centre de Sociologie des Innovations (CSI) aux Mines de Paris et des logiciels Leximappe, Candide, Calliope puis des versions successives de Réseau-lu [5] , a quasiment entièrement disparu dans sa forme originale [6] pour maintenant prendre la forme des fameuses « cartes de liens » sur lesquelles on aura l'occasion de revenir. Pour autant, il est utile de s'attarder sur la pratique historique des mots-associés tant ils semblent différer des usages actuels des réseaux en scientométrie (Wagner et Leydesdorff, 2005; Leydesdorff et Rafols, 2009; Van Eck et Waltman, 2010; Chen, 2006) ou en cartographie des controverses (Venturini, 2012; Marres, 2015a; Bruns et Burgess, 2011; Graeff et al., 2014) [7].

5. La figure 1.1 montre une capture d'écran de ce logiciel dont l'appartenance à la famille des mots-associés mériterait d'être discutée plus longuement.

6. Coulter et al. (1998) en proposent une des applications les plus récentes (sur la recherche en ingénierie logiciel).

7. Sachant que les mots-associés sont nés d'une critique de l'analyse des citations, il est cocasse de les voir maintenant intégrer la parfaite suite logiciel du scientomètre.

Pratiquement, l'analyse par mots-associés se décompose en trois étapes : (i) indexation des « mots-clés » et calcul des indices de proximité, (ii) analyse des clusters ainsi créés, (iii) plongement des clusters au sein d'un diagramme stratégique.

Il faut noter que les éléments des cartes produites par cette méthode étaient constitués initialement de mots-clés (ou d'une sélection de mots-clés, les plus fréquents et les plus rares étant éliminés *a priori* par commodité), les mots-clés présents dans la notice d'une publication scientifique à des fins d'indexation. On comprend bien l'intérêt pragmatique d'un tel choix, mais il semble difficile à tenir théoriquement. Certes, on peut toujours le défendre en considérant que l'activité des documentalistes fait partie intégrante de la même chaîne de traduction et d'inscription littéraire, mais ce sont bien les mots des acteurs scientifiques que l'on souhaite tracer et pas leur interprétation par des tiers usant de méthodologies d'indexation ayant leur objectif propre.

Différents indices étaient utilisés pour mesurer la capacité des mots à s'enrôler les uns les autres. Ainsi, Callon et al. (1986) introduit une première mesure d'association plus tard appelée « indice d'inclusion » qui s'exprime

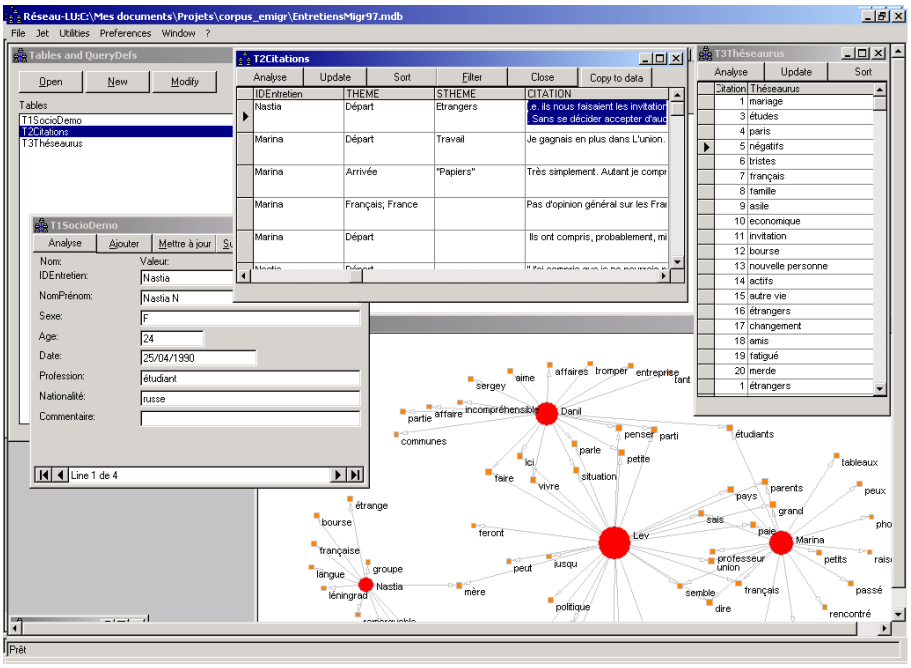


FIGURE 1.1: Capture d'écran du logiciel Reseau-lu, le réseau visualisé est de nature socio-sémantique. La figure est empruntée à (Mogoutov et Vichnevskaia, 2006).

sous la forme suivante :

$$S_{ij}^{inc} = \frac{n_{ij}}{min(n_i, n_j)}$$

où n_{ij} désigne le nombre de documents conjointement indexés par les mots-clés i et j et n_i désigne le nombre d'articles indexés par le mot-clé i. Cette mesure étant asymétrique, elle génère une structure hiérarchique des champs de recherche qui prend la forme d'un arbre, une fois débarrassé de ses branches redondantes, et dans laquelle les mots les plus fréquents se trouvent au sommet (voir figure 1.2).

Une autre mesure introduite dans le même article est l'indice d'association spécifique défini, en conservant les mêmes notations, par l'équation suivante :

$$S_{ij}^{spe} = \frac{n_{ij}}{n_i n_j / n}$$

Contrairement, à l'indice d'inclusion, les mots de faible fréquence ne sont pas « écrasés » par cette mesure, et l'indice d'association permet de construire des réseaux de proximité qui révèlent des connexions existantes entre mots périphériques et relais (de faible fréquence). L'indice d'association compare un nombre de cooccurrences observées au nombre de cooccurrences attendues si les deux mots-clés étaient distribués de façon aléatoire sur l'ensemble des documents. Callon et al. (1983) définissent même le coefficient de spécificité S_{ij}^{spe} comme le rapport entre la probabilité qu'un document déjà indexé par i soit aussi indexé par j (n_{ij}/n_i) divisé par la probabilité qu'un document soit

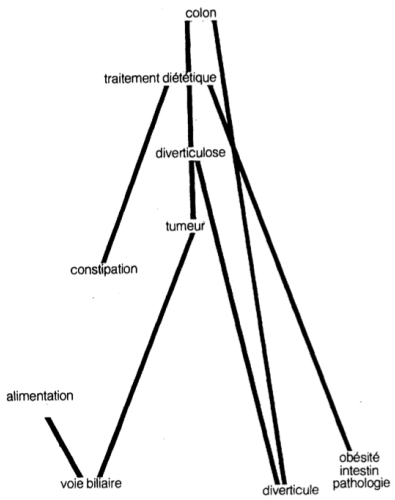


FIGURE 1.2: Détail d'une carte de proximité Leximappe extrait de (Callon et al., 1986). Le « pôle central », au sommet de l'arbre, porte sur les recherches d'un traitement diététique approprié pour les maladies du colon. Les mots-clés à un niveau intermédiaire (comme « tumeur ») sont appelés des « médiateurs » (relay), les mots-clés en bas de l'arbre (comme voie biliaire) sont dits périphériques.

indexé par j (n_j/n). Ce n'est que plus tard (Callon et al., 1991) que l'indice d'équivalence (parfois nommé « e-coefficient ») est introduit :

$$S_{ij}^{eq} = \frac{n_{ij}^2}{n_i n_j / n}$$

et que l'on appelle parfois coefficient d'inclusion mutuelle. Comme le précédent, il permet de calculer des cartes de proximité entre mots.

Quelle que soit la métrique employée, des cartes d'inclusion ou de proximité sont construites à partir de ces mesures et regroupent l'ensembles des associations dont la force est supérieure à un seuil donné. Différentes procédures ont été mises au point pour découper le réseau en sous-réseaux ou clusters pertinents (les cartes d'inclusion étaient réputées bien reconstruire les grands thèmes d'un domaine, les cartes de similarité permettent quant à elles de mieux caractériser les relations entre idées mineures). Outre la créativité des mesures de similarité et des procédures de clustering (qu'on ne décrira pas ici) mises en place, on notera que la méthode des mots-associés s'est rapidement enrichie d'une autre forme de visualisation que celle du réseau. Le « diagramme stratégique » vise à se doter d'une représentation globale de la structure d'un champ scientifique dans un espace bi-dimensionnel dans lequel les clusters sont disposés en fonction de leur densité sur l'axe vertical (importance relative du nombre de liens internes) et de leur centralité (vis-à-vis des autres clusters)[8]. Le diagramme stratégique est particulièrement utile pour saisir la dynamique des clusters qui se déplacent d'un cadran à l'autre. On distingue ainsi les champs centraux et développés en haut à gauche, qui s'opposent aux clusters périphériques et sous-développés du cadran inférieur gauche, etc.

8. Les deux notions sont directement héritées de l'analyse de réseaux sociaux américaines et notamment de Ronald Burt. Par contre, les méthodes de clustering (par exemple modèles de bloc développés à la même époque par Breiger et al. (1975)) ne sont pas repris par la méthode des mots-associés.

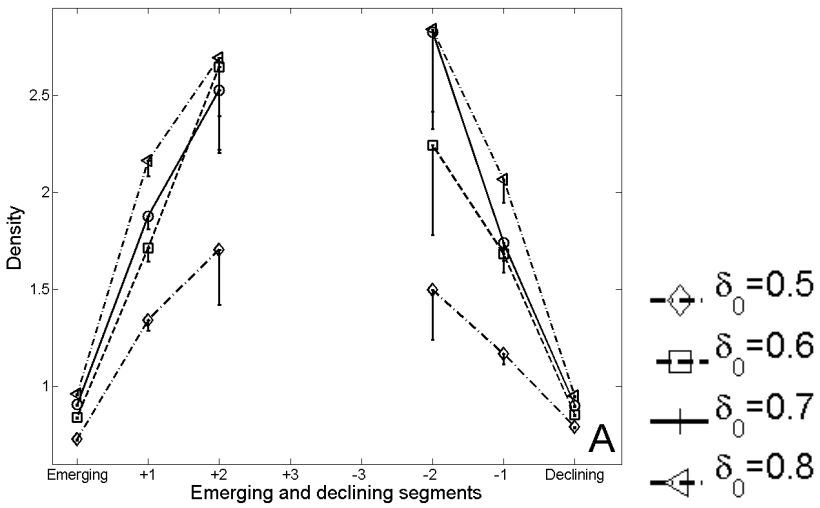


FIGURE 1.3: Corrélation entre la densité normalisée des clusters et leur âge après leur naissance (« Emerging ») ou leur espérance de vie avant disparition (« Declining ») calculée pour différentes valeurs de seuil inter-temporel (δ). On devine la forme prise par le cycle d'un nouveau champ scientifique, encore peu structuré à sa naissance, gagnant en densité au fil des années, et risquant de disparaître lorsque les mots qui le composent se dispersent dans le réseau sémantique.

Partant du constat de la grande mobilité des termes qui se lient et se délient au sein de réseaux hybrides, la méthode tend à positionner des champs de recherche au sein d'un diagramme qui cache largement la richesse des associations sous-jacentes. C'est là tout le paradoxe de la pratique de l'analyse de mots-associés. On pourra objecter que le diagramme stratégique vise justement à identifier les champs en croissance susceptibles de transformer la structure du domaine. Certes, et l'article de Callon et al. (1991) est d'ailleurs remarquable de précision et d'inventivité dans ses multiples tentatives de capturer et décrire les transformations de cluster à l'interface entre recherche publique et R&D à l'œuvre dans la chimie des polymères, mais il n'en demeure pas moins que les dynamiques auxquels ils s'intéressent se déploient sur des échelles qui sont finalement assez compatibles avec les travaux et les hypothèses défendues par les tenants de l'analyse citationnelle (Garfield et al., 1972; Chen et al., 2010; Boyack et Klavans, 2010). À tel point que les deux méthodes après s'être opposées se retrouvent de plus en plus comparées (Leydesdorff et Zaal, 1988; Healey et al., 1986) ou même couplées (Zitt et Bassecoulard, 2006; Zitt et al., 2011). Les méthodes et les usages ont énormément évolué également et même si Shi, Foster, et Evans (2015) reconnaissent explicitement l'influence de la sociologie de la traduction sur leur choix de modélisation, les réseaux hybrides mélangeant auteurs, drogues et pathologies qu'ils construisent n'ont plus grand chose à voir avec les diagrammes stratégiques de Michel Callon et encore moins avec les diagrammes narratifs de Teil et Latour (1995).

Notre travail de reconstruction de la « phylomémie des sciences » (Chavalarias et Cointet, 2013) consiste en une sorte d'écho tardif mais relativement fidèle à la pratique des mots-associés. Certes notre méthode était différente sur de nombreux points : la métrique que nous avions utilisée pour construire le réseau de similarité entre termes, les termes eux-mêmes qui provenaient

d'une indexation du contenu des abstracts, et enfin la nature des clusters détectés qui résultaient d'une analyse structurelle de réseau. Mais c'est en suivant le même objectif de caractérisation des transformations des champs scientifiques (Callon et al., 1991) que nous avions reconstruit des généalogies de clusters de mots. Par la suite, nous avions mesuré leur longévité et leur probabilité de disparition avec l'idée de les corréler à leur densité interne, empruntant au créateur de l'analyse par mots-associés l'une des métriques qui rentraient dans la composition du diagramme stratégique. Nous avons notamment montré, à grande échelle, figure 1.3, combien la force des liens entretenus localement au sein d'un cluster pouvait être un prédicteur efficace de la longévité d'un champ scientifique.

1.1.2 L'école de lexicométrie française

L'école de lexicométrie française est éminemment plurielle. Selon les courants et les époques on a ainsi pu l'appeler : statistique textuelle, analyse de données en linguistique, et plus récemment textométrie ou logométrie (Ollivier, 2010). Nous nous concentrerons spécifiquement dans cette section aux recherches menées à partir des années 60 depuis le laboratoire de « Lexicologie politique » de l'Ecole Normale Supérieur de St. Cloud, en nous appuyant largement sur le travail de reconstruction historique de Loiseau (2016). S'intéressant aux textes et discours politiques, son principe premier est que l'analyse du vocabulaire employé dans les textes permet de révéler les idéologies cachées dans les discours, prolongeant ainsi un programme de recherche débuté par Ferdinand Brunot au début du XX^{ème} siècle.

La théorie sous-jacente sur laquelle s'appuie ce courant, fortement influencé par les transformations politiques de mai 68 et à forte connotation marxiste, peut s'énoncer comme suit. Le langage est fortement politique et le discours est un lieu de luttes idéologiques. Voyant dans le lexique une fenêtre privilégiée pour saisir un ordre économique, social et politique, la lexicométrie fait appel à des méthodologies variées pour analyser le vocabulaire : analyse fréquentielle, analyse des correspondances (voir section suivante 1.1.3), cooccurrences, etc. que l'on retrouve notamment dans le logiciel *Lexicloud* développé sous l'impulsion d'André Salem (et qui est encore utilisé en France par des historiens et politologues sous le nom de LEXICO actuellement distribué dans sa troisième version [9]).

9. http://lexi-co.com

On retiendra donc essentiellement de la « méthode lexicométrique » la croyance forte dans une approche comparative des textes que l'on retrouvera dans de très nombreuses méthodes ultérieures. Le sens des mots, jugé trop instable, importe finalement peu (et il est de tout de façon inconcevable

d'en capturer les subtilités avec une méthode d'analyse automatique dans les années 60). Dès lors, les lexicomètres préfèrent s'attacher à la surface des textes pour énumérer et surtout comparer les fréquences des mots en fonction du locuteur ou de la partie du discours. On peut citer Bonnafous et Tournier (1995) pour qui la recherche lexicométrique est

> « *chargée d'examiner, à partir de corpus de textes soumis à comparaison, comment les termes échangés dans l'espace public autour des enjeux de pouvoir rendent compte des luttes d'appropriation ou de dépossession symboliques qui se jouent dans le milieu de l'échange.* »

Si des écarts de fréquence d'usage de termes entre locuteurs traduisent des différences d'orientation politique, un des enjeux méthodologiques clés de la lexicométrie est d'être capable de détecter et de mesurer ces écarts de fréquence au sein d'un corpus. C'est là que les « spécificités » entrent en jeu. Elles visaient à quantifier la « surprise » qu'il y a d'observer un terme à une certaine fréquence dans une sous-partie d'un corpus. La fréquence observée était alors comparée à une fréquence théorique attendue dont on suppose qu'elle suit une loi hypergéométrique (Lafon, 1980; Salem, 1988; Labbé et Labbé, 1994). Le résultat de la méthode des spécificités se présente sous la forme d'une liste de termes en sur- ou sous- représentation dans un segment du corpus (que l'on considère une partition temporelle ou dépendant du locuteur). Entre autres méthodes nées au sein de cette école, on peut également mentionner la théorie des segments répétés (Salem, 1987) (qui correspond à la recherche de colocation en linguistique), la recherche de rafales (Brunet, 2006) (« bursts » d'occurrences dans un segment de texte qui s'oppose à une distribution plus régulière (rafalité/régularité)).

Parmi les corpus remarquables traités par l'approche lexicométrique, on retrouve l'analyse comparative et historique de discours de campagne électorale (Prost, 1974), la parole syndicale (Launay, 1980), les tracts de mai 68 (Tournier, 2007), les motions du Parti Socialiste au congrès de Metz (Bonnafous, 1983), ou plus récemment aux primaires socialistes de 2011 (Marchand et Ratinaud, 2012), etc. Mais les principes lexicométriques ont également été appliqués en littérature (Brunet, 2009) et en histoire (Lemercier et Claire, 2010; Bonin et Dallo, 2003), plus rarement en sociologie. Ses hypothèses très fortes sur le texte en sont sans doute la cause. Celles-ci ont en effet été très violemment critiquées notamment par les tenants d'une approche pragmatique qui lui reprochent, en excluant de prime abord la question du sens, de réduire le texte à un matériau inerte qui ferait doublement insulte aux locuteurs et à l'analyste (Chateauraynaud, 2003b). En dépit de ces critiques, l'approche comparative et la recherche des spécificités est extrêmement présente dans Prospero (et dans bien d'autres logiciels d'analyse texutelle). Certes, les entités mobilisées sont (re)travaillées de façon beaucoup plus raisonnée par le chercheur, mais l'approche par contraste (qui oppose un corpus à son anti-corpus) reste un ressort important de Prospero qui oppose ainsi des périodes, des

locuteurs, etc. Force est de constater que dans un espace théorique clivé, les méthodes et les procédures statistiques continuent de circuler assez librement. Les méthodes beaucoup plus récentes d' « ideological scaling » (Himelboim et al., 2013; Barberá et al., 2015) *a contrario* semble embrasser les ambitions de la lexicométrie politique, c'est à dire qu'elles visent à identifier de façon automatique dans le texte des idéologies ou accointances politiques sans chercher à réellement comprendre le sens des textes mais en s'appuyant sur des méthodes d'apprentissage.

1.1.3 L'analyse des correspondances

Mais la première des approches à se développer en France, et qui s'est également révélée l'une des plus populaires, est l'analyse des correspondances. Développée par Benzécri (1973) dans les années 70 et trouvant sa source dans les travaux de statitique de Guttman, Hirschfield ou Fisher dans les années 30 et 40 [10], l'analyse des correspondances propose un prolongement de l'analyse en composantes principales à des variables catégorielles, permettant de projeter dans un espace (typiquement bi-dimensionnel) des mots ou des catégories d'individus (profession, âge, etc.).

10. Benzécri lui-même se montre très modeste quant à son rôle dans l'analyse statistique de nuages de points, se contentant de décrire l'apport de la recherche française sur le sujet comme l'élaboration d' « une philosophie statistique nouvelle » (Benzécri, 1976)

L'analyse des correspondances a été très largement employée par Bourdieu et ses collègues dans les années 70 avant de se diffuser plus largement dans le monde anglo-saxon (Phillips, 1995). La Distinction de Bourdieu (1979) est clairement le travail le plus connu s'appuyant en grande partie sur cette méthode (voir reproduction figure 1.4). Il montre une homologie structurelle entre l'espace des styles de vie conçus comme des capitaux culturels, ou économiques et l'espace des positions sociales. L'espace de l'analyse des correspondances devient sous la plume de Bourdieu un véritable champ de forces guidant les acteurs dans le choix du sport qu'ils pratiquent, leurs goûts musicaux, leurs préférences culinaires, etc. Il faut noter tout de suite que Bourdieu (et par la suite ses disciples), s'il usait beaucoup de l'analyse des correspondances pourtant conçue comme une méthode inductive d'analyse des données langagières [11], s'en servait surtout pour analyser des enquêtes par questionnaire dans lesquels les enquêtés étaient interrogés sur leurs goûts avec force détails et nuances.

11. C'est Benzécri lui-même qui l'écrit « C'est principalement en vue de l'étude des langues que nous nous sommes engagés dans l'analyse factorielle des correspondances » (Benzécri, 1973) et encore plus tard comme veille à le rappeler le manuel d'Alceste (Reinert, 1998) : « L'analyse des correspondances a été initialement proposée comme une méthode inductive d'analyse des données linguistiques » (Benzécri, 1981).

L'analyse des correspondances recouvre en réalité trois grandes techniques d'analyse qui sont fortement liées : l'analyse en composantes principales (ACP), l'analyse factorielle des correspondances (AFC) et enfin l'analyse multiple de correspondances (ACM) dont relève la Distinction. Mais avant de rentrer dans les détails techniques du fonctionnement de l'analyse des correspondances, nous souhaitions faire un pas de côté pour nous interroger

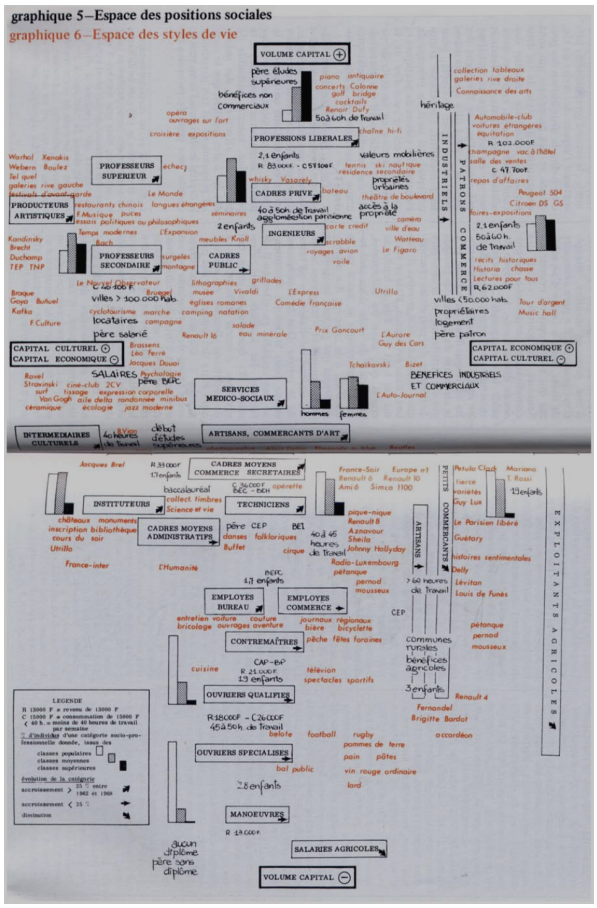


FIGURE 1.4: Espace socio-culturel de La Distinction, photographie empruntée à (Mercklé, 2010) et elle même tirée de (Bourdieu, 1979, p. 140-141) les indicateurs de style de vie sont les variables actives tandis que les statistiques sur les activités professionnelles sont ajoutées comme variables supplémentaires. L'axe horizontal correspond à la composition du capital culturel et économique des individus, l'axe vertical correspond au volume total de capital détenu par les individus (qu'il soit culturel ou économique). De l'avœu même de Bourdieu, la représentation ci-contre n'est pas un véritable « diagramme plan d'analyses des correspondances » mais une reconstruction, un modèle simplifié, fondé sur des données et des résultats partiels obtenus à l'occasion d'autres enquêtes.

sur les questions que posent La Distinction à l'heure des données numériques massives et de l'intelligence artificielle. Ce qui nous intéresse ce n'est pas tant de savoir si les structures identifiées par Bourdieu opèrent toujours, mais plutôt de nous interroger, à un niveau beaucoup plus prosaïque, sur les formes que pourrait prendre un tel programme de recherche alors que des millions d'individus témoignent maintenant « ouvertement » de leur style de vie sur toute une série de plateformes numériques.

C'est une lecture possible du travail que nous avons mené avec Irène Bastard et Dominique Cardon sur la carte des partages sur Facebook créée dans le cadre du projet Algopol [12] . La carte des partages regroupe près de 600 noms de domaines régulièrement partagés par un échantillon de plus de 10000 utilisateurs du réseau social qui ont partagé au total 2 289 765 liens pointant vers près de 139 964 noms de domaine différents. Mais la distribution des fréquences de citations est si hétérogène qu'en se limitant aux 600 domaines principaux, on capture en fait près de 72% du nombre total de partages. Le réseau de co-citations [13] décrit ainsi l'espace des sources d'information ou des partages d'information qui aurait pu figurer dans les questionnaires

12. Une description plus précise du projet et de l'application qui nous a permis de récolter les données sera donnée au dernier chapitre, section 3.2.3.

13. Deux domaines sont liés sur la carte lorsque de nombreux utilisateurs ont régulièrement partagé des urls provenant de ces deux sites. Plus précisément il s'agit d'une mesure indirecte telle que celles décrites dans le second chapitre, section 2.2.3

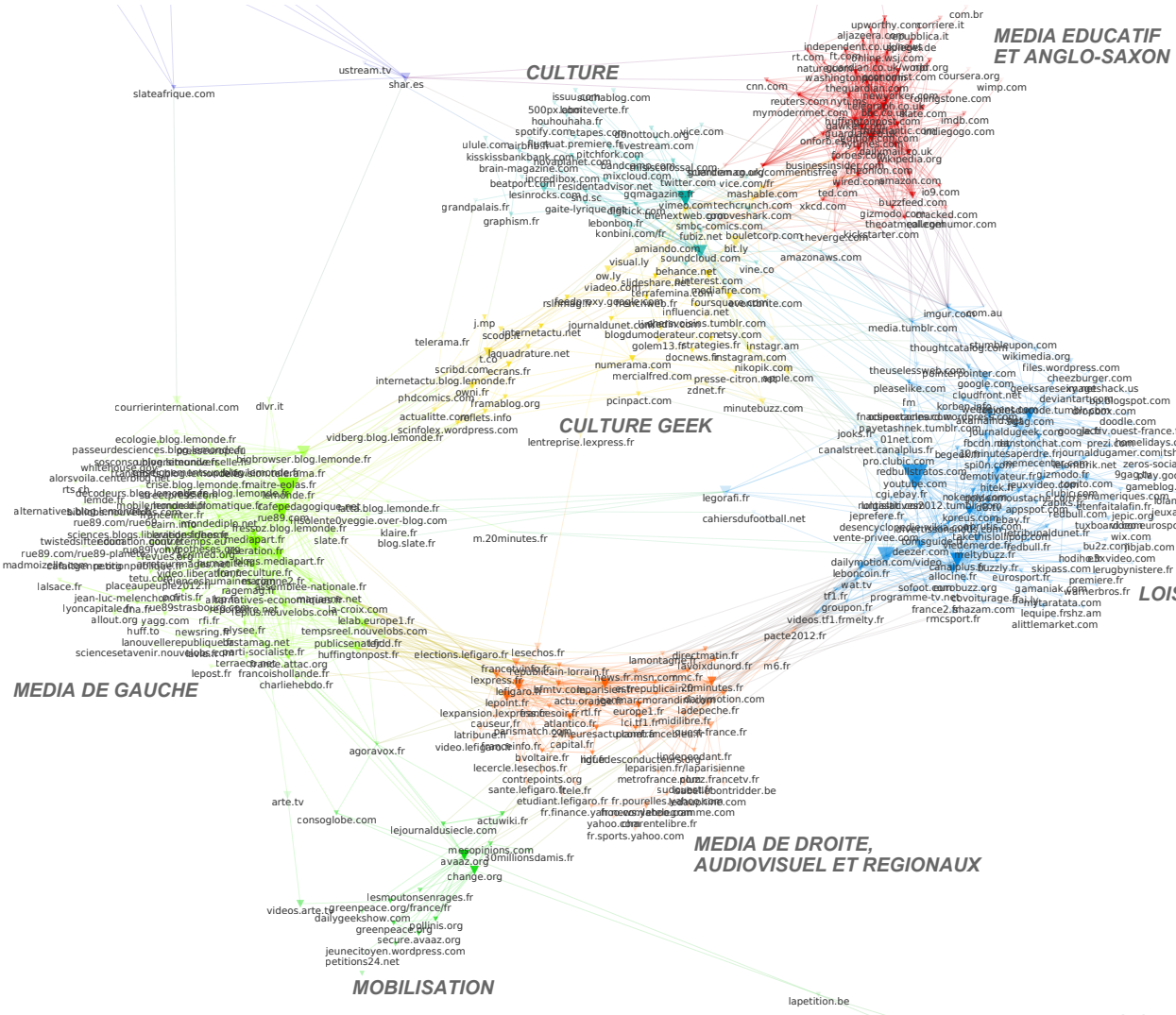


FIGURE 1.5: Carte des partages des enquêtés Algopol. Pour en simplifier la lecture, on n'a représenté que les clusters de partage centraux (les clusters périphériques correspondant à des clusters de sites web étrangers).

14. En terme de lecture de journaux par exemple, Bourdieu montre sans grande surprise combien le niveau d'éducation est fortement corrélé avec la fréquence de lecture et le type de quotidien (qui peuvent être dits d'omnibus, sportifs ou politiques à prétention généraliste).

15. À nouveau, il s'agit bien ici d'une logique d'adressage et non une carte du web classique dans laquelle les liens correspondraient à des liens hypertextes, les sites sont ici liés car co-cités, ce sont des relations liées aux pratiques des internautes.

de la Distinction si l'étude avait été faite 40 ans plus tard [14]. L'espace qui en émerge [15] révèle une structuration finalement très classique (la culture, les loisirs, la politique, etc.) qui permet d'associer à tout un chacun son profil de consommation (il faudrait parler de profil de partage pour être entièrement rigoureux) des sources documentaires sur Internet sans avoir à demander aux enquêtés de répondre à un long formulaire. Dans les faits, les participants devaient tout de même remplir une questionnaire au moment du téléchargement de l'application, mais un questionnaire dont les questions étaient extrêmement simples et portaient uniquement sur leurs données socio-démographiques.

Nous avons produit la carte des partages dont une version limitée aux

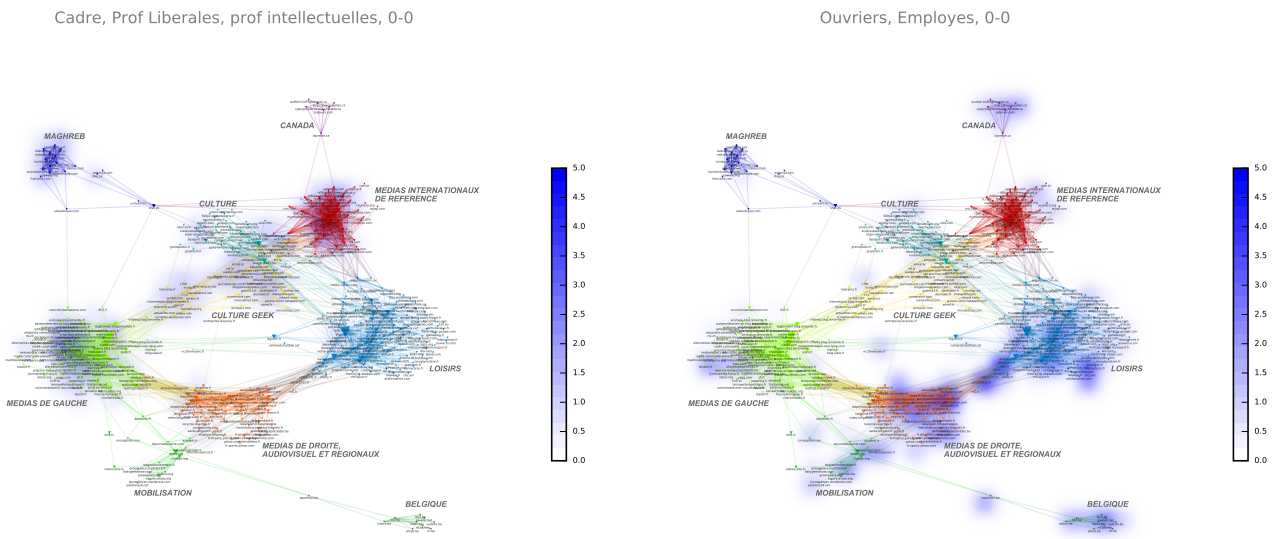


FIGURE 1.6: Carte de co-citations des domaines partagés sur Facebook. Les « Heatmap » représentent les zones de présence préférentielle d'une catégorie d'enquêtés. À gauche, les cadres et professions libérales ou intellectuelles qui se concentrent sur le partage de contenu relevant des médias de gauche, de la culture (geek) et des médias internationaux de référence, à droite, les ouvriers et les employés de notre échantillon d'enquêtés, qui partagent relativement plus de contenus venant de médias de droite et régionaux ou de la sphère des loisirs.)

clusters centraux est présentée figure 1.5 [16]. Les clusters de sites web ainsi formés dessinent une partition naturelle et en définitive très classique des habitudes de consommation des internautes qui, dans notre échantillon, se distribuent entre les médias de gauche (en vert clair - *Le Monde*, *Médiapart*, *Télérama*, *France Culture*, mais aussi quelques blogs (*Maître Eolas*), et sites politiques institutionnels (*Elysee.fr*) ou partisans (*Francoishollande.fr*) etc.), les médias de droite, audiovisuel et régionaux (en orange sur la carte - *BFM TV*, *Direct Matin*, *La Montagne*, *Le Parisien*), les sites relevant de l'univers des loisirs (en bleu), les sources culturelles (en turquoise - *Spotify*, *le Grand Palais*, *les Inrocks*) ou dévolus à la culture geek (en jaune - *numerama*, *ecrans.fr*, *zdnet*), les « médias internationaux » (en rouge - *Coursera*, *IMDB*, mais aussi *TED*, *Al Jazeera* ou *CNN*) et enfin, les sites de mobilisation pour des causes variées (en vert clair - *Avaaz* ou *30 millions d'amis*) [17] .

16. Les clusters périphériques correspondent en réalité à des ressources provenant du web francophone non français : on retrouve ainsi des sites belges, maghrébins ou canadiens. On reviendra dans une section ultérieure 3.3.3 sur ce phénomène de « débordement » des frontières du corpus.

Autre point important, les données récoltées grâce à l'application Algopol permettent de construire le fond de carte en suivant les actions de partage des individus, mais elles permettent aussi de qualifier les individus : soit parce que l'information était déjà renseignée dans leur profil, soit parce qu'elle leur a été (re-)demandée, ou encore parce que ces propriétés relèvent directement de métriques construites par et sur la plateforme. On peut par exemple comparer la distribution des « goûts » en fonction de la catégorie socio-professionnelle des enquêtés (figure 1.6) ou de leur sociabilité (figure 1.7). On s'aperçoit ainsi que les comptes Facebook avec le moins d'amis semblent partager relativement plus de contenu relevant de l'univers de l'associatif et du militantisme

17. On peut se référer à l'analyse qualitative rédigée en ligne par Irène Bastard pour plus de détails : http://algopol.humanum.fr/appresults/le-web-vu-de-facebook/

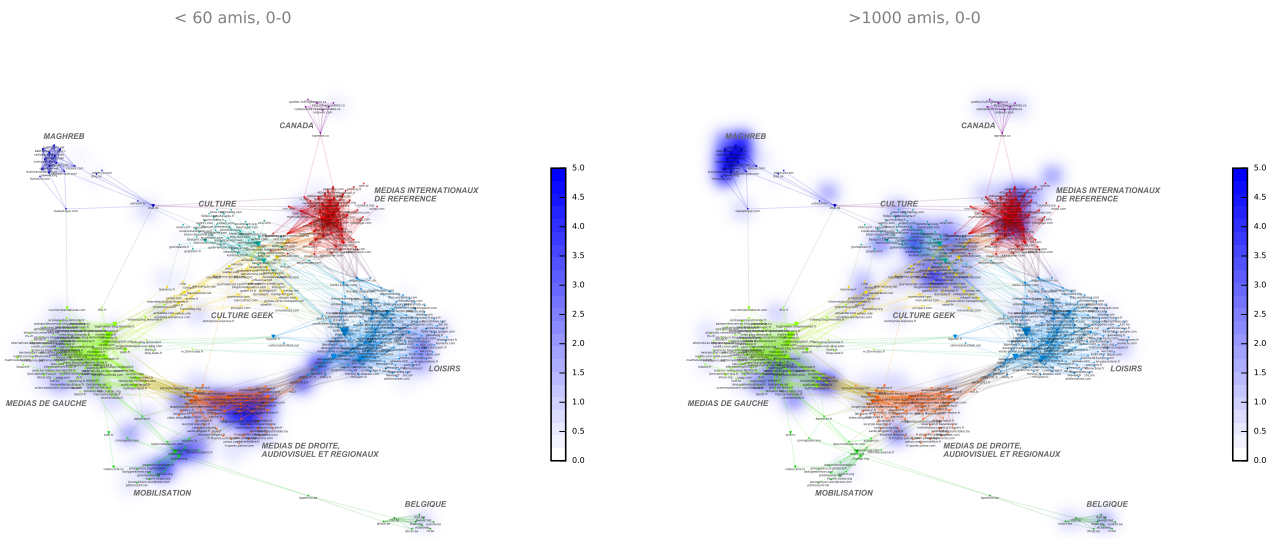


FIGURE 1.7: Ces heatmaps sont obtenues en contrastant les zones de partage préférentiel en fonction de classes d'enquêtés rassemblés par nombre d'amis.

(catégorie assez hétérogène regroupant avaaz.org, greeenpeace.org, mais aussi 30millionsdamis.fr).

Plus important, la méthode mise en œuvre ici découple les catégories de la carte des partages projetées en premier lieu. Il en est de même de l'analyse des correspondances de Bourdieu qui - même si c'est sujet à débat (Mercklé, 2010) - usait de variables décrivant, domaine par domaine, les pratiques et les préférences, comme variables actives, et les positions sociales comme variables supplémentaires qui ne rentrent donc pas dans la composition des facteurs. Mais plutôt que de projeter une catégorie sociale en un point de l'espace, la méthode des heatmaps ici illustrée permet de distribuer une catégorie dans l'ensemble de l'espace des préférences. C'est à dire qu'elle ne réduit pas une catégorie d'individus (ou même un individu unique !) en un point mais l'autorise à être en plusieurs lieux à la fois, avec différents niveaux d'intensité. Certes ce mode de représentation est coûteux au sens où il ne permet pas aisément de multiplier les variables projetées au sein d'un diagramme unique. Il faut finalement prévoir autant de « calques » de heatmaps qu'une modalité donnée a de variables, mais c'est le prix à payer pour éviter cette opération d'abstraction qui voudrait assigner des individus à des positions figées et définitives [18]. Il s'agit en réalité d'une critique plus large qui peut-être adressée à l'analyse des correspondances dans son entier et même aux autres approches « géométriques » [19]. Concevoir une forme d'analyse de données géométrique suppose presque nécessairement d'assigner des objets à des positions. Et

18. Cette option est rarement exploitée mais les méthodes d'analyse des correspondances permettent de construire des « ellipses de confiance » autour des variables qui décrivent leur variabilité selon les différents axes. Mais même entourée d'une ellipse, le modèle assigne alors un centre unique à une variable dont la distribution pourrait tout aussi bien être multi-modale.

19. Même dans le cas des plongements de mots dont on décrira les principaes dans la section 1.2, la prise en compte de la polysémie des mots par exemple implique une déformation profonde du modèle original (Arora et al., 2016; Li et Jurafsky, 2015) au risque d'en perdre l'attrait original qui réside dans l'existence même d'un espace partagé.

ce qu'on gagne en compréhension globale d'un système et de l'articulation de ses variables les unes aux autres, on le perd en compréhension de la variabilité propre à chaque variable. C'est là un constat paradoxal pour une méthode dont la popularité s'est aussi fondée sur la promesse d'un modèle alternatif aux représentations unidimensionnelles jugées trop réductrices dans un contexte post-mai 68 (Desrosières, 2008).

Mais revenons à nos facteurs et précisons le type de transformation en jeu dans les différentes formes d'analyse des correspondances. Dans l'analyse en composantes principales (ACP) tout d'abord, n variables quantitatives sont représentées dans un nouvel espace de dimension généralement réduite (typiquement 2 ou 3). Les nouvelles dimensions de l'espace peuvent alors être décomposées comme des combinaisons linéaires des variables originales. La longueur et la largeur des sépales et des pétales de différentes variétés de fleurs sont ainsi réduites en deux dimensions dans le classique jeu de données d'iris de Fisher (1936). L'objectif est de réduire la dimensionnalité du système en projetant les données sur des axes d'inertie maximale et non corrélés les uns avec les autres. Techniquement, la matrice de variance-covariance construite à partir des observations de départ est diagonalisée dans une base orthonormée qui fournit les axes des facteurs. Avant de poursuivre, signalons que l'ACP est parfaitement équivalente (en tout cas pour la méthode de base appelée Torgerson (Hastie et al., 2011)) aux méthodes de positionnement multidimensionnel (MDS - *MultiDimensional Scaling*) qui réalisent une opération de réduction de dimensionnalité pour représenter une matrice de proximité entre objets.

Plus formellement, une analyse en composantes principales vise à projeter un nuage de points dans un espace de départ à n dimensions dans un nouvel espace à m dimensions ($m \leq n$) de telle manière que l'inertie portée par les premières dimensions du nouvel espace soit maximale, et ce sous contrainte d'indépendance des nouvelles dimensions calculées. Si on considère un ensemble de points pondérés par un poids p_i et une distance entre points que l'on notera δ, alors l'inertie du nuage de points vaut : $I = \sum_i p_i \delta(i, G)^2$. Elle peut également être interprétée comme la somme des variances de l'ensemble des variables étudiées. On peut montrer que le premier axe de projection qui préserve le maximum de l'inertie totale peut-être directement obtenu en diagonalisant la matrice de variance-covariance de la matrice du nuage de points centrés. Cette dernière matrice étant symétrique et positive à coefficient réelle, le théorème spectral garantit l'existence d'une base orthonormée dans laquelle elle peut-être diagonalisée. Les vecteurs propres obtenus correspondent au repère que nous recherchions, et les valeurs propres associées sont proportionnelles à la part de variance expliquée par chacun des axes.

L'analyse factorielle des correspondances prolonge les principes de l'analyse en composantes principales à des variables catégorielles. L'objectif de l'analyse

factorielle est de projeter des données dans un espace de faible dimension tel que les corrélations entre les différentes modalités de deux variables qualitatives soient représentées de la façon la plus fidèle possible. L'analyse des correspondances multiples est simplement une extension de l'analyse factorielle des correspondances à plus de deux variables qualitatives.

Mais essayons de donner une description technique plus précise du fonctionnement de l'analyse factorielle des correspondances. Partant d'un ensemble d'individus caractérisés par deux variables catégorielles - par exemple un ensemble d'individus dont on connaît la couleur de cheveux et des yeux - on construit la table de contingence notée N qui énumère le nombre d'individus caractérisés par un couple de propriétés : combien de personnes ont les yeux bleus et sont blonds, ou sont roux avec les yeux verts, etc. ? Sommer la ligne i de la table fournit le nombre total d'individu dont la première caractéristique vaut i (par exemple le nombre de personnes qui ont les cheveux blonds). La somme de la colonne j indique le nombre total d'individus dont la seconde caractéristique vaut j (yeux marrons). Sommer tous les éléments du tableau fournit le nombre total d'individus.

La table N peut-être normalisée en ligne ou en colonne. On peut en déduire la table des profils-lignes, pour lesquels les coordonnées originales de la matrice N sont divisées par la somme de leur ligne ; ou la table des profils-colonnes, pour laquelle les coordonnées originales de la matrice N sont divisées par la somme de leur colonnes. Plus formellement on construit les tables N^- et $N^|$ dont les coordonnées s'expriment sous la forme suivante : $N^-_{ij} = \frac{N_{ij}}{N_{i\bullet}}$ et $N^|_{ij} = \frac{N_{ij}}{N_{\bullet j}}$ où $N_{i\bullet}$ et $N_{\bullet j}$ correspondent à la somme en ligne et en colonne de la table de contingence originale ($N_{i\bullet} = \sum_k N_{ik}$ et $N_{\bullet j} = \sum_k N_{kj}$). [20]

20. Il est également possible d'exprimer N^- et $N^|$ sous la forme d'un produit matriciel : $N^- = D_r^{-1}N$ et $N^| = ND_c^{-1}$ où D_r (resp. D_c) est la matrice diagonale dont les coordonnées correspondent aux sommes des lignes (resp. des colonnes).

Le principe de l'AFC consiste ensuite à trouver les dimensions pour lesquelles la dispersion de ces deux matrices est maximum. Et pour trouver cet espace, on réalise simplement une analyse en composantes principales sur l'une des deux matrices N^- ou $N^|$ (elles sont équivalentes de ce point de vue). Partons de la matrice des profils-lignes N^-. On se dote de la distance du χ^2 entre deux profils-lignes [21] (autrement dit entre deux modalités de la première variable) i et i' qui s'écrit comme suit :

21. L'un des avantages d'un tel choix est que les catégories peuvent être librement combinées ou re-découpées sans altérer le résultat final. Si par exemple on se contente de regrouper les couleur des yeux en deux catégories : clairs et foncés et non en quatre (bleu, vert, noisette et marron), la distance entre les couleurs de cheveux restent totalement inchangée.

$$d_{\chi^2}(i,i') = \sqrt{\sum_j \frac{1}{N_{\bullet j}}(N^-_{ij} - N^-_{i'j})^2} \tag{1.1}$$

Il faut noter ici que la distance du χ^2, de par sa pondération par $1/N_{\bullet j}$ donne un point plus important aux variables les plus rares. On définit ensuite l'inertie du nuage de points (pondérés par leur poids $N_{i\bullet}$) comme la somme

des distances quadratiques à son barycentre G :

$$I^- = \sum_i N_{i\bullet} d^2_{\chi^2}(i, G^-) \tag{1.2}$$

où G^- désigne donc le barycentre pondéré des profils lignes. Ses coordonnées sont donc :

$$G^-_j = \sum_i N_{i\bullet} N^-_{ij} = \sum_i N_{i\bullet} \frac{N_{ij}}{N_{i\bullet}} = N_{\bullet j}$$

On note E_{ij} le nombre de cooccurrences espérées entre les modalités i et j sous hypothèse d'indépendance des deux variables qualitatives. On l'estime simplement à $E_{ij} = \frac{N_{i\bullet} N_{\bullet j}}{N_{\bullet\bullet}}$. On utilise le test d'indépendance du χ^2 pour comparer ces effectifs théoriques aux effectifs observés et définir le score : $X^2 = \sum_{ij} \frac{(N_{ij} - E_{ij})^2}{E_{ij}}$ (voir section 2.2.2 pour plus de détail).

Si on développe [22] maintenant l'expression de l'inertie 1.2 avec la formule 1.1 de la distance du χ^2, on obtient en définitive l'identité suivante :

$$I^- = \frac{X^2(N)}{N_{\bullet\bullet}}$$

Autrement dit l'inertie du nuage de profils-lignes est égal au test d'indépendance du χ^2 à un facteur près. Il est facile de montrer qu'il en est de même pour le nuage de profils-colonnes :

$$I^| = \frac{X^2(N)}{N_{\bullet\bullet}}$$

Partant de cette observation, on se retrouve exactement dans le cas de l'analyse en composantes principales. Les profils-lignes jouent le rôle d'individus (ou de fleurs) avec un poids égal à $N_{i\bullet}$, et avec une distance qui est celle du χ^2. Enfin l'analyse multiple des correspondances est une extension directe de la méthode à N variables catégorielles. Il est intéressant de remarquer d'ores et déjà la forte proximité de l'analyse des correspondances avec les métriques de similarité sémantiques que l'on introduira au chapitre suivant. La distance employée par Benzécri est clairement de nature « indirecte » (cf. section 2.2.1), mais elle ressemble, au moins dans sa forme, à la façon dont une similarité directe entre deux termes qui cooccurrent peut être calculée avec l'aide d'un test du χ^2 (même si une telle mesure s'exprime légèrement différemment dans un tel contexte - voir section 2.2.2).

Mais ce ne sont sans doute pas les raffinements mathématiques de la méthode de Jean-Paul Benzécri qui ont séduit tant de sociologues. Selon Phillips (1995) une première explication de l'attrait pour l'analyse des correspondances vient de la possibilité qu'elle offre de travailler directement sur des variables

22. $I^- = \sum_i N_{i\bullet} d^2_{\chi^2}(i, G^-) = \sum_{ij} \frac{N_{i\bullet}}{N_{\bullet j}} (N^-_{ij} - N^-_{i'j})^2 = \sum_{ij} \frac{(N_{ij} - \frac{N_{i\bullet} N_{\bullet j}}{n})^2}{N_{i\bullet} N_{\bullet j}} = \sum_{ij} d^2_{\chi^2}(i, j) = \frac{X^2(N)}{N_{\bullet\bullet}}$

catégorielles sans avoir à poser d'hypothèses causales préalables sur des relations entre variables. De ce point de vue l'analyse des correspondances de Benzécri est purement exploratoire, *a contrario* des méthodes d'analyse de facteurs (« factor analysis ») qui se développent sous l'impulsion de psychologues aux Etats-Unis [23] et qui requièrent de définir un modèle causal *a priori*. L'avantage de l'analyse des correspondances était en effet de proposer une technique purement descriptive, sans théorie économique sous-jacente (Desrosières, 2008). Comme on l'a déjà souligné, sa multi-dimensionnalité était conçue, juste après mai 68, comme une garantie de pluralisme qui permettrait enfin d'échapper à la monotonie des structures hiérarchiques et des représentations uni-dimensionnelles. Autre argument capital : l'analyse des correspondances fournit une représentation visuelle intuitive des individus et des variables projetées dans le même espace. L'attrait principal de l'analyse des correspondances est bien là. Un nuage de points aux corrélations complexes et multiples se (re-)présente dans un espace unique dont on peut tirer des interprétations globales quant à sa structure (quelles variables sont corrélées ou non, de quelles variables les facteurs sont-ils « chargés » ?) et locales quant à la distance entre modalités qui se lit visuellement dans le plan factoriel.

23. Pour une explication plus technique des différences mathématiques (finalement assez mineures) entre les deux, voir (Fabrigar et al., 1999)

Bien sûr, cette pluralité a un coût. Une limite classique de l'analyse des correspondances tient au taux de variance expliqué par les deux (voire trois) facteurs principaux finalement conservés dans la représentation finale. Il s'avère souvent famélique (il n'est pas rare de voir des taux de variance expliquée inférieurs à 10%) et les distances entre variables projetées en 2d sont nécessairement d'autant plus « faussées » que les valeurs propres définissant ces axes sont faibles. La projection des attributs et des individus dans un même espace est une propriété très séduisante mais elle ne va pas sans poser des difficultés puisque, rigoureusement, seules les distances entre points-ligne ou entre points-colonne ont un sens. Par contre, la distance entre un individu et un attribut n'a pas de sens géométrique.

Une grande partie du travail pour l'analyste, une fois l'analyse des correspondances réalisée, est de donner une interprétation aux différents axes. En nommant ce qui est mesuré le long de chaque dimension, l'analyste fournit finalement un modèle *a posteriori* qui explique le positionnement relatif des variables et des individus dans l'espace. Le travail de Boltanski et al. (1984), qui est probablement l'un des premiers à utiliser la méthode de Benzécri pour traiter un matériau textuel, illustre parfaitement la façon dont un modèle explicatif peut être tiré de l'interprétation d'une analyse factorielle. Les lettres de dénonciation envoyées au Monde bénéficient ainsi d'un codage très fin empruntant au modèle actantiel de Greimas et sur lequel nous aurons l'occasion de revenir ultérieurement (voir section 3.1.1). Elles sont ensuite projetées dans un espace factoriel qui fait apparaître en variable supplémentaire [24] les jugements de normalité. L'espace factoriel est défini par des variables ayant

24. Rappelons que des variables supplémentaires sont projetées dans l'espace calculé par les variables actives sans entrer dans la composition des facteurs

trait au contenu des lettres, à leur graphie, aux origines sociales des dénonciateurs, etc. Fort de leur schéma de codage et de cette l'analyse factorielle, les auteurs parviennent à caractériser les critères de normalité d'une dénonciation publique de façon très fine et proposent même un modèle pour expliquer ce qui pousse certains auteurs à produire des dénonciations jugées comme anormales alors même que ces derniers visent justement à remplir ces critères de normalité.

1.1.4 La méthode Alceste

Le développement du logiciel Alceste débute au milieu des années 80 au sein même du laboratoire de Jean-Paul Benzécri. Avec Alceste, l'analyse des correspondances s'enrichit d'une méthode à part entière pour traiter des documents textuels en tant que tel sans avoir à les « questionner » ou à les re-coder au préalable. Mise au point par Max Reinert, la méthode Alceste propose une analyse automatique des « mondes lexicaux » contenus dans un corpus textuel hétérogène (Reinert, 1993). Encore maintenant les méthodes introduites à l'époque par Max Reinert comme la classification hiérarchique descendante (CHD) sont régulièrement utilisées, et ce notamment grâce au logiciel Iramuteq qui a récemment pris la relève d'Alceste [25].

Héritant du scepticisme de Benzécri à l'encontre des approches « idéalistes » purement hypothético-déductives ou fondées sur des modèles linguistiques hérités de Chomsky, l'analyse des correspondances et Alceste sont finalement comme le remarque Beaudoin (2016) très proches de l'approche de Bloomfield et Harris qui pensaient pouvoir reconstruire de façon entièrement inductive les lois de la grammaire avec des hypothèses purement distributionnelles (voir section 2.2).

Les algorithmes qui composent Alceste sont multiples (détermination des unités de contexte, classification, contrôle de robustesse, projection sur les axes factoriels après catégorisation, etc.) et suivent parfois une méthodologie un peu confuse. Mais une des propriété principales sur laquelle il est intéressant de revenir est le modèle de language qui sous-tend l'analyse dans Alceste. Max Reinert se définit comme une sociolinguiste et s'appuie, à ce titre, sur une théorie de l'énonciation originale. Elle se distingue clairement de l'analyse de discours de l'époque à laquelle Reinert reproche sa tendance à décontextualiser des fragments de textes, qui devraient avant toute chose être conçus comme des « actes de paroles ».

Dans la théorie de Reinert, le corpus est conçu comme une suite d'énoncés élémentaires indissociables de leur locuteur. Le discours doit donc être saisi

25. Notons que si l'analyse des correspondances est encore pratiquée en France, elle semble n'être jamais parvenue à se distinguer de l'analyse multivariée à l'étranger, ce qui a donné lieu à un certain nombre de controverses entre statisticiens anglo-saxons et les disciples de Benzécri (Armatte, 2008)

comme un acte, un acte sémiotique dont on ne peut comprendre le sens mais dont les répétitions trahissent les postures énonciatives des locuteurs. La répétition dite indicielle se traduit par des coupures dans le texte que le logiciel reproduit par le découpage en unités de contexte.

Ce découpage en unités élémentaires définit le rythme de l'énonciation [26] et fonde la spécificité d'Alceste par rapport à une analyse des correspondances classique qui saisirait les groupements de mots sans interroger leur configuration dans le texte. Au sein de chacune de ces unités, la répétition iconique se traduit dans la répétition de « mots-pleins » la co-présence des mots-pleins traduisant le « fond associatif » ou « fond topique » dans le texte. Après avoir découpé le texte en unités de contexte (séquence d'unités de contexte élémentaires composés de quelques formes « actives » séparés par des formes dîtes « supplémentaires »), une classification hiérarchique descendante est appliquée sur les tableaux croisant unité de contextes et formes actives [27].

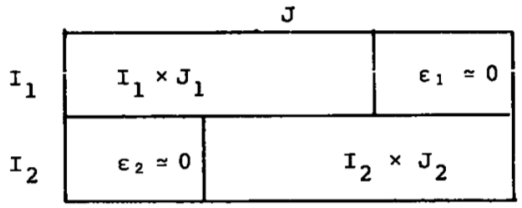


FIGURE 1.8: Méthode de classification hiérarchique descendante de Reinert. On cherche à ordonner les lignes et les colonnes de la table ci-contre (image empruntée à (Reinert, 1983)) de telle manière qu'une partition de l'ensemble des unités de contexte I en deux parties I_1 et I_2 fasse apparaître leur vocabulaire préférentiel comme deux ensembles aussi peu recouvrants que possible.

Cette méthode de classification est itérative. Elle comprend 3 étapes (Reinert, 1983). Une analyse des correspondances est d'abord menée sur le tableau croisant unités de contextes et formes. Les unités de contextes sont alors ordonnées en fonction de leur projection sur le premier facteur. Elles sont alors divisées en deux classes qui maximisent l'inertie inter-classes (voir figure 1.8). Certaines unités de contexte peuvent alors changer de classe jusqu'à ce que l'inertie inter-classe (mesurée par un test de χ^2) ne puisse plus être améliorée. Les trois étapes sont répétées sur la plus grande classe restante jusqu'à avoir obtenu le nombre de classes souhaité [28].

Une fois les classes identifiées, la structure des mondes lexicaux peut être interprétée à travers l'analyse du vocabulaire le plus spécifique de chaque monde ou *via* l'extraction des unités de contexte les plus représentatives (Reinert, 1990). Hormis le choix de le taille des unités de contextes élémentaires (ou au moins de leur intervalle de taille) et du nombre de classes final, l'analyste intervient finalement très peu dans cette première phase (même si ses choix sont déterminants pour le résultat (Dalud-Vincent, 2010)). Il est principalement appelé à contribution, une fois les mondes lexicaux révélés pour en proposer une interprétation. Et c'est sans doute là qu'un vrai dialogue s'établit, tant les mondes lexicaux invitent l'utilisateur d'Alceste à ré-interroger sa lecture des textes (Reinert, 2003) dans un mouvement que Reinert qualifie

26. Reinert (2007) écrit lui même : « Le tableau de données modélisant cette lecture artificielle contient en lignes, les différents segments de texte, et en colonnes, les mots pleins. Il modélise grossièrement l'activité énonciative dans sa capacité d'animer rythmiquement des contenus ».

27. les formes actives peuvent être composées de noms, verbes, adjectifs, adverbes, elles se distinguent des formes supplémentaires généralement composées de mots outils - tels que les pronoms, conjonction, certains adverbes et verbes fréquents.

28. Le choix du nombre de classes est donc entièrement déterminé par l'analyste, qui peut s'il le souhaite également paramétrer lui-même la longueur moyenne des unités de contexte élémentaires.

d'abductif.

Les écrits de Reinert, mélangeant théorie sémiotique et procédures mathématiques, sont parfois un peu confus. Il résume néanmoins parfaitement le principe du fondement topique des énoncés de façon métaphorique dans la conclusion de (Reinert, 1999) en établissant un parallèle entre les activités humaines structurées par des « habitus » chez Bourdieu, et les discours, traces langagières de ces activités, qui sont structurés par des systèmes de « lieux » (ou topoï) « agissant comme des attracteurs pour le locuteur » et se définissant toujours de façon relative. Un « lieu » se définit en effet toujours par opposition à d'autres lieux [29] et sans lesquels il perd son identité. Les mondes lexicaux émergent donc à force de répétitions des mots-pleins mis en relation via les actes d'énonciation des acteurs (modélisés par les unités de contexte).

29. La méthode CDH épouse parfaitement ce principe.

Mais en dépit de cette théorie linguistique omniprésente dans les écrits de Reinert qu'il hérite là du pragmatisme de Pierce (Reinert, 2001) mâtiné d'influence lacanienne, la voie que propose Reinert (et Benzécri *a fortiori*) ne semble nullement incompatible avec nombre de théories sociologiques. Méthode « tout terrain », Alceste a été utilisé sur des corpus littéraires, en marketing, mais aussi en psychologie clinique ou pour analyser des discours politiques (Marchand et Ratinaud, 2012). C'est aussi un logiciel qui semble pouvoir s'adapter à différents types de modèles de l'action sociale. Après tout, si on se réfère à son acronyme « Analyse des Lexèmes Coocurrents dans les Enoncés Simples d'un Texte », Alceste se présente bien comme un moyen d'enquêter sur les motifs émergents des cooccurrences entre éléments textuels au sein d'énoncés. L'analyse de mots-associés n'est pas si loin et les modèles sémantiques vectoriels que nous aborderons plus tard encore plus proches. Elle semble au moins compatible avec la sociologie pragmatique (Boltanski et al., 1984), avec la théorie des champs de Bourdieu [30] mais aussi d'analyse des entretiens pour faire de la sociologie interactionniste (Zalio, 2007) ou même récemment des corpus de tweets pour analyser une controverse (Smyrnaios et Ratinaud, 2017).

30. Bourdieu (2000) concevait même une homologie naturelle entre analyse des correspondances et théorie des champs, paradoxalement à même « d'expliquer » (et non de décrire) les stratégies des agents à partir de la distribution des pouvoir et des intérêts.

1.1.5 Prospero et l'approche pragmatique

Prospero est en fait un acronyme qui signifie : PROgramme de Sociologie Pragmatique, Expérimentale et Réflexive sur Ordinateur. Développé par Francis Chateauraynaud au GSPR [31] avec le concours de l'informaticien Jean-Pierre Charriau, ce logiciel se présente comme une technologie littéraire pour les sciences humaines qui permet de suivre des dossiers complexes composés de séries temporelles de textes correspondant au déploiement ou à la résolution d'une controverse ou d'une affaire publique. Le texte est alors conçu comme

31. Le Groupe de Sociologie Pragmatique et Réflexive a été créé dans les années 1990 à l'EHESS sous la direction de Francis Chateauraynaud pour mener le programme de sociologie pragmatique portant sur des affaires et les controverses.

un outil pour accéder aux « actes, [...] prises de position, des opérations argumentatives et normatives effectuées par les acteurs » (Chateauraynaud, 2003b, chap 6)

Prospero s'oppose très fortement aux approches lexicométriques et plus largement à l'analyse de données. Chateauraynaud est ainsi très critique de la séparation qu'elles orchestrent entre les représentations délivrées par les algorithmes et les interprétations du chercheur qui arrive nécessairement *a posteriori* et doit se soumettre à une AFC comme si elle consistait un fait établi. Chez Chateauraynaud, l'analyste doit s'immiscer à toutes les étapes de l'analyse, et en particulier dans les activités de codage. Plutôt que de soumettre un ensemble de documents homogènes à une procédure standardisée comme en lexicométrie, Prospero encourage le chercheur à enrichir son corpus de textes (conçu comme un objet hétérogène et dynamique susceptible d'évoluer dans le temps et qui doit capturer les différents aspects de l'affaire étudiée), retravailler sans cesse son codage (notamment lorsque de nouveaux textes sont ajoutés), s'intéresser au contenu des textes plutôt que de rester à distance. A l'impératif inductif, il répond que toute procédure d'analyse, aussi élémentaire soit-elle, requiert la mise en place d'un modèle. Sans codage réfléchi préalable aucune propriété pertinente ne peut émerger automatiquement des données.

En tant qu'entreprise de sociologie pragmatique, Prospero porte une attention particulière à la façon dont les acteurs se justifient et émettent des critiques : comment des arguments sont produits dans une discussion, fût-elle composée de textes épars. Chateauraynaud défend cette idée contre le relativisme qu'accompagne la critique de « l'ordre des discours » en sciences sociales qui ne verrait dans les succès de telle ou telle position dans un débat que le règne des croyances (Chateauraynaud, 2011, chap 2). Il réaffirme au contraire la réalité des épreuves de vérité auxquels les arguments et les acteurs doivent sans cesse se soumettre. L'analyse de controverse est donc avant tout pour Chateauraynaud celle des argumentations mobilisées, et possiblement contestées par les acteurs. S'appuyant sur Tarde, il admet l'importance de croyances au sein des affaires mais continue à défendre les arguments et la logique comme leviers premiers pour modifier le curseur de la « foi » des acteurs :

> *« [...] il ne s'agit pas là, en raisonnant, de faire engendrer la conclusion par les prémices, comme on le suppose dans les écoles. [...]. En fait l'utilité du raisonnement réel, pratique, consiste non pas à faire naître des propositions nouvelles, induites ou déduites [...] mais bien à modifier notre opinion, - j'ajoute : ou l'opinion d'autrui principalement »* (Tarde, 1904, chap 4)

Or capturer une argumentation n'est pas chose aisée, un énoncé ne forme un argument qu'en fonction d'un certain contexte. C'est aussi la raison pour laquelle Prospero est si bien ajusté à des corpus qui prennent la forme d'« affaires », dans lesquelles les acteurs confrontent leurs arguments et s'efforcent

sans cesse de faire preuve de pédagogie, si bien que l'analyste n'a plus qu'à cueillir la structure argumentative des énoncés exprimée de façon si explicite. Dans le modèle prospérien, on est donc très attentif à des « formules » particulières telles que « Au nom de X, il n'y a pas de raison que Y », « Pour X, il faut Y », etc. C'est sans doute là une limite de la méthode, qui s'accommode *a priori* plus difficilement de textes non polémiques.

Sept classes d'objets textuels différents forment les briques de base du méta-langage qu'utilise Prospero pour « lire un dossier ». Les quatre classes principales pour coder le texte sont : les entités, les qualités, les épreuves, les marqueurs (Chateauraynaud, 2003b, chap. 13). Ces classes sont indexées en fonction de dictionnaires construits manuellement même si certains dictionnaires peuvent être mis en mémoire pour que leur cadre d'analyse soit réactivé (moyennant adaptation éventuelle) dans un autre dossier. Les entités ressemblent fort à ce que les linguistes appellent des entités nommées et regroupent des personnes ou des groupes, des objets, des animaux, des concepts, des institutions, des procédés, des lieux, des périodes, des émotions, etc. Des qualités (*a priori* des adjectifs) peuvent par la suite être associées à ces entités (ainsi Céline peut être qualifié d'antisémite ou d'admirable dans un même dossier). Des épreuves (souvent des verbes) indiquent le type d'action ou de transformation ou de jugement entre entités[32]. Les marqueurs sont généralement des adverbes qui vont nuancer l'argumentation.

32. Chateauraynaud ne manque pas de souligner la naïveté de ce qu'il appelle des « cartes de liens » de ce point de vue qui se contentent d'associer des entités sans jamais interroger les modalités de ces associations.

Saisis à travers cette grille patiemment conçue, à force d'allers-retours avec le texte d'origine, les textes sont indexés par ce bestiaire d'opérateurs textuels. Autre opération analytique cruciale permise par le logiciel : la gestion d'êtres fictifs qui réunissent sous un même nom l'ensemble des acceptions possibles d'une personne publique, d'un parti politique, voire de l'ensemble des personnages politiques, etc. Les êtres fictifs sont les « piliers des structures narratives » et à ce titre, peuvent regrouper toutes ses désignations possibles aussi hétérogènes soient-elles. L'être fictif « @TONTON » peut ainsi regrouper les entités suviantes : « président de la république », « François Mitterand », « tonton » (Chateauraynaud, 2003b, chap. 14). Naturellement, ces regroupements s'avèrent plus ou moins adaptés en fonction de la source des données. Compte tenu de la sensibilité de la méthode au corpus, celui-ci doit être intimement connu du chercheur (et par conséquent, rester de taille raisonnable). Avec Prospero, on peut aussi conclure sur l'appartenance d'un texte à un ordre de discours donné (comme le registre de controverse, la figure de dénonciation ou même la logique de territoire) grâce à des noyaux lexicaux (pré-établis ou à construire). Enfin les formules permettent d'explorer systématiquement les modes d'argumentation rencontrés dans un dossier. Il est par exemple possible de lister les principes sur lesquels les acteurs s'appuient en construisant la formule : AU NOM DE/DU + ENTITE. Les formules servent de fil directeur pour interroger/enquêter sur un corpus.

Fort de cette grille de codage très riche, les corpus peuvent ensuite être interrogés à travers différents prismes : l'approche réseau, l'approche temporelle et l'approche que l'on qualifiera de fréquentielle. Malgré la multiplicité des points de vue, on verra que le principe statistique qui domine est celui de l'approche comparative ou contrastive.

Le réseau d'une entité est formé par l'ensemble des entités qui entrent en contact avec elle au fil des énoncés dans lesquels elle figure. Mais différents types de « contact » sont possibles selon que les entités soient directement liées au travers d'une épreuve, ou qu'elles soient associées de façon plus lâche au sein d'un même texte. Dès lors, on peut apprécier comment le réseau égo-centré [33] d'une entité (mais aussi son anti-réseau, composé des entités avec lesquelles elle cooccurre significativement moins - par exemple dans l'affaire Sokal (Chateauraynaud, 2003a) : l'être-fictif postmodernisme est incompatible avec les entités « signes » et « calculs ») varie selon que l'on segmente le corpus en fonction d'un locuteur, d'une période temporelle, etc.

33. Il faut bien insister ici sur le fait que les réseaux de Prospero sont toujours liés à une entité, ils n'ont pas l'ambition de saisir la structure globale du dossier. Il est possible de calculer le réseau global des personnes se retrouvant dans les mêmes énoncés par exemple(Chateauraynaud et Debaz, 2011, chap 3), mais la visualisation du réseau est alors sous-traitée à des logiciels tiers comme Pajek sans que sa topologie ne soit l'objet d'une analyse.

La segmentation, c'est en réalité l'opération primordiale dans Prospero dont la pratique vise essentiellement à générer des vues comparatives d'un segment de corpus par rapport à un autre. On peut découper ces segments de différentes manières : séparer des textes en fonction de leur date, de leur(s) auteur(s) selon qu'ils contiennent ou non (on parle alors « d'anti-corpus ») certaines entités, ou qu'ils relèvent d'un ordre du discours donné. Cette capacité à contraster une partie du corpus avec une autre permet d'en extraire à tâtons - en expérimentant sur différentes variables (pour découper les corpus mais aussi pour en décrire les différences) - les lignes de force d'une affaire, les oppositions et les accords, les moments de dormance et de relance [34]. Il faut souligner que les scores et autres mesures mises en œuvre dans Prospero s'appuient sur une mathématisation contrôlée (Ollivier, 2010). On se contente ainsi pour construire ces listes d'entités spécifiques de mesurer des écarts à la manière d'un taux de croissance corrigé des différences de volume entre sous-corpus (Chateauraynaud, 2003b), ce qui n'est pas sans poser de problèmes statistiques dont Chateauraynaud préfère s'accommoder plutôt que d'avoir à interpréter des indices plus complexes tels que ceux manipulés par les lexicomètres (voire section 1.1.2). Ces résultats sont toujours rendus sous forme de listes ordonnées (cf figure 1.9), les visualisations étant réduites à portion congrue dans Prospero pour au moins deux raisons : (i) la volonté de permettre à l'analyste de naviguer librement entre les niveaux et à tout moment de revenir au texte original et donc de se contenter de visualisations au plus près du contenu textuel, (ii) la crainte déjà mentionnée plus haut, avec la visualisation de données, de figer les positions et de mettre en équivalence des contextes d'énonciation pourtant multiples.

34. Il faut pour rendre entièrement justice à l'outil, ajouter qu'un module d'analyse temporelle permet de détecter automatiquement des textes qui dans une série d'énoncés font figures de pionniers ou de repreneurs dans un dossier car ils introduisent les premiers ou combinent des entités de façon inédite.

Prospero a été utilisé, surtout au GSPR, pour suivre de très nombreuses

affaires se déployant pour certaines sur des dizaines d'années et sur des sujets aussi variés que les polémiques intellectuelles (affaire Sokal, Céline), les risques sanitaires collectifs (prion, nucléaire, amiante, pesticides), les affaires judiciaires (MNEF, sans-papiers). Naturellement, Chateauraynaud ne peut faire l'économie, compte tenu de son objet d'étude, de comparer son approche à la « cartographie des controverses » inventée par les STS et pérennisée comme méthode pédagogique en France aux Mines puis à Sciences Po sous l'influence de Bruno Latour. L'un des arguments les plus saillants que nous retenons dans cette critique est celle du risque que l'impératif de cartographie des acteurs et des arguments ne les « arrachent aux processus et aux milieux dans lesquels ils prennent corps » (Chateauraynaud, 2013).

Mais Prospero, par rapport à d'autres logiciels d'analyse textuelle en sociologie et notamment par rapport à l'approche inductive assumée de la lexicométrie, se singularise sans doute avant tout par la place toute particulière qu'il donne à l'interprétation. Celle-ci n'arrive pas en conclusion, « une fois les données traitées par des algorithmes indépendants des orientations théoriques », mais est portée par l'utilisateur à toutes les phases de l'analyse. Le rôle assigné au logiciel « est celui d'un dispositif d'expression des stratégies interprétatives développées par ses utilisateurs pour rendre intelligibles les structures et les évolutions en oeuvre dans leurs corpus. Il est le témoin, ou le garant, de multiples voies d'accès à l'objectivité et non le générateur autonome de l'objectivité elle-même » (Chateauraynaud, 2003b, introduction).

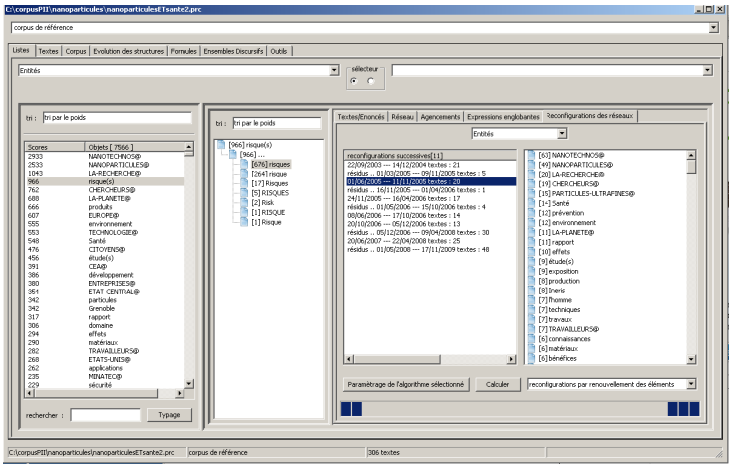


FIGURE 1.9: Interface de Prospero 2. L'entrée dans les corpus se fait essentiellement *via* des listes d'entités ordonnées par score (généralement de fréquence). On perçoit également bien la nature extrêmement modulaire du logiciel qui permet de changer de perspective analytique à l'envi (formules, évolution des structures, ensembles discursifs). Capture d'écran obtenue sur le carnet de recherche : https://socioargu.hypotheses.org/850 qui illustre aussi le caractère collectif de l'expérimentation que constitue Prospero (et son écologie de logiciels)

On ne peut que suivre Chateauraynaud quant au rôle premier de l'argumentation dans le déploiement de controverses publiques, et être admiratif de la minutie avec laquelle il identifie et caractérise le processus argumentatif avec un modèle de la prise de parole très riche. Si l'analyse comparative des formes prises par différentes affaires consiste en une promesse riche, force est néanmoins de constater que cette attention a un coût insurmontable dès lors que l'on souhaite déplacer la méthode vers d'autres objets. Prospero

est dédié à l'analyse d'affaires, et ses catégories seraient mal ajustées pour lire d'autres corpus, correspondant à des « styles argumentatifs » différents, voire à de simples juxtapositions et répétitions d'affirmations dont le caractère contradictoire ne serait pas « verbalisé » par les acteurs. On peut penser à un corpus de publications scientifiques, dans lequel deux sous-communautés développeraient des recherches dans des directions parfaitement contradictoires tout en s'ignorant l'une l'autre sauf à faire appel à des références différentes (Shwed et Bearman, 2010) et parfois à une terminologie divergente [35].

35. A vrai dire, un corpus de publications scientifiques peut effectivement être absorbé par Prospero comme le mentionne ce rapport pour l'ANSES rendant compte d'un travail d'enquête sur les risques sanitaires liés aux perturbateurs endocriniens (Chateauraynaud et al., 2013). Mais le matériau textuel réuni (7000 publications) est si massif et exotique par rapport aux textes que Prospero peut traiter, qu'il ne sert finalement qu'à enrichir des collections et des dictionnaires existants.

Une autre limite de Prospero tient à sa sensibilité à la taille des corpus. C'est une difficulté beaucoup plus prosaïque, mais l'analyste étant sans cesse engagé dans une lecture au plus près des dictionnaires et des textes permettant d'accéder à un méta-langage éprouvé, l'utilisation du logiciel sur des corpus dont le volume dépasse ses capacités de lecture est exclu par conception.

On peut néanmoins s'interroger sur la possibilité de concilier une modélisation du contenu textuel qui soit à la fois plus riche que la seule « mise en sac » de mots glanés dans une série d'énoncés suffisamment automatisée pour bénéficier pleinement d'un traitement par la machine. En proposant d'abord de faire des comptages à grande échelle, on ouvre l'analyse à d'autres acteurs qui participent aussi des/aux controverses (pourquoi résumer l'opinion des « simples » consommateurs à un diagramme google trend (Debaz, 2013) lorsque les prises de position des « experts » bénéficient d'un traitement aussi raffiné ?). Naturellement, cela requiert au moins pour le moment de diminuer le degré d'exigence du modèle argumentatif de Chateauraynaud, la question étant alors d'évaluer le coût réel d'une telle simplification : réduire formellement les qualités aux seules formes adjectivées ; construire des êtres fictifs automatiquement en utilisant les outils de linguistique computationnelle (reconnaissance d'entités nommées, sémantique distributionnelle) ; apprendre des ordres discursifs avec des algorithmes d'apprentissage qui se fondent sur quelques exemples de prototypes pré-étiquetés ; etc. Le programme de sociologie pragmatique ne gagnerait-il pas en lâchant du lest : déléguer à la machine pour gagner en robustesse quitte à revenir à un modèle moins informatisé par moment ?

Sur un mode autrement plus modeste, nous avons, dans un projet en cours avec Ian Gray, analysé le corpus « diplomatique » des comptes-rendus quotidiens de l'Earth Negociation Bulletin durant les COPs [36] des 20 dernières années. Le matériau textuel en question est extrêmement codifié. Un guide de bonnes pratiques est d'ailleurs distribué aux rédacteurs des bulletins pour qu'ils suivent un certain nombre de règles de rédaction qui garantissent une homogénéité des comptes-rendus dans le temps et entre rédacteurs. Le modèle que nous avons conçu se situe à l'intersection d'une perspective « prospérienne » et de l'analyse de sentiment.

36. Les « Conference Of the Parties » réunissent l'ensemble des Etats signataires (les « Parties ») de la CCNUCC () depuis 1995

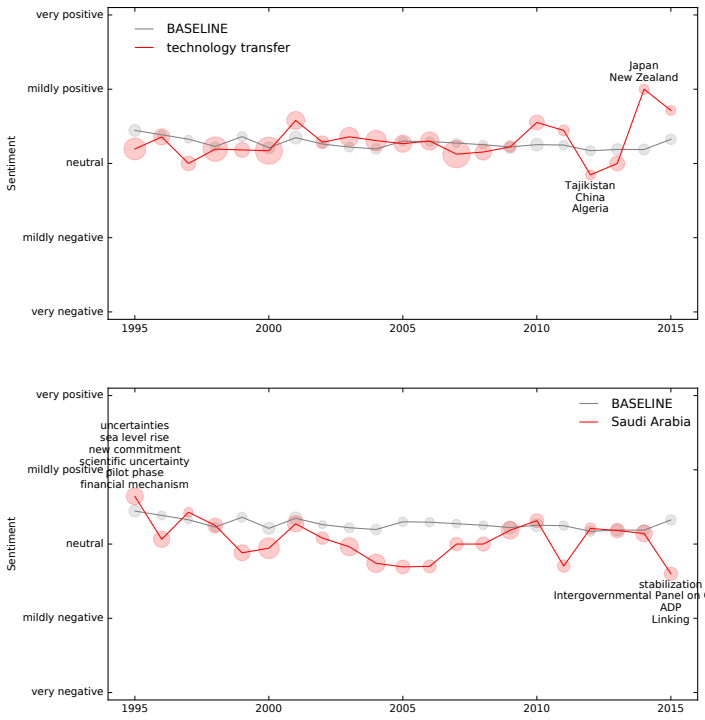


FIGURE 1.10: Modalité de l'ensemble des paragraphes mentionnant le terme « technology transfer ». La polarité d'un sujet est mesurée à travers les types de verbes qui les introduisent. Les pays les plus associés à une modalité négative ou positive de technology transfer sont ajoutés à la série temporelle lors des COPs 18 à Doha et 20 à Lima [haut]. De façon duale, on voit sur le diagramme suivant que les interventions de l'Arabie Saoudite sont généralement associées à l'usage de verbes plus « négatifs » que la moyenne (évolution « moyenne » en gris). Les sujets pour lesquels elle déploie un vocabulaire négatif lors de la COP de Paris sont stabilization, ADP, etc. [bas]

Les différents paragraphes des bulletins sont constitués de phrases assez standardisées dans lesquels un pays ou un conglomérat de pays prend position (à travers un verbe) sur un sujet donné. Les bulletins peuvent donc être codés de façon relativement systématique comme des triplets qui mettent en jeu un acteur, une modalité d'action et un sujet de discussion. Par exemple, si on considère le paragraphe suivant trouvé dans le bulletin du 1er décembre 2015 de la COP21 à Paris : « Le Président Rafael Correa Delgado, Équateur, a plaidé pour un libre accès aux technologies d'atténuation et la création d'un tribunal international pour la justice environnementale », on en tire aisément et de façon entièrement automatique les deux triplets suivant : (« Équateur » ; « plaider pour », « libre accès aux technologies d'atténuation ») et (« Équateur » ; « plaider pour », « création d'un tribunal international pour la justice environnementale ») qui dans la nomenclature de Prospero peuvent être modélisées comme des formules : ENTITE + EPREUVE + ENTITE.

En ordonnant les verbes selon une échelle de modalités minimaliste [37] il est possible de saisir à une échelle agrégée et de façon systématique les points d'accord et de désaccord autour des grands sujets de discussion des négociations figure 1.10. Voilà une stratégie de réduction à laquelle Chateauraynaud aurait probablement fort à redire. Il n'empêche, malgré son caractère très réducteur, cette mise en équation des verbes d'action permet de construire des

37. ((oppos*, object*, warn*, etc.) : negatif, (endorse*, criticize*, caution*) : positif, (conclude*, describe*, clarifi*, etc.) : neutre

récits généraux sur la nature contestée ou non des grands sujets de discussion, la stratégie adoptée par les pays, etc. Libre à l'analyste par la suite de contrôler la véracité de ces récits en revenant aux données primaires. C'est sans doute à ce moment là que l'usage de la visualisation de données, ou plus précisément le design d'interface s'avère indispensable. À titre d'exemple, et sur les même données, Pablo Ruiz Fabo a récemment mis en ligne[38] , parallèlement au travail que nous avons mené, une interface qui permet d'un seul regard, de lire l'ensemble des phrases mettant en jeu un acteur avec un verbe d'action donné. Ce qui permet d'un seul regard de contrôler le sens véritable des énoncés dans leur contexte. Plus qu'une simple mise à plat des données primaires, l'interface offre également à l'utilisateur un moyen de « régler le degré d'autonomie » laissé aux algorithmes de traitement automatique de la langue.

38. L'adresse de l'interface est http://apps.lattice.cnrs.fr/ie/uidev. La mise en relation entre entité s'appuie sur une analyse syntaxique et sémantique beaucoup plus sophistiquée que la nôtre (Fabo et al., 2016).

1.2 *Des Topic Models aux Plongements de Mots par réseaux de neurones*

Nous nous intéresserons maintenant à des méthodologies d'analyse beaucoup plus récentes nées dans les années 2000 voire 2010. Si l'école française d'analyse textuelle était, dans la plupart des cas, fortement « sociologisée » par conception, les nouvelles approches qui émergent de communautés d'informaticiens, de physiciens ou plus globalement du domaine de l'intelligence artificielle, requièrent un travail spécifique pour s'accorder aux pré-requis de la recherche en sciences sociales. N'étant pas homologuées à la conception, tout un travail de conceptualisation et d'adaptation des pratiques doit être mené pour les déplacer de leur espace de naissance vers de nouvelles applications.

Nous excluons de notre panorama l'analyse par réseau de proximité qui sera introduite plus longuement dans le chapitre 2 ainsi que les méthodes de nature purement hypothético-déductive qui font essentiellement appel à des approches d'apprentissage supervisé. Partant de catégories pré-définies par un « expert » et assignées à un certain nombre de cas (documents, individus, etc.), ces méthodes visent à prédire ces catégories pré-étiquetées sur un nombre limité d'objets à de nouveaux exemples en fonction d'une production textuelle donnée (Evans et Aceves, 2016). Par exemple, si on postule un lien entre réception critique et réussite économique d'un film , il est possible - à partir de données passées du box office et en analysant les critiques reçues par les films - de construire un modèle qui prédit les revenus d'un film à partir des critiques qu'il a reçues la semaine précédant sa sortie (Joshi et al., 2010). De la même façon, on peut chercher à prédire l'évolution de la bourse en fonction de « l'humeur » de Twitter (Bollen et al., 2011). Nous faisons le choix de mettre de côté ces méthodes pour nous concentrer sur les méthodes plus inductives, qui visent réellement à explorer un corpus de textes avant d'en dégager une

théorie ou une interprétation d'un phénomène social. Nous traiterons de façon détaillée deux méthodes extrêmement populaires : les topic models et les modèles de plongement de mots dont les applications sont clairement de nature exploratoire et l'approche inductive. Une troisième section listera sans les décrire de façon exhaustive les autre familles de méthodes alternatives qui sont peut-être moins directement pertinentes pour les sciences sociales ou simplement moins mûres.

1.2.1 *Topic Models*

Les « topics models » recouvrent un ensemble de modèles probabilistes récents extrêmement populaires pour la caractérisation d'un corpus de textes sous la forme d'ensembles sémantiquement cohérents appelés « topics » (ou thématiques, même si la traduction de topic model en modèle thématique ou modèle de sujet n'est pas très répandue). On parle de topic model car on postule l'existence d'un modèle génératif de nature bayésienne qui, dans la cas du modèle le plus répandu appelé LDA pour Latent Dirichlet Allocation (Blei et al., 2003), produit des documents (i) correspondant à un mélange de topics (ii) à chacun desquels est associée une certaine distribution de probabilités sur l'ensemble du vocabulaire. Plus visuellement (voir figure 1.11) les topics models proposent un modèle mixte de classification des mots dans les topics dont on fixe le nombre en amont (ici 3 dans la partie droite du schéma), et des documents auxquels sont assignés une distribution de ces topics (partie gauche du schéma).

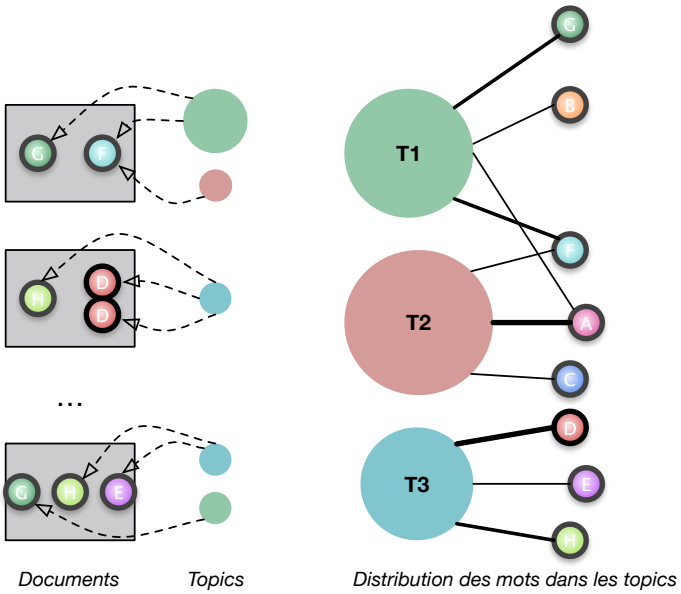


FIGURE 1.11: Représentation d'un topic model. De la simple observation de la distribution des mots dans les documents d'un corpus (partie en gris sur le schéma), on en déduit un processus génératif qui étant donné une certaine distribution de topics pour chaque document (illustrée par la « colonne » intitulée Topics) est susceptible de re-générer la distribution de mots originale sachant que chaque topic (dont le nombre est fixé *a priori*) est défini comme une distribution de probabilités particulière sur l'ensemble des mots du corpus.

Historiquement, les topic models développés par Blei depuis les années 2000 consistent en un raffinement de modèles d'analyse sémantique plus anciens dits de sémantique latente probabilistes (PLSA ou parfois PLSI) introduits par Hofmann (2000). Ces derniers héritent eux-même de modèles plus simples d'analyse sémantique latente (Deerwester et al., 1990) (Latent Semantic Analysis LSA, également appelé LSI pour Latent Semantic Indexing) qui visent à identifier les dimensions sémantiques latentes cachées d'un corpus de textes en effectuant, un peu à la manière de l'analyse des correspondances, la décomposition factorielle de la matrice documents - termes.

Pour être plus précis, l'analyse des correspondances n'est pas entièrement équivalente à l'analyse sémantique latente. Dans l'AFC, les profils-lignes et colonnes sont d'abord normalisés et « équipés » de la métrique du χ^2 avant d'être diagonalisés. Sans que les matrices soient tout à fait les mêmes, la recherche de valeurs propres est néanmoins isomorphe dans les deux méthodes. Dans le cas de la LSA, on effectue directement une décomposition en valeur singulière (SVD) de la matrice brute document terme tandis que l'analyse des correspondances passe par le calcul de la matrice de co-variance (qui est positive et symétrique à valeur réelle, donc diagonalisable dans une base orthonormée). Même si les deux décompositions sont équivalentes (Shlens, 2014), de façon plus cruciale, dans le cas de l'analyse des correspondances, l'inertie du système est compressée sur quelques facteurs principaux qui permettent de produire une représentation géométrique du système. La prétention géométrique de l'analyse sémantique latente est beaucoup plus limitée puisque l'espace vectoriel final est encore composé de quelques centaines de dimensions. L'objectif de la LSA est de calculer efficacement des distances sémantiques entre textes, entre termes ou entre termes et textes, pas de représenter ces variables dans un espace commun [39].

39. *A contrario*, les modèles de plongement de mots que l'on décrira dans la prochaine section partagent cet objectif.

Dans le cas de la LDA, les topics sont conçus comme des variables latentes qui expliquent la composition des mots observés dans chaque document. Dans sa version de base, ces variables latentes sont totalement indépendantes les unes des autres, c'est à dire qu'il n'existe *a priori* aucune corrélation entre topics. Néanmoins, les topic models peuvent être complexifiés pour intégrer une dépendance dans la dimension temporelle (Blei et Lafferty, 2006), ou vis-à-vis de variables exogènes (Rosen-Zvi et al., 2004; Blei et Lafferty, 2007). Techniquement on utilise des méthodes d'inférence de paramètres pour maximiser la vraisemblance du modèle en fonction des données observées (deux familles d'algorithmes sont employés : à base d'échantillonnage (Gibbs par exemple) ou les méthodes variationnelles (Blei, 2012)).

Les textes (typiquement quelques milliers de documents (*a minima*)) sont donc modélisés comme des sacs des mots dont l'ordre est ignoré. Le vocabulaire est généralement constitué par l'ensemble des monogrammes présents

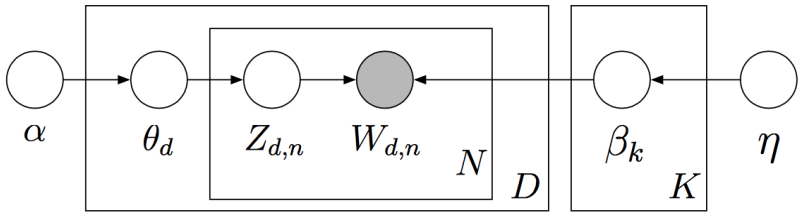


FIGURE 1.12: Modèle « graphique » de LDA (extrait de (Blei, 2012)) représentant les relations entre les différents variables aléatoires du système. Les mots (variable grisée car empiriquement observée) résultent d'un processus génératif qui dépend du « mélange » de sujets présents dans un document donné. Le plus grand rectangle désigne l'ensemble des documents D présents dans le corpus. Chaque document est composé d'un certain nombre de mots N. K désigne les hyper-paramètres du modèle.

dans le corpus à l'exception des mots trop fréquents ou trop rares. Même si de nombreuses recherches ont proposé des critères formels pour déterminer le nombre de topics, aucune ne s'est réellement imposée, et l'analyste doit généralement choisir le nombre final de topics. Il est aussi classique qu'il exclue manuellement les topics qu'il ne peut pas interpréter et qu'il juge de fait impertinents. Le manque de pertinence des topics est d'ailleurs une des limites connues du modèle (Chuang et al., 2012; Newman et al., 2010). La difficulté à savoir comment paramétrer le modèle avec un nombre de topics donné en est une autre (Schmidt, 2012).

Les topic models ont bénéficié d'un très fort intérêt depuis leur création, alors que son créateur en promouvait l'universalité :

> « *In addition to scientific applications, such as genetics or neuroscience, one can imagine topic models coming to the service of history, sociology, linguistics, political science, legal studies, comparative literature, and other fields where texts are a primary object of study.* » (Blei, 2012) [40]

Et dans les faits, les sciences humaines et sociales se sont effectivement largement saisies de l'algorithme [41], notamment outre-atlantique, où le journal Poetics lui a même consacré un numéro spécial (Mohr et Bogdanov, 2013). Les corpus analysés couvrent un très grand éventail de sources : articles de presse, tweets, publications scientifiques, rapports administratifs, fictions, etc. Le très grand nombre d'implémentations de la méthode dans différents environnements et langages de programmation n'est sans doute pas entièrement étranger à leur popularité [42].

On remarquera néanmoins l'inconfort relatif dans lequel cette méthodologie place l'analyste. La plupart des sociologues et historiens qui en font usage soulignent le soulagement produit par le passage à l'échelle qui permet de raisonner sur des ensembles de termes plutôt que sur un terme unique, toujours trop ambigu. La capacité des topics models à générer des catégories de façon entièrement émergente est également souvent soulignée comme une forme de libération (DiMaggio et al., 2013). Pour autant l'analyse par topic model déplace le moment de l'interprétation en aval du processus calculatoire (Mohr et Bogdanov, 2013), le calcul est premier, le travail herméneutique second. Pour autant, il faut à nouveau être attentif aux pratiques réelles. Dans les faits, l'interprétation s'inscrit dans un processus itératif. Une difficulté

40. « En plus des applications en sciences comme la génétique ou les neurosciences, on peut imaginer que les « topic models » se mettent au service de l'histoire, la sociologie, la linguistique, les sciences politiques, le droit, la littérature comparée, et tout autre champ dont le texte constitue l'objet d'étude premier. »

41. Templeton (2011) a même recensé l'ensemble des travaux en sciences sociales s'appuyant sur les topic models il y a quelques temps.

42. Citons Mallet, le plus répandu, développé à UMASS et à l'University of Pennsylvania (McCallum, 2002) mais aussi TMT développé à Stanford (Stanford Topic Modeling Toolbox) également en Java, la librairie Gensim sous python, etc.

interprétative oblige le chercheur à reformuler ses hypothèses, changer le nombre de topics, filtrer différemment les termes analysés. Les concepteurs de la méthode encouragent même à un usage décomplexé de l'outil pour les non-statisticiens en transformant un outil mathématique en instrument de mesure expérimental :

> « *finding the right lens is different than evaluating a statistical model based on a population sample. The point is not to estimate population parameters correctly, but to identify the lens through which one can see the data most clearly.* » (DiMaggio et al., 2013) [43]

43. « trouver la bonne lentille n'a rien à voir avec l'évaluation d'un modèle statistique à partir d'un sondage de la population. L'objectif n'est pas d'estimer correctement les paramètres de la population correctement mais d'identifier le prisme à travers lequel les données puissent être perçues le plus clairement possible. »

Un problème, majeur, que l'on retrouve dans de nombreux témoignages de sociologues vient de la pauvreté des représentations finales, qui se réduisent souvent à des listes ordonnées [44]. La visualisation de la distribution des topics sur les documents demanderait également le développement d'interfaces ad-hoc (Blei, 2012). Enfin, mais on y reviendra plus tard, le résultat d'un topic model est très rapidement réduit à des liste de mots. En réalité, il fournit un véritable modèle génératif de documents. La question reste ouverte, mais serait-il possible, en sciences sociales, d'articuler le modèle génératif complet (de mots dans les topics et de topics dans les documents) avec des théories du social. En l'état, les topics, listes de mots auxquelles sont associés des scores [45] fournissent une représentation assez crue des grandes thématiques qui structurent un corpus de documents. Il est impossible de comprendre l'articulation d'un terme avec un autre au sein d'un topic. Dans le modèle de base, les topics sont entièrement déconnectés les uns des autres (et par construction, cela fait partie des hypothèses de la LDA). Bref, leur usage semble nécessiter de faire preuve d'une très grande imagination sociologique. Mohr, Wagner-Pacifici, Breiger, et Bogdanov (2013) n'en manquent pas lorsqu'ils usent d'un topic model comme composante d'un modèle d'énonciation plus large pour analyser un corpus de rapports gouvernementaux (voir section 1.3.1). Plus récemment, Schmidt (2015) s'appuie sur la classification en topics des sous-titres provenant d'un corpus de films pour reconstruire leur structure narrative qui prend la forme de séquences types de thèmes se succédant les uns aux autres.

44. La méthode poussant peut-être l'abstraction mathématique le plus loin est finalement la plus proche, du point de vue de la présentation des résultats, de Prospero qui plaide pour une mathématisation contrôlée.

45. Ces listes sont d'ailleurs parfois polluées par des termes très généraux et vides de sens, c'est un autre des problèmes techniques de la méthode. Notons par ailleurs qu'un mot peut *a priori* figurer dans plusieurs listes avec différentes probabilités.

Une autre limite que nous avons identifié avec Alix Rule et Peter Bearman dans notre travail sur les discours de l'État de l'Union tient à la nature de modélisation dynamique permise par les topic models (Rule et al., 2015). Certes, comme on l'a déjà précisé, il est possible de complexifier le modèle graphique de départ (figure 1.20) pour rajouter des variables complémentaires dont les topics vont dépendre (Rosen-Zvi et al., 2004), le temps pouvant être l'une d'entre elles (Wang et al., 2012). L'algorithme ToT en est un exemple parfait (Wang et McCallum, 2006). Il permet d'identifier des clusters de termes dont la présence dans un corpus daté varie au cours du temps. Pour autant, ToT est incapable de capturer la façon dont la composition d'un ensemble de termes formant un sujet à un moment donné se transforme au fil du temps. Et si le modèle de Blei et Lafferty (2006) cDTM (continuous Dynamic Topic

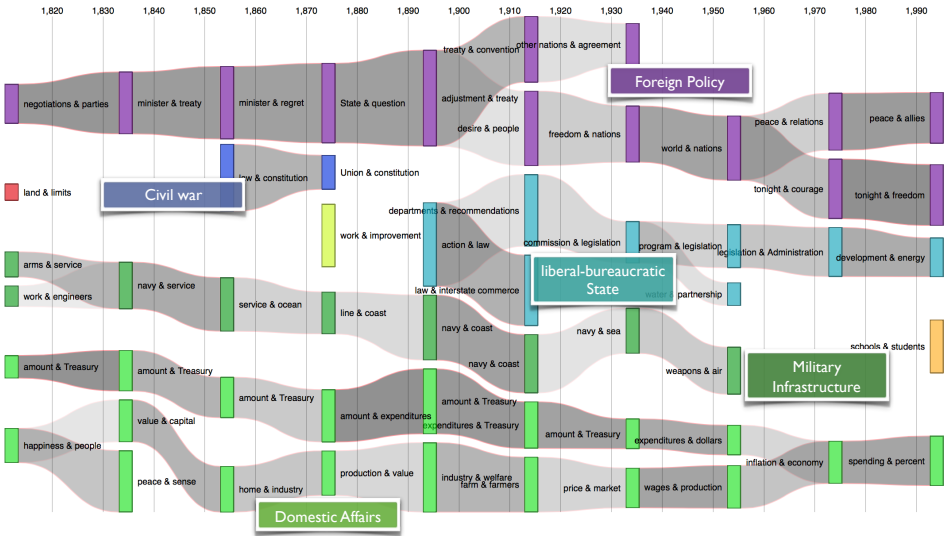


FIGURE 1.13: Le diagramme alluvial de l'histoire des discours de l'État de l'Union. Les clusters (représentés par les rectangles de couleur) émergent de l'analyse du réseau de proximité sémantique des 300 mots les plus fréquents calculé tous les 20 ans. Les relations inter-temporelles entre périodes adjacentes sont indiquées par les flux grisés dont l'opacité est proportionnelle à la stabilité du cluster. La topologie créée par ces événements dynamiques permet d'induire des « rivières discursives » correspondant aux grandes tâches de la gouvernance américaine qui ont traversé les siècles (étiquetées manuellement). De haut en bas on distingue, (en vert clair) une lignée discursive d'ordre économique qui, depuis le milieu du XX^{ème} siècle a vu une conversation sur la politique fiscale d'une part et les questions de production industrielle et agricole d'autre part fusionner, (en vert) un fil de discussion sur l'infrastructure militaire essentiellement navale qui s'est achevée à l'issue de la deuxième guerre mondiale, (en bleu clair) une discussion sur la régulation et le financement des infrastructures publiques qui persiste actuellement sous la forme d'une discussion sur l'État Providence, (en violet) la politique étrangère, de nature exclusivement bilatérale aux XVIII^{ème} et XIX^{ème} siècle, s'internationalise au XX^{ème} siècle avant de se diviser à l'époque contemporaine en une conversation sur la sécurité intérieure d'un côté, et le rôle de régulation des États-Unis dans le monde en tant que super-puissance de l'autre. Des discussions temporellement plus bornées ont aussi structuré les discours de l'État de l'Union comme celle sur le contrôle du territoire intérieur (en rouge), celle sur la politique sociale (en orange), ou la guerre de sécession (en bleu foncé).

Modeling) a cette ambition, il requiert à nouveau de fixer le nombre de topics à l'avance dans une super-structure thématique entièrement figée qui contraint les thématiques (scientifiques, le corpus des articles publiés dans la revue *Science* tout au long du XX^{ème} siècle est analysé) à évoluer linéairement (il existe un topic pour les neurosciences, un autre pour la physique atomique, mais aucune chance n'est laissée à un nouveau topic qui couplerait les deux d'émerger).

Gao et al. (2011) faisant le constat de ces limites de LDA (« these methods do not reveal the dynamics and interconnections among the detected topics » [46]) semblent être les seuls à avoir proposé une solution qui reconstruit des évolutions de clusters en incluant des événements dynamiques critiques tels que la naissance de nouveaux topics, leurs divisions, fusions ou disparition. Dans leur étude des dépêches de presse sur Obama collectées sur Bing durant 16 jours, ils proposent une solution assez technique à base de processus de Dirichlet hiérarchiques (HDP), qui les oblige, une fois le processus dynamique reconstruit à ajouter une phase d'analyse fondée sur les cooccurrences entre mots pour donner réellement corps au contenu des topics et rendre interprétable la dynamique médiatique. La stratégie que nous avons appliquée aux discours de l'État de l'Union est à la fois bien plus parcimonieuse et transparente. Elles s'appuie sur le calcul du réseau de proximité sémantique entre mots pour construire des clusters thématiques (qu'on décrira plus tard, section 2.2). La structure de co-occurrence entre couples de mots, absente des topic models, est directement exploitée pour reconstruire les clusters sémantiques mais aussi leur connexions inter-temporelles. La représentation des clusters dynamiques sous forme de diagramme alluvial et leur interprétation est alors immédiate [47].

46. « ces méthodes ne peuvent pas révéler les dynamiques et les connexions entre thématiques détectées »

47. Une visualisation interactive (http://bit.ly/2ioWI1Y) permet d'interroger la composition de chaque cluster mais aussi les transformations subies par les clusters à chaque changement de période

1.2.2 *Modèles de plongement de mots*

Les modèles de plongement de mots (« Word embedding ») rassemblent toute une série d'algorithmes récents qui assignent à chaque mot d'un corpus un vecteur dans un espace de quelques centaines de dimensions. Les mots dont le sens est similaire se retrouvent proches dans cet espace (du point de vue de la mesure du cosinus). Par similaire, on entend en réalité qu'ils sont interchangeables sans que l'énoncé ne perde en plausibilité. Il peut s'agir de synonymes ou d'antonymes, mais aussi d'une couleur qui en remplace une autre, un jour qui est équivalent à un autre jour, etc. En effet, on se retrouve là dans un paradigme distributionnel : le sens d'un terme résulte entièrement du profil statistique de ses contextes d'apparition (voir section 2.2 au chapitre suivant).

Par rapport à des modèles antérieurs comme l'analyse sémantique latente (voir section précédente 1.2.1) ou le clustering de Brown (Brown et al., 1992) qui produisent également un espace vectoriel dans lequel les mots se retrouvent « projetés », les modèles de plongement se distinguent avant tout par leurs excellentes performance sur un certain nombre de tâches. Plus précisément, les modèles de plongement de mots sont devenus très populaires depuis quelques années grâce à leur performance sur des tâches classiques d'estimation de la similarité entre deux mots ou de recherche d'analogie. Dans les tâches de similarité, la capacité des algorithmes à estimer correctement la similarité entre deux termes est comparée à une évaluation humaine (des corpus spécialisés compilent ce type de score[48]). Pour tester les analogies, l'algorithme doit répondre à des questions du type : « A est pour A* ce que B est pour B* ». Typiquement des corpus[49] listent des quadruplets du type : « PARIS est pour la FRANCE ce que TOKYO est pour Le JAPON » (Levy et al., 2015). Paris, France et Tokyo sont les entités soumises à l'algorithme qui doit répondre Japon. Mais de façon plus cruciale encore, les représentations vectorielles apprises avec ces modèles ont été incorporées comme « ingrédient magique » (Luong et al. (2013) parlent de « secret sauce ») à des algorithmes classiques de traitement automatique de la langue. Les résultats ont été spectaculaires et des gains de performance importants ont été obtenus pour un grand nombre de tâches : analyse de sentiment (Dos Santos et Gatti, 2014), analyse des grammaires de dépendance (Chen et Manning, 2014), etc.

48. WordSim Similarity, SimLex-999, etc.

49. Google's analogy dataset, MSR's analogy dataset

On a un temps cru que l'on avait affaire à un nouveau type d'approche car word2vec, la première famille de ces algorithmes à avoir obtenu des performances exceptionnelles, apprenait les positions des vecteurs grâce à un réseau de neurones à deux couches, ce qui semblait le distinguer des modèles traditionnels de sémantique distributionnelle. Pour autant on a depuis montré que sous certaines conditions de paramétrage (modèle skip-gram avec échan-

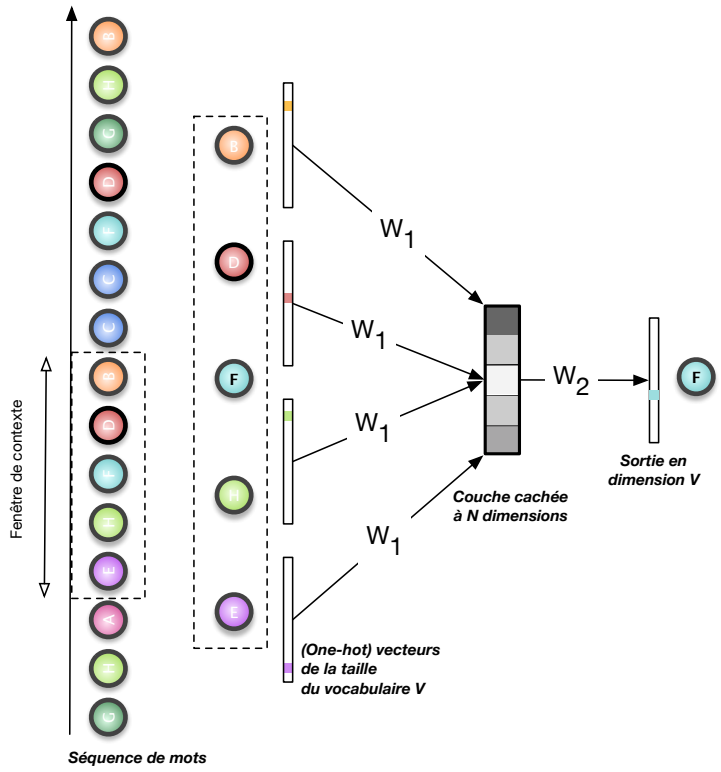


FIGURE 1.14: Modèle CBOW, les matrices W_1 et W_2 (de tailles respectives $\mathcal{V} \times N$ et $N \times \mathcal{V}$) sont apprises par le réseau de neurones qui, étant donné les mots (ici *B*, *D*, *E* et *H*) apparaissant dans le contexte d'un mot donné (*F*), doit prédire qu'il s'agit bien du mot en question (*F*).

tillonnage négatif (SGNS)), word2vec entretient en réalité des liens très forts avec des méthodes plus classiques telle que la factorisation matricielle de la matrice d'information mutuelle (Levy et Goldberg, 2014) qui est à une translation près l'opération déjà réalisée par l'analyse sémantique latente. Glove (Pennington et al., 2014), un algorithme développé un an après word2vec, propose une alternative compétitive en se fondant sur une procédure de calcul plus classique qui ne fait pas appel à des modèles neuronaux.

Au fond, les raisons qui expliquent les performances de cette classe de méthodes ne sont pas encore très bien comprises. On sait qu'elles ne sont pas nécessairement liées à l'utilisation d'un modèle particulier (Levy et al., 2015) mais semblent particulièrement sensibles aux (nombreux) paramètres du modèle. On mentionnera outre la dimension et la taille du contexte : le sous-échantillonnage des mots fréquents, l'utilisation d'une fenêtre de contexte dynamique (qui donne plus de poids aux mots immédiatement proches qu'aux mots plus distants), etc. Autrement dit, ces modèles offrent des performances inédites, mais sans que l'on comprenne très bien ce qui produit cette plus-value qualitativement. On fera abstraction de cette incertitude et on essaiera plutôt d'imaginer ce que les propriétés des plongements de mots pourraient offrir comme applications possibles aux sociologues.

Mais avant de nous intéresser à ces applications potentielles, on décrit dans les paragraphes suivants le fonctionnement du modèle phare des plongements de mots : word2vec. Si Bengio et al. (2003) ont les premiers introduit l'idée d'une architecture à base de réseaux de neurones pour « apprendre » des vecteurs sémantiques, les modèles CBOW et skip-gram (Mikolov et al., 2013b,a) ont grandement simplifié les choses en 2013. Ces modèles ont également bénéficié d'une diffusion rapide grâce à la librairie word2vec partagée par Google [50]. Le fonctionnement du modèle CBOW (Continuous Bag-of-Words) est illustré figure 1.14. Il s'agit simplement d'un réseau de neurones à deux couches avec une couche cachée dont le nombre de neurones N est défini par la dimensionnalité du plongement vectoriel final souhaité (typiquement de l'ordre de 100). Le modèle CBOW vise à prédire un mot (F dans notre exemple) en fonction des termes qui l'entourent (E,H,D et B). En entrée chaque mot du contexte est représenté comme un vecteur de dimension égale à la taille totale du vocabulaire notée $\mathcal{V}$. C'est un encodage de type « one-hot » : les coordonnées d'un mot valent 0 partout sauf à l'emplacement même du mot où le vecteur vaut 1. Ces « one-hot vecteurs » sont multipliés par la matrice W_1 (de taille $\mathcal{V} \times N$) qui transforme les vecteurs d'entrée en leur version réduite à N dimensions avant de les sommer. Enfin, le vecteur qui résulte de cette opération est à nouveau transformé par une dernière multiplication par la matrice W_2 (de taille $N \times \mathcal{V}$). Le vecteur final, de la dimension du vocabulaire, doit permettre d'inférer le mot le plus probable, à savoir F.

50. On remarquera que Tomas Mikolov, qui a développé word2vec lorsqu'il travaillait pour Google a depuis été embauché par Facebook et ce en dépit d'une confiance toute relative dans l'application des réseaux profonds à convolution pour des applications en TAL.

Les colonnes (respectivement lignes) de la matrices W_1 (respectivement W_2) fournissent les représentations des mots du vocabulaire dans l'espace réduit. En pratique on se contente souvent de la représentation fournie par la matrice d'entrée W_1. Le modèle alternatif skip-gram est le symétrique exact de CBOW, il consiste à optimiser un réseau de neurone dont la fonction est de prédire le contexte connaissant le mot central (prédire B, D, E et H sachant F).

Word2vec modélise un terme à partir de ses seuls contextes d'apparition. On retrouve à nouveau une hypothèse du sens fortement distributionnelle. Dès lors, on comprend pourquoi les mots partageant la même famille de sens (synonymes, antonymes, mais aussi couleurs, villes, etc.) sont proches dans l'espace réduit. Mais la proximité spatiale peut aussi très bien rapprocher des mots relevant du même registre de langue par exemple, tout simplement parce qu'ils co-occurrent singulièrement les uns avec les autres sous la plume des mêmes personnes...

Mais le plus simple est sans doute de montrer par des exemples le type d'information que capture cette mesure de proximité. Nous avons ainsi testé ces modèles de plongement de mots sur le corpus de discours de l'État de l'Union. L'exemple nous semble intéressant puisque ces méthodes ont la réputation de nécessiter des corpus de données de très grande taille [51]. Malgré

51. Dans la plupart des évaluation du modèle, les vecteurs sont d'ailleurs appris sur des corpus tels que Wikipedia, ou google 5gram (chaque 5 gram consistant en une fenêtre de contexte pour le terme central).

FIGURE 1.15: La carte générale est composée des 7349 termes des discours de l'État de l'Union dont la fréquence est supérieure à 5. Un algorithme de réduction de dimensionnalité de type t-SNE a été utilisé pour projeter les plongements de mots (originellement appris avec un modèle CBOW en 100 dimensions et sur une fenêtre de contexte de 10 mots) en deux dimensions. Les étiquettes thématiques en grisé ont été rajoutées manuellement.

la modestie du corpus (à peine 1 763 622 mots si on prend en compte les discours jusqu'en 2015), les voisinages obtenus sont tout à fait raisonnables, et même après une nouvelle réduction en 2d avec l'algorithme t-SNE qui tente de préserver les distances locales (Maaten et Hinton, 2008), les positions des mots semblent cohérentes. La figure 1.15 montre ainsi les 50 principaux voisins du mot *land* , et leur localisation dans une représentation réduite de l'intégralité des 7349 mots [52] les plus fréquents des discours de l'État de l'Union. On observe que les plus proches voisins de *land* renvoient à des univers thématiques connexes mais légèrement différents comme : la question des frontières (*river*, *territory*, *shore*, *possession*, etc.), de la propriété (*ownership*, *private*, *public domain*, *title*, etc.), ou des ressources (*soil*, *mine*, *natural ressource*, etc.). Par définition, les plongements de mots visent à assigner à chaque mot une position unique. Mais on peut aisément contourner le caractère atomiste du modèle. On voit ainsi qu'il suffit de regarder les voisinages d'un terme et

52. Pour être précis, le vocabulaire a été construit après avoir passé le texte en minuscule et identifié les bigrammes (*public domain* et trigrammes (*within their limit*). Seuls les entités apparaissant plus de 5 fois ont été conservé dans le vocabulaire final.

la distribution spatiale de ce voisinage (c'est à dire les similarités existantes entre voisins) pour saisir son caractère polysémique, ou en tout cas capturer les univers sémantique distincts auxquels il renvoie.

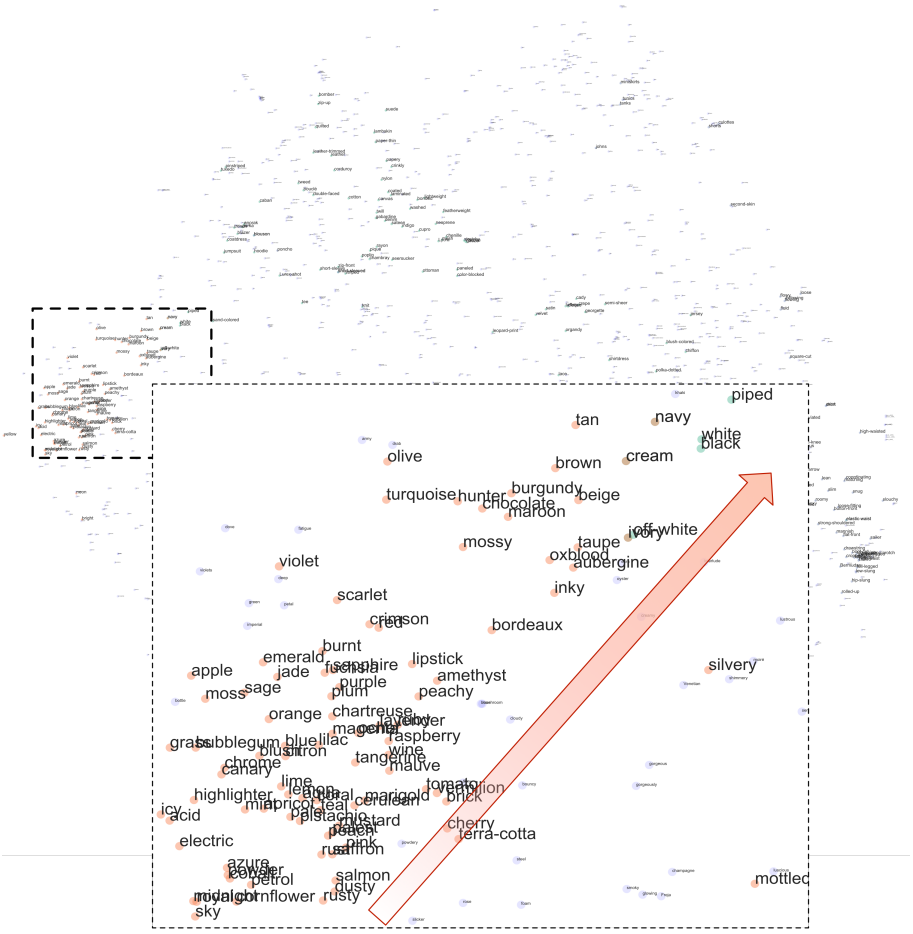


FIGURE 1.16: Plongement sémantique des 2000 mots les plus fréquents (modèle original entraîné sur 9878 mots au total) de la base de données des comptes-rendus de défilé de Vogue projetés en 2d grâce à t-SNE. La carte intégrale (contenant tous le vocabulaire) est consultable en suivant ce lien : http://bit.ly/2lL6aCY). La version réduite peut également être consultée sans les inserts mais en haute résolution à l'adresse suivante : http://bit.ly/2n1PV5w. Les trois ensembles colorés correspondent au champs lexicaux des couleurs (également présenté en encart), des coupes/styles (figurant en turquoise au centre sur la carte : *ankle-grazing*, *fluid*, *below-the-knee*, *fitted*, *flaring*, *loose*, *narrow*, *second-skin*, etc), et des matières (en gris à droite sur la carte : *silk*, *velvet*, *viscose*, *neoprene*, *chenille*, *featherweight*, *papery*, etc.) qui ont été identifiées automatiquement.

Dans le cadre d'un autre projet, un nouveau réseau de neurones a été entraîné sur un corpus de plus de 12 000 compte-rendus de défilés de mode parues dans Vogue depuis 15 ans. Après avoir calculé la position des mots du corpus, nous avons simplement demandé au modèle de nous indiquer les 50 mots les plus proches d'un vecteur moyen constitué des trois couleurs primaires *red*, *yellow* et *blue* . Comme escompté les termes qui émergent décrivent d'autres couleurs (*yellow* , *navy* , *orange* , *off-white*) des ambiances lumineuses (*sky* , *electric* , *acid*) et autres nuances liant couleurs et textures (*inky* , *mottled* , *burnt*). Une représentation de l'ensemble du vocabulaire est présentée figure 1.16. Autre propriété remarquable, s'il est possible d'extrapoler tous les éléments d'une classe à partir de quelques exemples (à proximité de deux villes, on trouvera d'autres villes, une année est proche d'autres années, etc.), les éléments ainsi identifiés ne se déploient pas de façon isotrope dans l'espace.

Ils se structurent selon des directions privilégiées. Ainsi, partant des seules couleurs identifiées, on pourrait reproduire leur structure propre, mais on voit déjà sur la carte globale (figure 1.16) qu'elles s'organisent selon un axe allant des couleurs plus hivernales (partie supérieure droite : *cream*, *brown*, *aubergine*) aux couleurs les plus estivales (partie inférieure gauche : *azure*, *salmon*, *canary*)

Plus étonnant, certaines relations sémantiques entre concepts semblent déjà endogénéisées dans les positions relatives des mots dans l'espace. Des relations d'analogie (appelées aussi similarités relationnelles) peuvent notamment être reconstruites directement à partir d'opérations vectorielles. Ainsi si on effectue l'opération (vectorielle) *roi* − *homme* + *femme* on se retrouve dans un point de l'espace très proche du vecteur *reine*, *Obama* + *Russia* − *USA* et le vecteur *Putin* sont également très proches [53]. Ces analogies peuvent être de différents types : syntaxique (singulier/pluriel, masculin/féminin, présent/passé), mais aussi sémantique. Ainsi les capitales peuvent être aisément reconstruites en effectuant une simple translation depuis le nom des pays, ainsi que les PDG depuis le nom des entreprises qu'ils dirigent. Des relations plus complexes sont également possibles du type : la viande est au boucher ce que le pain est au boulanger, etc.

Ces opérations vectorielles offrent déjà un certain nombre d'applications. Ainsi les stéréotypes racistes ou sexistes ressurgissent dans le discours et peuvent être révélés quantitativement. Caliskan-Islam et al. (2016) montrent ainsi combien un certain nombre de préjugés culturels, raciaux ou de genre sont déjà inscrits dans la structure du langage. Joseph et al. (2017) explorent la structure des stéréotypes détectables sur Twitter. À titre expérimental, nous avons, à nouveau sur le jeu de données des discours de l'État de l'Union, cherché à mesurer le produit scalaire entre une série de concepts et le vecteur élémentaire « good - bad » qui peut s'apparenter à l'échelle archétypale de l'analyse de sentiments. Sans rentrer dans les détails, nous pouvons mesurer cette projection à différentes périodes durant les deux derniers siècles (cf. figure 1.17) pour mesurer l'évolution de la positivité/négativité des mots au cours de l'histoire [54]. On peut tirer les observations suivantes. La guerrre est connotée très négativement, sauf à l'approche de la seconde guerre mondiale, alors que Roosevelt s'efforçait de convaincre les américains de la nécessité pour les États-Unis de rentrer dans le conflit. L'emploi, tout comme la protection sociale, sont au contraire connotés plutôt positivement avant de subir une dépression importante dans les années 80. C'est sans doute l'effet d'un discours néolibéral de détricotage de l'État providence que l'on observe ici et la crise économique de 2008 ne fait qu'accentuer le caractère pessimiste des discours sur l'emploi. L'adjectif *foreign* semble opérer une transition de phase à l'orée du XX$^{\text{ème}}$ siècle où il devient fortement positif, ce qui est compatible avec le changement de doctrine de la politique étrangère américaine amorcée par la politique interventionniste de Wilson dans les années 1910.

53. Pratiquement, la mesure du cosinus est utilisée pour trouver les mots les plus proches du vecteur roi - homme + femme mais Levy et al. (2014) ont montré que d'autres mesures sont possibles et sont même plus efficaces pour identifier des analogies comme COSmul. Ils montrent d'ailleurs également que des modèles beaucoup plus simples de type mots-contexte (on parle alors d'une représentation vectorielle explicite) dans des espaces de très hautes dimensions (et en calculant le poids de chaque composante en calculant une information mutuelle exactement à la manière de l'une des distances distributionnelles que l'on introduit dans le chapitre suivant !) permettent de reconstruire des similarités relationnelles aussi efficacement que les modèles de plongement en moindre dimension.

54. On notera que sans faire appel à des modèles de plongements de mots récents, mais en usant de modèles d'analyse sémantique latente traditionnels, Sagi et Dehghani (2014) utilisent la même idée pour mesurer les 5 dimensions morales d'un texte se distribuant selon les oppositions : *care/harm*, *fairness/cheating*, *loyalty/betrayal*, *authority/subversion*, *purity/degradation*).

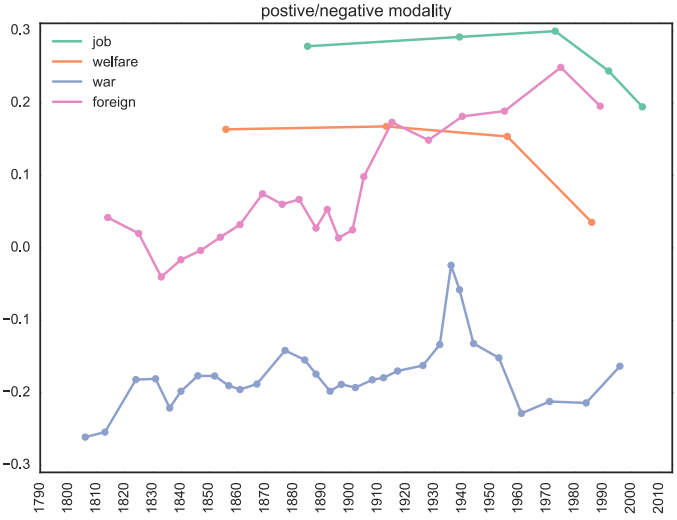


FIGURE 1.17: Modalité des termes *job*, *welfare*, *war* et *foreign* mesurée en calculant le produit scalaire entre chacun des mots (ou pus précisément son vecteur durant une période donnée) avec le vecteur *good* − *bad*. Un nouveau point est tracé toutes les N occurrences.

Alors que certains groupes de recherche mêlant spécialistes de l'apprentissage et sociologues ou historiens investissent cette nouvelle voie prometteuse aux États-Unis, il est encore trop tôt pour prédire comment ce type de modèle sera pris en charge par la sociologie ou les sciences politiques. Les plongements lexicaux rappellent fortement les modèles géométriques de l'analyse des correspondances de Benzécri (voir section 1.1.3) mais ils s'en distinguent d'au moins deux façons. Premièrement, l'objectif est certes de produire un modèle du sens géométrique, mais si l'analyse des correspondances espère concentrer la variance dans deux dimensions, les modèles de plongement jouent clairement dans une autre catégorie et réduisent les mots à des vecteurs denses mais dont la dimension est tout de même importante (100, 200 voire 500 dans certains cas). La représentation géométrique n'est donc pas l'objectif premier [55]. Il s'agit en premier lieu d'aboutir à une représentation vectorielle qui soit la plus efficace possible pour reconstruire des similarités sémantiques entre objets. Par ailleurs, les plongements de mots ne construisent pas non plus de classes à la manière des topic models. Il est néanmoins théoriquement possible après plongement des mots d'utiliser un algorithme de clustering pour construire des catégories sémantiques. On peut à titre d'exemple consulter ce travail en cours qui montre une visualisation interactive du corpus du New York Times (comprenant un résumé de tous les articles publiés depuis 1900) dont la clusterisation avec l'algorithme HDBScan (Campello et al., 2013) fournit des classes sémantiques raisonnables. Enfin, ces modèles s'appuient sur un modèle de language pour le moins minimal. Comme on l'a vu, les modèles word2vec ne s'appuient pas sur des considérations théoriques très profondes vis-à-vis de l'acte d'énonciation mais résultent d'un raisonnement de nature purement ingénierique qui vise à optimiser la prédiction d'un mot connaissant son contexte.

55. Si le recours à un algorithme de réduction de dimensionnalité est fréquent, il ne fait nullement parti de la méthode originale.

Pour autant, il est possible que l'efficacité de ces modèles ne soit pas entièrement fortuite. La métaphore de l'apprentissage est prise au sérieux par certains qui souhaitent exploiter des hypothèses cognitives pour améliorer les modèles de plongement de mots. Hu et al. (2016) s'appuient ainsi sur des résultats de neuropsychologie de l'apprentissage des langues chez l'enfant pour et proposent d'intégrer des informations grammaticales pour améliorer les modèles de plongement de mots tandis que Ruan et al. (2016) pensent pouvoir évaluer des relations de similarité sémantique en imageant l'activité neuronale. Des travaux récents, qui prolongent des tentatives plus anciennes de faire se correspondre structures d'activité neuronale et espace sémantique latent (Mitchell et al., 2008), observent une homologie structurelle entre plongements sémantiques et activité neuronale d'individus soumis à des stimuli visuels (Güçlü et van Gerven, 2015, *via* IRM) ou en train de lire (Wehbe et al., 2014, *via* MEG).

En tout état de cause, les relations d'analogie qui sont reconstruites par ces modèles constituent encore un mystère pour les théoriciens et susciteront sans doute nombre d'applications dans les années à venir. Il faut également noter que le modèle géométrique des plongements de mots ouvre des perspectives passionnantes pour mesurer la vitesse d'évolution du sens des mots et ainsi interroger les processus d'évolution culturelle ou de glissements sémantiques (Kulkarni et al., 2015; Hamilton et al., 2016). Nous présenterons une application possible pour détecter des périodes saillantes dans un corpus historique, section 2.1.4 puis comme outil de délinéation de corpus dans le troisième chapitre, section 3.3.3.

1.2.3 *Autres méthodes : de l'analyse de sentiment à la lecture distante*

Les deux sections précédentes n'épuisent pas, loin de là, l'ensemble des approches contemporaines. Cette section vise à les présenter brièvement pour au moins donner une idée de leur diversité. On mentionnera donc dans cette section les méthodes digitales venant de la DMI (Digital Methods Initiative), l'analyse de sentiments (appelée aussi de façon quasiment équivalente analyse d'opinion), l'analyse stylistique plus balbutiante mais dont les premiers résultats sont déjà très prometteurs, le renouveau des approches lexicométriques promues par les Google ngrams, mais encore la lecture distante de corpus littéraire de Franco Moretti, etc.

Le DMI (« Digital Method Initiative ») est un groupe de recherche constitué à l'université d'Amsterdam dans le département de Media Studies autour de Richard Rogers. Leur recherche, dès le début des années 2000[56], interroge le web comme un nouveau média. Le web devient un objet d'étude à part

56. Par exemple, le « issuecrawler », outil de crawling de sites web, a été développé en 2004 (McNally, 2005; Bruns, 2007), il permettait grâce à une interface en ligne de construire un corpus web et de l'analyser directement en ligne.

entière ce qui appelle ainsi au développement de nouvelles méthodes pour les sciences sociales adaptées à ces données nativement numériques (« natively digital ») (Rogers, 2009).

Ils introduisent notamment le concept de « re-purposing » (Marres, 2012) qui propose de re-distribuer les dispositifs des médias sociaux en un outil d'enquête des dynamiques sociales et politiques. Cette approche est très proche théoriquement et en pratique des STS, de la théorie de l'acteur-réseau et plus précisément de la cartographie des controverses. Malgré son ancienneté et la pléthore d'outils qui ont été développés par la Digital Method Initiative [57], il est difficile de dégager un socle méthodologique commun de l'aveu même de ses fondateurs (Marres, 2015b). Ces outils, dont une caractéristique importante est la modularité, empruntent aussi bien à l'analyse des mots-associés, à la visualisation de graphe qu'à des applications de collecte de données en ligne. Actuellement, une partie importante de ces outils visent en effet à construire des jeux de données en simplifiant l'utilisation des APIs les grandes plateformes du web (notamment Youtube, Twitter ou Wikipedia) pour enquêter sur des processus sociaux ou politiques.C'est là une application directe du concept de repurposing que nous aurons l'occasion de ré-interroger dans le dernier chapitre (voir section 3.2.3).

57. Leurs seuls noms (« lippmanian device », « issue crawler », etc.) témoignent d'ailleurs de la prééminence du cadre théorique de l'ANT dans leur conception

Un autre type d'analyse de contenu textuel est l'analyse de sentiment (*sentiment analysis*) parfois confondue avec l'analyse d'opinion ((*opinion mining*). Les promesses d'application liées à une telle lecture des traces sociales du Web 2.0 ont soutenu une recherche très forte dans le domaine. Malheureusement, les résultats ont souvent déçu, notamment car nombre de modèles s'appuient en réalité sur de simples lexiques confondant de naïveté (Boullier et Lohard, 2012). Le domaine de recherche reste néanmoins très actif, et incorpore progressivement des modèles syntaxiques plus complexes (Duric et Song, 2012; Agarwal et al., 2011), intégrant des modèles sémantiques (Titov et McDonald, 2008) et capturant des échelles d'émotions plus riches (Kim et al., 2013; Gonçalves et al., 2013) que le modèle dichotomique classique distinguant sentiment positif et négatif. D'autres recherches visent à mieux comprendre au sein de la phrase les entités nommées sur lesquelles des jugements sont portés (Ruiz et Poibeau, 2015).

Disposer d'un moyen de mesurer l'attitude d'un locuteur par rapport à un sujet ou mesurer la polarité globale d'un document, voilà des outils d'analyse qui pourrait intéresser un sociologue désireux de décoder un corpus d'expressions publiques en ligne par exemple. C'était d'ailleurs une modalité de codage de l'analyse de contenus américaine dès les années 50 si on s'en réfère à Lasswell, Lerner, et de Sola Pool (1952, 37) :

> « *The word direction in content analysis refers to the attitude expressed toward any symbol by its user. Such expressions of attitude are usually categorized as favorable,*

unfavorable, or neutral. Various or related polarities—e.g., positive-negative, friendly-hostile—are sometimes used. The actual recording operations usually use the symbols + (plus), − (minus), and 0 (zero) for the three categories. It is important that the rules for classifying individual observations into these three categories should be very clear and explicit in order to maximize reliability in recording the data and validity in the inferences from data to conclusions. »[58]

Si l'exercice est déjà complexe pour un codage manuel, il est extrêmement ardu pour la machine qui face à des contenus courts et un vocabulaire moins classique ont toutes les chances d'échouer. Pour autant l'analyse de sentiment est particulièrement utilisée pour l'analyse de tweets, mais toujours de façon relative, afin de calculer des tendances ou d'établir des comparaisons, tant la précision des méthodes à l'échelle d'un exemple est aléatoire. La raison de cette difficulté est relativement simple et bien connue : mesurer les sentiments requière une modélisation de la sémantique et des aspects figuratifs du langage (et notamment de l'ironie qui est un élément critique pour résoudre cette tâche (Bosco et al., 2013; Ghosh et al., 2015)), etc.

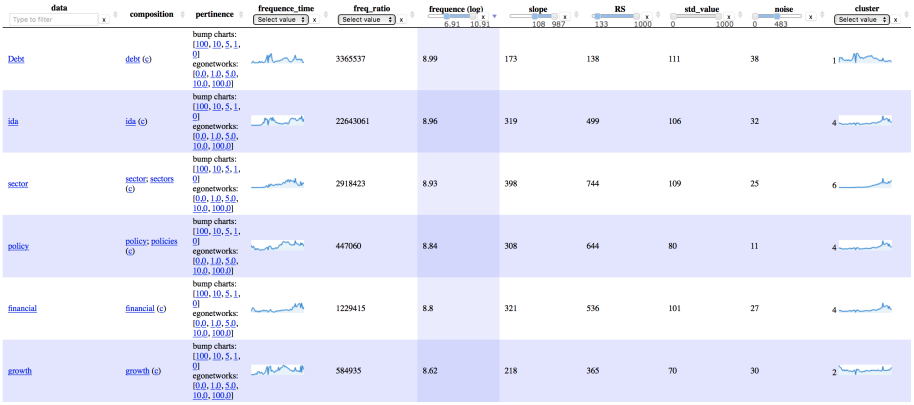
Moretti (2004) défend dans le champ de la littérature[59] le concept de lecture distante (« distant reading »). Non sans provocation il appelle ses collègues à cesser de lire pour enfin saisir dans son entièreté la nature de la littérature victorienne. La lecture distante relève avant tout d'une approche particulière, d'une certaine attitude par rapport à l'analyse de corpus littéraires. À proprement parler les méthodes qu'il met en œuvre sont assez classiques : il s'agit essentiellement de techniques lexicométrique ou d'analyse de réseaux sociaux liant les acteurs d'une pièce de théâtre par exemple (Moretti, 2011).

Plus récemment, il a appliqué la même stratégie à la lecture des rapports annuels de la Banque mondiale et décrit les évolutions du « répertoire langagier » de l'institution (Moretti et Pestre, 2015). « Bankspeak » décrit ainsi les évolutions stylistiques (augmentation des figures de « nominalisation » par exemple), lexicales (émergence d'un vocabulaire spécifiquement lié au management) ou syntaxiques (augmentation du nombre de noms par rapport au nombre de verbes) majeures de l'institution depuis sa création en 1946. L'article parvient à caractériser des transformations de fond de la doctrine politique de la Banque mondiale grâce à un travail très fin sur les catégories qui, à la manière des formules prospériennes, sont extrêmement malléables : l'analyste peut librement émettre en suivant ses intuitions des hypothèses sur la prolifération ou la disparition d'une forme réthorique dont le profil d'évolution est alors calculé automatiquement. Par exemple après avoir tracé le profil d'évolution de certains termes savamment sélectionnés, Moretti et Pestre analysent la dynamique des formes acronymes dans le corpus. On voit là combien les analystes font ici équipe avec la méthode d'analyse textuelle : ils l'interrogent, se laissent surprendre, reformulent leurs hypothèses jusqu'à extraire des motifs remarquables. Les classes sémantiques et syntaxiques qui

58. « La direction d'un terme en analyse de contenu réfère à l'attitude exprimé à l'égard d'un symbole par le locuteur. De telles expressions d'attitude sont généralement catégorisées comme favorables, défavorables ou neutres. D'autres échelles de polarité sont parfois employés : positif-négatif, amical-hostile. Le codage effectif de ces opérations s'appuie normalement sur les codes + (plus), − (moins) et 0 (zéro) pour chacune des trois catégories. Il est important que les règles de classification des observations individuelles dans ces trois catégories soient très claires et explicites et ce afin de maximiser la fiabilité du codage des données et la validité des inférences qui seront faites des données vers les conclusions. »

59. Déjà Krippendorff (2004) évoque les études littéraires comme le premier espace académique où des questions d'analyse quantitative de contenu textuel ont été débattues. Plus précisément au XVIII^{ème} siècle, une large controverse publique et académique a éclaté pour déterminer si une série d'hymnes anonymes (« Songs of Zion ») très populaire était nocive à l'église orthodoxe suédoise. Repérage de symboles, discussion de leur sens dans leur contexte d'apparition, comparaison avec des corpus de chansons officiels, toujours d'après Krippendorff (2004) de nombreuses idées et techniques en analyse de contenus sont nées à cette époque.

sont énumérées ne sont pas closes *a priori* mais ouvertes à la créativité des "lecteurs".



data	composition	pertinence	frequence_time	freq_ratio	frequence (log)	slope	RS	std_value	noise	cluster
Debt	debt (c)	bump charts: [100, 10, 5, 1, 0] egonetworks: [0.0, 1.0, 5.0, 10.0, 100.0]		3365537	8.99	173	138	111	38	1
ida	ida (c)	bump charts: [100, 10, 5, 1, 0] egonetworks: [0.0, 1.0, 5.0, 10.0, 100.0]		22643061	8.96	319	499	106	32	4
sector	sector; sectors (c)	bump charts: [100, 10, 5, 1, 0] egonetworks: [0.0, 1.0, 5.0, 10.0, 100.0]		2918423	8.93	398	744	109	25	6
policy	policy; policies (c)	bump charts: [100, 10, 5, 1, 0] egonetworks: [0.0, 1.0, 5.0, 10.0, 100.0]		447060	8.84	308	644	80	11	4
financial	financial (c)	bump charts: [100, 10, 5, 1, 0] egonetworks: [0.0, 1.0, 5.0, 10.0, 100.0]		1229415	8.8	321	536	101	27	4
growth	growth (c)	bump charts: [100, 10, 5, 1, 0] egonetworks: [0.0, 1.0, 5.0, 10.0, 100.0]		584935	8.62	218	365	70	30	2

FIGURE 1.18: Interface « distant reading », profils temporels d'un millier d'expressions issues des rapports de la Banque mondiale. L'interface complète contenant les 1000 groupes nominaux principaux est consultable à cette adresse : http://bit.ly/2l9aMHF

La souplesse de construction des catégories est inhérente à l'approche *distant reading* dont l'ambition est de circonscrire le style d'une production (littéraire ou non, Bankspeak en est l'illustration). Si les outils qu'il met en œuvre sont relativement simples - après tout, il s'agit essentiellement de comptage et de quelques représentations de réseau - l'ensemble de l'édifice de Moretti s'appuie en réalité sur une théorie de la littérature et de son évolution extrêmement forte dont l'origine prend ses sources, comme l'indique le titre de son premier ouvrage sur le sujet (Moretti, 2005), sur la géographie (les cartes), la théorie de l'évolution (pour ses arbres) et l'histoire quantitative (pour ses graphes). Bien que faisant une large part à la sérendipité, L'édifice théorique sur lequel s'appuie la méthode de la lecture distante est donc extrêmement solide.

On est bien loin d'autres approches fréquentistes comme les *culturomics* (Michel et al., 2011) dont l'ambition est d'exploiter les profils d'apparition de mots ou groupes de mots (les fameux n-grammes) dans le corpus des livres numérisés par Google en adoptant une posture où la taille des corpus analysés (« high-throughput data ») cache l'absence de point de vue théorique. Le résumé de l'article est d'ailleurs sans ambiguïté quant aux hautes ambitions de leurs auteurs qui pensent poser les bases d'une « nouvelle science » [60] :

> « *We show how this approach can provide insights about fields as diverse as lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology* » (Michel et al., 2011) [61]

C'est en partant de ce modèle exploratoire de nature abductive que j'ai développé l'interface [62] *distant reading* de CorText , dont une capture d'écran est reproduite figure 1.18. Elle permet de sélectionner, trier et naviguer à travers le vocabulaire employé dans tout corpus textuel indexé temporellement (dans cette illustration, le corpus de rapports de la Banque mondiale). Elle permet

60. « a great cache of bones from which to reconstruct the skeleton of a new science »

61. « On montre que cette approche peut produire des connaissances en lexicographie, pour comprendre l'évolution de la grammaire, la mémoire collective, l'adoption de technologies, la poursuite de la célébrité, la censure et l'épidémiologie historique »

62. Un des paris de départ de cette recherche (qui n'est pas entièrement finalisée) était précisément de faciliter le travail de construction des catégories lexicales, une interface étant le moyen idéal pour l'analyste d'opérer et de visualiser différents types de regroupements à l'envi.

de sélectionner les types grammaticaux pour n'afficher que les adverbes, les verbes ou les adjectifs, trier les listes par fréquence, ne sélectionner que les termes dont le profil temporel a un coefficient de régression linéaire positif ou négatif, ou filtrer exclusivement les termes dont le profil est fortement non uniforme (fort coefficient de « burstiness »), ou appartenant à un cluster de profils temporels de telle ou telle forme (Chen et al., 2015).

Une idée structurante de l'interface est en effet de suivre un principe de regroupement des entités textuelles qui ne procède pas d'une similarité sémantique mais plutôt de la similarité ou de la dissimilarité de leur profils temporels. Au-delà de cette vision panoptique des profils d'évolution de l'ensemble de termes de la Banque mondiale, l'interface permet de multiplier les perspectives de lecture possibles (voir figure 1.19). En premier lieu, parce que Moretti et Pestre font en réalité sans cesse varier les niveaux de lecture. Mais aussi car l'ambition de cette interface est de prolonger l'expérience de *distant reading*, en rendant plus fluide et naturelle la comparaison des catégories (et idéalement leur recomposition) que l'analyste puisse d'un regard rapprocher car leurs profils d'évolution se ressemblent ou parce qu'elles sont suffisamment proches sémantiquement. Multiplier les perspectives et les points de départ pour lire les corpus, c'est également le principe défendu par Prospero même si l'on déborde ici très largement du contexte des « affaires » publiques.

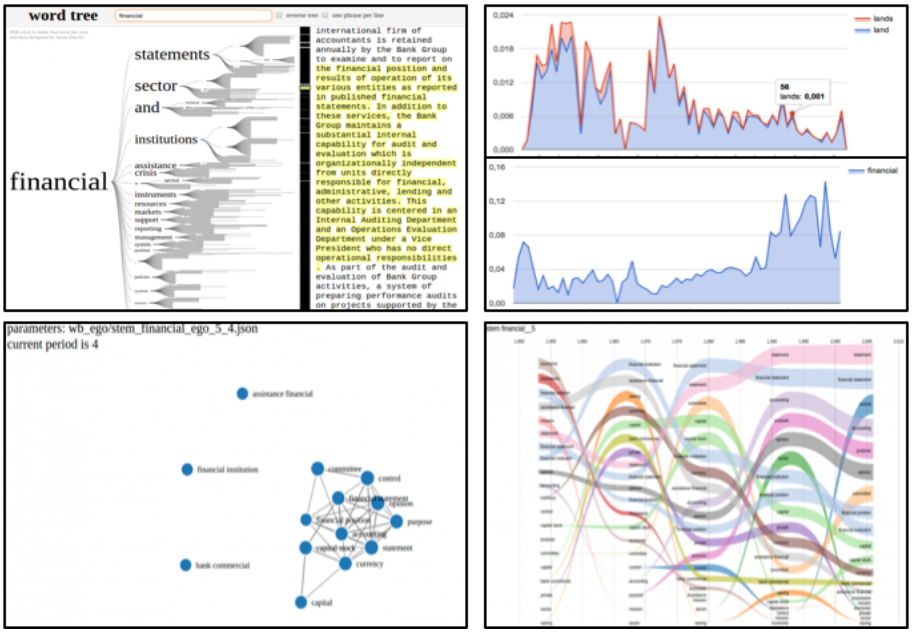


FIGURE 1.19: Interface de visualisation complémentaire de l'outil *distant reading* de CORTEXT : différentes « vues », toutes à l'échelle du mot pour permettre de consulter, l'ensemble de ses contextes d'apparition au sein d'un « treecloud » (en haut à gauche), de sa dynamique d'occurences (en haut à droite), de la structure de son réseau sémantique égo-centré (en bas à gauche), de ses plus proches voisins au fil du temps (en bas à droite)

Établir un index raisonné de l'ensemble des méthodes existantes serait fastidieux. D'une part, il serait difficile d'en fixer la limite. En effet, au delà

de la sociologie, nombre de disciplines prétendent maintenant apporter des connaissances sur le monde social et s'appuient pour se faire sur l'analyse de données textuelles. Le cas du programme de recherche *culturomics* présenté quelques paragraphes plus tôt constitue un exemple typique. Entre autres exemples de recherches provenant de ces espaces hybrides, on peut par exemple citer les travaux d'analyse stylistique qui proposent, en s'appuyant sur des approches à la frontière de la psychologie et de la sciences des données, de modéliser les relations de pouvoir entre individus et leur évolution à travers des traces textuelles d'interactions (Danescu-Niculescu-Mizil et al., 2012), ou d'interroger le couplage des dynamiques de diffusion d'innovation linguistiques et d'évolution de communautés en ligne (Danescu-Niculescu-Mizil et al., 2013).

D'autre part, certaines directions sont encore peu investies. C'est notamment le cas de l'analyse narrative quantitative dont l'ambition comme le soulignent Franzosi et al. est de « se concentrer de façon systématique sur les acteurs, leurs actions et de façon critiques, les situations spatio-temporelles dans lesquelles se situent leurs interactions »[63], . De nombreux auteurs ont proposé un traitement de la structure narrative des textes. Bearman et Stovel (2000) représentent et comparent des récits biographiques sous la forme de réseaux narratifs. Dans un registre plus léger Jurafsky et al. (2014) s'appuie sur différentes dimensions langagières (du temps auquel les verbes sont conjugués aux sentiments véhiculés par le vocabulaire en passant par le type de pronoms utilisés) pour qualifier la structure narrative de critiques de restaurant en ligne. Autre exemple : l'analyse des biais médiatiques qui relève typiquement des sciences politiques, bien qu'ancienne dans ses questionnements (Gentzkow et Shapiro, 2006) redevient un sujet de recherche à l'heure où les réseaux sociaux sont accusés de créer des bulles informationnelles (Barberá et al., 2015).

63. « focusing systematically on actors, their actions, and, criti- cally, their spatio-temporally situated interactions »

1.3 *Typologie générale*

1.3.1 *Les grandes étapes des méthodes d'analyse de contenu*

Fort de cette description quasi-systématique des principales méthodes d'analyse de texte en sciences sociales, nous pouvons désormais essayer d'en inférer une architecture générique. Difficile *a priori* de concilier les fortes contraintes de réflexivité de Prospero, avec les réseaux de neurones des plongements de mots aussi insondables que des boîtes noires, mais nous tenterons tout de même d'en proposer un modèle commun, quitte à insister ultérieurement sur ce que le schéma final échoue à capturer. On décompose le processus analytique figure 1.20 en deux étapes principales : (i) le codage du contenu

textuel (ii) la transformation *via* une procédure quantitative de la collection de documents ainsi codés sous une nouvelle forme supposée plus aisément exploitable par l'analyste. Toutes les méthodes peuvent être décrites par la succession de ces deux étapes, mais elles ne revêtent pas la même importance selon les cas. Identiquement, en fonction de l'approche, l'utilisateur se retrouve dans une situation de délégation plus au moins forte vis-à-vis de la machine. Ainsi l'essentiel de l'apport analytique de Prospero se situe dans la première phase de codage des textes, phase durant laquelle l'utilisateur est aussi le plus mobilisé pour construire et valider les regroupements les plus pertinents. *A contrario*, la phase d'indexation du texte dans Alceste est entièrement automatisée et si l'utilisateur doit définir quelques paramètres simples comme la fréquence minimale ou maximale des entités à intégrer ou la longueur caractéristique des unités de contexte, il ne participe réellement au processus analytique qu'en bout de chaîne pour faire sens des mondes lexicaux identifiés.

Mais détaillons plus avant ces deux étapes. Le point de départ consiste simplement en un corpus défini comme un ensemble de documents textuels auxquels sont optionnellement associés un certain nombre de méta-données (par exemple la date de publication, l'auteur, la source, etc.). Ces documents textuels qui peuvent prendre la forme de longues pages de rapports, d'un paragraphe de résumé, d'un simples tweet de moins de 140 caractères subissent alors une procédure de « codage ». Différents aspects du texte de départ peuvent être sélectionnés et conservés pour l'analyse ultérieure. Pour reconstruire un schéma actantiel, on identifiera les êtres qui peuplent un texte : quelles personnes, objets ou organisations sont agencés au sein d'un même document ? Dans une perspective pragmatique, on essaiera de détecter dans un discours des marqueurs textuels qui renvoient à différentes modalités d'argumentation. En lexicométrie, on aura pré-défini un dictionnaire de formes qui permettent d'indexer un concept donné. Les approches plus inductives s'appuieront sur une indexation automatique soit de l'ensemble du vocabulaire d'un corpus donné (c'est le cas des modèles à plongement de mots, topic models ou d'Alceste [64]), soit sur une base grammaticale donnée (l'analyse de sentiment sera par exemple particulièrement sensible aux adjectifs et adverbes). L'analyse de style opérera un décompte des mots vides, qui sont justement ignorés par la majorité des méthodes de traitement automatique de la langue. À la recherche de dynamiques narratives on s'intéressera exclusivement aux personnages qui interviennent dans une même scène ou qui se donnent la réplique dans une tragédie de Shakespeare.

64. Pour ces trois familles d'analyse de texte, on peut avoir une intervention minimale sur le codage du contenu de départ et indexer l'ensemble des monogrammes dont la fréquence dépasse un seuil donné, ou user de moyens d'indexation additionnels plus ou moins sophistiqués consistant à identifier des n-grammes par des algorithmes de colocation, regrouper les termes partageant le même lemme (Sennrich et Haddow, 2016; Boyd-Graber et al., 2014; Mayaffre, 2005), éliminer les mots vides (« stop-words »), etc.

De façon plus générique et quelles que soient les dimensions du texte que l'on souhaite capturer dans le codage, deux options sont envisageables : les méthodes supervisées (à base de dictionnaires, ou de listes déjà établies) ou les méthodes dites non-supervisées que l'on paramètre à dessein [65]. Natu-

65. Quelle fréquence minimale doit avoir un mot pour être indexé, les verbes sont-il inclus, doit-on également considérer les bigrammes, etc ?

rellement il est tout à fait imaginable de faire appel à des méthodes mixtes, c'est d'ailleurs à un codage semi-automatisé du texte que nous avons fait appel dans les deux projets que nous avons menés avec Sylvain Parasie sur les commentaires publics en ligne : une liste d'entités textuelles pertinentes (au sens statistique et syntaxique du terme) a été automatiquement extraite des forums de la Voix du Nord et du blog du LA Times avant de donner lieu à un codage plus complexe qui a déjà été décrit section 3.1.2.

L'utilisation de dictionnaires permet de construire des catégories très fines sur lesquelles l'analyste a un contrôle intégrale. C'est résolument le choix fait par le logiciel Prospero dont les dictionnaires sont établis entièrement manuellement et circulent au gré des enquêtes. Mais on peut également faire appel à des dictionnaires pré-existants comme l'ont fait Klingenstein et al. (2014) qui utilisent la structure hiérarchique du « Roget's thesaurus » pour pré-catégoriser des délibérés de justice à différentes échelles. Dans les deux cas, on est néanmoins en droit de s'interroger sur les conséquences d'une telle stratégie de codage. Dans le premier cas, parce qu'un être fictif « importé » d'une affaire aura peut-être un sens entièrement différent dans un autre contexte (à moins que l'analyste ne s'épuise à sans cesse contrôler et mettre à jour son codage), et dans le deuxième cas, lorsque les ressources employées n'épousent pas parfaitement le langage d'une époque ou d'un domaine (comment un dictionnaire publié au début du XX^{ème} siècle pourrait fidèlement indexer la langue employée à la cour du Old Bailey au XVII^{ème} siècle ?). Dès lors, faire appel à des procédures issues du Traitement Automatiques des Langues comme l'extraction d'entités nommées, ou l'analyse morphosyntaxique (pour n'identifier que les groupes nominaux par exemple), voire l'analyse syntaxique (pour lier un sujet à une action) semble une option raisonnable pour traiter de grandes quantités de textes provenant de sources hétérogènes de façon robuste. L'idée n'est pas nouvelle, Franzosi (1989) appelle à l'usage de méthodes informatisées pour le codage de matériaux textuels en sociologie depuis bientôt trente ans ! Et à nouveau, faire appel à une méthode entièrement automatique n'est pas synonyme d'aveuglement techniciste pourvu, comme le souligne Grimmer et Stewart (2013), qu'elle soit toujours accompagnée d'une lecture au plus près du texte (« close reading ») et de procédures de validation.

Continuons à décrire le second étage du modèle. La partie inférieure du schéma décrit les différentes transformations possibles une fois le corpus textuel « réduit » en un modèle simplifié composé d'une séquence d'objets jugés pertinents lors de l'opération de codage. La plupart du temps, il ne s'agit même pas d'une séquence ordonnée mais d'un simple ensemble et on parle alors d'un modèle de sac de mots (« bag of words »). Chaque document (représenté par un rectangle) est alors caractérisé par un certain nombre de « termes » [66].

66. Ces termes peuvent potentiellement être composés de plusieurs mots. On parle alors de « n-grammes »

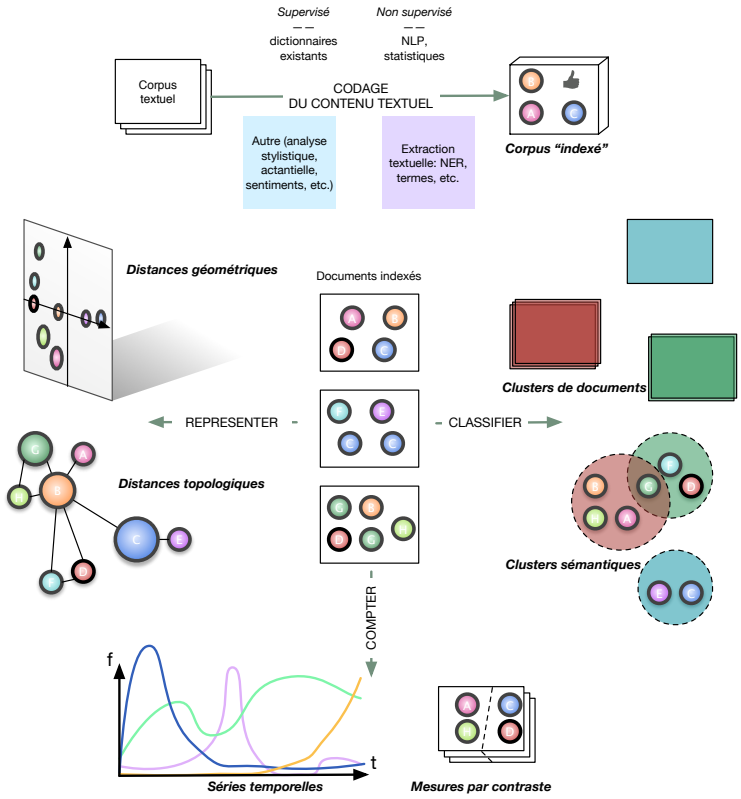


FIGURE 1.20: Schéma synthétique des différentes étapes des méthodes d'analyse de contenu. On débute par une phase de codage du contenu textuel brut avant de faire subir au corpus ainsi indexé une série de transformations (de classification, comptage, projection géométrique) dans une seconde phase. La pratique du logiciel Prospero nous rappelle également qu'il est tout à fait possible de concevoir ce schéma comme un cycle que l'analyste est libre de répéter à sa guise.

Classiquement trois grands types d'opération de « transformation » des données de départ sont possibles. La plus ancienne et la plus simple consiste simplement à dénombrer les marqueurs textuels identifiés. Une fois une fréquence assignée aux éléments indexés dans les textes, il est possible, à la manière de ce que propose les logiciels de lexicométrie ou Prospero, de caractériser la spécificité de documents écrits par tel ou tel auteur (ou type d'auteurs), dans telle ou telle revue, ou à telle ou telle période. Le résultat prend alors souvent la forme de listes de termes dont la fréquence dans une sous-partie du corpus est particulièrement importante ou faible relativement au reste du corpus. Il est également possible de mesurer à l'échelle des termes un certain nombre de propriétés dynamiques comme la propension d'un terme à apparaître sous forme de rafales chère aux lexicomètres, ou des mesures plus complexes. Par exemple, Klingenstein et al. (2014) montrent l'émergence de la violence comme catégorie politique au XIX^{ème} siècle en Angleterre, en mesurant une divergence de Shannon-Jensen croissante entre le traitement des crimes violents et non-violents à la cour criminelle de Old Bailey. En plus de l'approche contrastive, la simple analyse de l'évolution temporelle de la fréquence des entités identifiées a été travaillée de multiples façons entre le décompte de structures grammaticales complexes patiemment construites du de la lecture distante, et les métriques sur les séries temporelles permettant de

comparer les profils d'évolution de mots indexés dans l'intégralité de la base Google Books développées au sein du projet « culturomics » (Michel et al., 2011).

Une deuxième classe de méthodes vise à classer les documents ou les termes pour construire des clusters de documents ou des clusters sémantiques. Ainsi Prospero vise par des métriques simples qui comparent la fréquence d'apparition de différentes categories pré-codées à regrouper des textes mobilisant les mêmes être fictifs, ou au contraire identifier les textes ou groupes de textes qui s'opposent. Alceste, via la procédure de clustering descendant hiérarchique, construit des mondes lexicaux qui agrègent en premier lieu des unités de contexte élémentaires (donc des extraits de documents). Le Topic Model a un fonctionnement mixte qui permet d'inférer conjointement des « topics » comme une distribution de probabilités sur les termes et des documents comme des mélanges (probabilistes également) de topics [67]. L'ensemble des ces méthodes d'analyse qui structurent un matériau textuel sous la forme d'un ensemble de thématiques, sont typiquement utilisées dans des travaux sur le cadrage des problèmes publics (Snow et al., 1988). L'analyse de corpus permet alors de décrire la variété des cadres d'analyse possibles et leur dynamique (Sagi et al., 2013; DiMaggio et al., 2013; Fligstein et al., 2014).

67. Un document peut donc dans ce cadre potentiellement relever de plusieurs topics

Une autre opération classique vise à représenter les termes ou plutôt leurs distances relatives. Différentes mesures de proximité ont été proposées (en linguistique computationnelle, en scientométrie, etc.) telle que la mesure d'inclusion de Callon ou la distance du χ^2 chère aux adeptes de l'analyse des correspondances [68]. Une fois les distances entre objets établies, on peut soit tâcher d'en proposer une représentation géométrique dans un espace bi-dimensionnel qui soit le plus fidèle possible à la complexité de la matrice de proximité originale, soit utiliser le formalisme des réseaux (et les méthodes de représentations afférentes) pour analyser la position des termes vis-à-vis de leur environnement topologique. Cette dernière solution est naturellement privilégiée par l'analyse de mots-associés, même si la visualisation des réseaux n'est pas véritablement le résultat final de l'analyse, le diagramme stratégique semblant être le point de départ privilégié pour l'interprétation. Les méthodes de cartographie des controverses (notamment du DMI) sont par contre friandes de ce mode de représentation. Enfin, les modèles de plongement sémantique relèvent très clairement d'une approche géométrique du sens (même si la dimension de l'espace est très importante), propriété qu'ils partagent avec l'analyse des correspondances. Dans le cas des plongements de mots, la position est première et les distance entre éléments découlent d'une certaine façon de la procédure d'apprentissage. C'est le contraire de l'analyse des correspondances qui infère la position (dans un espace de faible dimension) des éléments à partir d'une matrice de distances entre éléments déjà calculée.

68. il faut noter qu'à part les modèles LDA, les autres approches requièrent de définir une mesure de proximité entre documents ou entre termes sous une forme ou une autre

Il faut également souligner que classification et représentation sont parfois complémentaires l'une de l'autre. Ainsi Iramuteq propose de visualiser les mondes lexicaux au sein d'une analyse des correspondances. L'approche réseau (par mots-associés ou plus largement) permet de mesurer et visualiser le rôle d'un terme dans son environnement mais aussi d'identifier et de caractériser des clusters et leur interaction, etc.

Dans certains cas, le modèle de langue retenu dans la première étape peut embarquer des hypothèses sociologiques très fortes. L'article de Mohr, Wagner-Pacifici, Breiger, et Bogdanov (2013) en offrent un exemple frappant. La « grammaire des motifs » du théoricien de la littérature Kenneth Burke (1969) sert de cadre de référence pour analyser des rapports bureaucratiques qui définissent la stratégie des affaires intérieures américaines. Au cœur du modèle de Burke se trouve notamment le pentagone dramatistique (« dramatistic pentad ») décrivant tout récit comme composé d'une action, d'une scène, d'un agent, d'un agencement qui permet l'action (agency) et d'un motif. Mohr et al. (2013) proposent d'en reproduire (en partie au moins) la structure en faisant appel à une stratégie mixte. Un algorithme d'extraction d'entités nommées permet d'abord d'identifier les agents comme « Germany », « Eastern Europe » ou « Department of Defense ». L'action est décrite à l'aide d'une analyse syntaxique qui lie chaque agent à un répertoire d'actes (correspondant en fait simplement à des verbes). Plus original, la scène du pentagone de Burke, censée fournir le contexte d'action des agents, est produite par un « topic model » appliqué au texte brut. Seulement une fois le texte lu à travers ces trois grilles d'analyse, l'analyse du « style discursif de l'État » peut commencer à proprement parler. On voit comment dans cet exemple particulier, la méthode de classification offerte par les topic models (située en phase 2) est intégrée et enrichie par d'autres procédures automatisées pour nourrir la complexité du codage réthorique (phase 1). Du point de vue de notre schéma figure 1.20, les niveaux ont eu le temps de se mélanger et le cycle d'analyse aura été répété plusieurs fois....

1.3.2 Bilan

Pour récapituler, nous avons consigné dans le tableau 1.1 les sept approches que nous avons pris le temps de décrire précédemment : lest mots-associés, la lexicométrie, l'analyse des correspondances, Alceste, Prospero, les topic models et les plongements de mots. Comme on l'a déjà trop répété, ces familles ne sont pas entièrement comparables. Certaines s'inscrivent dans des hypothèses sociologiques fortes dont la méthodologie est le reflet parfait (l'archétype en étant Prospero) ce qui rend leur extension à d'autres types de données ou d'autres formes d'interrogation problématique, d'autres méthodes

(typiquement les plongements de mots ou même les topic models) viennent d'univers plus lointains et les intégrer dans une enquête sociologique demande alors un bonne dose d'imagination. Mais leur neutralité théorique les rend aussi très malléables, capables de circuler et de s'adapter à d'autres contraintes.

Nous décrivons donc ces différentes approches en fonction d'une grille d'analyse qui mentionne les logiciels existants, le formalisme mathématique ou statistique sous-jacent. Les grands principes structurants de chaque approche sont également renseignés (s'appuie-t-on sur un principe fréquentiel, contrastif, géométrique, etc.), la colonne « sortie » décrit le type d'objets attendus (listes, matrices, positions) et les formes d'inscription qu'elles produisent (diagramme 2d, carte, tableaux, etc.). Nous décrivons également l'épistémologie des différentes approches, avec la difficulté qu'elle dépend nécessairement de l'usage qui en est fait[69]. Enfin, nous essayons de qualifier le type de théorie du langage (comment l'acte d'énonciation est modélisé ?) avant de citer les lieux, fondateurs ou réalisations remarquables de chaque approche. La dernière colonne essaye de décrire le champ d'application de chacune des méthodes au sein des sciences sociales.

69. Notons au passage l'absence d'approches déductives, exclues par construction de notre panorama.

TABLE 1.1: Typologie des méthodes d'analyse de corpus en sciences sociales

Approche	Logiciel(s)	Formalisme mathématique	Inscriptions	Epistémologie	Type de données	théorie du langage	Fondation	Modèle sociologique
Mots-associés	Leximappe, Candide, Calliope (Sci^2, VOSViewer, CiteSpace, TI.exe)	Analyse de cooccurrences, détection de communautés, MDS (Tlab.it)	Diagramme stratégique, cartes de réseau	inductif, forte délégation à la machine	Jeux de données moyens à grand, provenant de sources variées (scientométrique)	ANT (inspiré de Greimas) progressivement diluée	cartographie des sciences - Michel Callon	sociologie de la traduction, puis cartographie des sciences / issue mapping
Lexicologie politique (lexicométrie traditionnelle)	LEXICO, HYPERBASE	Fréquentielle, AFC, Spécificités, rafalités, segments répétés	listes ordonnées et AFC	inductif	multiple sources (discours, rapports, etc.), mais surtout politiques	les différences fréquentielles du vocabulaire cachent des écarts idéologiques	l'Ecole Normale Supérieur de St. Cloud, André Salem	discours politiques, corpus historiques ou littéraires
Analyse des Correspondances	FactoMiner, Prince	AFC, ACM	Espace factoriel	inductif *a priori* mais abductif par l'usage	données individus/attributs	Questionnaires (Bourdieu), Modèle actantiel (Boltanski)	Benzécri à la conception, popularisé par Bourdieu	champs de force chez Bourdieu
Alceste	Alceste, Iramuteq, TXM, Rtemis, Tlab.it	CDH (+ Analyse des correspondances)	classification dans les mondes lexicaux (+ projection AFC)	inductif, forte délégation à la machine	corpus de taille raisonnables, de toute origine (textuels ou catégoriels)	rythmé par la scansion des UCE dans Alceste	Reinert, Alceste	« fond topique » de Reinert, mais forte adaptabilité
Sociologie pragmatique des affaires publiques	Prospero (accompagné de Marlowe, Tiresias et Chéloné)	fréquentiel, contrastif	listes ordonnées, réseaux d'entités, dictionnaires	abductive, délégation minimale à la machine	sources hétérogènes, petit corpus, expression publique	métalangage de l'argumentation (linguistique pragmatique et schéma actantiel Greimas)	Chateauraynaud & Jean-Pierre Charriau, au GSPR	sociologie pragmatique, analyse des controverses
Topic Model	Gensim, Mallet, TMT, topicmodels R	modèle génératif bayésien, mélange topics et mots	Classification probabiliste, listes ordonnées	inductif, l'analyste doit essentiellement choisir le nombre de topics	prétention de traiter toute forme de texte, mais grands corpus	sac de mots	LSA puis David Blei	Aucun, mais facilement adaptable
Plongement de mots	Glove, gensim, Deeplearning4j	réseau de neurones, espace sémantique vectoriel	espace sémantique continu, relations d'analogie, pas de visualisation	inductive, forte délégation	sources variées mais impératif de traiter de très grands corpus (?)	sac de mots, modèle cognitif ?	Mikolov (Google & Facebook) et Sémantique dsitributionnelle	Aucun, mais appliqué aux études des inégalités, etc...

2

Cartographie Hetérogène de Réseau

La cartographie de réseau sémantique regroupe un certain nombre de propriétés qui en font, à nos yeux, un excellent modèle pour l'analyse de corpus textuels en sciences sociales. Fondée, comme beaucoup d'autres méthodes, sur l'hypothèse distributionnelle selon laquelle le sens des mots émerge de leur contexte d'apparition, elle propose une forme de voie moyenne entre l'ascèce bayésienne des topic models et l'expressivité graphique de l'analyse des correspondances. Pour autant, ce n'est pas un choix dogmatique et figé, et ce mémoire regorge d'exemples où, en fonction du type de matériau textuel, on privilégiera d'autres approches mieux ajustées à la question de recherche : des plongements de mots (voir section 3.3.3) à une simple régression (voir section 3.2.2) en passant par l'analyse fréquentielle (voir section 1.2.3). De plus ces cartes doivent être pensées non comme une fin en soi mais comme un point de départ pour poser de nouvelles questions à un corpus, interroger son évolution, articuler de nouvelles variables, etc. Ce chapitre ne vise donc pas à décrire une méthodologie qui prétendrait répondre de façon universelle à toute question de recherche posée à un corpus de textes. Ce que nous décrivons ici c'est la série de transformations mathématiques et graphiques auxquelles on soumet le texte pour le visualiser sous une forme qui permet de l'interroger. On entend simplement présenter les choix et principes de construction de cartes qui font office de points d'étape, des « médiateurs » (Latour, 1993), et jamais la finalité de l'enquête sociologique.

On peut entièrement reprendre à notre compte l'analyse faite par Mohr et Bogdanov (2013) sur les topic models (et qui serait d'ailleurs tout aussi valide pour décrire l'analyse des correspondances) :

> « *One implication is that well informed interpretive work — hermeneutic work — is still required in order to read and to interpret the meanings that operate within a textual corpus even when one is peering through the lens of a topic model. It is not the need for a deep understanding of one's textual corpus that has changed, it's the place where this*

style of knowledge comes into play. »[1]

Comme pour les topic models, dans le cas de la cartographie de réseau, le travail herméneutique de compréhension en profondeur des textes, le moment où le chercheur apporte sa subjectivité est largement déplacé en aval de la chaîne analytique. Il n'est plus besoin *a priori* d'avoir une compréhension historique et contextuelle approfondie des textes pour construire les bonnes catégories ou sélectionner les bons symboles comme l'exigent l'analyse de contenus ou la lexicométrie traditionnelle mais une telle connaissance reste indispensable pour proposer des interprétations pertinentes des résultats obtenus.

Nous adhérons à l'ambition originale de Lasswell selon laquelle :

« *Content analysis is a technique which aims at describing, with optimum objectivity, precision, and generalizability, what is said on a given subject in a given place at a given time.* »[2](Lasswell et al., 1952, p. 34)

Mais nous pensons aussi, qu'aussi objectif soit-il, le décompte de mots ne peut suffire à répondre de façon satisfaisante à l'impératif de départ, à savoir « décrire ce qui est dit en un lieu et un moment donné ». Or, c'est bien là le cœur de l'école d'analyse de contenu américaine que de compter les « symboles »[3]. Et la méthode est encore très répandue : qu'il s'agisse de tracer l'évolution de la fréquence des termes dans une base de données de livres numérisés (Lin et al., 2012)[quand bien même le type grammatical des mots serait connu], de comparer la fréquence d'usage de mots dans les réseaux sociaux en fonction de propriétés psychologiques du locuteur (Schwartz et al., 2013), ou d'indexer un concept politique donné comme la violence en fonction d'un vocabulaire prédéterminé (Klingenstein et al., 2014). Dans tous les cas, les résultats supposent de réifier *a priori* le sens des mots dont on suit l'évolution temporelle, comme si toutes les occurrences étaient entièrement inter-changeables et que chaque mention d'un terme renvoyait invariablement à un concept précis et connu de tous.

L'un des objectifs premiers de l'analyse de corpus textuels par cartographie de réseau (d'ailleurs partagé avec beaucoup d'autres méthodes) est donc, *via* l'analyse des configurations relationnelles entre termes, d'accéder à un nouveau niveau de description qui dépasse l'échelle nécessairement subjective des occurrences de mots simples décontextualisés. À cette condition seulement, la représentation des attitudes peut-elle être objective et généralisable (au risque d'avoir troqué la précision du comptable pour l'intuition du géomètre mais nous aurons le temps d'y revenir)

La cartographie de réseau emprunte naturellement beaucoup à l'analyse de mots-associés. Pour autant, elle s'en distingue notamment par le mode de représentation qu'elle privilégie. Alors que les mots-associés ont transformé les relations de mots en indicateurs stratégiques, la cartographie de réseau

1. « Une des conséquences, c'est que le véritable travail d'interprétation - le travail herméneutique - est toujours requis pour saisir et interpréter les significations qui traversent un corpus textuel même lorsqu'on le regarde à travers le prisme d'une topic model. Ce n'est pas tant la nécessité d'une compréhension approfondie des corpus de texte qui a changé que le lieu même où ce type de connaissance entre en jeu. »

2. « L'analyse de données est une technique visant à décrire de façon générique et avec autant d'objectivité et de précision que possible ce qui est dit en un lieu et un moment donné sur un sujet donné »

3. Pour être précis, il ne s'agissait pas chez Lasswel de dénombrer explicitement des mots mais des symboles. L'analyse de contenu est définie en effet chez lui comme l'analyse de la distribution spatio-temporelle des symboles. Mais il se réfère en réalité toujours à des mots ou des clusters de mots construits manuellement qui, dans sa perspective, « symbolisent » des « attitudes ». L'ambiguité est d'ailleurs partiellement levée dans le manuel que publie plus tard son étudiant en se référant à la notion de contenu explicite dans la définition qu'il donne de l'analyse de contenu qu'il définit comme « a research technique for the objective, systematic, and quantitative description of the manifest content of communication. » (Berelson, 1952, p 18)

continue de donner à voir les relations locales sans que l'espace ainsi décrit ne soit *a priori* ordonné. C'est également à ce titre qu'elle se distingue de l'analyse des correspondances dont l'espace géométrique résulte de facteurs censés expliquer la position de l'ensemble des variables. Les réseaux, même s'ils sont presque toujours visualisés en deux dimensions, peuvent être spatialisés de différentes manières. Et si le non-déterminisme des algorithmes de spatialisation peut être source d'anxiété de prime abord, il a l'avantage de relativiser l'importance des positions absolues et rappelle au lecteur (de la carte) que les distances topologiques sont premières. Pour autant, cette représentation graphique est également un véritable guide pour : repérer les groupes de nœuds les plus denses d'un seul regard, identifier les nœuds en position de ponts structuraux, etc.

Comme les topic models, on cherche à construire des clusters de mots, en faisant appel à des outils de classification fondés sur la structure topologique des réseaux, mais ceux-ci gagnent en relief par rapport à des simples listes ordonnées : la topologie du réseau de similarité permet de saisir conjointement l'articulation entre termes au sein d'un cluster, leur caractère central ou périphérique dans la constitution de ces groupes, mais aussi l'articulation entre clusters. Enfin, les algorithmes de spatialisation de réseau permettent une lecture « géométrico-topologique » directement interprétable pour l'analyste qui rend visible les agrégats sans jamais cacher la richesses des motifs locaux que l'on peut toujours suivre grâce aux liens apparents.

Mais avant de reprendre point par point les éléments de la typologie déjà déployée au chapitre précédent pour positionner cette nouvelle approche parmi les autres méthodes d'analyse textuelle, listons les trois grandes étapes de la construction d'une carte sémantique [4] qui seront décrites ci-après :

1. sélectionner les termes à cartographier (partie 2.1),
2. mesurer la similarité sémantique entre éléments lexicaux (partie 2.2),
3. analyser et cartographier la structure du réseau de similarité (partie 2.3).

4. En toute rigueur, il pourrait s'agir de tout forme de variable même catégorielle comme des auteurs, des sources citées, etc. On se concentrera néanmoins sur la cartographie de réseau sémantique qui se compare le plus directement aux autres méthodes.

2.1 *Extraction lexicale*

Contrairement aux autres méthodes récentes, l'approche cartographique requiert de faire des choix assez drastiques quant aux éléments lexicaux participant à l'analyse. C'est une contrainte essentiellement technique (difficile de faire figurer tous les mots d'un corpus sur une page A4 de manière lisible passés quelques centaines de termes) qui ne doit pas apparaître comme une limite analytique. Pratiquement, cartographier quelques centaines de termes, pourvu qu'ils soient pertinents, suffit souvent à produire une « image »

fidèle des grandes thématiques au sein d'un corpus. Difficile de définir *a priori* un nombre idéal, mais généralement, il apparaît une taille critique au-delà de laquelle l'ajout de nouveaux éléments ne modifie plus sensiblement la structure des cartes, c'est à dire que les thématiques déjà présentes sont robustes à l'ajout de nouveaux éléments, ou que ces nouveaux éléments précisent la nature ou la structure des thématiques déjà identifiées à une résolution si fine qu'elle ne rend compte que de structures secondaires pour l'analyste.

La cartographie de réseau se fonde sur un modèle du texte minimaliste de type « sac de mots ». Chaque document du corpus est ainsi indexé par une série de « mots-clés ». Est-ce défigurer un corpus que de le réduire aux occurrences de quelques centaines d'entités (quel que soit le soin que l'on ait mis à les sélectionner) flottant de texte en texte ? Sans aucun doute. Les finesses de l'énoncé sont perdues à jamais, mais c'est une réduction parfaitement adaptée pour décrire et identifier les grandes thématiques discutées dans un corpus que l'on retrace à travers les situations de présence et de co-présence de termes. Comme l'énoncent fort justement Grimmer et Stewart (2013) dans leur revue des outils d'analyse textuelle pour les sciences politiques,

> « *All quantitative models of language are wrong — but some are useful.* »[5]

C'est résolument notre perpective ici, modéliser le contenu textuel de chaque document à l'aide de quelques descripteurs remarquables dont les éléments soient suffisamment discriminants pour reconnaître un document dans un corpus, mais également suffisamment partagés pour servir de points de comparaison entre textes. En somme nous ne sommes pas plus dupe que Lasswell, Lerner, et de Sola Pool (1952, p52) vis-à-vis de notre indexation initiale :

> « *It should be frankly recognized that content analysis is a procedure of deliberate simplification. Such deliberate simplification is painful. Coders are pained at losing some of the richness of the meaning of the text. Invariably a conscientious reader will bring to his supervisor some passage full of innuendo, metaphor, or double entendre, and protest that the scheme being used distorts its real meaning. Toward such difficulties the analyst may well be ruthless. When the questionable passage has become but one of fifty-seven checks in category xyz, the analyst will have neither time nor cause to consider its individuality. Content analysis is a statistical procedure, and, like any statistical procedure, it disregards the individuality of the particular case for the sake of discovering the uniformities in the mass.* »[6]

Les méthodes d'analyse de texte adoptent traditionnellement trois attitudes par rapport à la définition de ces éléments de base. Première option, le codage est purement manuel et les éléments sont choisis et regroupés dans une même catégorie en fonction d'une grille pré-définie par l'analyste (ce qui n'empêche pas certains problèmes comme le souligne déjà Franzosi (1989)). La deuxième option consiste à exploiter des listes ad-hoc de termes pour indexer les termes d'intérêt (comme dans Prospero ou dans des travaux plus récents qui s'appuient sur des dictionnaires pré-existants (Klingenstein et al.,

5. « Tous les modèles quantitatif du langage sont faux - mais certains sont utiles. »

6. « Il devrait être franchement reconnu que l'analyse de contenu est une procédure de simplification délibérée. Une telle simplification délibérée est douloureuse. Les codeurs sont peinés de faire perdre au texte une partie de la richesse de son sens. Invariablement, un lecteur consciencieux se plaindra auprès de son supérieur que quelque passage rempli d'insinuations, de métaphores ou de doubles sens ait été dévoyé de son sens premier par la grille de codage employée. Confronté à de telles difficultés, l'analyste ferait aussi bien de se montrer sans pitié. Lorsque le passage problématique n'est plus qu'un parmi cinquante-sept exemples dans la catégorie xyz, l'analyste n'aura jamais le temps ni le besoin de les envisager individuellement. L'analyse de contenu est une procédure statistique, et, comme toute procédure statistique, elle ignore la singularité du cas particulier pour mieux découvrir l'uniformité dans la masse. »

2014)). Enfin, la troisième attitude possible consiste à extraire les termes d'intérêt automatiquement à partir du contenu textuel brut mais sans que leur sélection ne pose problème. Dans ce dernier cas, les mots partageant la même racine ou le même lemme peuvent être regroupés. Un algorithme de colocation est parfois utilisé pour identifier les bi-grammes ou les tri-grammes les plus significatifs statistiquement. Une fois cette liste constituée, une pratique courante consiste également à éliminer les mots vides (*stop words*)[7] et les mots trop rares[8]. Ainsi les algorithmes de plongement de mots ou de topic models peuvent faire appel à une certain nombre de procédures de normalisation du texte de départ, mais le filtrage des entités pertinentes est toujours extrêmement limité : dans l'immense majorité des cas, ces algorithmes se contentent d'éliminer les termes les moins fréquents.

Pour la génération de cartes sémantiques, la question demeure : comment identifier automatiquement les termes les plus remarquables d'un corpus ? Si l'idée d'une procédure partant du contenu textuel brut est séduisante, comment garantir que les termes choisis soient en nombre raisonnable et offrent une couverture satisfaisante des thématiques du corpus.

7. Dans le cas, un peu à part des algorithmes de plongement de mots, la qualité des distances mesurées est en fait moindre lorsque ces mots outils sont éliminés. Sans que leur rôle dans le modèle soit encore bien compris, il semble qu'ils jouent le rôle d'une forme de « glue ». Pour autant (et cela illustre bien combien l'adage de Diesner (2015) : « Small decisions with big impact on data analytics »), Levy et al. (2015) expliquent qu'échantillonner aléatoirement le vocabulaire en éliminant les mots avec un probabilité proportionnelle à leur fréquence est bénéfique pour la performance du modèle. Pré-traiter le texte en normalisant les termes (en les remplaçant par leur lemme par exemple) ne semble pas non plus nécessairement être une opération bénéfique pour les modèles de plongement de mots.

8. Lorsque la fréquence tombe à quelques occurrences seulement, on s'expose à noyer le résultat dans un bruit statistique.

2.1.1 *Filtrage grammatical*

Avant même que de définir une mesure de pertinence en tant que telle, une première étape de filtrage consiste à ne conserver parmi les termes candidats que ceux correspondant à un type grammatical donné. C'est une étape classique dans la la littérature sur l'extraction terminologique (Milios et al., 2003) appelée aussi ATR (*Automatic Term Recognition*). Par défaut, on effectue un premier filtrage qui permet de ne conserver que les groupes nominaux. Techniquement, on reconnait un groupe nominal comme une séquence de mots dont les étiquettes grammaticales correspondent à une série d'adjectifs et de noms potentiellement séparés par une conjonction de coordination ou une préposition comme dans l'expression « origin of life ».

On pourrait se doter d'un autre critère[9], mais les groupes nominaux semblent une hypothèse raisonnable dès lors que notre objectif est de cartographier - pour l'exprimer de la façon la plus générique possible - de grandes thématiques comme des situations de présences et de co-présences d'objets au sein d'un corpus de documents. Nous cherchons finalement à remplacer les fameux mots-clés chers à Michel Callon (parce qu'il ne sont pas toujours présents, mais aussi parce qu'ils induisent un « biais d'indexation » potentiel (Whittaker, 1989)), qui prennent d'ailleurs couramment la forme de groupes nominaux.

9. On a déjà vu que, dans certains contextes, l'usage de verbes peut-être utile pour - par exemple - caractériser le positionnement de pays prenant la parole dans une négociation (voir section 1.1.5).

À ce filtrage grammatical, on peut rajouter, sans rentrer dans les détails, toute une série d'opérations de standardisation qui visent à regrouper sous le même libellé des termes apparaissant sous des formes variées : mots acceptant plusieurs orthographes (avec ou sans tiret par exemple), termes composés de mots dont l'ordre importe peu (« oil and gas », « gas and oil ») etc.

2.1.2 *Mesures de pertinence*

Mais un critère grammatical ne saurait suffire à délimiter l'ensemble des termes pertinents qui permettront de décrire les cadres d'un débat, circonscrire les questions de recherche d'une communauté scientifique, ou suivre les évolutions de l'agenda médiatique. Dès que le corpus est un peu conséquent, on est confronté à un nombre très important de groupes nominaux qui doivent à nouveau faire l'objet d'une sélection.

Une première approche consiste à les filtrer en fonction de leur nombre d'occurrences pour ne conserver que les plus fréquents. Malheureusement, beaucoup des mots les plus fréquents sont aussi les moins intéressants. La fréquence des termes dans un corpus donné hérite en grande partie de leur fréquence dans le langage courant, ce qui disqualifie la fréquence brute comme un critère de sélection pour caractériser de façon spécifique un ensemble thématique donné.

Une seconde approche consiste à comparer les fréquences des termes au sein du corpus avec leurs fréquences dans un corpus tiers. L'idée qui a déjà fait l'objet de nombreux développement (Gelbukh et al., 2010) vise donc à détecter les mots dont la fréquence au sein d'un corpus donné est exceptionnellement haute comparativement à ce que l'on peut mesurer « habituellement ». Mais on voit immédiatement que ce principe se heurte rapidement à une question pratique difficilement soluble : par rapport à quel corpus de référence doit-on jauger les fréquences calculées dans un corpus donné ? Prenons l'exemple d'un ensemble de publications scientifiques sur la biologie de synthèse. Quel corpus de référence choisir pour contraster la fréquence des mots au sein de ces articles ? Vaut-il mieux utiliser les statistiques fournies par l'indexation de Google Books ? On imagine bien que des termes en apparence bien anodins comme *method*, *measure*, *experimental protocol* seront largement sur-représentés dans notre corpus d'articles scientifiques alors qu'ils ne correspondent finalement qu'à un vocabulaire scientifique généraliste qui s'avérerait bien peu discriminant pour décrire la structure des recherches en biologie de synthèse. Si nous choisissions un corpus de référence composé de publications scientifiques choisies au hasard, alors, on imagine à nouveau que des mots spécifiques à la biologie mais pas nécessairement à la biologie de synthèse

comme *DNA*, *blood*, *cell*, *body* seront à nouveau sur-représentés dans notre corpus. Devrait-on plutôt apprécier les fréquences dans notre corpus à l'aune de celles d'un corpus de référence composé exclusivement d'articles de biologie, mais on ferait alors une hypothèse très forte sur la nature des recherches qui sont menées dans la communauté, et on s'expose cette fois à sur-représenter les termes liées à l'ingénierie ou à la modélisation. En somme, il semble impossible de caractériser un corpus en se servant d'un point de vue extérieur (ou plutôt sans que l'utilisation de ce point de vue extérieur de joue le rôle d'un point de vue), tout simplement parce que la définition même de ce qui compose l'extérieur demanderait déjà de connaître les frontière de l'intérieur. Il faut donc nous résoudre à une approche internaliste sans tomber dans la naïveté d'un filtrage basé sur le seul nombre d'occurrences.

Un des algorithmes d'extraction terminologique les plus populaires et les plus robustes KEA (Witten et al., 1999) s'appuie sur deux heuristiques simples. Tout d'abord, les meilleurs candidats pour indexer un document donné sont ceux qui apparaissent fréquemment dans le document en question. Cette propriété est mesurée en calculant le tf.idf des termes de chaque document. L'idée générale du tf.idf est de donner du poids aux termes fréquents localement *via* la composante fréquentielle de l'expression (tf) sans que les profils ne soient dominés par les termes globalement très fréquents qui sont négativement pondérés par la fréquence inverse de document [10].

10. Pour un terme t, $\text{idf}(t) = log(\frac{N}{N_t})$ où N désigne le nombre total de documents du corpus et N_t le nombre de documents dans lesquels figure le terme t

Un autre critère auquel le tf.idf est combiné par Witten et al. (1999) est la position de la première occurrence du mot au sein du document ; un terme employé dès les premières phrases a plus de chance d'être pertinent pour caractériser ce document. Malheureusement cette heuristique souffre d'exceptions : un terme placé au début d'un texte peut parfois être trop générique pour indexer le document de façon satisfaisante. Le tf.idf permet d'identifier des mots qui sont effectivement pertinents à l'échelle d'un texte, mais qui peuvent dans d'autres contextes être tout à fait anecdotiques. Or, nous cherchons ici à établir une liste de termes pour construire une cartographie globale du corpus et pas nécessairement à identifier ce que chaque texte a de singulier (au risque de nous retrouver avec des terminologies non recouvrantes). Contrairement à la tradition en recherche d'information (Manning et al., 2008), notre objectif n'est pas de rechercher un document particulier, mais avant tout de saisir l'univers sémantique partagé entre documents.

Si on se fixe pour objectif l'obtention d'une liste de termes pour l'ensemble des documents d'un corpus, alors l'hypothèse que les termes les plus pertinents apparaissent à plusieurs reprises au sein des documents qui les mentionnent se traduit de façon directe au travers d'une mesure de type gf.idf. Contrairement au tf.idf, ce ratio met en rapport le nombre total d'occurrences d'un terme et le nombre de documents distincts dans lequel il apparait. Un

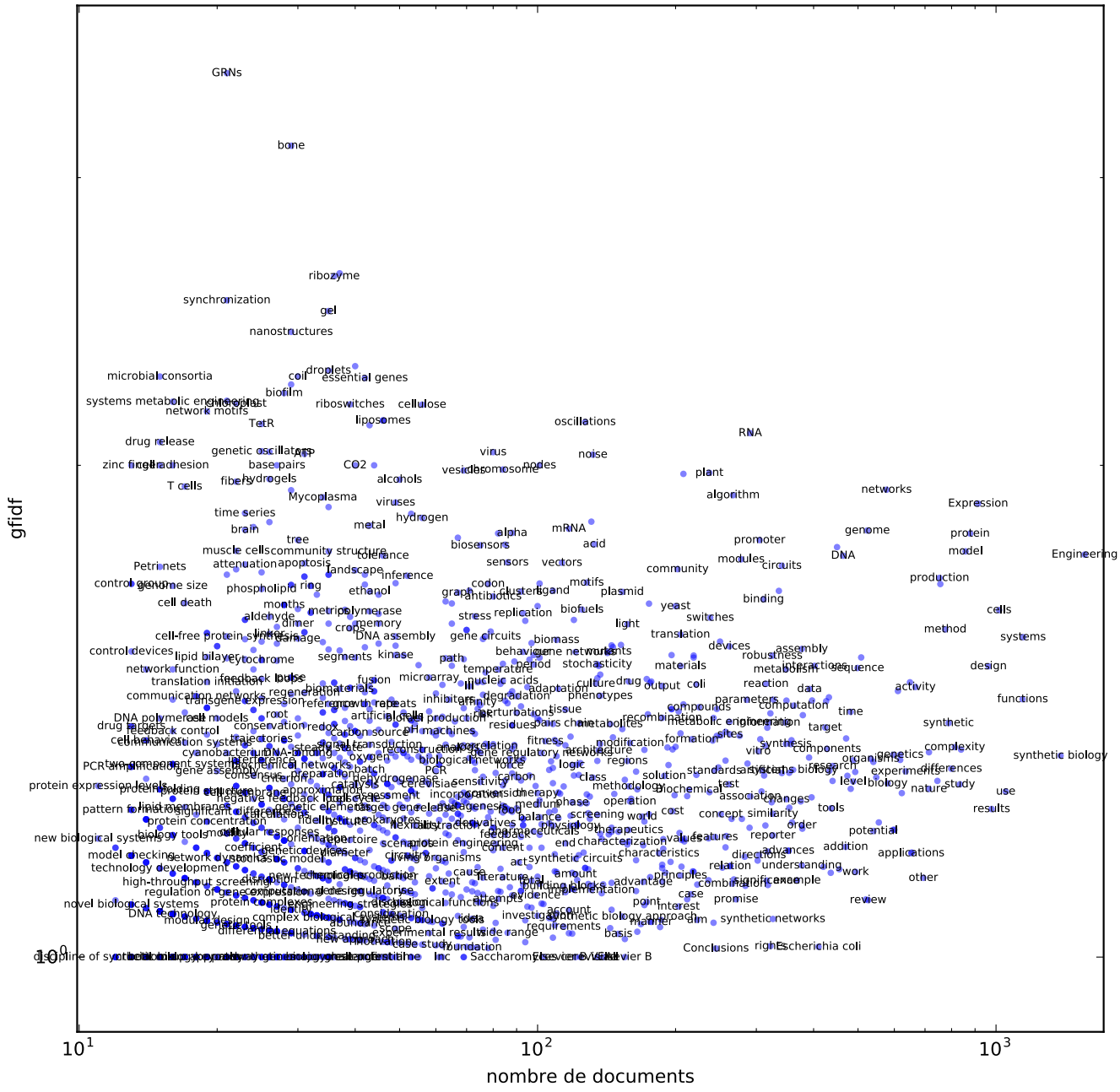


FIGURE 2.1: Les 1615 groupes nominaux les plus fréquents (apparaissant dans strictement plus de 10 articles distincts) d'un corpus d'abstracts portant sur la biologie de synthèse positionnés ont été identifiés. Ils sont positionnés en fonction de leur fréquence (en abscisse), et de leur gf.idf.Pour ne pas surcharger la visualisation nous avons volontairement limité le nombre de termes étiquetés.

ratio important indique que la distribution du terme se concentre sur un nombre réduit de documents [11].

Appliqué à un corpus d'abstracts de publications scientifiques (voir figure 2.1), ce score semble relativement efficace pour distinguer les termes pertinents indépendamment de leur fréquence. Ainsi pour les termes les plus

11. Pour autant, il est tout à fait possible qu'un seul document regroupe l'ensemble de ces occurrences. Alternativement, on peut faire appel à une mesure d'entropie pour mieux caractériser la distribution du nombre d'occurrences d'un mot par document (Dumais, 1991)

fréquents *high*, *mutliple*, *synthetic* [12] mais aussi *nature*, *review*, *Conclusions*, *Elsevier* ont un gf.idf très faible proche de 1. Les termes tout aussi fréquents comme *cell*, *method*, *networks* ou *algorithm* apparaissent en moyenne près de deux fois dans les articles où ils figurent et semblent effectivement plus pertinents. Avec une fréquence moindre, les termes tels que *magnitude*, *Motivation*, *era*, *new field* ou encore *discipline of synthetic biology* ont un gf.idf beaucoup plus faible. Au contraire, *T cells*, *droplets* ou *synchronization* sont de très bons candidats.

12. Étiquetés comme noms dans certains cas, ces deux adjectifs se retrouvent dans la liste des termes candidats

La mesure de gf.idf a malheureusement tendance à sur-classer les termes très fréquents au sens où plus un terme est fréquent et plus probable est l'éventualité de le voir apparaître plusieurs fois dans le même document. Pour s'en prémunir, on peut par exemple mesurer le ratio entre le nombre de documents uniques d dans lesquels apparaît un terme et le nombre théorique d^*de documents sous hypothèse d'une distribution totalement aléatoire de ses f occurrences au sein de N documents. Le nombre attendu de documents distincts valant [13] alors : $d^* = N - N(\frac{N-1}{N})^f$. La figure 2.2 montre combien cette nouvelle mesure permet de corriger les excès du gf.idf

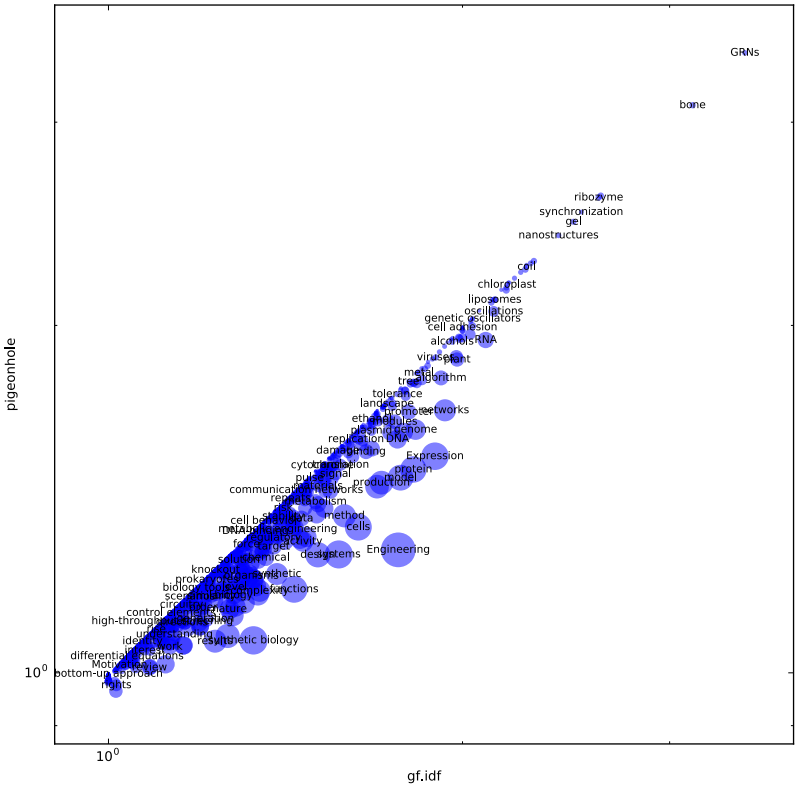


FIGURE 2.2: La formule du gf.idf est une bonne approximation du ratio entre nombre de document distincts observés divisé par le nombre théorique de documents distincts attendus que nous avons appelé "pigeonhole". Pour autant, le correctif est nécessaire pour les termes plus fréquents dont le gf.idf est sinon surévalué. Ainsi, on observe notamment que le score de termes très fréquents comme *engineering*, *function* ou *method* était sensiblement sur-évalué par rapport à des termes moins fréquents. La taille des termes est proportionnelle à leur fréquence totale sur le graphique. Seule une sélection (aléatoire) de termes est étiquetée pour clarifier la représentation.

Une difficulté subsiste néanmoins : comment traiter des corpus composés de documents de longueur réduite comme un ensemble de tweets pour lesquels les termes apparaissent presque certainement une seul fois par documents ?

13. Dans le script d'extraction lexicale de CorText cette mesure est appelée *pigeonhole* en référence à la version anglaise du principe des tiroirs de Dirichlet qui décrit la probabilité de voir un boulin occupé par deux pigeons dans un pigeonnier. Il est facile d'en établir l'équation en remarquant que la probabilité qu'un casier d'un pigeonnier comportant N cases soit vide alors que f pigeons y nichent vaut $(\frac{N-1}{N})^f$ (la probabilité que l'un des f pigeon ait choisi un autre casier valant simplement $\frac{N-1}{N}$). L'espérance du nombre de boulins inoccupés vaut alors simplement : $N(\frac{N-1}{N})^f$

2.1.3 *Unithood et termhood*

Avant d'aller plus loin, il est utile d'emprunter à Kageura et Umino (1996) les définitions des concepts de *unithood* et de *termhood* qui mesurent deux dimensions souhaitables pour qualifier un terme de pertinent. La *unithood*, que nous appellerons simplement caractère unitaire ou unité, est une mesure de la stabilité d'un terme composé de plusieurs mots, de sa capacité à renvoyer à une « unité conceptuelle ». On se demande simplement si une séquence de mots constitue une unité linguistique valide [14]. À titre d'exemple (emprunté à (Wong, 2009)), dans l'expression *E. coli food poisoning*, *E. coli* et *food poisoning* sont des termes fortement unitaires, au contraire de *E. coli food*. Différentes méthodes existent pour évaluer de façon efficace le caractère unitaire de termes extraits à partir de leur seules fréquences d'apparition : la C-value de Frantzi et al. (2000) est l'un des algorithmes les plus populaires. Par contre la meilleure façon de mesurer nature terminologique d'une séquence (*termhood*) d'un terme est encore un sujet de recherche discuté dans la littérature.

14. Plus précisément, le caractère unitaire d'un terme correspond à la « force ou la stabilité dune combinaisons et de collocations syntagmatiques » (« the degree of strength or stability of syntagmatic combinations and collocations »)

La *termhood* mesure à quel degré une unité linguistique est spécifique d'un domaine donné [15]. Pour donner un exemple simple, le terme « litterature review » est certainement très répandu dans un corpus de publications scientifiques. Pour autant, il ne permet pas de décrire un sous-champ de recherche particulier, sa distribution est sans doute indépendante des grandes thématiques qui parcourent le champ (on fait des états de l'art dans tous les champs de recherche). Autrement dit, on s'attend à ce que « litterature review » ait une *termhood* basse lorsqu'un corpus d'articles scientifiques est indexé. *A contrario*, une méthode spécifique d'une des branches de la biologie de synthèse comme le *gene editing* devrait obtenir un score de *termhood* élevé.

15. « the degree that a linguistic unit is related to domain-specific concepts », (Kageura et Umino, 1996)

Un des algorithmes les plus populaires pour capturer cette propriété a été introduit par Mihalcea et Tarau (2004). Le fonctionnement de Textrank est basé sur l'analyse de cooccurrences au sein des documents. Un document est transformé en un réseau dont les noeuds sont composés par ses termes. Deux termes sont liés s'ils co-occurrent dans une fenêtre de quelques mots. Une mesure inspirée du page-rank est calculée pour chacun des noeuds du réseau ainsi extrait. Seuls les termes avec le « textrank » le plus élevé sont conservés pour indexer le document. La définition du TextRank a plus tard été raffinée en introduisant des mesures alternatives liées à la topologie du réseau (Boudin, 2013) (betweenness centrality, closeness centrality, etc.) ou en imposant des contraintes grammaticales sur les liens du réseau (Blanco et Lioma, 2012).

Toutes ces méthodes partagent la même idée de départ : un réseau de similarité entre termes est construit à l'échelle de chaque document. L'objectif

est d'identifier automatiquement une série de termes qui constituent un index convaincant d'un document. Mais notre objectif n'est pas celui de la recherche d'information traditionnelle. L'extraction terminologique vise classiquement à identifier une série de termes qui soient pertinents pour caractériser chacun des documents d'un corpus donné. Notre objectif est de produire une cartographie de l'ensemble du corpus. De plus, nombre de ces méthodes s'avèreraient inutilisables confrontées à des documents courts.

Nous avons donc proposé une stratégie alternative qui s'appuie sur la recherche des termes dont la distribution est biaisée par rapport à la distribution des principales thématiques du corpus. La difficulté à laquelle on se confronte est que nous ne connaissons rien à ce stade des dites thématiques. S'appuyant sur la même stratégie, les concepteurs de VOSVIEWER ont développé une méthode qui construit cet ensemble de thématiques grâce à un algorithme standard de PLSA avant de mesurer la spécificité des termes par rapport aux topics identifiés (Van Eck et al., 2010).

Nous avons souhaité mettre en place une méthode plus simple qui ne requiert pas d'opérer une classification a priori des termes en topics mais qui s'appuie sur le principe suivant : identifier les termes dont la distribution des occurrences est fortement corrélée à celle de certains autres termes candidats. En somme nous cherchons à identifier les termes spécifiques en nous fondant sur les propriétés relationnelles avec les autres termes candidats. Si un terme n'est pas informatif (comme « litterature review ») alors il peut apparaître dans une grande variété de contextes, et le nombre de cooccurrences qu'il entretiendra avec n'importe quel autre mot sera proche du nombre théorique de cooccurrences espéré sous hypothèse d'indépendance. *A contrario*, un terme caractéristique d'un sous-domaine du corpus étudié comme *degree distribution* relève du vocabulaire de l'analyse de réseaux complexes qui consiste en l'une des méthodologies mobilisées en biologie de synthèse. La distribution de ses occurrences dans le corpus ressemblera fortement à celle d'autres termes relevant du même champ tels que *scale-free network*, *network topology*, *connected component*, etc. Apparaissant souvent dans les mêmes documents, *degree distribution* aura un nombre important de cooccurrences avec ces mots là. Sans pré-calculer des topics ni même analyser la structure des réseaux de cooccurrences induits par chaque document, on peut directement estimer la termhood d'un terme au sein d'un corpus en comparant son profil de cooccurrences agrégé sur l'ensemble des documents au profil théorique qu'aurait pu avoir un terme de même fréquence mais dont les cooccurrences seraient distribuées de manière totalement aléatoire.

Il existe de nombreuses façons de calculer cette écart à l'indépendance. Une mesure assez simple que nous avons employée à de nombreuses reprises pour ce type de tâche (Cointet et al., 2012a; Omodei et al., 2014), et implémentée

dans CorText est inspirée de la mesure du χ^2. Etant donné un mot i faisant partie d'un vocabulaire de termes candidats $\mathcal{V}$, la pertinence de i est mesurée comme suit :

$$s(i) = \sum_{j \in \mathcal{V}, N(i,j) > \hat{N}(i,j)} \frac{(N(i,j) - \hat{N}(i,j))^2}{\hat{N}(i,j)}$$

où $N(i,j)$ désigne le nombre de cooccurrences entre i et j et $\hat{N}(i,j)$ le nombre de cooccurrence que l'on aurait dû observer si les cooccurrences de i et j étaient distribuées de façon entièrement aléatoire, c'est à dire : $\hat{N}(i,j) = \frac{\sum_k N(i,k) \sum_l N(l,j)}{\sum_{kl} N(k,l)}$. D'autres choix seraient possible comme de comparer les distributions de contexte de chaque terme à l'aide d'une distance de Kullback Liebler.

Une option alternative plus rigoureuse statistiquement consiste à employer le test de ratio de vraisemblance qui, partant d'un modèle binomial du nombre de cooccurrences attendues, compare l'hypothèse d'indépendance et de corrélation de deux termes [16]. Dunning (1993) a ainsi montré que le test de log vraisemblance $Cor(i,j)_{lgl} = -2log(\lambda)$ permet de distinguer entre des cooccurrences significatives et non-significatives statistiquement [17]. Pratiquement, on calcule le ratio λ entre la valeur maximum de la vraisemblance sous hypothèse d'indépendance divisée par la valeur maximum de la vraisemblance sous hypothèse d'indépendance.

16. On reconnait, en passant, la mesure de spécificité utilisée par les lexicomètres du premier chapitre (section 1.1.2) pour mesurer la spécificité d'un terme dans un sous-corpus donné !

17. Si Dunning a le premier suggéré l'emploi de ce test pour détecter des bigrammes dans un texte, l'analyse statistique des tableaux de contingence à l'aide de tests de rapport de vraisemblance est beaucoup plus ancienne (Wilks, 1935)

Il est utile de détailler un peu les équations qui permettent d'exprimer ce rapport de vraisemblance - les même formules seront à nouveau mobilisées pour calculer la similarité sémantique entre deux termes dans la partie 2.2. Plus formellement, on note n_{ij} le nombre de cooccurrences entre deux termes i et j, n_i et n_j le nombre total de cooccurrences de i et de j, et enfin n le nombre total de cooccurrences indexées. On distingue entre deux hypothèse H_1 et H_2. Dans le premier cas (H_1), on émet une hypothèse d'indépendance des distributions des deux termes i et j. Ainsi, la probabilité de voir i apparaître ne dépend pas de j et on la note $p = p(j|i) = p(j|\neg i)$ Dans l'autre cas (H_2) les occurrences de i sont corrélées à celles de j et on a des probabilités conditionnelles p_1 et p_2 différentes : $p(j|i) = p_1 \neq p_2 = p(j|\neg i)$

En exprimant le nombre de cooccurrences observées conditionné à un nombre fixe de cooccurrences totales pour les deux mots, on peut exprimer la vraisemblance d'observer n_i, n_j et n_{ij} occurrences de i, j et de i et j conjointement selon les deux hypothèses sous les formes suivantes [18] :

$$\begin{cases} H_1(p) = & p^{n_{ij}}(1-p)^{n_i - n_{ij}} \binom{n_i}{n_{ij}} p^{n_j - n_{ij}} (1-p)^{n - n_j - n_i + n_{ij}} \binom{n-n_i}{n_j - n_{ij}} & (H_1) \\ H_2(p_1, p_2) = & p_1^{n_{ij}}(1-p_1)^{n_i - n_{ij}} \binom{n_i}{n_{ij}} p_2^{n_j - n_{ij}} (1-p_2)^{n - n_j - n_i + n_{ij}} \binom{n-n_i}{n_j - n_{ij}} & (H_2) \end{cases}$$

On s'intéresse au ratio des maximum de vraisemblance que l'on note

18. Les vraisemblances sont composées du produit de deux probabilités : celle d'observer précisément n_{ij} cooccurrences de i avec j parmi toutes les co-occurrences n_i de i sachant que la probabilité conditionnelle vaut p (hypothèse H_1) ou p_1 (hypothèse H_2) (distribution binomiale de paramètres B(n_{ij}, n_i, p) ou B(n_{ij}, n_i, p_1)), et que conjointement sur les $n - c_i$ fois où i n'est pas impliqué dans une cooccurence, $n_j - n_{ij}$ d'entre elles contiennent le terme j (la distribution binomiale s'exprime alors sous la forme B($n_j - n_{ij}$, $n - n_i$, p) ou B($n_j - n_{ij}$, $n - n_i$, p_2))

classiquement $\lambda = \frac{max_p H_1(p)}{max_{p_1,p_2} H_2(p_1,p_2)}$. Or il est facile en calculant les dérivées partielles de H_1 (par rapport à p) et H_2 (par rapport à p_1 et p_2) de montrer que les maxima s'obtiennent pour des probabilités valant respectivement :

$$p = \frac{n_j}{n}, p_1 = \frac{n_{ij}}{n_i}, p_2 = \frac{n_j - n_{ij}}{n - n_i}$$

Le ratio se trouve « largement » simplifié sous la forme suivante (porté classiquement au log) :

$$\begin{aligned} 2log(\lambda_{ij}) = n_j log(\tfrac{n_j}{n}) - (n - n_j) log(\tfrac{n - n_j}{n}) - n_{ij} log(\tfrac{n_{ij}}{n_i}) - (n_i - n_{ij}) log(\tfrac{n_i - n_{ij}}{n_i}) \\ - (n_j - n_{ij}) log(\frac{n_j - n_{ij}}{n - n_i}) - (n - n_j - n_i + n_{ij}) log(\frac{n - n_i - n_j + n_{ij}}{n - n_i}) \end{aligned} \quad (2.1)$$

$2log(\lambda_{ij})$ suit la distribution du χ^2, si bien, que l'on peut se référer aux tables classiques et affirmer que les occurrences de i et j sont corrélées significativement ou non en garantissant un seuil de probabilité d'erreur. Dunning (1993) fait également remarquer qu'en remplaçant la distribution binomiale d'occurrences par une distribution normale de même moyenne et variance, le rapport de vraisemblance peut être approximé comme le test de Pearson de χ^2 déjà rencontré précédemment et que l'on détaillera section 2.2.2.

Si l'on part d'une modélisation de la table de contingence sous la forme d'une distribution multinomiale (et non pas sous la forme de deux distributions binomiales en s'appuyant sur les contraintes qui pèsent sur la somme des lignes et des colonnes de la table de contingence), le calcul, bien qu'un peu plus complexe, mène à une forme plus simple (Evert, 2005) du rapport de vraisemblance qui est alors classiquement noté $G_2 = -2log(\lambda_{ij}) = 2\sum_{k,l \in \{1;2\}^2} O_{kl} log(\frac{O_{kl}}{E_{kl}})$ où O_{kl} et E_{kl} correspondent aux valeurs observées et attendues des cellules du tableau de contingence entre les mots i et j et dont nous donnons une représentation et les valeurs précises un peu plus tard (voir figure 2.9). Les deux expressions sont pourtant entièrement équivalentes.

Mais revenons à notre question première qui visait à concevoir un score pour chaque terme qui quantifie dans quelle mesure sa distribution est biaisée par rapport au reste du vocabulaire. La solution est maintenant toute proche : une fois calculé pour chaque terme candidat i la significativité de sa corrélation $-2log(\lambda_{ij})$ avec les autres termes [19], on peut en calculant simplement le pourcentage des termes dont la corrélation est significative par rapport à l'ensemble des contextes possibles apprécier la « spécificité relationnelle » des termes candidats. Le ratio ainsi calculé offre une mesure agrégée simple de la pertinence d'un terme. Une représentation du résultat d'un tel calcul est fournie figure 2.3 toujours sur le même corpus de résumés d'articles en biologie de synthèse. Le résultat semble relativement satisfaisant, au sens

19. Nous verrons prochainement que ce même score fournit une très bonne mesure de similarité sémantique pour construire nos réseaux sémantiques

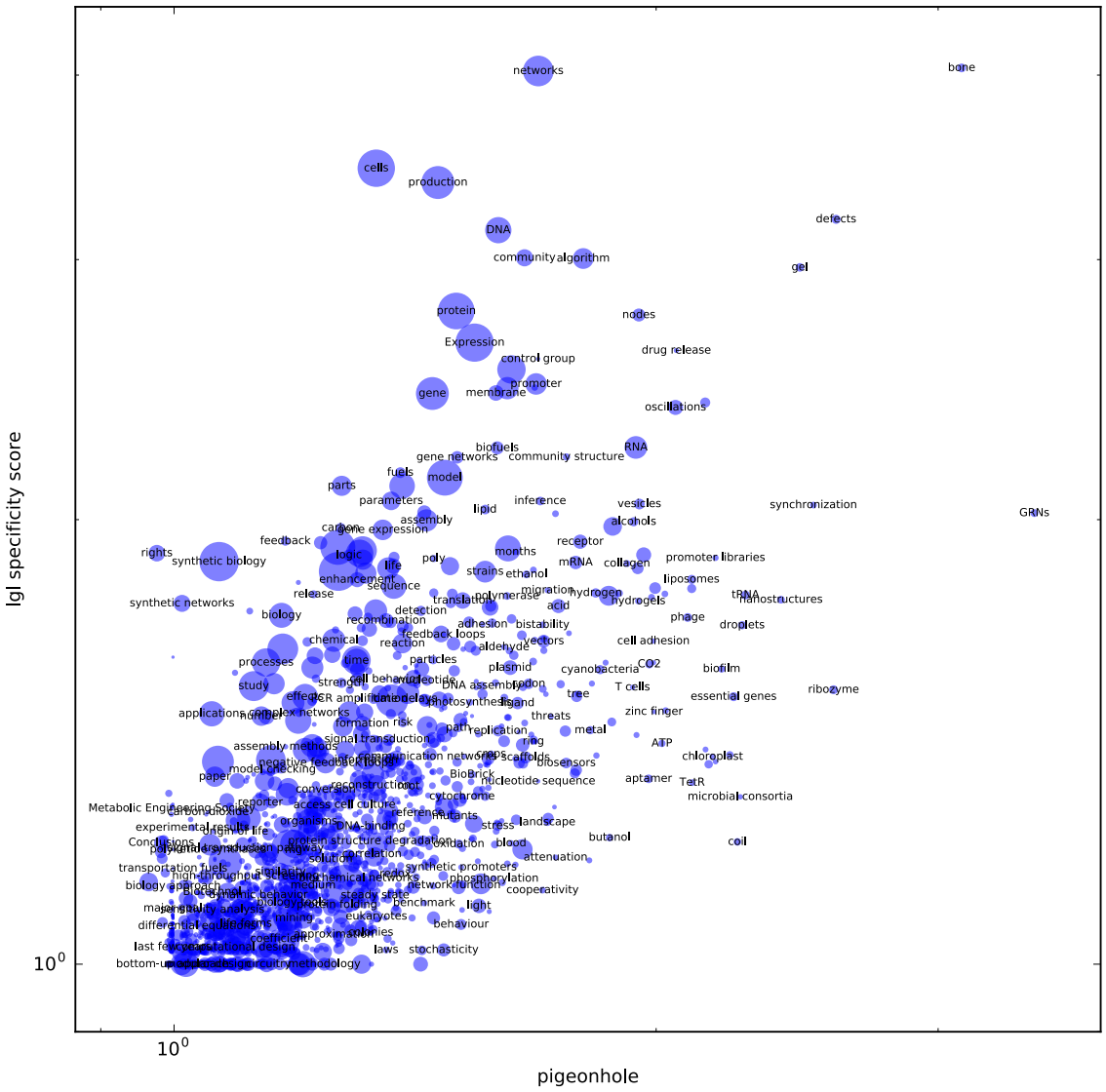


FIGURE 2.3: Corrélation entre mesure fréquentielle de spécificité (*pigeonhole*) et mesure relationnelle de spécificité (*lgl specificity score*, notée G^2)

où notre score est aussi bien capable d'identifier des termes très fréquents (comme *networks* ou *genome*) ou relativement rares (*drug concentration*, *time series*) que d'éliminer des termes trop génériques qu'ils soient fréquents (*have*, *example*) ou pas (*biochemistry*, *new design*). Sans être redondant avec la mesure de multiplicité déjà introduite (*pigeonhole*), la spécificité semble donner des résultats cohérents.

Difficile de se doter d'une procédure d'évaluation systématique de ces listes. C'est sans doute cette difficulté qui explique la grande variété et la prolifération des méthodes existantes (Zhang et al., 2008). En l'absence d'un gold standard (ou de gold standards compte tenu de la variété des objectifs visés : recherche

d'information, cartographie de corpus, classification de documents, etc.) par rapport auxquels on pourrait aisément juger leur mérite respectif, impossible de calculer précision et rappel comme pour d'autres tâches en traitement automatique de la langue.

2.1.4 *Profils sémantiques*

La méthode d'extraction automatique de mots-clés que nous venons d'exposer vise en priorité à cartographier la structure sémantique d'un corpus de documents. Dans un premier temps, les termes les plus caractéristiques du corpus sont utilisés pour indexer les documents. Par suite leurs cooccurrences sont calculées pour extraire les principaux champs sémantiques qui structurent le corpus (voir section 2.3.2). La figure 2.4 prolonge le principe du schéma général de l'analyse textuel de la figure 1.20. Partant d'une série de documents déjà étiquetés par leur termes caractéristiques, il est possible de construire un réseau de similarité entre termes mais aussi entre documents. Dans un tel réseau, les documents sont liés en fonction de leur proximité sémantique calculée comme la proportion de termes qu'ils partagent par exemple.

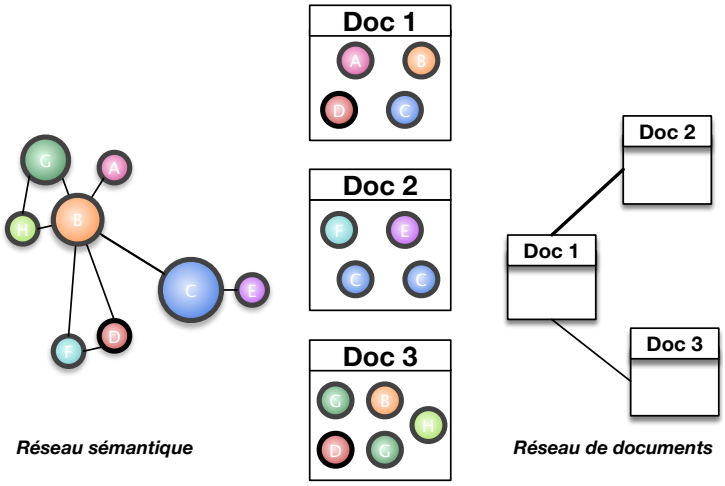


FIGURE 2.4: Transformation d'un corpus de documents déjà indexés en un réseau sémantique dont les termes sont liés s'ils ont conjointement été utilisés dans le ou les mêmes documents [gauche], et un réseau de documents dans lequel les documents sont liés en fonction de la ressemblance de leur contenu [droite]. Les données de départ [centre] sont de nature bipartite, au sens où deux types d'entités sont mises en relation : des termes au sein de documents. La projection de ce réseau bipartite peut dès lors produire un réseau homogène composé uniquement de documents (à droite) ou exclusivement de termes (à gauche).

C'est en adoptant cette perspective duale que j'ai commencé à travailler sur les discours de l'État de l'Union avec mes collègues : Alix Rule et Peter Bearman. La question qui nous a intéressés était notamment celle des périodes historiques. Pour les historiens et socio-historiens ou même pour un sociologue travaillant sur des affaires traversées par différents régimes discursifs, une question importante que les big data en général et l'analyse de corpus en particulier peut éclairer sous un jour nouveau tient précisément à ces comparaisons longitudinales (Bearman, 2015). En reformulant, la question se pose ainsi : comment détecter dans une série de textes des moments de rupture, séparant des périodes plus homogènes ?

Une première hypothèse qui peut-être posée est que différentes périodes auront une signature lexicale différente. On espère donc, en comparant le vocabulaire employé à différents moments, faire émerger des clusters temporels cohérents. Contrairement à l'approche lexicométrique classique qui produit des listes de termes dont l'écart de fréquence est maximal entre deux partitions du corpus, nous souhaitons ici obtenir une mesure globale, macroscopique, des variations du vocabulaire à différents pas de temps.

Il s'agit donc d'établir dans un premier temps le profil lexical de l'ensemble des documents publiés au même moment, et ce à chaque pas de temps. Différentes métriques permettent de décrire ce profil sous la forme d'un vecteur. Une des plus populaires est le tf.idf qui modélise le profil lexical comme un vecteur de la taille du vocabulaire complet (calculé sur le corpus complet) et dont les coordonnées sont définies par la fréquence locale (en ne considérant que les publications du pas de temps courant) d'un terme pondérée par sa fréquence inverse de documents [20].

20. Malgré la très grande popularité du tf.idf, sur toute une série de tâches, le modèle de log entropie semble offrir une meilleure alternative (Lee et al., 2005).

Dans un second temps, on mesure la différence entre deux pas de temps comme une simple mesure de similarité entre leur profil vectoriel respectif. Classiquement on utilise un cosinus, même si, à nouveau, d'autres mesures de similarité [21] sont envisageables. On compare ici les vocabulaires comme des ensembles de mots indépendamment des nuances de sens que leur ordonnancement est susceptible de créer. Une fois cette matrice de similarité produite, l'enjeu est naturellement d'extraire des groupes d'années connexes les plus cohérentes, l'hypothèse de contiguïté des pas de temps simplifie grandement le calcul, même si elle impose une contrainte technique. Le problème ne peut pas se réduire à des problèmes classiques [22] de ré-ordonnancement des lignes et colonnes d'une matrice.

21. Les traditionnelles mesures de Dice, euclidiennes, par bloc mais aussi la distance de Google(Cilibrasi et Vitanyi, 2007) (calculée à partir des résultats de requêtes sur le moteur de recherche), ou des distances passant par une première modélisation du texte *via* un espace sémantique latent, peuvent aussi être utilisées (Gomaa et Fahmy, 2013).

22. Le problème est classique mais fait toujours l'objet de recherches de nos jours (Behrisch et al., 2016).

La solution que nous avons finalement adoptée dans notre travail sur les discours de l'État de l'Union (Rule et al., 2015) peut être décomposée en deux étapes. D'abord, comme on vient de le décrire, on a construit la matrice de dissimilarité Δ dont les coordonnées pour chaque couple d'années (y, y') correspondent à une mesure de dissimilarité qui vaut 1 moins la mesure du cosinus entre les profils respectifs des discours prononcés pour chaque année y et y'. Une fois cette matrice de dissimilarité D calculée sur un vocabulaire composé d'un millier de mots [23], nous avons défini une mesure de qualité d'une partition temporelle scindant le corpus en deux périodes successives en faisant l'hypothèse que la partition idéale résulte en la création de deux blocs temporels les plus "homogènes" possibles du point de vue de leur vocabulaire. Posé en ces termes, si on note $[y_0 : y_f]$ l'extension temporelle du corpus et $\hat{y}$ l'année à laquelle débute la seconde période, nous cherchons à minimiser l'hétérogénéité que l'on définit simplement comme la moyenne des

23. Le résultat final est robuste quelque soit la taille du vocabulaire retenu dès qu'il est composé de quelques centaines de termes

dissimilarités entre années contribuant à la même période :

$$H^* = \underset{y^* \in [y_0 : y_f]}{\arg\min} \frac{1}{(y^* - y_0)^2 + (y_f - y^* + 1)^2} \sum_{y \in [y_0 : y_f]} \left\{ \begin{array}{c} \sum \\ y' < y^* \; si \; y < y^* \\ y' \geq y^* \; si \; y \geq y^* \end{array} \right. \Delta(y, y')$$

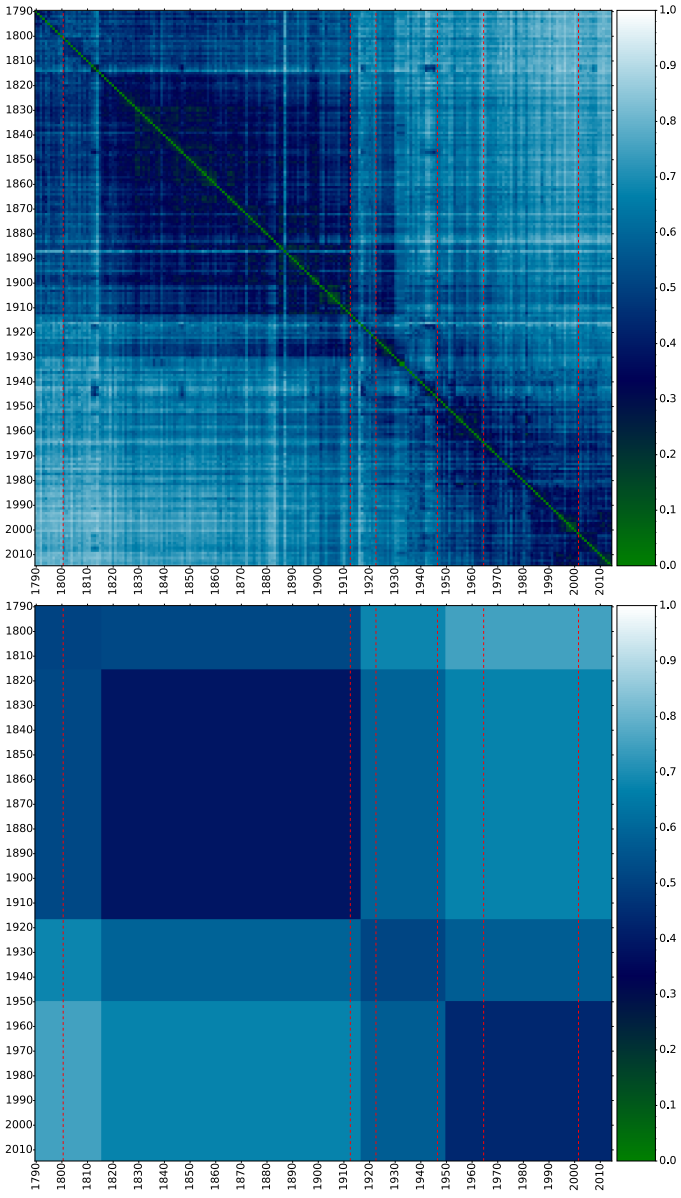


FIGURE 2.5: [haut] Matrice de dissimilarité entre discours de l'État de l'Union prononcés entre 1790 et 2014. Les lignes rouges pointillés (aux années : 1801, 1913, 1947, 1956 et 2002) correspondent à des changements remarquables dans la façon dont le discours est délivré (à l'oral, en version écrite, radiodiffusé puis télévisé) sans que la structure d'ensemble ne s'en trouve apparemment affectée. [bas] Matrice dont les blocs (réorganisés selon les quatre périodes suivantes : $p_1 = [1790 : 1815]$, $p_2 = [1816 : 1916]$, $p_3 = [1917 : 1949]$, $p_4 = [1950 : 2014]$).

On remarque que l'on peut aussi définir l'hétérogénéité locale d'une période donnée comme la moyenne des dissimilarités inter-annuelles au sein de ladite

période qui s'écrit en notant P la dite période et $|P|$ sa longueur :

$$H_P = \sum_{(y,y')\in P^2} \frac{\Delta(y,y')}{|P|^2} = \sum_{y\in P} \frac{1}{|P|} \sum_{y'\in P} \frac{\Delta(y,y')}{|P|} = \sum_{y\in P} \frac{1}{|P|} \Delta(y,P)$$

où $\Delta(y,P)$ est la dissimilarité entre le profil vectoriel d'une année y et le profil vectoriel moyen d'une période P. L'hétérogénéité d'une période s'exprime donc aussi naturellement comme la moyenne des dissimilarités des années qui la compose avec son vocabulaire moyen.

On en déduit alors que l'optimisation ci-dessus est équivalente au problème suivant :

$$H^* = \underset{y^* \in [y_0 : y_f]}{\arg\min} \; H_{[y_0:y^*-1]}(y^* - y_0)^2 + H_{[y^*:y_f]}(y_f - y^* + 1)^2$$

c'est à dire, que l'on recherche une partition temporelle telle que la somme des hétérogénéités des deux périodes ainsi définies soit aussi faible que possible en pondérant chaque période de façon quadratique par rapport à leur durée. Visuellement (cf partie inférieure figure 2.5), cela revient à minimiser la clarté des blocs diagonaux de la matrice. [24]

Sur la base de données des discours de l'État de l'Union, 1917 formait une année charnière à partir de laquelle émergeait le discours moderne des tâches de la gouvernance américaine. En re-divisant les deux périodes obtenues, on obtient 4 périodes déjà perceptibles lorsqu'on observait la matrice de départ [25]. Hormis ces changements sur des temps longs, aucun changement notable de vocabulaire n'a été mesuré selon que le discours soit délivré oralement ou par écrit, télévisé ou radiodiffusé [26]. Nous n'avons pas non plus noté un renouvellement de vocabulaire plus significatif que la normale ($\Delta = 0.28 \pm 0.09$) lorsqu'un nouveau président arrive au pouvoir ($\Delta = 0.30$) ou même lors des changements de majorité ($\Delta = 0.31$). En somme, le vocabulaire, tel que nous le caractérisons [27], semble être un objet culturel relativement stable qui a très peu changé pendant le XIX$^{\text{ème}}$ siècle et se stabilise de nouveau après la confusion qui a régné durant la première moitié du XX$^{\text{ème}}$ siècle. Plus récemment, nous avons repris ces résultats originaux en incluant les deux derniers discours de Barack Obama en 2015 et 2016 - sans que la périodisation obtenues n'ait changée d'aucune manière (voir figure 2.6).

Ce travail a également bénéficié d'un autre prolongement qui mobilise les modèles de plongement de mots et ici plus précisément le modèle doc2vec (Le et Mikolov, 2014). Ce modèle permet d'apprendre les vecteurs de documents conjointement avec les vecteurs de mots [28]. Dans notre cas, cette propriété permet simplement d'attribuer à chaque discours [29] une position dans l'espace. La matrice de distance sémantique que l'on en extrait est présentée figure 2.7. Nous n'observons plus, comme précédemment (figure 2.5, haut) l'évolution

24. Si la méthode mise en œuvre suit plus volontiers une intuition graphique qu'une mesure statistique bien référencée, notre méthode ressemble tout de même fortement à une évaluation par mesure "d'inertie intra-classe" (Lebart et al., 1979).

25. Si nous avons procédé par bissections temporelles successives dans ce travail, rien n'empêche de réécrire la formule de départ avec N périodes et de se livrer à une optimisation globale qui trouve simultanément les bornes de l'ensemble d'entre elles. Sans intuition pré-existante sur le nombre de périodes, certains critères statistiques endogènes tels que la statistique de Gap peuvent s'avérer utiles (Tibshirani et al., 2001), c'est d'ailleurs un paramétrage possible implémenté dans CorText que de déterminer automatiquement le nombre optimal de périodes distinctes.

26. Rappelons néanmoins à nouveau que nous travaillons uniquement à partir de groupes nominaux qui sont sans doute moins sensibles à des changements d'ordre stylistique.

27. En nous limitant aux seuls groupes nominaux, nous perdons l'essentiel des effets de style et de réthorique qui répondent à d'autres types de dynamiques (Teten, 2003).

28. De ce point de vue, on retrouve l'ambition de l'analyse des correspondances qui permet de projeter dans un même espace l'ensemble des objets et de leurs attributs.

29. Par mesure de simplicité nous supposons à nouveau qu'un discours et un seul a été prononcé chaque année.

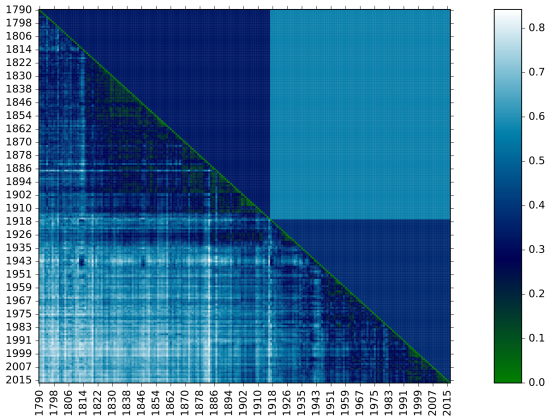


FIGURE 2.6: Matrice de dissimilarité du vocabulaire des discours de l'État de l'union prolongée jusqu'en 2016. Seule la partition séparant les discours en deux ères a été re-calculée pour à nouveau converger vers 1917

d'un vocabulaire à travers les siècles mais de façon plus profonde la dérive sémantique qu'opèrent les discours dans le temps. En somme, cette méthode neutralise dans notre mesure de transformation, les effets de variation et de renouvellement du vocabulaire quand celui-ci indexe peu ou prou les mêmes objets ou les mêmes concepts. Par exemple, qu'un président adopte une convention linguistique donnée (*railway*) ou une autre (*railroad*) n'a pas d'influence notable sur la mesure de différences inter-annuelles du fait de la très forte proximité des deux concepts dans l'espace (*railway* est d'ailleurs le plus proche voisin de *railroad*). Mais de façon plus large, lorsque les *railroads* de la ruée vers l'or sont remplacés par les *highways* du monde contemporain, le modèle sémantique a déjà endogénisé le fait que les deux termes renvoient à des infrastructures publiques de transport [30]. Aussi, parler de l'un ou de l'autre n'affectera pas la position des discours de façon très sensible, les deux termes étant extrêmement proches dans l'espace sémantique final qui agrège l'ensemble des années. Il est remarquable d'observer qu'avec ce modèle beaucoup plus réaliste, on retrouve à nouveau une structure extrêmement nette séparant l'ère pré-moderne et moderne. Sans avoir calculé l'année précise de bascule, il apparaît visuellement que la transition se situe à nouveau durant la première guerre mondiale.

30. Les plus proches voisins de *railroads* en partant du terme le plus proche sont : *housing*, *waterway*, *aviation*, *building*, *program*, *system*, *construction*, *rehabilitation*, *road*, *job training*, *transportation*, *hospital*, *navigation*, *project*, *governmental*, *ocean*, *banking*, *credit*, *route*, *agency*, *airport*, *education*, *commerce*, *assistance*, *fortification*, *research*, *university*, *defense*, *center*, *railway*, *transit*, *mean of communication*, *railroad*

2.2 *Mesurer le sens*

Le titre de cette section peut sembler provocatrice. Pour autant, si notre ambition est de reconstruire de façon automatique la structure conceptuelle d'un corpus textuel, il faut bien nous doter d'un moyen de rendre dénombrable la « charge sémantique » des éléments lexicaux que nous venons patiemment de sélectionner. Mais dès lors que l'on parle de mesure, on est en droit de s'interroger quant à la métrique propre du sens des mots. Existe-t-il un mètre étalon propre à la sémantique ? Quels instruments permettent-ils de calculer

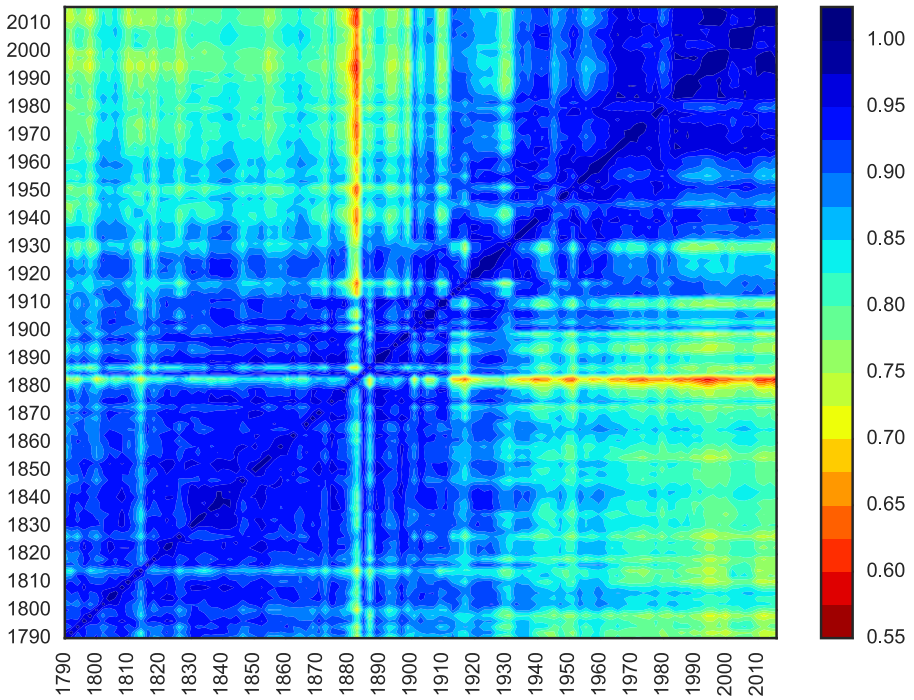


FIGURE 2.7: Matrice de transformation sémantique des discours de l'État de l'Union. Chaque discours annuel est réduit à un point dans l'espace vectoriel. La proximité (mesurée par la métrique du cosinus) entre deux discours prononcés à deux moments distincts est non seulement sensible aux variations de vocabulaire mais plus profondément aux variations de "sens" des discours. Le modèle Doc2Vec (paramétré avec les options suivantes sous gensim (Řehůřek et Sojka, 2010) : Dbow, 100 dimensions, seuil de fréquence fixé à 10, fenêtre de 5 mots, hierarchical softmax) a été calculé en étiquetant chaque paragraphe des discours par son année.

le sens d'un terme ?

La tâche semble ardue, perdue d'avance d'après les plus sceptiques qui vont jusqu'à douter de l'existence même d'une telle notion. Pour autant, le secret a déjà été éventé au chapitre précédent : à défaut d'une échelle absolue et universelle du sens, l'approche distributionnelle permet de comparer la significations des termes les uns avec les autres. On montrera donc dans cette partie la manière dont les linguistes ont conceptualisé et opérationnalisé (non sans que cela n'occasionne certaines controverses (Poibeau, 2014)) une définition relationnelle du sens des mots permettant de mesurer la similarité sémantique entre couples de termes en comparant leurs contextes.

2.2.1 *L'hypothèse distributionnelle*

Prolongeant le principe propositionnel de contexte de Frege (Conant, 1998), Wittgenstein (1953, 1953) nous enjoint à porter attention à l'usage réel des termes dans le langage pour en saisir le sens : « Bedeutung ist der Gebrauch ». Autrement dit, le sens des mots est entièrement défini par leur contexte d'apparition (et pour être plus précis la question du sens n'est valide que posée en ces termes). Cette même intuition a été formulée en linguistique dès le milieu du XX$^{\text{ème}}$ siècle par Harris (1954) et plus tard popularisée et

explicitée par le fameux adage de Firth (1957) :

« *A word is characterized by the company it keeps.* »

Certains linguistes défendent une hypothèse un peu moins forte. Selon Ploux et Victorri (1998), le sens précis d'un mot dans un énoncé résulte en effet d'une interaction « entre un apport sémantique constant associé à cette unité et le contexte d'énonciation de cette unité » (Venant, 2008). Partant de cette hypothèse de construction géométrique du sens, un « espace sémantique » associé à chaque mot peut être construit qui positionne les multiples sens d'un mot au sein d'un espace propre [31].

Mais une fois posée l'hypothèse que le sens des mots est co-constitutif de leur contexte, c'est à dire qu'il émerge des mots se trouvant dans son environnement, il faut bien concéder la circularité de l'argument puisque le sens d'un mot se trouve alors dépendre du sens des mots qui l'entourent. Reste que cette définition permet de distinguer entre des termes qui sont sémantiquement proches car partageant les mêmes contextes, ou sémantiquement très différents car s'entourant de termes différents. L'information que nous pouvons tirer de cette hypothèse est donc de nature purement relationnelle. Partant du principe que des termes partageant les mêmes propriétés distributionnelles ont des sens similaires, il est possible de mesurer une similarité sémantique entre termes. Pour autant, ce principe de permet pas d'assigner au sens d'un terme une position dans un espace métrique absolu.

Avant de préciser la nature de ces métriques, il faut distinguer entre deux modalités distinctes de modélisation du contexte. Comparer les contextes d'apparition entre mots permet en effet de mesurer une similarité d'ordre syntagmatique ou paradigmatique (Sahlgren, 2006) selon la façon dont la notion de contexte est comprise. Les relations syntagmatiques (du grec *suntagmatikos* - qui signifie arrangé, mis en ordre) lient ainsi des entités qui co-occurrent au sein d'un texte. On dira de deux mots qui sont en relation syntagmatique, qu'il peuvent aisément être combinés l'un à l'autre au sein d'une même phrase. Les relations paradigmatiques (qui vient du grec : *paradeigmatikos* - renvoyant à ce qui sert de modèle), quant à elles, ont trait à la notion de substituabilité d'un mot par un autre. Deux mots entretiennent une relation paradigmatique lorsque ils partagent les même contextes, sans nécessairement apparaître conjointement dans les mêmes textes.

Sans que les relations entre le courant de linguistique distributionnel et l'école de linguistique structurale soient si directes, Saussure donne une définition très claire des rapports syntagmatiques et associatifs [32]

> « *Le rapport syntagmatique est in praesentia : il repose sur deux ou plusieurs termes également présents dans une série effective. Au contraire le rapport associatif unit des termes in absentia dans une série mnémonique virtuelle* » Saussure (1967, p. 171)

La figure 2.8 montre néanmoins que la notion de rapport associatif de Saussure

31. Dans ce modèle, l'espace sémantique est démultiplié, chaque mot générant son propre espace riche de toutes ses significations possibles. La même hypothèse distributionnelle aboutit donc paradoxalement à des modèles topologiques opposés des espaces sémantiques vectoriels classiques que nous nous apprêtons à définir.

32. Le terme paradigmatique ne fut introduit que tardivement pour remplacer ce que Saussure appelait à l'époque des rapports associatifs. Il est cocasse de ce point de vue de noter, que les mot associés, qui s'appuient sur une sociologie associative font appel à des mesures de type syntagmatique !

est plus large que la notion de substibuabilité. Dans sa conception les rapports associatifs peuvent renvoyer

— à des termes partageant la même racine (on dirait le même lemme maintenant) : *enseigner, enseignons*,

— à des termes dont les « signifiés » sont en relation d'analogie. Il s'agit du second axe : *apprentissage*, *éducation*, etc.

— dans des séries basées sur un suffixe partagé : *clément*, *justement*, etc.

— ou même des termes renvoyant à une même « communauté des images acoustiques » (notamment susceptibles de générer des jeux de mot)

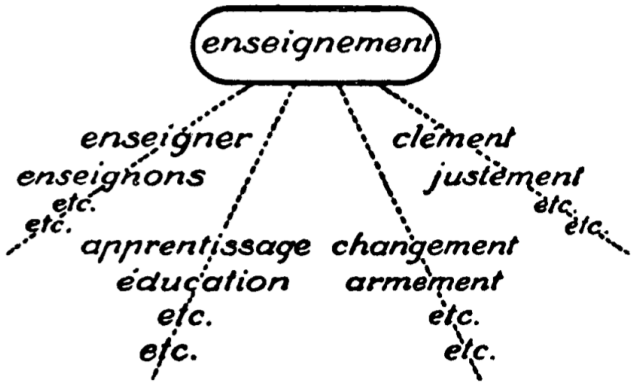


FIGURE 2.8: Illustration de la famille associative du terme *enseignement* extraite de (Saussure, 1967, p175) dans lequel il est précisé que les « termes d'une famille associative ne se présentent ni en nombre défini, ni dans un ordre pré-déterminé ».

Ici on entend bien définir les rapport paradigmatiques non pas par l'ensemble des « groupes formés par association mentale » qui comme le souligne Saussure (1967, p. 173) « renvoient à des rapports divers », mais bien à capturer les relations de substituabilité entre signifiés. Il est néanmoins intéressant de noter que les modèles de plongement de mots, semblent très performants pour construire un espace vectoriel dans lequel plusieurs types de rapport sont conservés. Les analogies de type grammatical (singulier/pluriel, présent/passé, verbe/substantif) correspondent ainsi à des opérations vectorielles simples (axe 1). Les relations de similarité sémantique entre termes sont également bien reconstruites (axe 2). Dans le même espace, on observera également que l'ensemble des adverbes, ou des termes complexes par exemple (parce qu'ils correspondent alors à un certain registre de langue) se retrouvent dans un même sous-espace (axe 3). Des mesures de perplexité (Shahaf et al., 2015) pourraient être imaginées pour tenter de modéliser l'axe 4 et ainsi construire automatiquement (mais aussi ordonner !) les familles associatives de Saussure.

2.2.2 *Mesures syntagmatiques (dites directes)*

Débutons avec les mesures de similarité syntagmatiques que l'on appellera également des mesures directes de similarité au sens où elles dépendent exclusivement des statistiques de co-occurrences observées entre deux termes.

En réalité, nous avons déjà rencontré un certain nombre d'entre elles dans le chapitre précédent. Les indices d'inclusion ou d'équivalence ou le coefficient d'association spécifique sont toutes des mesures directes d'une relation syntagmatique entre mots-clés. Elles ne sont pas toutes équivalentes pour autant, l'indice d'inclusion permet par exemple d'extraire des relations d'inclusion entre mots qui ne sont pas symétriques comme l'est l'indice d'équivalence. Parmi les autres mesures de similarité directes qui ont plus tard été introduites en scientométrie, on peut mentionner le cosinus [33], et diverses mesures ensemblistes telles que l'indice de Dice ou Jaccard (Sternitzke et Bergmann, 2008). Les mesures introduites précédemment doivent être conçues comme de simples « scores » qui permettent d'estimer la force d'une association entre deux termes mais sans garantie statistique quant à la significativité de cette association. La mesure du χ^2 vise justement à palier ce manque. Si elle n'est pas entièrement inconnu des scientométres (Leydesdorff et Welbers, 2011; Mogoutov et al., 2008), on retrouve surtout cette mesure dans la littérature en traitement automatique du langage (Evert, 2005). Le test du χ^2 de Pearson est le test classique pour mesurer l'indépendance des lignes et des colonnes d'une matrice de contingence (Manning et Schütze, 1999) et se doter d'un moyen de mesurer le degré de confiance que l'on a dans une telle association. On la retrouve également sous une forme détournée lorsque Jean-Paul Benzécri mesure l'inertie d'un nuage de points pour construire une analyse factorielle comme on l'a vu dans le chapitre précédent (section 1.1.3).

33. On parle ici du cosinus « simple » tel qu'introduit par Salton qui s'exprime selon cette forme : $S_{ij} = \frac{n_{ij}\sqrt{N}}{\sqrt{n_i n_j}}$

Pratiquement, une matrice de contingence récapitule les effectifs croisés obtenus à l'intersection de deux variables. On peut par exemple construire la matrice de contingence par rapport à une variable qui mesure la présence ou l'absence de deux mots A et B (voir illustration sure la figure 2.9). Et on conclura qu'il y a un lien entre A et B si les lignes et les colonnes de la matrice de contingence (voir figure 2.9) ne sont pas indépendantes.

Le test de χ^2 de Pearson est le test standard pour évaluer s'il y a indépendance ou non entre les occurrences de A et de B. Il consiste à calculer une valeur X^2 que l'on obtient en sommant pour chaque cellule de la matrice de contingence le carré de la différence entre l'effectif observé et l'effectif attendu si les distributions de A et de B étaient parfaitement indépendantes, après avoir pris soin de normaliser chaque carré par l'effectif attendu. Il est aisé de calculer les effectifs de la matrice de contingence avec l'hypothèse d'indépendance de A et B à partir des probablités marginales de A et B. Ainsi, pour la première cellule, on observe $n(A, B)$ cooccurrences entre A et B. Le nombre attendu en postulant l'indépendance des deux variables est obtenu grâce aux probabilités marginales de A et B et peut se calculer selon l'équation suivante : $\hat{n}(A, B) = \frac{n(A)n(B)}{n}$. Si on note O_{ij} les composantes de la matrice de contingence observée, et E_{ij} les composantes de la matrice de contingence attendues sous hypothèse d'indépendance (la figure 2.9 fournit l'expression

de chacune des cellules des deux matrices), alors

$$X^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

X^2 suit asymptotiquement une distribution du χ^2 à 1 degré de liberté, ce qui permet d'associer directement sa valeur à la p-value d'un test d'indépendance entre les occurrences des deux termes [34].

34. Par commodité, la mesure de X^2 est parfois simplifiée (Mogoutov et al., 2008)en ne conservant que la première partie de la somme. Celle-ci concentre de toute façon l'essentiel de la contribution à X^2 lorsque la cooccurrence est significative. On obtient alors l'expression (conservatrice) simplifiée suivante : $X^2 = \frac{(n(A,B) - \frac{n(A)n(B)}{n})^2}{\frac{n(A)n(B)}{n}}$

	A	$\neg A$	
B	$n(A,B)$	$n(B)$- n(A,B)	$n(B)$
$\neg B$	$n(A)$- n(A,B)	n -$n(A)$-$n(B)$ +n(A,B)	n- $n(B)$
	$n(A)$	n-$n(A)$	n

	A	$\neg A$	
B	$\frac{n(A)n(B)}{n}$	$\frac{(n-n(A))n(B)}{n}$	$n(B)$
$\neg B$	$\frac{(n-n(B))n(A)}{n}$	$\frac{(n-n(A))(n-n(B))}{n}$	n- $n(B)$
	$n(A)$	n-$n(A)$	n

FIGURE 2.9: Tables de contingence observée (à gauche notée O) et théorique (à droite, notée E) de deux mots A et B qui co-occurrent $n(A,B)$ fois. $n(A)$ et $n(B)$ désignent le nombre total de cooccurrences de A et B respectivement. n le décompte total de cooccurrences (impliquant A, B ou un autre terme). La seconde table théorique correspond aux effectifs théoriques calculés sous hypothèse d'indépendance de A et B.

Notons l'existence d'autres mesures statistiques tels que le t-test ou le ratio de vraisemblance déjà décrit précédemment (équation 2.1) ou le test exact de Fisher, qui permettent également de tester l'hypothèse d'indépendance des occurrences de deux termes.

Pour un exposé plus complet des mesures de similarité directes, on pourra se référer au travail de Pecina et Schlesinger (2006) qui recensent près de 82 mesures différentes. Mais ce qu'il importe de retenir ici, comme l'ont montré Tan et al. (2002), c'est qu'il n'existe pas nécessairement une mesure idéale de similarité entre deux variables. Leur étude porte sur les formes de dépendance entre n'importe quelle source de variable, que l'on s'intéresse à des co-occurrences entre termes ou à une corrélation entre sexe et réussite scolaire. Dans tous les cas, si l'on liste l'ensemble des propriétés souhaitables pour calculer ces corrélations : invariance de la mesure aux effets de volume, nullité de la mesure en cas d'indépendance des deux variables, conditions de symétrie, (etc.) on se rend compte qu'aucune des mesures « classiques » existantes [35] ne peut toutes les satisfaire. Il s'agit donc bien de choisir la mesure la moins insatisfaisante en fonction d'un objectif donné. Voilà qui explique aussi sans doute la prolifération de ces mesures dans la littérature.

35. 21 d'entre elles sont testées dans le travail de Tan et al. (2002) : ϕ coefficient, information mutuelle, index de Gini, cosinus, indice de Jaccard, mesure d'association croissant de façon monotone avec le nombre de cooccurrence, et décroissant avec la taille des effectifs etc.

2.2.3 *Mesures paradigmatiques (dites indirectes)*

Dans le second groupe on retrouve des mesures dites indirectes qui correspondent à des associations paradigmatiques entre termes. L'hypothèse qui est

faite est que la comparaison des contextes d'apparition des termes, capturée à travers leur similarité distributionnelle est corrélée à leur similarité sémantique. Harris (2012) le traduit de façon limpide dans le chapitre introductif intitulé « structure distributionnelle » :

> « *[...]if we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with difference of distribution.* »[36]

Si on renverse l'assertion, il s'agit finalement comme l'affirment Miller et Charles (1991) d'assimiler la similarité sémantique à des relation d'inter-substitutions :

> « *for words in the same language drawn from the same syntactic and semantic categories, the more often two words can be substituted into the same contexts the more similar in meaning they are judged to be.* »[37]

Ce que Church et al. (1994) précisent encore en définissant finalement la similarité distributionnelle sous la forme suivante :

> « *The distributional similarity of two words is the extent to which they can be inter-substituted without changing the plausibility of the sentence.* »[38]

Le modèle distributionnel est souvent assimilé à un modèle géométrique du sens (*word space model*) au sens où il offre une représentation spatiale du sens des mots. La similarité sémantique entre deux mots peut être représentée dans un espace à N dimensions, dans lequel deux mots similaires sont également proches (ce qui sous-entend également qu'un mot occupe une position unique, une hypothèse atomiste sur laquelle nous reviendrons). Modèle distributionnel et représentation géométrique en haute dimension sont donc intimement liés.

Rubenstein et Goodenough (1965) ont été parmi les premiers à tester empiriquement le modèle distributionnel en prouvant qu'il permettait de retrouver des relations de synonymie entre mots. Depuis de très nombreuses mesures de similarité distributionnelle ont été introduites. Etant donnés deux vecteurs qui décrivent le profil de cooccurrence de chaque terme avec l'ensemble des termes d'un corpus, on peut calculer la similarité entre ces deux termes de multiples manières : comme une simple distance euclidienne ou un cosinus, une distance entre leur deux distributions de probabilité (divergence Jenson-Shannon ou distance de Kullback-Leibler par exemple). Weeds et Weir (2005) introduisent et comparent une vingtaine de ces mesures en proposant un modèle systématique de construction de telles mesures.

De façon stylisée, une mesure de similarité distributionnelle peut être décomposée en deux étapes. Dans la première phase, on associe à chaque terme candidat i un profil vectoriel V_i qui caractérise ses contextes d'apparition. Il peut s'agir dans sa version la plus élémentaire d'un vecteur de la taille du vocabulaire et dont les coordonnées correspondent directement aux co-occurrences de i avec des termes tiers. Mais il semble arbitraire de donner

36. « [...]si l'on considère que la différences de sens entre les mots ou morphèmes A et B est plus grande que celle entre A et C, alors on trouvera plus souvent que les distributions de A et B diffèrent plus que celle de A et C. Autrement dit, la différence de sens est corrélée à la différence des distributions. »

37. « pour des mots de la même langue et appartenant à la même catégorie syntactique et sémantique, le plus souvent deux mots peuvent être substitués dans les mêmes contextes, les plus similaires seront jugés leur sens. »

38. « La similarité distributionnelle de deux mesure à quel degré ils peuvent être substitué l'un à l'autre sans modifier la plausibilité de la phrase. »

plus de poids à un terme plutôt qu'à un autre au seul prétexte qu'il est plus fréquent (et le nombre de cooccurrences est mécaniquement fortement corrélé à la fréquence). Une solution alternative consiste à pondérer la contribution individuelle de chaque dimension de notre profil vectoriel en fonction de la présence « préférentielle » de ce terme dans le contexte du terme candidat. Les coordonnées de V_i peuvent donc être définies en employant l'une des mesures de similarité directe que l'on a décrites dans la section précédente. Par exemple Fung et McKeown (1997) étaient les premiers à utiliser l'information mutuelle entre un terme et le reste du vocabulaire pour définir un profil vectoriel dont les coordonnées valent alors $V_i(j) = I(i,j) = log(\frac{n(i,j)n}{n(i)n(j)})$.

Une fois ces profils construits, dans une deuxième étape, et en suivant la définition première de la similarité comme mesure de substituabilité, différentes métriques peuvent être envisagées pour mesurer à quel degré le profil d'un terme i, ressemble, au moins sur son support (c'est à dire, pour les termes qui se retrouvent (préférentiellement) dans son contexte), au profil contextuel d'un terme j. On peut privilégier une approche ensembliste et définir la similarité sémantique entre deux termes i et j comme un coefficient de Dice ou de Jaccard entre les supports de i et j. On peut également concevoir n'importe quel type de distance distributionnelle entre les deux profils mesurés et en suivant (Weeds et Weir, 2005) adopter une approche « différentielle » qui mesure précisément les différences de prévalence de tous les terme tiers dans le contexte respectif des deux termes candidats.

C'est cette dernière mesure calculée en définissant les profils contextuels comme des informations mutuelles que nous avons privilégiée dans nombre de projets cartographiques (Rule et al., 2015; Raimbault et al., 2016; Weisz et al., 2017).

$$s_{d^I}(w_1, w_2) = \frac{\sum_{c, I(w_1,c)>0} (min(I(w_1,c), I(w_2,c))}{\sum_{c, I(w_1,c)>0} I(w_1,c)}$$

Sa particularité est d'être asymétrique. Si $s_{d^I}(w_1, w_2)$ est fort, de l'ordre de 0.5 par exemple, l'inverse n'est pas forcément vrai, tout simplement car la relation de substituabilité n'est pas symétrique[39]. Notons à nouveau que ces mesures de similarité sémantique semblent entretenir des liens très forts avec les méthodes de plongement de mots. Levy et Goldberg (2014) ont ainsi montré que le modèle word2vec, dans sa version la plus populaire (skip-gram avec échantillonage négatif (SGNS)) revenait en réalité à factoriser un matrice mot - contexte en utilisant l'information mutuelle comme métrique (à une constante près).

39. On peut aisément remplacer le mot chat par animal, remplacer animal par chat risque de générer des phrases moins plausibles, voire risibles. Notons qu'avec ce type de mesure de similarité le degré entrant d'un terme mesure sa capacité à épouser la distribution de contexte d'autres mots. Les mots les plus génériques (ce qui ne veut pas dire qu'ils soient forcément les plus fréquents) ont ainsi le plus fort degré entrant.

2.2.4 *Quelle(s) métrique(s) ?*

Les mesures que nous avons décrites capturent des couples de termes dont les contextes sont substituables. Cela ne veut pas dire que substituer les mots eux-mêmes ne changera pas le sens de la phrase. Ainsi, deux termes très proches par rapport à ces métriques distributionnelles peuvent très bien être antonymes. Par exemple les adjectifs *chaud* et *froid* apparaissent généralement dans des contextes très similaires alors même que leur sens est diamétralement opposé. C'est une des raisons pour lesquels il est délicat de choisir entre les différentes métriques. Il est possible d'évaluer leur performance, mais encore faut-il savoir par rapport à quelle tâche : hyponymie (Weeds et al., 2004), synonymie (Rubenstein et Goodenough, 1965), pseudo-désambiguisation (Weeds et Weir, 2003), hypernymie (Lenci et Benotto, 2012), test d'association, mais aussi antonymie (Sahlgren, 2006), etc. Les mesures donnent des résultats différents en fonction de ces différentes corpus de références. Dès lors, il n'est pas évident d'en choisir une plutôt qu'une autre. Sahlgren (2006) insiste beaucoup sur les différences entre mesures de nature syntagmatique et paradigmatique[40]. Notre expérience intuitive montre *a contrario* que les deux mesures aboutissent à des résultats relativement équivalents qualitativement. Lorsque le nombre d'événements de cooccurrence est plus faible (de l'ordre de quelques unités), les mesures directes sont naturellement plus fragiles[41] Les mesures paradigmatiques comparent des profils de cooccurrences plus riches ce qui garantit une plus grande robustesse *a priori*. Mais pourvu que les données soient suffisamment larges et en dépit du résidu de bruit généré par quelques cooccurrences potentiellement plus rares, on ne trouve généralement, du point de vue de la cartographie, pas de différence qualitative majeure entre les résultats obtenus avec l'un ou l'autre des types de mesure. La métrique idéale est finalement sans doute celle qui résulte en une carte permettant l'interprétation la plus riche.

40. Pour être précis, il montre que la différence tend à se réduire lorsqu'on augment la taille des contextes servant de base pour calculer les cooccurrences. Pour autant, il conclut bien qu'elles sont inconciliables.

41. Rappelons que le test du χ^2 exige pour être valide par exemple que les cellules de la matrice de contingence soient toutes au moins égales à 5. Avec des effectifs plus faibles, on s'expose naturellement à plus de bruit statistique.

Mais le choix de la métrique n'est pas le seul paramètre important : la taille du contexte dans lequel on peut considérer que deux termes co-occurrent est également déterminant. La taille du contexte est également une variable déterminante dans les autres méthodes. Là où les *topic model* et plus largement *LSA*, conçus dans la tradition de la recherche d'information, envisagent l'ensemble du document comme contexte, les méthodes dites géométriques (plongement de mots ou « méthode Alceste ») définissent des contextes de taille souvent beaucoup plus réduite (classiquement, une fenêtre glissante de 5 à 10 mots dans le premier cas et la fameuse unité de contexte élémentaire d'une dizaine de mots dans l'autre cas). Cette précision est loin d'être cosmétique. Un contexte de taille réduite ne permettra que de capturer les relations syntaxiques entre mots. S'il est un peu plus large, on peut espérer mieux modéliser les relations sémantiques. À l'échelle du document, on

capture uniquement les grands sujets. C'est une interrogation que partage même Prospero dans lequel les liens entre entités sont pondérés selon qu'ils participent à la même « épreuve » ou non. Dans nos travaux, la taille adéquate du contexte s'avère aussi dépendre de la nature des textes, si les paragraphes des comptes-rendus de l'ENB (voir section 1.1.5) sont des séparateurs naturels, un verset de la bible est potentiellement insuffisant et on peut prouver qu'une fenêtre glissante couvrant plusieurs versets permet d'améliorer la qualité de la description finale (exprimée sous la forme d'un compromis modularité - couverture, voir carte2.10) .

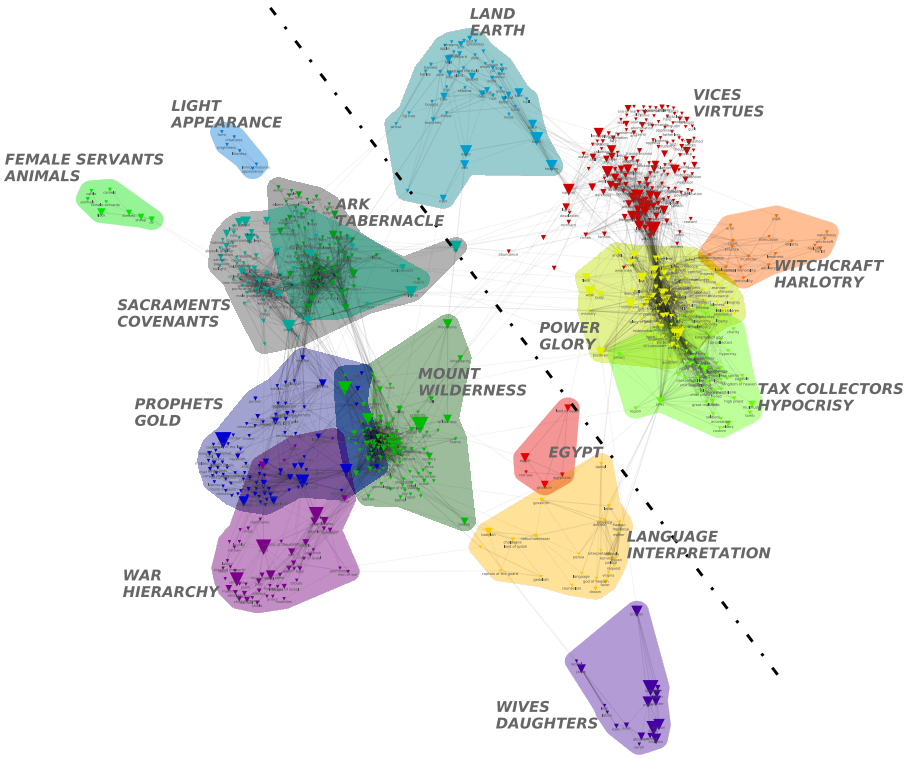


FIGURE 2.10: Carte sémantique de la bible (version modernisée de la bible de King James). Les cooccurrences sont calculées en considérant une fenêtre glissante de 5 versets et une mesure de similarité distributionnelle s_{dl}. La diagonale distingue grossièrement entre les considération morales (à droite) et le récit biblique à proprement parler (à gauche). Certains liens, bien que logiques, ne peuvent que surprendre. Ainsi, sans rentrer dans les détails, *female servant* se retrouve en situation d'équivalence structurelle avec *donkey* ((cluster *Female servants & Animal*). Les vices de nature sexuelle ont une position singulière (cluster *Witchcraft & Harlotry*) .

Donnons tout de suite un exemple pour illustrer cette stabilité relative. Nous avons testé neuf mesures de similarité sur un jeu de données assez simple composé de près de 40 000 recettes différentes provenant de Yummly, une grande plateforme de recettes de cuisine californienne [42]. Chaque recette est composée d'un certain nombre d'ingrédients (11 en moyenne). Nous nous servons de ce jeu de données de test pour construire différents réseaux qui sont tous composés du même nombre de nœuds : les 200 ingrédients les plus fréquents [43]. Un réseau est produit pour les neuf mesures de similarité suivantes : cooccurrences brutes, indice d'inclusion, indice de Jaccard [44], test de χ^2, coefficient d'association spécifique, test du rapport de vraisemblance G^2, information mutuelle, cosinus (mesure très classique en scientométrie,

42. Plus précisément, le jeu de données que nous utilisons a été publiquement partagé à l'occasion d' une compétition sur https://www.kaggle.com/c/whats-cooking intitulée « What's cooking ? ».

43. Les ingrédients ont été normalisés préalablement pour regrouper singuliers et pluriels (cucumber = cucumbers) ou regrouper les ingrédients manifestement identiques (kernel corn=corn kernel=corn kernels).

44. La similarité entre deux ingrédients i et j s'exprime comme le rapport entre le nombre de recettes mentionnant conjointement les deux ingrédients divisés par le nombre total de recettes mentionnant au moins l'un des deux ingrédients : $s = \frac{n_{ij}}{n_i + n_j - n_{ij}}$.

notamment pour l'analyse des réseaux de co-citations (Leydesdorff, 2008)) et encore deux autres mesures de similarité distributionnelle fondées sur l'information mutuelle (s_{d^I}) et le test du rapport de vraisemblance ($s_{d^{G^2}}$). Dans chaque cas, on ne conserve dans le réseau que les 1400 liens les plus forts[45]. Le réseau étant filtré, certains nœuds peuvent être isolés dans le réseau de similarité final, auquel cas ces nœuds ne figurent pas dans la représentation finale[46]. Les cercles colorés figurent les différents clusters détectés (cf. section 2.3) pourvu qu'ils soient de taille suffisante (*i.e.* composés de plus de trois ingrédients). Les neuf cartes finales sont compilées sur la figure 2.11. Si l'on excepte la première ligne qui correspond à des métriques plus anciennes, mesures directes (seconde ligne) et indirectes (troisième ligne) produisent des clusters d'ingrédients qualitativement comparables en dépit d'un rendu visuel différent.

45. Ce choix est parfaitement arbitraire (même s'il permet une bonne lisibilité du réseau.) mais nous reviendrons plus tard (section 2.3.1) sur les procédures de filtrage de réseau

46. Nous avons choisi de n'afficher que les 1400 liens principaux afin qu'un minimum de nœuds soient isolées tout en garantissant une densité de liens « raisonable ».

Livrons nous à une rapide description des résultats obtenus en commençant par les cartes figurant sur la première ligne. La première carte en haut à gauche est simplement obtenue en sélectionnant les 1400 couples d'ingrédients qui occurrent le plus souvent. Naturellement la structure est très hiérarchisée, le réseau héritant du caractère hétérogène (la fameuse loi de Zipf) de la distribution de fréquence des ingrédients dans les recettes[47]. *Salt*, l'ingrédient le plus répandu, est ainsi lié à 145 autres ingrédients, s'accaparant plus d'un cinquième des liens[48]. Autres ingrédients très populaires *onions*, *olive oil* et *pepper* sont également très connectés et se retrouvent aussi dans une position centrale d'où ils inondent le réseau de liens non spécifiques qui semblent « cacher » les structures plus intéressantes[49]. Ce premier exemple est intéressant car il correspond à l'application brute d'une approche de type réseaux sociaux à ces réseaux sémantiques. Or, la nature des relations entre entités est ici bien statistique, si bien que dénombrer de simples « interactions » ne donnera jamais à voir autre chose que la variabilité des fréquences d'apparition des mots. La seconde carte calculée grâce à l'indice d'inclusion de Callon souffre du même problème (*salt* accapare même plus de 27% des liens, un autre quart pour les seuls *garlic* et *onions*). Ce n'est pas une surprise pour un coefficient censé reproduire les relations hiérarchiques entre éléments. Il y excelle effectivement avec comme corollaire que de très nombreux ingrédients se retrouvent exclusivement liés au nœud central. On parvient néanmoins à déceler comme précédemment le cluster bleu foncé regroupant des ingrédients de patisserie ou le cluster vert qui semble réunir des ingrédients de la cuisine italienne. La troisième carte est un réseau dont les liens sont calculés avec un simple coefficient de Jaccard. La structure communautaire est toujours assez confuse mais tout de même plus claire que dans les deux cas précédents. On reconnait à nouveau le cluster composé d'ingrédients pour les desserts en bleu foncé, un cluster vert clair pour la cuisine italienne, un cluster violet contenant essentiellement des ingrédients de la cuisine asiatique, un cluster vert foncé de cuisine mexicaine, etc.). On observe également que les noeuds

47. Les ingrédients dans les recettes de cuisine, comme les groupes nominaux dans la langue ou les références citées dans un corpus de publications scientifiques, suivent une distribution non homogène souvent comparée à une loi puissance.

48. Par défaut, on considère que les réseaux sont orientés, dans le cas présent, la mesure de similarité étant symétrique, tous les liens sont donc « dupliqués » et seuls 700 connections apparaissent sur la carte finale.

49. en y regardant « de plus près », on reconnaît tout de même un cluster bleu correspondant aux recettes de gâteaux, un cluster vert clair pour la cuisine asiatique, mais l'ensemble est plutôt confus visuellement.

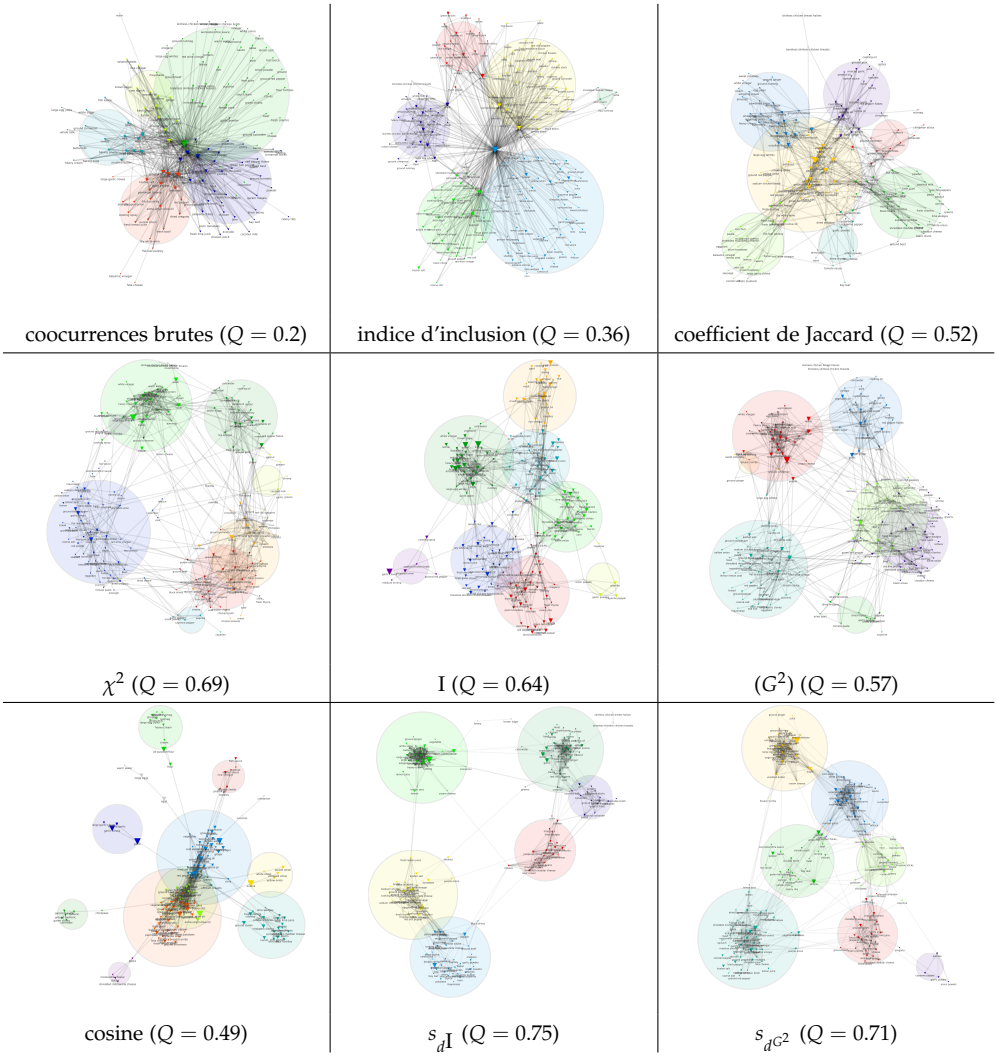


FIGURE 2.11: Les neuf réseaux liant les 200 ingrédients de recette de cuisine les plus fréquents en utilisant différentes mesures de similarité à nombre de lien fixe, les coefficients de modularité Q sont indiqués entre parenthèses.

de même taille ont tendance à être liés. En effet, le coefficient de Jaccard est tel que les ingrédients les plus utilisés ne peuvent être fortement connectés à des ingrédients plus rares, leur cooccurrence au numérateur étant bornée par le plus rare d'entre eux mais leur dénominateur étant au moins de l'ordre de grandeur de la fréquence de l'ingrédient le plus répandu. Sans donner lieu à un réseau étoilé comme pour l'indice d'inclusion où les ingrédients les plus fréquents attirent la majorité des liens, les mesures ensemblistes (Jaccard, Dice), parce que leur normalisation ne prend pas en compte la variabilité (potentiellement très forte) de la taille des éléments, imposent des contraintes très fortes sur les liens qui hypothèquent les chances de faire émerger des structures sémantiques réellement pertinentes.

La deuxième ligne du tableau rassemble le coefficient d'association spéci-

fique [50] et des mesures plus récentes qui s'appuient sur des modèles statistiques plus rigoureux tels le test de χ^2 ou le test de rapport de vraisemblance noté G^2. Dans la carte centrale, qui représente le réseau de similarité calculé à partir du coefficient d'association spécifique (ou l'information mutuelle à nouveau, ils donnent les mêmes réseaux, au poids des liens près), le nombre de cooccurrences est normalisé par le produit des fréquences de cooccurrrences respectives des termes, résultant en une distribution des liens sur la carte qui ne dépend plus de la fréquence respective des ingrédients comme précédemment. Bien que quelque peu confuse (le cluster bleu clair de cuisine asiatique au centre n'est pas très bien délimité par exemple), une structure modulaire apparaît bien, qui laisse à nouveau voir un cluster vert foncé décrivant les ingrédients de pâtisserie, mais aussi, un cluster rouge avec des produits typiquement italiens, etc. [51] La carte réalisée avec le test du χ^2 est également normalisée de façon à ce que les liens ne dépendent plus de la fréquence respective des ingrédients. Elle ressemble à la précédente malgré des frontières entre clusters plus confuses en apparence (sa modularité est pourtant relativement supérieure aux deux autre mesures de la même famille). La troisième carte produite en calculant le test du rapport de vraisemblance entre chaque couple d'ingrédients est sans doute la plus convaincante jusqu'à maintenant [52]. Les différents clusters agrègent l'essentiel des liens de la carte et fournissent un découpage plutôt convaincant de l'espace des recettes avec 4 clusters principaux : les ingrédients rentrant dans la composition de desserts (en haut à gauche en rouge), les ingrédients d'origine asiatique (en bleu en haut à droite) ou typiques de la cuisine mexicaine (en bas à droite en violet), les ingrédients de la cuisine italienne (en bas à gauche, cercle turquoise).

Enfin, la dernière ligne réunit des mesures de nature purement distributionnelle qui donnent les meilleurs résultats. Pour rappel, les métriques indirectes mesurent la similarité entre deux ingrédients en comparant les ingrédients tiers avec lesquels se retrouvent les deux ingrédients dans l'ensemble des recettes. Ainsi la mesure du cosinus consiste à mesurer le profil vectoriel brut de cooccurrences entre ingrédients, puis à calculer le produit scalaire normalisé entre tous les couples d'ingrédients candidats. Les profils de départ étant renseignés par le nombre brut de cooccurrences entre ingrédients, il en découle naturellement que le résultat du produit scalaire dépendra largement du nombre de cooccurrences des deux ingrédients candidats avec les ingrédients les plus communs (sel, poivre, oignons, etc.). C'est sans doute là l'explication de la relative confusion qui règne dans cette carte composée de petits clusters périphériques renvoient à des spécialités culinaires très précises (pâtes italiennes en bas à gauche en jaune, tortillas mexicaines à droite en bleu, pâtes asiatiques en haut en violet) mais aussi de clusters centraux sont structurellement trop recouvrants pour proprement distinguer différents types de cuisine. Les deux cartes suivantes, calculées en pondérant les dimensions des profils vectoriels des termes candidats en fonction du degré de familiarité

50. On peut remarquer que ce coefficient est une mesure d'information mutuelle au logarithme près.

51. C'est d'ailleurs la mesure employée par Teng et al. (2012) pour construire la carte des ingrédients des recettes de cuisine du site allrecipes.com

52. Le test du rapport de vraisemblance étant moins restrictif que le test du χ^2, on pouvait s'attendre à ce que ce réseau soit moins bruité.

entretenu entre le mot candidat et un terme tiers (via un mesure d'information mutuelle ou de rapport de vraisemblance) produisent des réseaux beaucoup mieux clusterisés pour lesquels chacun des clusters correspond bien à une spécialité culinaire reconnaissable. Sur la dernière carte, en partant de la partie supérieure et en parcourant les clusters dans le sens horaire, on observe le cluster d'ingrédients sucrés (en jaune en haut : *milk*, *white sugar*, *egg yolks*, etc.), deux clusters de cuisine asiatique (en vert et en bleu foncé) le second correspondant plus spécifiquement à la cuisine indienne (*tumeric*, *garam masala*, etc.), un cluster central vert qui semble regrouper les plats de la cuisine cajun (*green bell pepper*, *shrimp*, *garlic*, etc.) un cluster rouge réunissant des ingrédients typiques de la cuisine mexicaine (*corn tortilla*, *avocado*, *fresh lime juice*, etc.) et enfin un cluster turquoise en bas à gauche contenant les ingrédients de la cuisine italienne (*olive oil*, *mozzarella cheese*, *basil*, etc.).

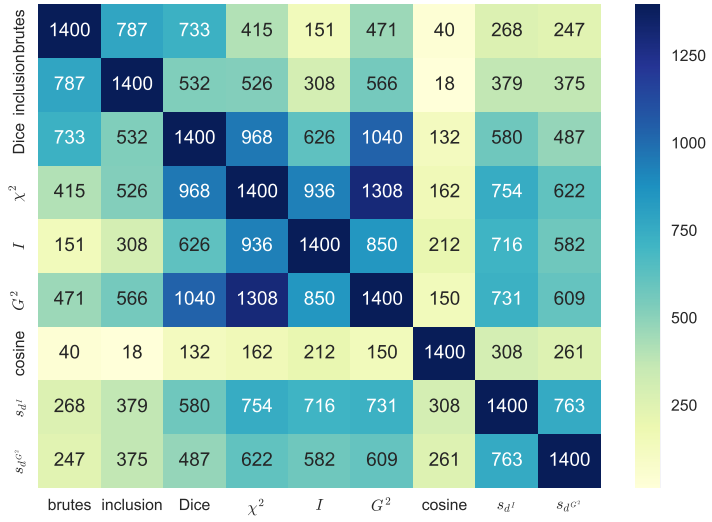


FIGURE 2.12: Nombre de liens partagés entre deux cartes. Si la carte produite avec la mesure d'inclusion ne génère que 18 liens en commun avec celle produite avec la mesure du cosinus, d'autres couples de distances produisent des réseaux beaucoup plus semblables. Mesure du χ^2 et test de vraisemblance G^2 se ressemblent beaucoup par exemple avec plus de 93% des liens partagés.

La comparaison de réseaux n'est pas chose aisée. Mais la tâche est ici facilitée du fait que nos neuf cartes réunissent le même nombre de nœuds et de liens. Nous avons donc décidé de représenter figure 2.12 le nombre de liens partagés entre les différents couples de réseaux. Différentes structures intéressantes apparaissent visuellement. D'abord, les mesures relevant des mêmes familles (et correspondant à chaque ligne de la figure 2.11) tendent à produire des réseaux comparables. Indice d'inclusion, coefficient de Dice et mesure brute de cooccurrences sont des mesures syntagmatiques de l'intensité de la corrélation entre deux entités. Les mesures de similarité de type χ^2, information mutuelle, et G^2, sont toutes des mesures statistiques de l'existence d'une relation syntagmatique pour un couple d'entités donné et génèrent des

réseaux très semblables[53]. Enfin, les trois dernières mesures correspondent à des relations paradigmatiques entre éléments (le cosinus de par sa normalisation discutable donnant les résultats les plus exotiques parmi toutes les mesures). Si les mesures syntagmatiques de type statistique entretiennent une proximité particulière les unes avec les autres, le recouvrement avec les réseaux fondés sur des mesures paradigmatiques est quasiment comparable (si on exclue à nouveau le cosinus) et la distinction théorique entre les deux types de relation entre termes *in praesentia* et *in absentia* s'estompe empiriquement. Cela ne signifie pas que toutes ces mesures se valent (la structure des deux derniers réseaux est visuellement beaucoup plus claire) seulement à cette échelle et sur ce jeu de données, il ne semble pas y avoir de différences qualitatives particulières entre similarités syntagmatiques et paradigmatiques.

53. On remarque que le réseau généré par l'indice de Dice a autant de liens partagés avec les mesures de la seconde ligne que de la première. Comme on l'a déjà commenté, cette ressemblance est logique, la force des liens mesurés par l'indice de Dice ne dépendant pas, contrairement aux autres mesures de la première ligne, des tailles respectives des entités en rapport.

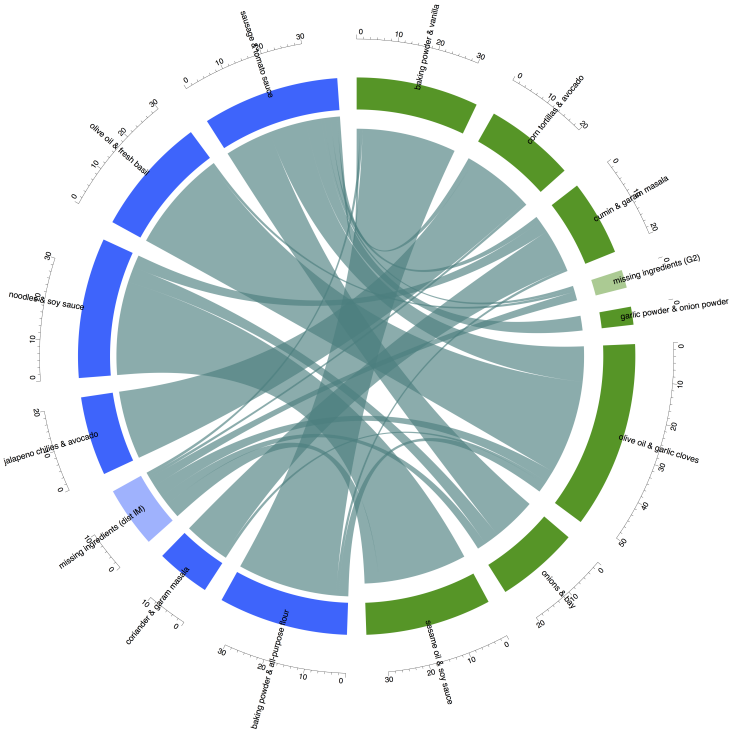


FIGURE 2.13: Comparaison entre les cartes obtenues avec une mesure distributionnelle de type G^2 (en vert) et information mutuelle (en bleu). Les nœuds manquant (*i.e.* ne figurant pas sur la carte car non connecté à tout autre ingrédient) sont figurés par une couleur moins opaque.

Mais ne s'intéresser qu'à la seule stabilité des liens est potentiellement trompeur. En effet, on peut facilement imaginer deux réseaux partageant la moitié de leur liens respectifs avoir des propriétés mésoscopiques tout à fait différentes. Dans la pratique, il semble que ça ne soit pas le cas. Pour mieux capturer les différences de structure entre couples de cartes, nous comparons la composition des clusters au sein d'un diagramme à cordes. Ainsi, la figure 2.13 compare les clusters respectifs de deux réseaux obtenus respectivement avec la mesure de similarité distributionnelle de type information mutuelle et ratio de vraisemblance. Les clusters sont étiquetés de façon automatique par les deux ingrédients les plus centraux de chaque cluster. On se rend compte qu'en dépit d'une homologie relativement faible des liens qu'entretiennent les

deux réseaux (763, soit un peu moins de 55% des liens totaux), leur structure communautaire est très semblable. Les différences notables sont l'absence d'un petit cluster réunissant les épices dans l'un des réseaux (étiqueté *garlic powder & onion powder*, en conséquence de quoi le cluster de cuisine cajun est beaucoup plus riche dans un cas que dans l'autre. Les autres différences sont vraiment minimes ou peu significatives (quelques ingrédients sont échangés entre le cluster de cuisine indienne et le cluster de cuisine asiatique, etc.)

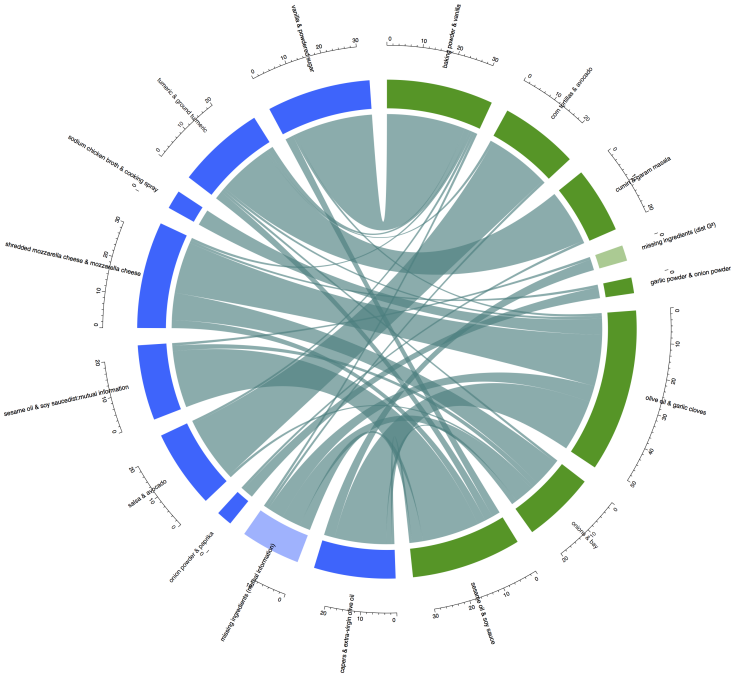


FIGURE 2.14: Comparaison entre les cartes obtenues entre la similarité paradigmatique distributionnelle de type G^2 (en vert s_{dG^2}) et la similarité syntagmatique calculée par information mutuelle (en bleu I) . Les nœuds manquant (*i.e.* ne figurant pas sur la carte car non connecté à tout autre ingrédient) sont figurés par une couleur moins opaque.

Au risque de fatiguer les yeux du lecteur, nous avons effectué la même comparaison entre une mesure de la seconde et de la troisième ligne, à savoir information mutuelle, et mesure de similarité distributionnelle s'appuyant sur le rapport de vraisemblance figure 2.14. Ce nouveau diagramme à corde permet d'apprécier la différence entre les clusters générés par une mesure syntagmatique et paradigmatique. À nouveau, la ressemblance entre les deux carte est frappante. En commençant par le haut du diagramme et en progressant dans le sens des aiguilles d'une montre, on s'aperçoit d'abord que la composition des clusters de desserts est presque parfaitement identique (*vanilla & powdered sugar*, et *baking powder & vanilla*). Il en va de même pour le cluster de cuisine mexicaine (*corn tortiall & avocado*) et *salsa & avocado*, la cuisine indienne fait preuve de la même stabilité (*cumin & garma masala* et *tumeric & ground tumeric*). Les clusters d'épices se font échos tout aussi parfaitement (*garlic powder & onion powder* et *onion powder & paprika*). La différence la plus notable apparaît au niveau du cluster de cuisine méditerranéenne (*olive oil & garlic cloves*) qui dans le réseau d'informations mutuelles se dédouble en un cluster typique de la cuisine italienne (*shredded mozarella & mozarella cheese*) [54]

54. Ce cluster est composé de fromages italiens comme le parmesan ou la mozzarella mais aussi d'ingrédients pour assaisonner des plats de pâtes (en rouge dans la carte d'origine 2.11) on retrouve également dans ce cluster *pasta*, *fresh basil*, *oregano*, etc.

et un second cluster plus typiquement grec (*capers & extra-virgin olive oil*), en bleu marine dans la carte d'origine). Par contre, les ingrédients du cluster de cuisine cajun (*onions & bay*) du réseau paradigmatique ont été absorbés dans le cluster de cuisine italienne ou ont été simplement exclus de la carte de similarité syntagmatique. Enfin, on retrouve de part et d'autre des clusters de cuisine asiatique qui sont très semblables (tous les deux étiquetés *sesame oil & soy sauce*).

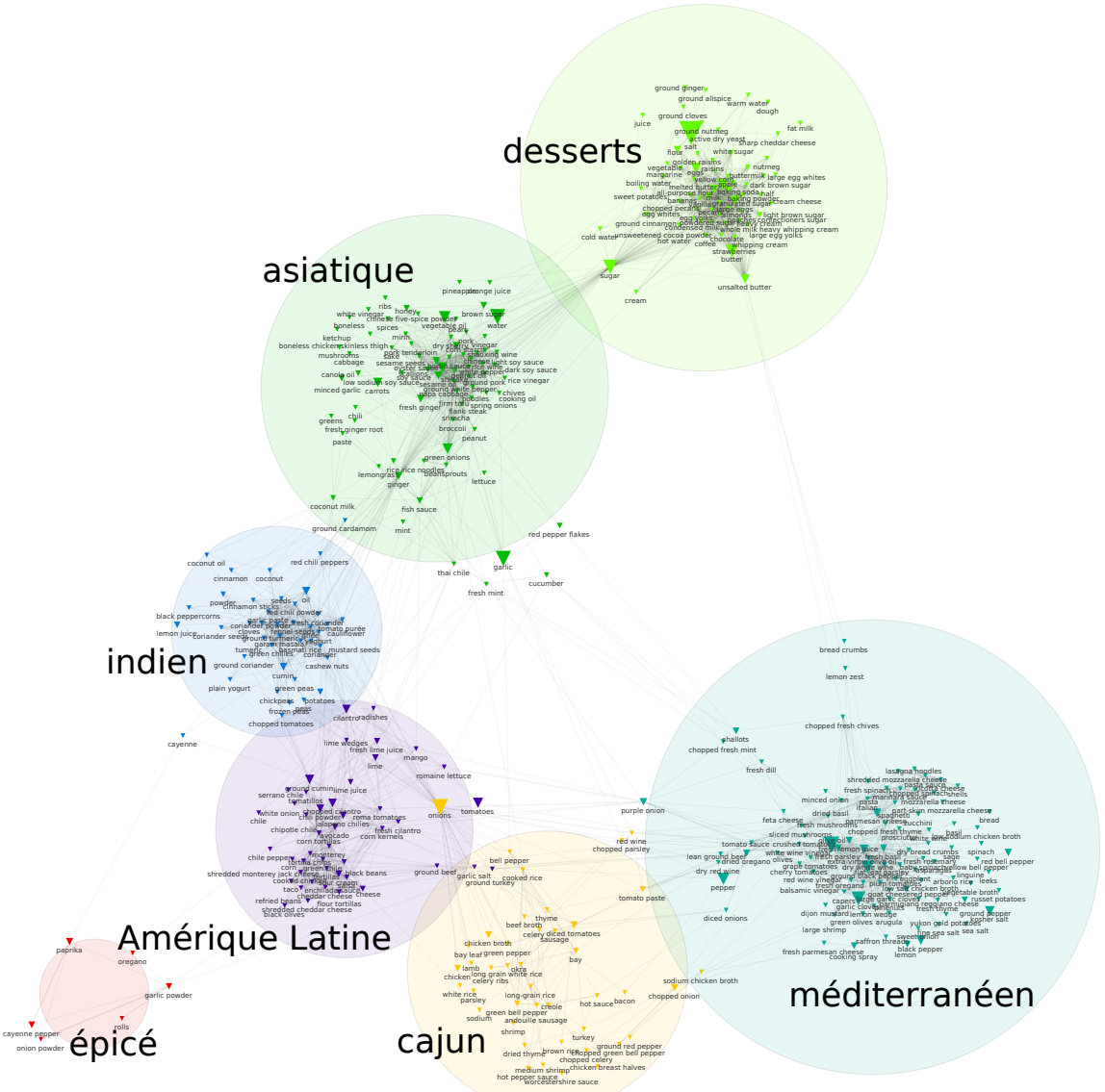


FIGURE 2.15: Carte des recettes - 400 ingrédients - similarité $s_{d^{G^2}}$.

Pour fournir une description plus précise de notre carte des ingrédients, mais aussi montrer qu'au-delà d'un seuil critique le nombre de mots n'a plus qu'une influence mineure sur la structure du réseau, nous avons calculé le même réseau avec 400 ingrédients au lieu de 200 (et deux fois plus de liens

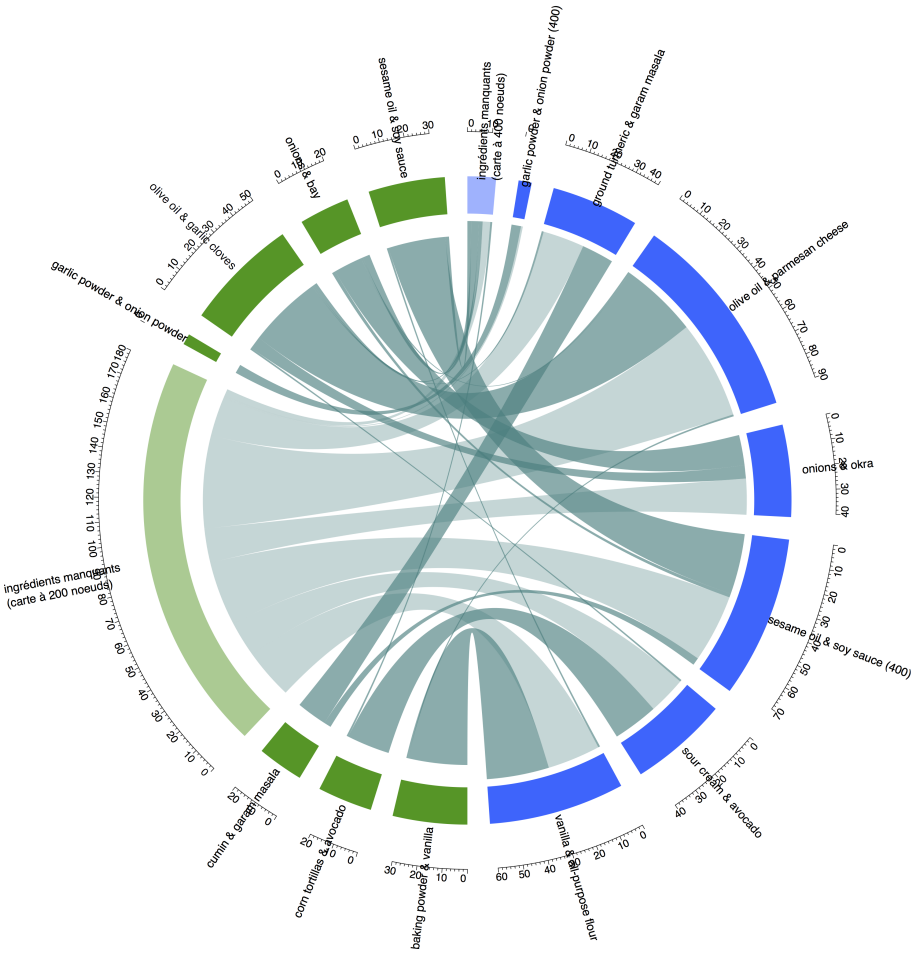


Figure 2.16: Comparaison entre les cartes obtenues avec 200 (à et 400 ingrédients) pour la mesure s_{dG^2}. Les clusters de la carte à 200 ingrédients se trouve sur la gauche du diagramme en vert. Naturellement, un grand nombre d'ingrédients sont aux abonnés absents comparé à la carte à 400 ingrédients. Pour le reste la composition générale est pratiquement inchangée.

pour conserver le même degré moyen). La carte qui en résulte (figure 2.15) a été calculée en utilisant la mesure de similarité indirecte basée sur le rapport de vraisemblance. On retrouve quasiment la même structure que dans la carte contenant 200 termes avec des clusters que nous avons qualifié de desserts, asiatique, méditerranéen, indien, sud-américains, épicé et cajun. Un diagramme à cordes comparant la composition respective des clusters des deux cartes le montre clairement figure 2.16

À l'inverse de l'analyse des correspondances, qui est beaucoup plus ouverte à la panmixie des variables, l'analyse de réseau de proximité nous contraint a une certaine forme d'orthodoxie calculatoire qui ne permet pas vraiment de mélanger des entités provenant de dimensions différentes. C'est que nous avons défini la similarité comme une mesure du caractère substituable d'une entité pour une autre. On peut néanmoins construire des réseaux constitués de nœuds de plusieurs types, des réseaux dites hétérogènes, avec des mesures directes de similarité comme la mesure du χ^2. C'est d'ailleurs ce que nous

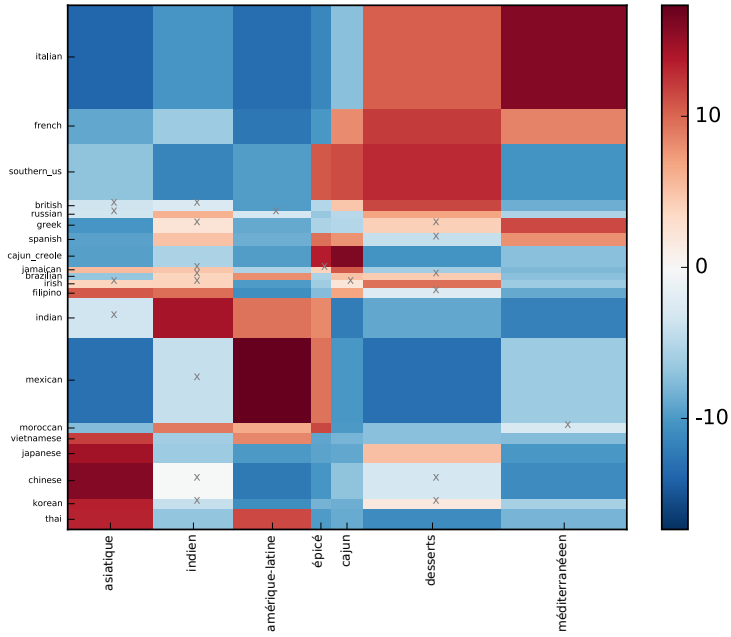


FIGURE 2.17: Matrice de contingence entre origine géographique des recettes (chaque recette était étiquetée en fonction de son origine dans la base de données de départ - les 20 origines géographiques les plus fréquentes sont ordonnées horizontalement) et clusters d'ingrédients (organisés par colonne et réétiquetés par les mêmes étiquettes que celles figurant sur la carte originale (figure 2.15)). La largeur des lignes et des colonnes correspond au nombre de recettes relevant respectivement d'une origine géographique donnée et d'un cluster d'ingrédients donné. La couleur des cellules mesure l'intensité de la corrélation positive (en rouge) ou négative (en bleu) entre les deux dimensions mesurées avec le coefficient de Cramer. De façon additionnelle, un test de Fisher exact est effectué pour évaluer si la non-indépendance des deux variables est statistiquement significative (p-value 0.05).

avons fait avec Sylvain Parasie dans le projet sur la Voix du Nord qui présentait un réseau croisant taille de commune et registres d'expression publiques (Parasie et Cointet, 2012). Or il n'est pas évident a priori de substituer un auteur par un mot, ou une citation par un éditeur, tout simplement parce que leur distribution de contexte ne s'exprime pas nécessairement dans le même espace. Il reste néanmoins naturellement possible de mesurer le degré de corrélation entre des modalités relevant de deux variables différentes. Appliqué à nos recettes de cuisine, cela veut simplement dire que l'on peut comparer la structure de la cartographie sémantique que l'on vient de calculer sur l'ensemble des recettes avec des catégorisations pré-existantes. Pour cela il suffit simplement, étant donné une partition donnée de l'espace sémantique, de « projeter » chaque recette sur le ou les clusters d'ingrédients auxquels elle ressemble le plus [55]. Même si notre méthode cartographique s'appuie uniquement sur les cooccurrences de termes, on voit qu'il est aisé, comme il est possible de le faire avec les topic models, d'en déduire une catégorisation duale des documents.

Nous pouvons ainsi vérifier comment les clusters émergeant des seules cooccurrences entre ingrédients correspondent ou non à l'origine géographique des recettes qui était déjà renseignée dans la base de données de départ. Nous avions ainsi tracé sur la figure 2.17 la corrélation existante entre les clusters de la carte 2.15 déjà décrite et les étiquettes pré-existantes. Les résultats confirment notre intuition première. Ainsi le cluster que nous qualifions d'asiatique est très fortement corrélé aux étiquettes chinoise, coréenne, thaï, japonaise, et vietnamienne. Le cluster que nous avions qualifié

55. Voir (Rule et al., 2015, SI) pour une description technique, la procédure d'appariement s'appuie sur la structure du réseau, les nœuds les plus centraux de chaque cluster ayant plus de poids pour assigner les recettes à leur(s) cluster(s) que des termes en position périphérique

de méditerranéen corrèle positivement avec les catégories pré-existantes de cuisine italienne, mais aussi grecque, espagnole et avec une intensité moindre, française). Les cuisines du sud des Etats-Unis, française, italienne, irlandaise, ou russe semblent plus portées sur le cluster de desserts. La cuisine philippine semble relever à la fois des clusters de cuisine asiatique et indienne.

2.3 *Cartographier*

En discutant la variabilité des mesures de similarité, une bonne partie du protocole de construction des cartes a été divulgué... Il n'y a de toute façon pas grand secret en la matière. On fait l'hypothèse que les termes pertinents [56] s'agencent au sein du réseau de similarité en groupes cohésifs qui définissent autant de thématiques qui structurent une discussion. La description de ces agrégats est une étape indispensable pour lire les réseaux de similarité. À la manière des textes qu'on peut aussi bien lire au plus près jusqu'à interroger le sens de chaque mot, que lire à distance (Moretti, 2004), les réseaux peuvent être lus à différentes échelles , les clusters formant l'échelle intermédiaire entre l'analyse fine du rôle de chaque terme dans la topologie jusqu'aux propriétés globales du réseau final (comme son diamètre par exemple ou la taille de sa composante connexe principale dont on illustrera l'intérêt sur des réseaux de collaboration dans le chapitre suivant (secion 3.3.1)). Mais avant de clusteriser nos réseaux, il nous faut lever le voile sur la procédure précise permettant de transformer la matrice de similarité calculée en un réseau de proximité à part entière.

56. Notons qu'un terme impertinent qui aurait eu le malheur de se trouver dans la liste de termes de départ, aura de fortes chances de ne pas figurer dans la carte finale. Par définition, si sa distribution n'est biaisée vers aucun autre terme, il risque d'être lié trop faiblement au reste du réseau pour participer à la topologie finale de la carte.

2.3.1 *Filtrer le réseau de similarité*

Nous avons décrit la façon dont des mesures statistiques couplées à une analyse morphosyntaxique permettent d'extraire un certain nombre de termes qui constituent autant de nœuds dans le réseau de similarité. Les liens entre nœuds sont *a priori* orientés et pondérés. Ces réseaux ne ressemblent pas tout à fait aux réseaux de l'analyse de réseaux sociaux classique (Wasserman et Faust, 1994) ni même aux « réseaux réels » issus de l'observation du web à grande échelle (Barabási, 2016, chapitre 2) simplement parce que contrairement à ces derniers, les réseaux de similarité sont généralement des réseaux denses. À titre d'exemple un réseau de cooccurrences entre ingrédients non filtré (représenté figure 2.19) serait composé de 19 241 liens, ce qui signifie que quasiment tous les couples d'ingrédients ont été utilisés au moins une fois dans une même recette. Un réseau de 200 nœuds complet contient en effet 19 900 liens. Avec une densité de 0.967, deux ingrédients déconnectés

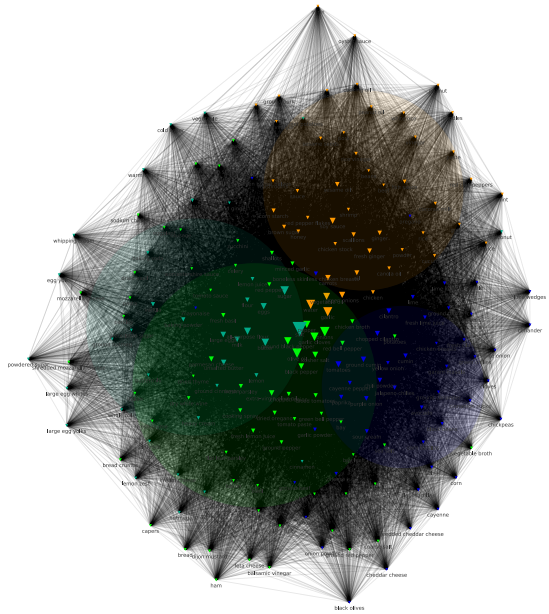


FIGURE 2.18: Carte de cooccurrences des ingrédients, avec un seuil fixé à 0

font figure d'exception (c'est le cas par exemple de la crème chantilly qui s'accommode probablement difficilement avec le piment de cayenne). La première conséquence de cette forte densité, c'est que toute forme d'analyse structurale et surtout visuelle est rendue très périlleuse comme l'illustre la figure 2.19 [57].

Alors que certaines mesures de similarité permettent de réduire le nombre de liens *a priori* (par exemple, on crée un lien entre deux nœuds i et j dans un réseau de similarité de type χ^2 que dans l'éventualité où le nombre de cooccurrences entre de i et j est supérieur au nombre attendu sous hypothèse d'indépendance des deux distributions), d'autres (et notamment l'ensemble des mesures paradigmatiques que nous avons testées) s'étendent entre 0 (profil de contexte à recouvrement nul) et 1 (égalité parfaite des profil de distribution de contextes) avec une densité avant filtrage potentiellement encore supérieure à celle des réseaux bruts de cooccurrence.

Il est donc naturel de filtrer ces réseaux pour mettre de côté les liens les plus faibles. Dans le cas des mesures directes fondées sur une mesure statistique de la dépendance d'un nœud avec un autre, on pourrait considérer par défaut que tous les liens dépassant un seuil donné (typiquemement sélectionner tous les liens dont la mesure du χ^2 est supérieure à 3.841 pour éliminer avec 5% de risque d'erreur l'hypothèse que les occurrences de deux nœuds liés sur la carte soient en réalité indépendantes). Il est entièrement légitime de procéder de la sorte, mais, en toute rigueur, cette méthode ne permet malheureusement pas d'interpréter le poids des liens comme une intensité de la relation qui les lie, car c'est en réalité le degré de certitude statistique de la dépendance entre les nœuds qu'ils connectent qui est mesurée. La mesure du χ^2 est ainsi

57. Malgré le manque de structure apparent du réseau, un algorithme de détection de communauté a été employé et a classé les ingrédients. Quelque soit la qualité des classes obtenues, il serait de toute façon impossible avec ce type de représentation d'aller au-delà dans l'analyse et de se livrer à un interprétation plus fine de la position de certains ingrédients ou des relations entre clusters.

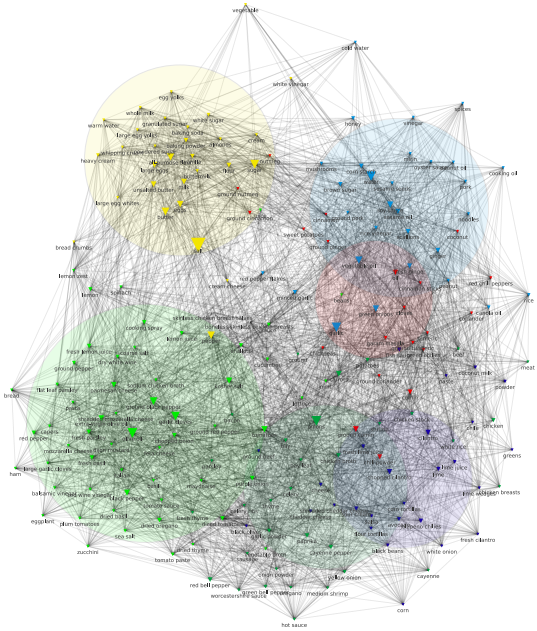


FIGURE 2.19: Carte des ingrédients, score de χ^2, avec un seuil fixé à 6.63 (risque d'erreur fixé à 1%), de nombreux liens correspondant à des corrélations de faible intensité entre ingrédients apparaissent sur la carte.

susceptible d'être moindre, « à dépendance égale », simplement parce que les termes en relations sont moins fréquents. S'ils sont très fréquents, alors, un écart même modéré par rapport au modèle d'indépendance permettra de conclure plus facilement à la dépendance des deux variables. On se retrouve potentiellement dans le cas de figure contre lequel McFarland et McFarland (2015) nous mettent en garde : avec les jeux de données massifs des « big data », on prend le risque de se retrouver submergé par des milliers de corrélations statistiquement significatives qui nous font perdre la vision globale du système. C'est d'ailleurs ce qui se passe sur le jeu de données modestes sur lequel nous expérimentons. La carte des ingrédients devient extrêmement confuse (voir figure **??**) lorsque l'on lie par exemple l'ensemble des couples d'ingrédients dont le score de dépendance du χ^2 est supérieur à 6.63 (avec donc 1% de risque d'erreur[58]). À titre d'exemple deux ingrédients comme « pasta » et « dry white wine » sont liés sur la carte alors qu'ils apparaissent conjointement dans 30 recettes, tandis que l'hypothèse d'indépendance prévoyait un chiffre de 20. Compte tenu de leur fréquence respective, cet écart pourtant faible est statistiquement suffisamment significatif pour créer un lien sur la carte ($\chi^2 = 7.7$)

En somme, par définition, les mesures de similarité fondée sur un test statistique du type χ^2, test du rapport de vraisemblance ou test de Fischer exact, ne peuvent pas être interprétées comme des mesures de l'intensité d'un lien entre deux termes. *A contrario*, les mesures comme l'information mutuelle ou les mesures distributionnelles, sans offrir de garanties statistiques, mesurent bien une intensité de corrélation qui permet d'ordonner les liens

58. On a pourtant pris soin de ne pas calculer de score pour les couples d'ingrédients dont le nombre de cooccurrences est inférieur à 5 selon la règle consacrée pour le calcul du χ^2 dans les tables de contingence.

d'un réseau des relations les plus fortes aux plus faibles indépendamment des fréquences absolues d'apparition des nœuds impliqués. En toute rigueur, il faudrait pour construire des réseaux de similarité syntagmatiques pondérés procéder en deux étapes

— Dans un premier temps sélectionner les couples de nœuds dont la corrélation est avérée statistiquement avec un niveau de risque donné,

— dans un deuxième temps évaluer l'intensité de cette corrélation à l'aide d'un score comme l'information mutuelle.

On se contente donc de travailler à partir de cet ensemble de mesures qui ne sont pas biaisées par rapport à la fréquence des nœuds pour proposer une stratégie de filtrage aussi générique que possible. On part de la situation extrême suivante qui décrit un réseau (quasi complet on l'a vu) où figurent tous les couples de nœuds avec une similarité positive et on élimine progressivement les liens les plus faibles de sorte à appauvrir progressivement le réseau. Si le corpus est structuré par une organisation thématique particulière composée des sous-ensembles sémantiquement distincts, alors le réseau [59] de similarité devrait être modulaire avec une densité supérieure de liens au sein de chaque module mais aussi des liens plus forts entre nœuds participant aux mêmes clusters qu'entre nœuds appartenant à des clusters distincts. Dès lors la détection de communauté est simplifiée (pour l'algorithme mais aussi visuellement) en augmentant le seuil de similarité.

59. On suppose ici que la mesure de similarité retenue capture bien l'intensité de la corrélation entre deux termes.

Pour autant, il semble aussi crucial, pour pouvoir formuler des hypothèses de plus haut niveau sur l'articulation entre ces différentes clusters, de ne pas détruire la structure globale qui lie les clusters les uns aux autres. Il s'agit donc de trouver un compromis entre la présence de connexions entre clusters de façon à pouvoir les situer relativement l'un à l'autre dans l'espace global et un assèchement maximal des liens du réseau qui révèle au mieux les frontières des clusters. Pour arriver à ce compromis, une stratégie efficace consiste simplement à trouver pour notre réseau de similarité le seuil critique à partir duquel il opère sa transition de phase, c'est à dire le seuil précis pour lequel la composante connexe principale qui le compose se délite en plusieurs continents déconnectés les uns des autres. La figure 2.20 illustre une telle recherche de seuil optimal. En pratique, on ne s'arrête pas au premier nœud qui se détache [60] mais à la première grappe de nœuds qui se détache. En définissant un seuil de similarité légèrement inférieur à ce seuil critique, on obtient généralement un réseau dont les clusters sont bien structurés mais encore connectés les uns aux autres ce qui permet d'en proposer une interprétation globale (fût elle aussi triviale que la cuisine asiatique est la plus proche du cluster des desserts car de nombreuses recettes sont sucrées).

60. il suffirait qu'un terme un peu exotique dans le contexte du corpus se soit glissé dans la liste des entités à cartographier pour abaisser dangereusement le seuil

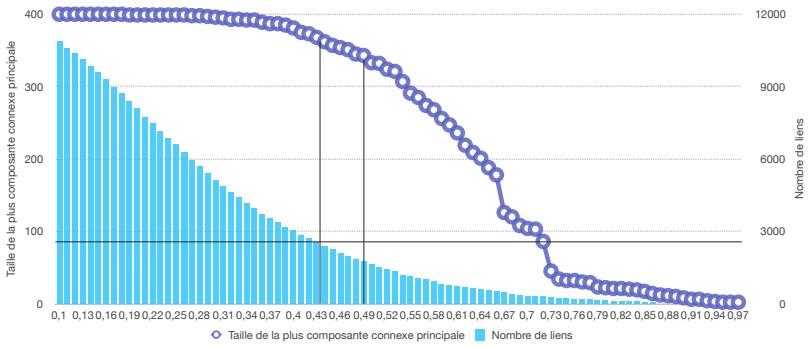


FIGURE 2.20: Evolution du nombre de nœuds et de liens de la composante connexe principale du réseau en fonction du seuil de similarité (en abscisse). On observe que la carte à 400 ingrédients présentée précédemment figure 2.15 et qui contenait 2800 liens correspond en réalité à une valeur de seuil légèrement inférieure au seuil « optimal » dont on a déterminé qu'il était atteint pour une valeur de 0.48 (second trait vertical). En conséquence, le réseau correspondant est un peu moins dense.

2.3.2 *Extraire les champs sémantiques*

Ce que nous cherchons à identifier en toute généralité, c'est la façon dont une « discussion » mobilise des champs sémantiques distincts. Nous empruntons ici la notion de « champ » ou « domaine sémantique » à la linguistique (même si elle a également été discutée en anthropologie (Ingold, 1996)) pour désigner un « segment de réalité » symbolisé par un ensemble de mots dont le sens est proche et se référant au même phénomène (Brinton, 2000). Sans rentrer dans les détails des débats qui animent le champs de la linguistique autour de la définition précise du concept [61], un champ sémantique réunit des termes qui entretiennent des relations de quasi hyponymie les uns avec les autres.

61. Il renvoie d'ailleurs aux notions de similarités paradigmatique et syntagmatique déjà discutées

Sans être entièrement figés, les champs sémantiques traduisent généralement dans le langage les systèmes de croyance propres à une population donnée en un temps donné. Nous préférons la définition plus fluide qu'en donne Ingold (1996, 1991 debate Language is the essence of culture, Part I) :

> « *semantic fields do not stand in relations of opposition to each other, nor do they derive their distinctiveness in this way, nor indeed are they securely bounded at all. Rather, semantic fields are constantly flowing into each other. I may define a field of religion, but it soon becomes that of ethnic identity and then of politics and selfhood, and so on. In the very act of specifying semantic fields, people engage in an act of closure [...]* » [62]

62. « les champs sémantiques ne s'opposent pas les uns aux autres, ils ne se définissent pas vraiment par leur différence, leurs frontières sont à peine sécurisées en réalité. Il est plus juste de dire que les champs sémantiques s'écoulent constamment les uns dans les autres. Je pourrais définir le champ de la religion mais aussitôt il devient celui de l'identité ethnique et plus tard de la politique et le l'individualité, etc. L'acte même de spécifier un champ sémantique est un acte de clôture. »

Cette définition a également l'avantage de remettre l'acte de langage au premier plan plutôt que d'imaginer ces structures comme des invariants qui surplomberait tout phénomène social.

Notre modèle est en définitive assez simple : les acteurs mobilisent des champs sémantiques qui renvoient à des artefacts culturels suffisamment partagés pour être reconnaissables par tous, mais qui peuvent être interprétés par chacun. Le vocabulaire dont ils usent et la façon dont les arguments mêmes lient ces structures sous-jacentes s'agrègent au sein d'un corpus pour constituer une construction collective des champs sémantiques actifs et de leur articulation. Pour prendre un exemple, les champs sémantiques peuvent

résulter de la description que donnent des biologistes de leur travail et aussi bien renvoyer à une discussion technique sur un nouvel outil de screening génétique, à la mention d'une hypothèse immunologique pour décrire le *modus operandi* de cellules cancéreuses ou à des concepts relevant de la théorie de l'évolution pour expliquer la prévalence d'une pathologie dans une population. Dans la littérature sur les mouvements collectifs, on peut rapprocher la notion de champ sémantique du concept de « cadre » (*framing*) à travers lequel peut se lire toute action collective qui offre ainsi un « schéma interprétatif » qui légitime et nourrit l'action :

> « *[...] collective action frames are action-oriented sets of beliefs and meanings that inspire and legitimate the activities and campaigns of a social movement organization.* »[63]

63. « les cadres de toute action collective sont des ensembles de croyances et de significations partagées, tournées vers l'action qui inspirent et légitiment les activités et les campagnes de tout mouvement social. » (Benford et Snow, 2000)

L'analyse des bases de données d'articles de presse avec les outils de cartographie sémantique présentés ci-dessus permettent typiquement de rendre compte de la façon dont ces cadres interprétatifs sont mobilisés par les acteurs et les commentateurs. A titre d'exemple, nous avions dans un travail maintenant assez ancien (Chavalarias et al., 2011) analysé près de 15 ans de discours médiatique sur *food security* et révéler les variations de cadrage que le concept traversait. Ainsi l'insécurité alimentaire se révélait tantôt, sous la plume des journalistes, une fatalité due à la multiplication des catastrophes naturelles, la conséquences de conflits armés locaux, un argument pour justifier le développement d'une agriculture plus durable dans les pays développés, ou la cause d'instabilités politiques au moment des printemps arabes.

Dans tous les cas, l'hypothèse principale que nous posons est simplement que les champs sémantiques qui structurent les discours individuels et collectifs peuvent être retracés dans le langage même, parce que le choix des mots et leur relation sont naturellement informés par la mobilisation de tel ou tel champ sémantique. Reconstruire ces structures sémantiques depuis l'observation d'un texte lui-même élaboré sous contraintes (à commencer par les contraintes grammaticales !) n'est évidemment pas chose aisée. Les champs sémantiques qu'on espère révéler peuvent ainsi être de nature variée correspondant à des vocabulaires recouvrants, se déployant à différentes échelles, etc.

Nous faisons néanmoins l'hypothèse que la structure des réseaux sémantiques offre une fenêtre privilégiée pour les saisir. Ces deux dernières décennies, un nombre impressionnant de méthodes ont vu le jour dans l'idée de révéler la structure modulaire des réseaux complexes (Lancichinetti et Fortunato, 2009; Xie et al., 2013). Méthodes de clustering de réseau, ou de « détection de communautés », elles mobilisent différentes définitions de ce que constitue un cluster. Compte tenu de notre définition des champs sémantiques comme regroupant des éléments partageant une même communauté de sens, on se limite à des algorithmes de clustering qui recherchent les aggrégats cohésifs au sein des réseaux de similarité (*i.e.* les sous-graphes du réseau à

l'intérieur desquels la densité de liens est plus importante)[64] .

Tandis que certains font appel à des méthodes algébriques qui pré-définissent *a priori* des communautés prenant tantôt la forme de cliques (Ploux et Victorri, 1998), quasi-cliques (Abello et al., 2002) ou de percolations de cliques (Palla et al., 2005), la majorité des méthodes de détection de communautés dans les réseaux résulte en réalité de l'optimisation d'un score qui dépend d'un découpage communautaire donné. C'est typiquement le cas de la (très) grande famille des algorithmes visant à optimiser la modularité d'un graphe. La modularité mesure la qualité de la partition d'un graphe (Newman, 2006). S'étendant entre -1 et 1, elle est proportionnelle à la différence entre le nombre observé de liens entre nœuds appartenant à la même communauté et le nombre théorique de ces liens en supposant un modèle nul dans lequel les liens seraient agencés indépendamment de toute structure communautaire. Parmi les solutions à ce problème d'optimisation, l'algorithme dit de Louvain (Blondel et al., 2008) est un des plus populaires de par la qualité des partitions qu'il fournit et sa capacité à traiter rapidement des grands réseaux. C'est cette algorithme qui est utilisé pour dessiner les frontières des clusters dans ce manuscrit. Une autre stratégie commune pour résoudre ce problème d'optimisation est de faire appel à des méthodes spectrales qui impliquent d'analyser le laplacien normalisé du réseau. Cette stratégie n'est pas sans rappeler l'analyse des correspondances. Lelu (2011) nous rappelle d'ailleurs que l'une des premières applications de l'analyse des correspondances portait sur la matrice d'adjacence d'un graphe[65]. Au-delà des méthodes fondées sur l'optimisation de la modularité, il existe encore d'autres sratégies : Rosvall et Bergstrom (2008) ont par exemple développé une méthode fondée sur la compression de l'information pour décrire les flots d'information le long des arêtes du réseau.

Ces méthodes peuvent également être classées en fonction du type de partition qu'elles construisent : certaines permettent, et c'est le cas de l'algorithme de Louvain par exemple, de reconstruire une partition hiérarchique grâce à des clusters définis à différentes échelles. Certaines méthodes permettent également à un nœud d'appartenir à plusieurs clusters de façon simultanée. On dit alors que les communautés sont recouvrantes. C'est le cas par défaut des percolations de clique (Palla et al., 2005), mais quels que soient les principes sur lesquels ils se fondent (De Domenico et al., 2015; Lancichinetti et al., 2009), les algorithmes de détection de communautés proposent maintenant quasiment tous une version avec recouvrement.

Dans tous les cas, sans véritable consensus établi sur ce qui constitue une bonne définition d'une communauté et en l'absence de mètre-étalon pour comparer les résultats sur une tâche empirique, l'analyste est encouragé à la prudence lorsqu'il analyse la structure communautaire d'un réseau[66]. Sans

64. Il faut admettre que la très grande majorité des algorithmes de clustering recherchent des sous-ensembles cohésifs, mais les modèles de blocs qui proposent une catégorisation des nœuds selon des principes d'équivalence structurelle procèdent selon des principes différents par exemple (Peixoto, 2014).

65. En réalité, on peut montrer que la modularité peut s'exprimer à l'aide d'une matrice B qui se construit directement depuis la matrice d'adjacence A sous la forme $B_{ij} = A_{ij} - \frac{k_i k_j}{2m}$ où k_i désigne le degré d'un nœud i et m correspond au nombre total de liens dans le réseau (Newman, 2013). Comme suit : l'optimisation de la modularité par bissection successive du réseau fait apparaître le laplacien de la matrice qui s'exprimer comme suite : $L = I - D_r^{-1}A$ où D_r désigne la matrice diagonale dont les valeurs correspondent aux sommes des lignes de la matrice d'adjacence. Or, on se rappelle dans le chapitre précédent section 1.1.3 avoir déjà rencontré $D_r^{-1}A$ qui n'est autre que la matrice des profils-lignes de la matrice d'adjacence !

66. Plus généralement, on peut même être tenté d'épouser la perspective critique de Grimmer et King (2011) par rapport aux méthodes de clustering en général.

rentrer dans le relativisme à tout crin (pour les réseaux « bien formés », la différence entre un algorithme et un autre sera minime : un cluster peut se retrouver divisé en deux sous-parties, certaines terminaisons seront détachées d'un cluster principal mais les différences seront marginales et sans conséquences pour l'interprétation qualitative *a priori*), visualiser la topologie du réseau en même temps que son découpage en communautés aide à adopter une perspective critique sur le résultat de l'algorithme.

Comme les topic models ou la méthode Alceste, on cherche dans la méthode cartographique à regrouper les termes sous la forme de champs sémantiques. Comme on vient de le voir, ces champs émergent de l'analyse structurelle du réseau sémantique et plus précisément de la structure communautaire de ce graphe. Mais quand Alceste catégorise des textes[67] qui sont progressivement séparés en deux à l'aide d'un clustering descendant hiérarchique[68], alors que les topic models optimisent la distribution jointe de termes au sein de topics et de topics au sein de documents, l'analyse de réseaux sémantiques ne s'attache à catégoriser que les termes, sans avoir à formuler d'hypothèse particulière quant à la distribution de classes sémantiques au sein des documents. Enfin, pour parfaire la comparaison, les méthodes de plongements de mots proposent avant tout une géométrie plutôt que des classes sémantiques.

67. Plus précisément ce sont des segments de texte, les fameuses unités de contexte élémentaires « u.c.e. ».
68. Les mondes lexicaux (Reinert, 2008) ne sont qu'un produit secondaire de cette opération de classification des textes.

2.3.3 *Varier les focales*

Il faut ici bien noter que l'opération consistant à transformer une matrice de similarité entre termes en une carte relève d'une double réduction. D'une part, aucune méthode de réduction de dimensionnalité ne peut parfaitement traduire la richesse d'une matrice de similarité originale en 2 ni même en 3 dimensions. Mais de façon peut-être plus pernicieuse, comme on l'a déjà commenté dans le cas de l'analyse des correspondances, une représentation géométrique embarque nécessairement avec elle l'hypothèse implicite qu'à chaque terme on peut attribuer un point auquel est associé un sens bien précis. Or cette hypothèse rend très mal compte des situations de polysémie pour lesquels on aimerait pouvoir doter nos termes de capacité d'ubiquité.

Partant de mesures de distance entre modalités, l'analyse des correspondances aboutit à un espace géométrique de dimension réduite censé capturer toutes les propriétés du système d'origine : à la fois les relations entre variables deux à deux, les distances entre individus, etc. Naturellement, il est difficile de concilier toutes ces contraintes en si peu de dimensions et s'en suit un certain nombre de difficultés que l'on a déjà listées dans le chapitre précédent (axes expliquant une proportion modeste de l'inertie totale du système, distance trompeuse entre modalités et individus, qualité variable de la représentation

des individus dans le plan factoriel, etc.). Les méthodes de plongement de mots procèdent dans le sens inverse puisqu'elles consistent à directement inférer une géométrie, sous la forme de vecteurs de termes dans un espace à N dimensions (N étant faible mais tout de même plus confortable que dans le cas de l'analyse des correspondances). Le fait que l'espace encode également des propriétés de similarité sémantique n'est alors qu'une conséquence (heureuse) du modèle. Mais les plongements de mots réduisent également chaque mot à une position unique dans l'espace autrement dit à un sens unique. Naturellement, il existe des raffinements complémentaires du modèle tels que des procédures de détection des multiples prototypes sémantiques possibles d'un terme (Huang et al., 2012) mais l'opération est un peu alambiquée...

En somme il semble difficile pour l'ensemble des méthodes de concilier une théorie continue du sens qui offre une géométrie où effectuer des opérations sémantiques comme on opère une relation de Chasle sans, dans un même temps, réduire les termes à des briques élémentaires dénuées de toute nuance. C'est qu'il faut bien des briques de départ pour bâtir l'édifice ! Dès lors comment bâtir un mur droit lorsque certaines briques se plaisent à jouer les transformistes.

La cartographie de réseau permet, en partie, d'échapper à cette aporie et ce pour différentes raisons. La première raison, c'est que l'espace du réseau est un espace à géométrie variable [69]. On peut faire subir à une carte une rotation de 90° sans que la spatialisation des nœuds ne trahisse plus ou moins la topologie originale du réseau. En somme l'espace n'est pas dirigé par quelques variables dominantes et surtout les relations spatiales sont essentiellement locales. Les algorithmes de spatialisation de réseau, généralement fondés sur des modèles physiques tentent de réduire le stress du système en tâchant autant que possible de placer côte à côte des nœuds connectés. L'espace devient un guide de lecture pour la topologie, mais la présence des liens sur la carte reste indispensable pour suivre la façon dont les véritables distances entre nœuds matérialisées par les connexions peuvent parfois tordre l'espace [70].

L'analyse topologique des voisinages d'un nœud donné [71], en particulier si on considère les voisins auxquels il est connecté par des liens entrants, est source d'enseignement quant à la polysémie du terme. On pourrait d'ailleurs imaginer différentes façon de la quantifier : un terme dont le réseau égo-centré est fortement clusterisé (forte proportion de triangles fermés parmi les triades dont le nœud central est le sommet) ou au contraire dont les voisins relèvent de champs sémantiques distincts (polysémie). Dans le cas où un algorithme de détection de communauté avec recouvrement est utilisé, ces termes polysémiques pourront naturellement être catégorisés dans des communautés distinctes. Dans le cas contraire, même si l'algorithme de détection de communauté ne peut leur donner qu'une couleur, il se positionneront entre les

69. Et c'est sans doute la raison pour laquelle la métaphore géographique pourtant très répandue est en réalité extrêmement trompeuse (Plantin, 2013).

70. Vowviewer (ainsi que d'autres logiciels/algorithmes moins connus) a fait un tout autre choix, en préférant travailler la production de cartes de densité (Van Eck et Waltman, 2010) à partir de matrices de similarité (le coefficient d'association de Callon) denses (*i.e.* sans filtrage *a priori* des relations de similarité) en réalisant une opération de réduction de dimensionnalité. Les cartes en résultant prennent la forme de heatmaps, et les relations locales enter termes sont perdues.

71. Comme on l'a déjà mentionné dans la section 2.11 les deux dernières métriques indirectes que nous avons décrites sont asymétriques. Cette propriété permet à des termes aux sens multiples de cumuler des connections vers des termes similaires à l'aune de chacun de ces noyaux de sens.

deux clusters, formant un pont, point de passage obligé pour connecter des clusters distincts et pourront par exemple se distinguer par une *centralité de betweenness* particulièrement élevée (Jensen et al., 2015).

Mais la lecture d'une topologie n'est pas choses aisée, c'est la raison pour laquelle il est indispensable de se doter d'interfaces d'exploration pour naviguer au sein des réseaux à toutes ses échelles, interroger le voisinage d'un nœud, lister les éléments d'un cluster, apprécier la structure de haut-niveau qui lie les clusters les uns aux autres, annoter les champs sémantiques, et ce de la façon la plus fluide possible. On partage là l'aspiration des *datascapes* du médialab de Sciences Po qui prétendent « dissoudre toute discontinuité entre situations et aggrégations » pour rendre sa « continuité » à l'analyse sociologique (Venturini et al., 2017).

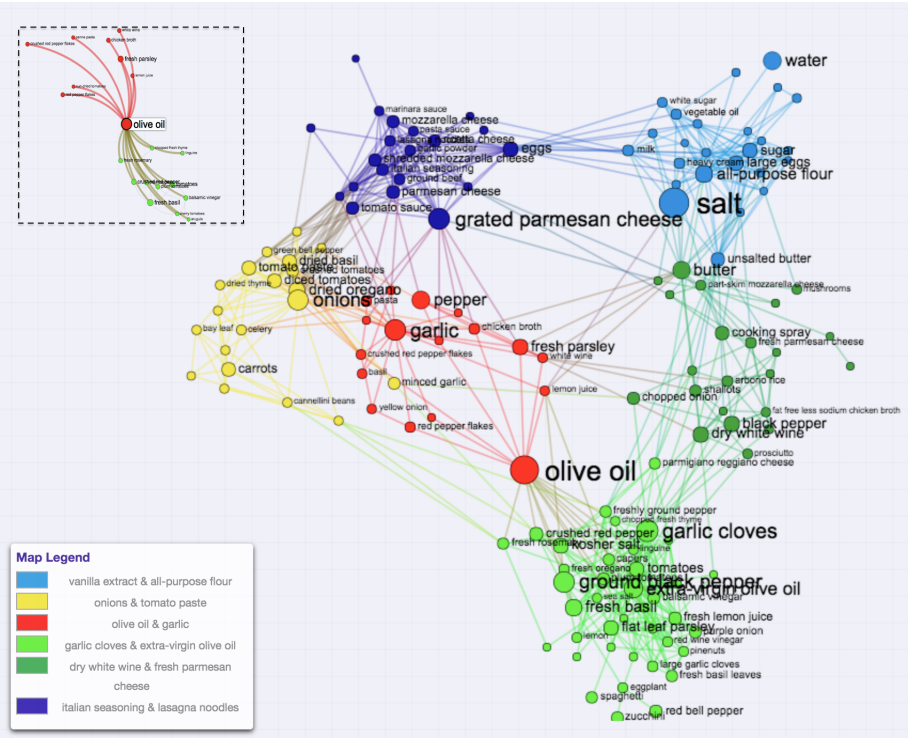


FIGURE 2.21: Carte des 150 ingrédients principaux de l'ensemble des recettes de cuisine italienne. Un corpus limité à ces seules recettes à été construit avant de générer cette nouvelle carte avec la mesure de similarité distributionnelle basée sur le rapport de vraisemblance. Le seuil de similarité a été déterminé automatiquement juste avant la « transition de phase » telle qu'expliquée section 2.3.1. L'utilisateur est invité *via* l'interface à se déplacer dans le réseau (pan, zoom), afficher plus ou moins de labels, identifier le réseau égo-centré autour d'un terme (voir le réseau ego-centré d'*olive oil* en insert en haut à gauche). Plus important encore, les étiquettes des clusters dans la légende en bas à gauche peuvent être éditées pour garder une trace de l'interprétation.

C'est là une des fonctions de CorText que d'assister des chercheurs en sciences sociales dans le choix de métriques, de solutions de filtrage, de méthodes de détection de communautés – rendre disponible ces outils d'analyse tout en pointant du doigt la multiplicité d'hypothèses qu'ils embarquent permet d'éviter un effet « boîte noire » qui condamne le chercheur à une seule option de modélisation sur laquelle il n'a pas de prise. La visualisation des cartes à travers des interfaces web participe également de cette ouverture qui doit permettre de rendre les méthodes et les résultats aussi transparents

que possible, mais aussi d'aider à leur diffusion et surtout accompagner le processus d'interprétation. À titre d'exemple, la visualisation de réseau sur la plateforme se fait maintenant à travers une interface web dédiée permettant de naviguer du texte d'origine aux agrégats. Celle-ci a été développée par André Spritzer à partir d'une première interface d'exploration de graphe développée à l'ISC-PIF et baptisée *graph-explorer* qui utilise le moteur de visualisation de réseau en javascript *sigma-js*. Une capture d'écran de l'interface originale (consultable en suivant ce lien) est présentée figure 2.21.

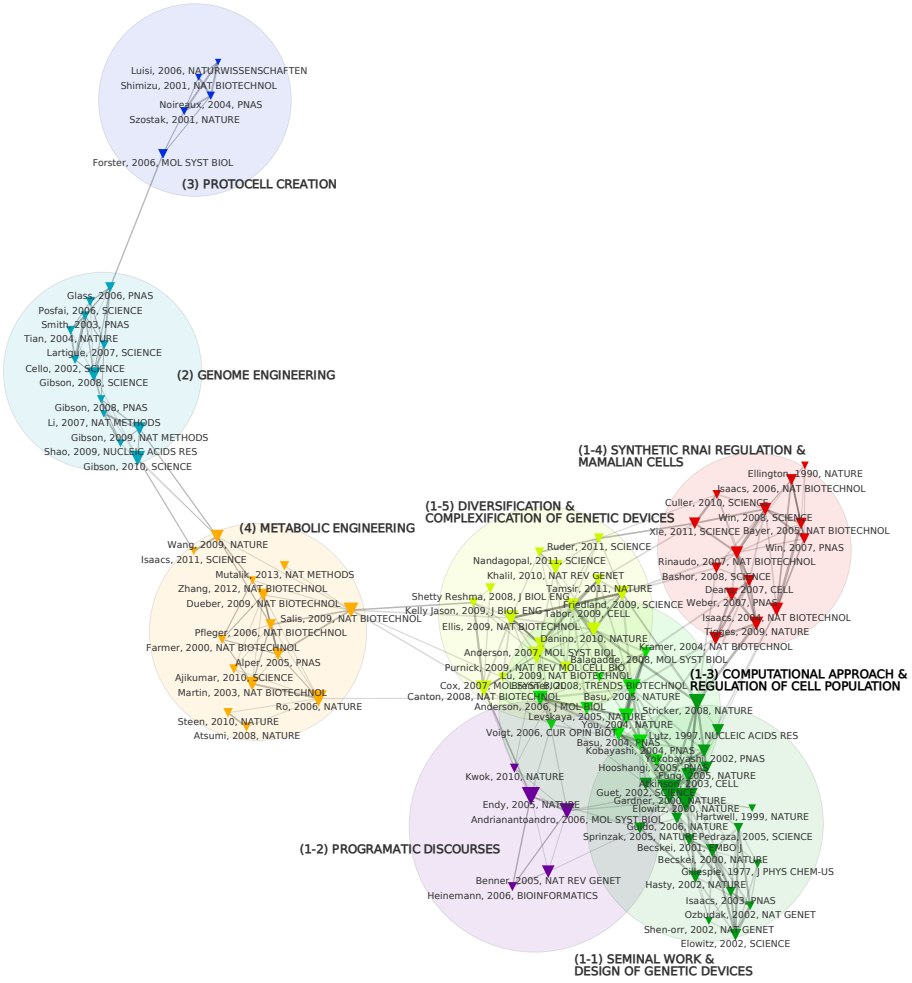


FIGURE 2.22: Carte du réseau de co-citations réunissant les 100 références les plus citées du corpus de biologie de synthèse. Comme dans les autres cartes, la couleur des nœuds dépend de leur cluster d'appartenance. La taille des cercles qui figurent ces clusters est proportionnelle au nombre d'articles qui se projette sur chacun des clusters. L'épaisseur des liens est proportionnelle à leur similarité (ici mesurée par un cosinus, choix classique dans l'analyse des réseaux de citations). Les étiquettes de cluster ont été ajoutées manuellement et décrivent les différentes communautés épistémiques qui composent la recherche en biologie de synthèse.

Le choix de la carte présentée au paragraphe précédent n'est pas entièrement anodin. Contrairement aux cartes d'ingrédients précédentes, nous nous sommes ici concentrés sur une portion particulière du corpus : l'ensemble des recettes de cuisine italienne. Il s'agit là d'une autre stratégie possible pour gagner en résolution et affiner une description. Plutôt que de nous contenter d'une nébuleuse d'ingrédients aux parfums de la Méditérannée, restreindre le jeu de données à ces seules recettes a permis de faire émerger de nouvelles

frontières plus fines : comme celle séparant les plats de pâtes ((notamment) en vert clair sur la carte 2.21) des recettes de risotto (en vert foncé au-dessus).

Concluons cette section par un exemple. Il ne s'agit pas d'un réseau sémantique mais d'un réseau de co-citations entre références citées calculé à partir de la base de données d'articles de biologie de synthèse. L'analyse de ce réseau reproduit sur la figure 2.22 nous a permis d'identifier les quatre grandes communautés épistémiques qui sous-tendent la recherche en biologie de synthèse (Raimbault et al., 2016). Les clusters du réseau de co-citations ont été détectés grâce à l'algorithme de Louvain. Visuellement on remarque que parmi les huit clusters détectés, cinq sont densément interconnectés. L'algorithme de Louvain a d'ailleurs généré une clusterisation hiérarchique qui, au niveau d'agrégation le plus élevé, regroupe les cinq clusters en question. C'est la raison pour laquelle nous avons étiqueté ces cinq clusters comme les sous-éléments de la même « école » que nous avons intitulée : « ingénierie des bio-briques ».

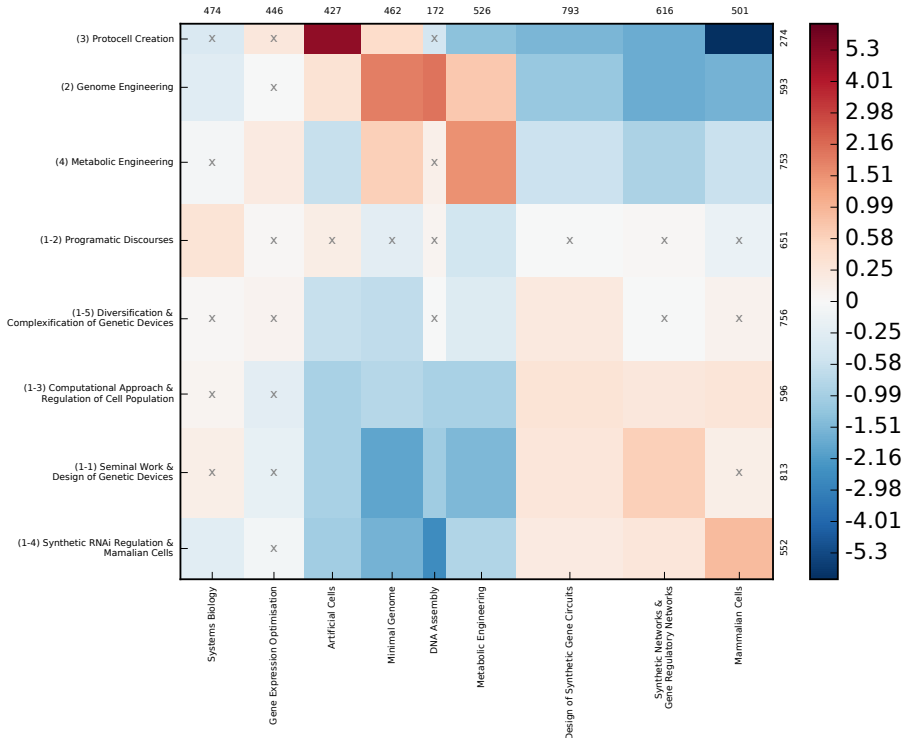


FIGURE 2.23: Matrice de contingence montrant les dépendances entre une partition sémantique (en colonnes - réseau lexical) et une partition disciplinaire (en ligne - réseau de co-citations) du corpus de Biologie de Synthèse. Le score que mesure la carte de couleur correspond à une forme de contribution locale de chaque cellule au score de χ^2 global de la matrice. Plus précisément et en reprenant les notations du premier chapitre déjà introduite dans la section sur l'analyse des correspondances : chaque cellule de la matrice vaut $\frac{N_{ij}-Eij}{\sqrt{E_{ij}}}$ où pour rappel $E_{ij} = \frac{N_{i\bullet}N_{\bullet j}}{N_{\bullet\bullet}}$. La présence d'une croix dans une cellule signifie qu'un test de Fisher exact avec une p-value de 0.05 ne permet pas de confirmer la corrélation des deux variables.

La figure 2.23 croise au sein d'une matrice de contingence dont la construction suit le même principe que celui rencontré plus tôt dans ce même chapitre (section 2.17) les clusters de co-citations que l'on vient de décrire (horizontalement) et des clusters sémantiques obtenus en cartographiant les 300 termes les plus pertinents du corpus (la carte sémantique associée peut être consultée

dans cette référence : (Raimbault et al., 2016)). On réalise que les clusters relevant de l'école des bio-briques ont un profil relativement semblable ce qui laisse croire qu'elles font appel au même vocabulaire et traitent de thématiques similaires (malgré certaines nuances visibles dans la matrice). Les trois dernières écoles ne partagent pas les thématiques de la première et se concentrent toutes fortement sur un ou deux clusters sémantiques propres. En revanche, le vocabulaire lié à la biologie des systèmes et à l'optimisation de l'expression génétique semble commun à l'ensemble de la communauté des biologistes de synthèse.

Voici un exemple de ce que l'on peut espérer apprendre d'une lecture cartographique de corpus. Comme on le voit, l'interprétation n'est pas donnée d'avance. Elle requiert une bonne connaissance de la source de données et plus largement du monde qu'elle représente (ici la recherche scientifique et ses pratiques de citation). Elle s'appuie aussi sur une série de regards successifs. À la manière d'un metteur en scène de cinéma, il faut trouver le bon éclairage, le bon angle et alterner champs et contre-champs pour transformer les données en véritables outils narratifs qui puissent ensuite accueillir une véritable interprétation sociologique.

3

Suivre les traces numériques

Ni personnes ni identités, les traces sont la matière première, tel est le constat posé par Boullier (2015) pour qui l'avènement des traces numériques marque l'entrée des sciences sociales dans un troisième âge tout entier fait de vibrations tardiennes. Néanmoins certains commentateurs jettent un regard plus pessimiste. La crise des sciences sociales empiriques, déjà prophétisée (Savage et Burrows, 2007) il y a 10 ans, n'a fait que se confirmer avec l'avènement des big data (Burrows et Savage, 2014). En effet l'utilisation de traces numériques comme matériau d'enquête du social a soulevé de nombreuses objections auxquelles ce chapitre tente de répondre. Nous avons tenté de rassembler ces objections sous la forme de cinq questions assez simples que nous détaillons ci-dessous : quelles populations, quels acteurs, quelles actions, quelles statistiques, quelles stratégies de recherche ?

— **quelles populations ? -** Les données issues du web et en particulier les données des réseaux sociaux souffrent intrinsèquement de très forts biais de sélection : qui est inscrit sur Twitter ? Qui ne l'est pas ? En conséquence, ces données sont souvent mal ajustées pour les enquêtes en sciences sociales (Ollion et Boelaert, 2015) et elles ne sont pas aussi « objectives » que leur taille voudrait le laisser croire. De plus l'activité en ligne est souvent distribuée de façon très hétérogène (Boyd et Crawford, 2011) : n'observe-t-on pas que les traces d'une portion congrue d'utilisateurs hyper actifs ? En privilégiant les traces d'activité, on prend le risque de rester totalement ignorant d'autres formes d'engagement plus passives (Tufekci, 2014).

— **quels acteurs ? -** Les individus dans les corpus web sont souvent démunis de toute épaisseur sociale ce qui rend leur analyse rédhibitoire à l'échelle individuelle et délicate aux échelles supérieures. Des données aussi simples et traditionnelles que la catégorie socio-professionnelle ou le sexe des acteurs sont généralement indisponibles. Dès lors, comment mener des analyses qui ne réduisent pas les acteurs à un individu moyen détaché de toute inscription sociale (Bowker, 2014; Beuscart, 2014) ?

— **quels comportement sociaux ? -** Il est hasardeux d'extrapoler des comportements réels à partir de comportements en ligne. Par exemple, Boyd (2006) explique combien les réseaux de sociabilité en ligne sont différents des réseaux personnels. Manovich (2011) nous met également en garde contre l'extrapolation littérale d'observations en ligne vers des pratiques sociales hors ligne : « Peoples' posts, tweets, uploaded photographs, comments, and other types of online participation are not transparent windows into their selves » [1]. Pour paraphraser, le web ne saurait être considéré comme une représentation homothétique de la société (Beuscart, 2014).

1. « Les publications, tweets, photographies, commentaires, et autres formes de prise de parole en ligne ne sont pas des fenêtres ouvertes sur l'intériorité de leur auteur »

— **quelles statistiques ? -** La taille des données n'est plus adaptée à nos outils statistiques pour lesquels tout phénomène devient significatif (McFarland et McFarland, 2015). L'intelligence artificielle répond aux contraintes de données toujours plus riches et nombreuses mais comment réconcilier les corrélations et l'excellence prédictives des méthodes d'apprentissage avec l'épistémologie des sciences sociales plus attachée à la recherche de relations de causalité ?

— **quelles stratégies de recherche ? -** L'accès aux données crée une dépendance des chercheurs vis-à-vis des plateformes qui fournissent (parfois) les données (Lazer et al., 2009; Boullier, 2015). Une des premières conséquences est que la collecte de données est réalisée plateforme par plateforme avec le risque que la recherche soit également tributaire de ce découpage en silo, les actions individuelles ou collectives étant alors artificiellement séparées de leur extension réelle (Tufekci, 2014). Certaines études mentionnent également le risque d'un « digital divide » pour l'accès aux données mais aussi du point de vue des infrastructures lourdes et des compétences requises pour les analyser (Manovich, 2011). La question des formes d'agencement inter-/pluri-/multi-disciplinaires à construire entre sciences sociales et ingénierie se pose (McFarland et al., 2015).

Mais avant de proposer des pistes pour répondre à certaines de ces questions ouvertes, nous prendrons acte dans la première partie (3.1) de ce chapitre des transformations induites par les nouveaux espaces numériques. On montrera, à travers plusieurs cas empiriques, comment de nouveaux acteurs émergent en ligne. On interrogera les options à notre disposition pour tâcher de leur redonner un peu d'épaisseur sociale et ne pas les réduire à de simples adresses IP. Plus directement en lien avec notre ambition d'analyse du monde social à travers ses traces textuelles, nous nous demanderons jusqu'où les technologies du web étendent la prise de parole, « liker » une photo relève-t-il encore d'un acte d'énonciation ? Alors que leurs actions sont enregistrées avec toujours plus de précision, comment caractériser les acteurs qui prennent la parole en ligne (Boellstorff, 2013) ?

Les espaces numériques sont l'arène de phénomènes sociaux tout à fait

originaux. On pourrait en faire un inventaire à la Prévert et mentionner les printemps arabes que certains ont parfois (et peut-être rapidement) qualifiés de « révolutions Twitter » (Cointet et Gallinari, 2017), la mise en place de dispositifs de participation collective par les foules (crowd-) pour financer des projets, résoudre des problèmes ingénieriques ou aider des astronomes à étiqueter des galaxies (Prpić et al., 2015), les mailing-listes comme nouveaux dispositifs collectifs de coordination du travail (Conein, 2004) ou encore les sites de rencontre qui transforment les relations amoureuses et sexuelles contemporaines (Bergström, 2011). Mais comment enquêter sur ces grandes métamorphoses sociales et politiques ? Nous défendons l'idée que pour pouvoir mener une enquête sociologique en ligne, il faut adopter une attitude modeste et commencer par s'évertuer à comprendre ces espaces et leur fonctionnement dans leur « état naturel ». On propose donc d'envisager les dispositifs numériques et les espaces sociaux qui ont émergé avec Internet comme des « milieux » dont nous pensons que les propriétés doivent être étudiées avant même que d'espérer comprendre les processus sociaux qu'ils supportent. Bien que nourris d'une philosophie de la circulation ouverte et égalitaire de l'information (Turner, 2010), les espaces numériques sont remplis d'ordres et de normes. Dresser des familles d'utilisateur sur Facebook est indispensable avant d'espérer pouvoir comprendre le rôle du réseau social dans les rassemblements Place Tahrir. On s'attardera donc dans la deuxième partie de ce chapitre (3.2) à comprendre le fonctionnement « métabolique » de ces espaces en changeant successivement de casquette. Tel un arpenteur, on tracera les frontières au cœur des espaces numériques. Plus tard, on se transformera en démographe pour estimer quelles populations habitent ces contrées. Et enfin, tel un géologue armé d'un sismographe, on interrogera les règles qui régissent la propagation des informations en ligne.

Enfin on se demandera dans la dernière partie (3.3) si les méthodologies décrites dans les chapitres précédents sont adaptées ou non à ces nouveaux espaces et ces nouvelles données. Les « big data » et leurs algorithmes menacent-ils de transformer l'épistémologie des sciences sociales, en remettant en cause certaines hypothèses et procédures fondamentales telles que l'usage des statistiques, la recherche d'explications causales, etc ? On interrogera notamment les problèmes d'échantillonnage que posent ou solutionnent le travail sur les traces numériques.

3.1 *Les pulsations de la vie numérique*

Le titre de cette partie est empruntée au livre de Cardon (2015b) traitant de « l'explosive expressivité numérique » du web faite « de tweets, de posts, de blogs, de photographies, de selfies, de check-in ». L'écosystème informationnel

dans lequel nous vivons a très largement été bouleversé durant ces dernières décennies. L'information circule de manière plus horizontale empruntant des canaux toujours plus personnalisés à travers les réseaux sociaux, les agrégateurs d'actualité, les forums… En simplifiant les conditions de sa production et de sa diffusion, Internet a donné la parole à tout un ensemble d'acteurs jadis inaudibles, rendant de plus en plus poreuse la séparation entre les producteurs et les consommateurs d'information.

Ce qui nous intéresse ici, c'est de saisir en quoi ce nouvel espace public transforme le programme (que nous avons commencé à décrire) d'analyse du social *via* ce qu'il convient maintenant d'appeler les « traces textuelles ». C'est une question finalement bien prosaïque : en quoi les concepts traditionnels d'acteurs, d'expression publique, de publics mêmes sont transformés au contact du numérique ? Quelles conséquences les sciences sociales doivent-elles tirer pour mener l'enquête dans ces nouveaux espaces ?

3.1.1 Nouveaux modes d'expression

Les plateformes en ligne offrent des capacités d'action toujours élargies à leurs usagers. Sans avoir disparu, le texte s'est laissé largement déborder par une multitude d'autres modalités d'expression et d'interaction : publier une photo, diffuser une vidéo, « liker » une publication, partager sa localisation, retweeter un lien, etc. La domaine de l'énonciation se trouve élargi quand un simple clic peut être interprété comme l'expression d'une subjectivité. Un premier enjeu de l'analyse des espaces numériques est donc d'intégrer la multiplicité de ces modes d'expression dans notre analyse.

C'est notamment le travail auquel nous nous sommes livrés dans le projet Algopol coordonné par Camille Roth. Facebook constitue une plateforme un peu particulière sur le web, au sens où l'espace qu'elle ouvre à ses utilisateurs est tout à la fois public et privé. Contrairement à d'autres plateformes comme Twitter sur lesquelles la quasi totalité des interactions sont publiques, Facebook entretient ainsi un « clair-obscur » au sens où les individus ne se dévoilent qu'à un public limité et contrôlé de proches (Cardon, 2008).

Gérée par Irène Bastard et développée par *Linkfluence* sous la direction de Stéphane Raux, l'application Algopol a permis de collecter les traces d'activité de 15 145 utilisateurs depuis leur inscription sur Facebook. Cherchant à caractériser les configurations de pratique sur la plateforme, nous nous sommes heurtés à la grande complexité des métadonnées livrées par l'API de Facebook. Cette complexité est d'abord de nature technique. L'API délivre les traces d'actions d'utilisateurs regroupées en fonction de catégories techniques et de

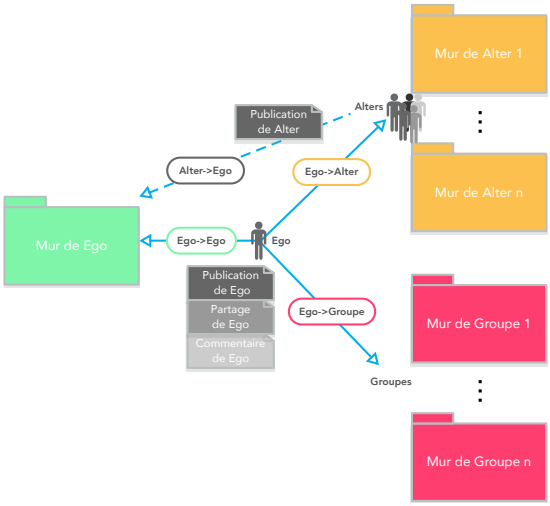


FIGURE 3.1: Modèle simplifié des actions possibles sur Facebook

formats qui sont avant tout censés faciliter la gestion du service. Au total, Facebook utilisent des centaines de catégories pour décrire la variété des actions possibles sur Facebook. Elles ont été entièrement refondues (« re-purposed » dans le langage de Marres (2012)) pour construire 92 actions élémentaires [2] qui ont encore été simplifiées pour correspondre à des classes d'action plus fidèles à notre propre expérience quotidienne de la plateforme : la description finale de l'activité des enquêtés que nous avons employée distingue ainsi simplement entre types d'activité (publications, partages, commentaires), espaces de publication (sur sa propre page, la page d'un ami ou sur une page de groupe) et provenance des publications (d'alter ou de ego) (voir illustration figure 3.1). En définitive, chaque compte est décrit sous la forme de 8 indicateurs compositionnels auxquels on adjoint le volume brut d'activité.

2. À titre d'exemple, pour illustrer la complexité de la nomenclature et la variété des actions à coder, on peut énumérer la famille des « APhotoTexteE » qui correspondent à des événements de type : Alter publie une photo avec du texte (un avis, un commentaire) sur le mur de Ego, ou la plus classique et non moins fameuse « ELienSecE » qui décrit des événements où Ego publie un lien sec (sans texte) sur son propre mur.

Les utilisateurs de la plateforme Facebook sont invités à prendre la parole d'une multitude de façons différentes. Ainsi, un internaute peut publier une photo, avec ou sans texte, partager un lien, rédiger un texte brut, etc. Les modes d'action sont multiples ains que les publics qui se mélangent au sein des réseaux des utilisateurs. Ainsi certaines actions comme le changement de photo de profil généreront beaucoup de likes, provenant de l'ensemble du réseau, et peu de commentaires. A contrario publier un statut qui prend la forme d'un texte sec sera beaucoup commenté relativement au nombre de likes qu'il reçoit (voir figure 3.2) . On le voit, même en considérant un seul et même individu sur une plateforme donnée, il faut rester vigilant avant de sommer les likes reçus et dénombrer les commentaires, différentes modalités d'énonciation co-existent et traduisent des pratiques tout à fait différentes.

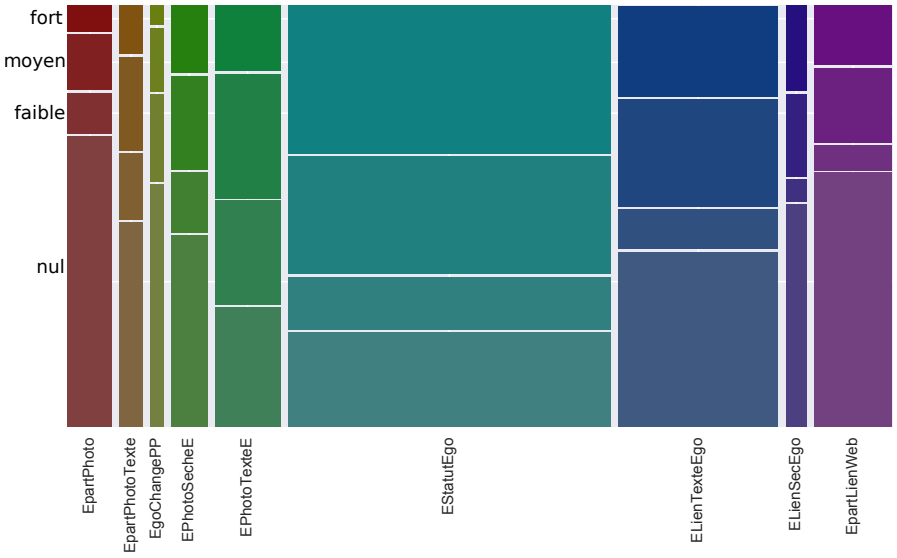


FIGURE 3.2: Diagramme mosaïque montrant la distribution de la proportion de commentaires reçus (sur le nombre de likes et de commentaires) par un statut en fonction de la nature de l'action de Ego. De gauche à droite : partage d'une photo, partage d'une photo accompagnée d'un texte, changement de sa photo de profil, publication d'une photo sèche, ou accompagnée d'un texte, publication d'un tete, publication d'un lien et d'un texte, publication d'un lien sec, ou partage l'un lien depuis le web. On voit que le statut sec est le plus susceptible de susciter une forme de réponse conversationnelle. Le partage de photo ou de lien décontextualisé, au contraire, génère une proportion très faible de commentaires, de même que le changement de photo de profil (qui est le statut qui génère le plus de likes provenant de l'ensemble du réseau).

3.1.2 *Nouveau langage*

Les acteurs prenant la parole dans l'espace public étant plus nombreux et plus divers, il était logique qu'ils s'expriment différemment. Un des enjeux liés au traitement de ces nouvelles données vient précisément du langage dont les acteurs font usage sur le web. L'analyse de corpus (même qualitative) s'est presque toujours penchée vers les mêmes types de textes : typiquement des corpus d'articles de presse, ou des discours politiques. Le défi que pose la profusion de nouvelles données n'est souvent pas tant lié à leur quantité (après tout si elles étaient trop nombreuses, on pourrait aussi bien les échantillonner) qu'à la langue dont elles usent : souvent moins codifiée et respectueuse de la syntaxe que dans les corpus textuels « classiques ».

Après avoir pris les ouvrages religieux comme référence, les ressources linguistiques sur lesquelles les modèles de traitement automatique des langues étaient entraînés compilaient des textes correspondant à un certain modèle de l'espace public. Bien sûr, ces ressources sont progressivement remplacées, mais il est intéressant de se souvenir de quelles briques elles étaient composées. À titre d'exemple, le corpus Brown[3], l'une des ressources les plus classiques utilisées en linguistique est composée de 500 textes en anglais provenant d'articles de presse, de publications religieuses, de livres de fiction et de manuels de cours. Dès lors un analyseur morpho-syntaxique, patiemment mis au point et évalué par rapport à ces corpus de référence, est rapidement mis en défaut lorsqu'on lui présente des tweets à la ponctuation douteuse, remplis d'emojis

3. pour être précis : « Brown University Standard Corpus of Present-Day American English »

et d'autres caractères exotiques. Plus récemment, une nouvelle génération d'analyseur syntaxique capable de traiter spécifiquement ce genre de contenu voit le jour. Ces nouveaux algorithmes sont distribués en open source qu'ils soient développés par des universités (Tweet NLP) ou des entreprises (spaCy, SyntaxNet).

La stratégie de codage que nous avons utilisée pour modéliser les commentaires laissés par des internautes sur le LA Times Homicide Report relève typiquement de cette transition. Plutôt que de chercher des ressources linguistiques adaptées à ce type de parole publique extrêmement hétérogène, remplie d'argots et de fautes de frappe, nous avons préféré travailler le codage manuellement en nous appliquant surtout à bien préciser un modèle d'énonciation publique adapté. Nous avons ainsi mobilisé le modèle classique de la prise de parole publique utilisé par Boltanski, Schiltz, et Darré (1984), à savoir décoder un énoncé comme la mise en relation d'un ensemble d'actants. Dans notre cas, les commentaires lient le locuteur, la victime de l'homicide, le ou les responsables (qui peuvent être un gangster local, la drogue, la police ou la fatalité) et enfin le public des internautes que l'on convoque pour solliciter sa compassion ou partager son indignation.

Plus précisément, nous nous sommes efforcés de caractériser ces énonciations à travers deux dimensions. La première tient précisément à la nature du lien entre locuteur et victime. Il s'agissait alors de se doter d'une méthode relativement fiable pour qualifier, lorsqu'il existait, le lien entre la victime et le locuteur (lien familial, amical, connaissance, etc.). L'autre dimension renvoie à la taille des actants mobilisés dans le discours. Fait-on exclusivement référence à des individus, au quartier, ou à des institutions telles que la police, voire à des entités collectives plus abstraites et générales comme la menace que la drogue fait peser sur la société ou le manque d'investissement public pour l'éducation ?

Appliquer ce modèle sur un tel matériau textuel aussi brut a demandé un travail particulièrement minutieux de définition de jeux d'expressions qui permettent de détecter avec un minimum de bruit et le meilleur rappel possible une modalité donnée. La table 3.1 récapitule certaines expressions que mon collègue, Sylvain Parasie, a repérées à partir d'une liste des 1500 expressions les plus fréquentes extraites automatiquement du corpus[4]. À titre d'exemple, on a repris une intervention ci-dessous et souligné les expressions caractéristiques (*cops*, *accountable for*, *citizen*) qui permettent d'indexer ce message sur la dimension des êtres mobilisés comme relevant des catégories « Insititutions » et « Problèmes publics ». Le commentateur ne précisant pas son lien avec la victime, ce message est étiqueté avec la modalité « absence de lien visible » du

4. On peut s'étonner de voir ressurgir les méthodes du TAL après les avoir accablées pour leur manque de robustesse en territoires inconnus. En réalité, les traitements effectués ici sont minimaux. On n'essaye pas de reconstruire la structure syntaxique des phrases (de toute façon souvent mal délimitées faute de ponctuation) ou d'extraire des entités nommées (bien camouflées derrière avec une casse flottante), mais simplement d'identifier les expressions les plus fréquentes qui seront susceptibles de couvrir une large part du corpus. Une lemmatisation a été réalisée quand elle était possible ; pour le reste on s'est contenté de mesures de collocation purement statistiques qui ne dépendent pas de la langue.

TABLE 3.1: Codage des commentaires du LA Times Homicide Report. Les prises de paroles sont codées selon deux axes principaux relevant du lien entre le locuteur et la victime et de la taille des êtres mobilisés dans les commentaires.

Axe	Modalité	Expressions
Lien entre le locuteur et la victime	Adresse à la victime	*you will forever be in our hearts / you were special / you were a brother / (. . .)*
	Lien Familial	*miss my husband / as a close family member of / was my great niece / (...)*
	Lien d'Amitié	*R.I.P. Bestfriend / He was a dear friend / (...)*
	Souvenirs Partagés	*good old memories / we were just chillin / (...)*
	Lien de Connaissance	*I lost someone very special / X was a terrific human / X went to my highschool / (...)*
	Absence de lien visible	
Êtres mobilisés	Parents, amis	*parents / love / little girl / momma / baby brother / care / (...)*
	Individus	*great person / great friend / sociopath / bastard / (...)*
	Gangs	*gang member / gang activity / rival gangs / bloods / 18th st hood / (...)*
	Quartiers	*residents / blocks / area / neighborhood / Watts / (...)*
	Métropole de Los Angeles	*Los Angeles County / LA area / City of LA / (...)*
	Institutions	*police department / officers / inmates / social workers / (...)*
	Problèmes publics	*political issues / prevent further crimes / social problems / positive change / (...)*
	Groupes sociaux et ethniques	*Whites / Black man / ethnicity / Caucasians / Latino communities / (...)*
	États-Unis	*Uncle Sam / U.S.A. / our country / America / (...)*
	International	*europe / Canada / other countries / mexico / (...)*

point de vue de cette dimension.

> « *Mighty Mike, disobeying orders from **cops** should not be enough of a reason to pull the trigger and kill the suspect. I guarantee you if **cops** were held **accountable for** their actions just like the average **citizen** they will not be so quick to pull the trigger.* » Jag, page de l'homicide de Jose Monteon homocide en réponse au commentaire de Mighty Mike

Etablir ces listes d'expressions est un travail exigeant car il demande de parcourir un nombre très important d'entre elles. Sélectionner l'une d'entre elles requiert parfois de contrôler son sens précis dans ses différents contextes d'apparition. Néanmoins, nous ne prétendons pas ici à un codage exhaustif ou dénué de toute erreur [5] pour la simple raison que notre unité d'analyse n'est pas le commentaire individuel mais les régimes de discours collectivement mobilisés à l'échelle des quartiers et que l'on assigne par la suite à différents publics. Autrement dit, la quantité de commentaires même si elle reste relativement modeste (un peu moins de 20 000 au total), nous permet d'envisager l'activité de codage avec moins d'intransigeance que si notre objectif était d'identifier un commentaire précis responsable d'un changement de régime dans l'ensemble de la dynamique du site web. Un être fictif boîteux, et Marlowe [6] risque de faire une erreur d'interprétation majeure en identifiant un texte comme pionnier alors qu'il n'est qu'un suiveur. Plutôt que de retravailler sans cesse un codage, on peut se contenter d'évaluer la qualité du codage de chacune de ses catégories [7] et s'accommoder du taux d'erreur observé pourvu qu'il ne mette pas en péril la validité des résultats obtenus [8].

5. À vrai dire, même une « lecture proche », commentaire par commentaire, ne permet pas toujours de lever toute ambiguïté quant au choix de telle ou telle modalité.

6. Marlowe est l'agent double de Propspero, un programme d'intelligence artificielle avec lequel le sociologue peut converser pour explorer son corpus (Chateauraynaud, 2003a).

7. Il serait tout à fait envisageable de mesurer cette erreur sous la forme d'une précision et d'un rappel *via* une évaluation manuelle des quelques centaines de commentaires.

8. Nous présentons certains résultats de ce projet dans la section suivante

3.1.3 *Des locuteurs inter-changeables*

Les traces issues du web donnent accès avec force détails à certains aspects des comportements individuels en ligne. On saura par exemple, à la seconde près, quel ami a liké le statut d'un utilisateur sur Facebook, suivi d'un commentaire bienveillant sur ses photos de vacances en retour. En revanche, d'autres types d'activité restent entièrement dans l'ombre. Il est ainsi difficile, sauf à travailler dans la Twitter data team, de savoir si un lien partagé apparaissant dans le flux d'un utilisateur a effectivement été cliqué [9]). Si les traces explicites sont souvent publiques et accessibles sur le web, l'audience reste une donnée rare. Les médias sociaux ont fondé leur philosophie sur la participation. Et pourtant beaucoup d'individus adoptent une attitude passive - ce qui menace jusqu'à l'équilibre économique des plateformes du web social (Kushner, 2016). Dès lors, il est très difficile d'estimer cette population et son comportement en ligne sinon sous des formes très agrégées (indicateur du nombre de vues reçues par une vidéo sur youtube par exemple). Les traces numériques indexent donc les comportements individuels de façon potentiellement très lacunaire.

9. C'est typiquement ce type de données qu'a utilisé une équipe de chercheurs de Facebook pour étudier les effets de bulle informationnelle (Bakshy et al., 2015).

Plus grave, de nombreux observateurs pointent le manque d'épaisseur sociologique d'usagers perçus à partir de leurs seules traces numériques : dans quels espaces sociaux, économiques ou politiques s'inscrivent-ils ? Comment faire sens de ces subjectivités si elles n'émanent pas d'individus mais d'un utilisateur moyen totalement désincarné dont on ne connait ni le niveau d'éducation, ni la classe sociale, ni même l'âge ou le sexe dans la plupart des cas (Beuscart, 2014) ? Si les individus étaient exclusivement définis par leur profil de consommation, en quoi ces traces pourraient prétendre capturer ce qu'est un individu dans la pluralité des mondes sociaux qu'il habite et dans sa capacité d'évolution :

> « *If I am defined by my clicks and purchases and so forth, I get represented largely as a person with no qualities other than "consumer with tastes"* » (Bowker, 2014)

Par contraste, en dépit de leur taille réduite, les discours de l'État de l'Union (déjà introduits aux chapitres précédents sections 2.1.4 et 1.2.1) sont des corpus normalisés permettant une analyse socio-historique rigoureuse de par la remarquable stabilité des institutions qu'elles représentent. Le discours de l'État de l'Union, tradition inscrite dans la constitution américaine, met en scène le même locuteur [10], le président des États-Unis dressant le bilan de l'année écoulée et de l'État de l'Union aux citoyens américains. Les compte-rendus des COPs rédigés par l'Earth Negociation Bulletin consistent en un autre corpus standardisé, au sens où sa forme, ses conditions de rédaction, et les objectifs présidant à sa rédaction, sont restés quasiment inchangés durant les 20 dernières années et permettent d'indexer précisément les objets de discussion autour desquels les négociateurs de chaque pays se sont focalisés.

10. Il s'agit du même locuteur, qu'il s'adresse au congrès ou qu'il donne un discours à la radio.

Dans ces deux cas, la situation d'énonciation est parfaitement connue.

Sur le web [11], la situation est à peu près opposée. De par le flou qui entoure la provenance des contenus en circulation, il est parfois difficile de savoir qui parle, ni devant quel public un contenu a été prononcé. Les contenus sont agrégés, triés et désagrégés avant de s'afficher sur nos écrans ou de peupler nos bases de données. L'acte d'énonciation ne peut que très rarement être rattaché à une individualité. Le souvenir de la situation d'énonciation est souvent perdu [12]. Quand Francis Chateauraynaud cherche à démêler le dossier Céline, il soumet une grande diversité de textes (pamphlets, critiques, compte-rendus de procès, etc.) à Prospero. Mais pour chacun d'entre eux, le contexte de publication (auteur, « lieu » de publication, date, etc.) est parfaitement renseigné. Dépourvu d'informations sur les situations dans lesquelles les traces du web ont été énoncées, est-on condamné lorsque l'on traite des données numériques, à toujours rester à la surface, traiter ces énoncés pourtant produits localement (au cœur d'une conversation ou en réaction à un événement particulier) comme une même masse informe ? Quelle valeur peut-on alors donner à ces traces si on ne peut réellement qualifier leur provenance ?

11. On entend, par cette expression, se limiter aux contenus « nativement numériques » par opposition aux contenus numérisés (Rogers, 2013).

12. Il y a bien quelques exceptions en la matière. Ainsi certains comptes bénéficient d'une petite pastille bleue rajoutée par Twitter les authentifiant comme des comptes officiels, mais la majorité des contenus publiés sur Twitter resteront pourtant écrits par de simples alias, qu'une description plutôt maigre pourra parfois enrichir. Il en est de même sur les forums de discussion, où hormis pour quelques « stars locales », l'anonymat règne en maître, etc.

Deux attitudes s'opposent pour répondre à cette difficulté. La première consiste simplement à l'ignorer et envisager la masse des inscriptions textuelles comme produite par un seul corps dont on prend le pouls. On dénombre les mots utilisés, on établit la distribution globale des grandes thématiques sans vraiment se soucier de savoir quels acteurs portent tel ou tel type de discours et avec, en corollaire, peu d'espoir de compréhension réelle de la forme prise par la discussion. Même si elle peut sembler assez grossière et à condition que le corpus soir relativement homogène, cette approche est très populaire et reste excessivement efficace pourvu que l'ambition reste de produire une description globale.

L'autre possibilité consiste à assumer une posture tardo-latourienne et postuler que les individus sont co-extensifs de leur trace. Nul besoin d'attacher à ces individus des catégories pour saisir leur dynamique, ils sont traces. Les algorithmes de recommandation partagent finalement les mêmes hypothèses athéoriques pour nous adresser leur conseil (parfois avec bonheur). Les catégories classiques des sciences sociales sont finalement déjà endogénisées dans les comportements individuels et donc dans les données (Kosinski et al., 2013; Sloan et al., 2015). Pourquoi s'embarrasser d'un modèle complexe de distribution des goûts en fonction des classes sociales, du genre, ou intégrant des modèles psychologiques complexes (Chavalarias, 2004), pour prédire les goûts de tout un chacun quand Netflix, entre autres, utilise un algorithme d'apprentissage automatique qui parvient à « maximiser la rétention » de ses abonnés (Gomez-Uribe et Hunt, 2016) avec de plus en plus de données implictes. Il en va de même pour Youtube dont les recommandations visent à

maximiser le temps de visionnage des utilisateurs sur la plateforme (Davidson et al., 2010) sans recourir à des notions d'âge, de sexe ou de niveau d'étude.

Entre ces deux options, on peut également, avec Bowker (2014), continuer à croire que les grandes catégories du social comme le sexe ou le capital économique, sont des forces qui continuent d'opérer en ligne, même si nos bases de données ne sont pas toujours volubiles à leur sujet :

> « *Just because we have big data does not mean that the world acts as if there are no categories. And just because we have big (or very big, or massive) data does not mean that our databases are not theoretically structured in ways that enable certain perspectives and disable others.* »

Dès lors le sociologue se fait maçon, et doit s'efforcer de consolider et épaissir, autant que faire se peut, les propriétés des locuteurs s'il veut pouvoir espérer comprendre les processus sociaux qui ont généré ces traces. C'est ce que nous avons entrepris dans les deux projets que nous avons menés sur les commentaires en ligne. Comme déjà commenté et illustré dans la table 3.1 dans le cas du projet sur le blog du LA Times, nous avons d'abord essayé de modéliser l'énonciation des internautes sous la forme d'un schéma actantiel. Mais nous nous sommes également efforcé, dans ces deux projets, d'inférer les caractéristiques des commentateurs à travers deux stratégies différentes : (i) en exploitant l'intensité de leur participation aux forums, (ii) en analysant l'alias dont ils usent pour signer leurs interventions.

La nature des informations discutées étant locale dans les deux cas (des élections municipales ou des homicides « au coin de la rue »), la première hypothèse que nous avons posée a simplement consisté à postuler que les commentaires avaient également été produits localement. Ainsi on suppose que ce sont les administrés d'une commune qui interviennent sur le forum de leur commune. On suppose également que l'essentiel des commentaires sur la page des victimes du LA Times homicide report [13] sont le fait d'habitants vivants dans le même quartier (ou à proximité immédiate) de celui où est décédée la victime. Nous excluons de cet ensemble les contributeurs dits métropolitains. Les « métropolitains » sont des usagers de la plateforme qui ont publié beaucoup plus de commentaires que les autres, concernant un nombre important d'homicides couvrant une large partie du conté de Los Angeles [14]. Le plus prolifique de ces contributeurs, qui se fait appeler Syscom3 a ainsi commenté sur les pages de près de 350 homicides couvrant plus d'une centaine de quartiers différents.

La caractérisation des locuteurs est vraiment minimale [15] : il s'agit d'une division binaire distinguant les commentateurs locaux dont on suppose qu'ils interviennent uniquement sur la page d'homicides dont ils sont proches, et des commentateurs métropolitains qui font des commentaires sur des homicides dans des quartiers qu'ils ne visitent probablement jamais. Malgré sa simplicité,

13. L'interface du site http://homicide.latimes.com offre de multiples voies d'exploration (par type d'homicides, par genre, âge des victimes, etc. mais l'entrée géographique prime malgré tout de par l'omniprésence de cartes).

14. Pratiquement nous avons défini un seuil arbitraire de 10 quartiers différents pour qu'un contributeur soit qualifié de métropolitain, 47 personnes rentrant dans cette catégorie (après contrôle manuel des homonymes manifestes (*anonymous*, *a friend*, etc.).

15. On ne souhaite pas nécessairement connaître la CSP de chaque commentateur. Ici le taux d'activité s'avère en réalité un excellent indicateur de l'engagement militant des individus au sein de la plateforme. Autrement dit, le simple nombre de traces laissées par chacun est déjà une caractéristique susceptible de nous guider dans l'interprétation.

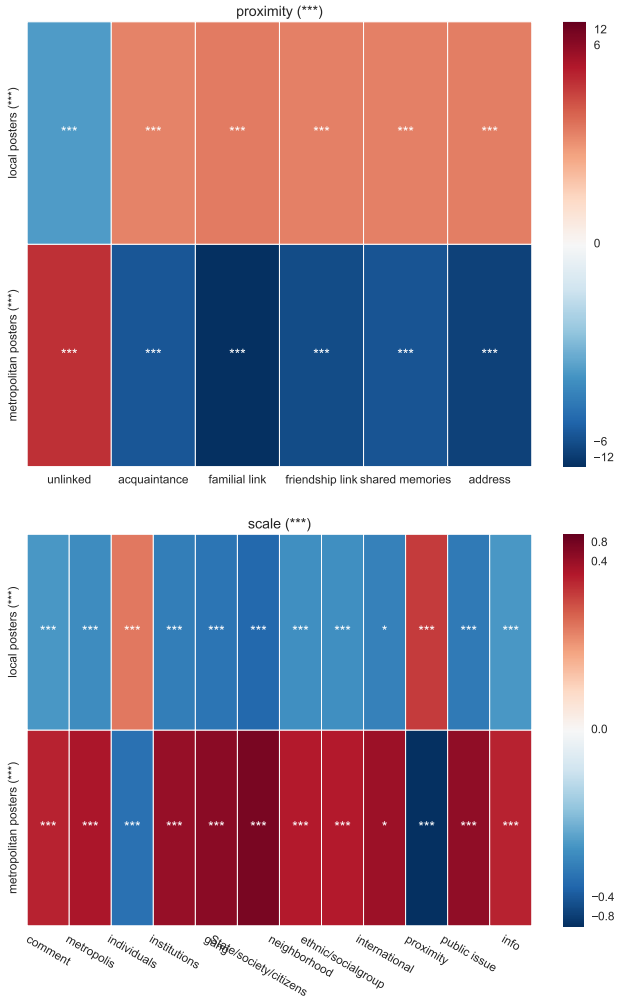


FIGURE 3.3: Matrices de contingence montrant les différences de registre de prise de parole entre commentateurs locaux et métropolitains. Les variables figurant sur l'axe horizontal correspondent à un codage du degré de proximité entre le locuteur et la victime [haut] et du niveau de discours [bas] selon la grille déjà présentée dans la section précédente (voir table 3.1). Les étoiles figurent à quel degré deux modalités sont corrélées ou non en fonction du test d'indépendance du χ^2 (p-value de 0.001 pour trois étoiles, 0.01 pour deux étoiles et 0.05 pour une seule étoile).

cette opération s'est avérée salvatrice pour distinguer deux populations bien différentes au sein de notre corpus de commentaires. En effet, les interventions de locaux et des métropolitains sur le forum ne sont pas comparables, ni par leur degré d'implication (par définition), ni par la nature de leur intervention. Les deux matrices de contingence figure 3.6 le montrent clairement.

On réalise à travers cette analyse combien, du point de vue de nos deux variables (proximité du locuteur à la victime et taille des êtres mobilisés dans les commentaires), les deux populations se différentient. On comprend ainsi, en s'appuyant sur la lecture des commentaires de quelques utilisateurs métropolitains, que ces derniers ont leur propre agenda politique (pas toujours compatible entre eux d'ailleurs) et se servent du LA Times Homicide Report comme d'une plateforme pour diffuser leurs idées. À titre d'exemple, Syscom3 accuse les gangs d'être responsables de cette violence urbaine et sélectionne systématiquement les homicides susceptibles d'illustrer son propos. L'inter-

vention de Jag citée dans la section précédente fournit une autre illustration à l'opposé du spectre politique.

Mais plus que la mise en évidence de deux registres de prise de parole différents, cette distinction nous permet par la suite de caractériser les publics auxquels nous avons affaire de façon beaucoup plus fine que si nous avions indistinctement mélangé l'ensemble des contributeurs. Nous montrons ainsi comment au sein de certains quartiers, en ne considérant que les contributeurs locaux commentant certains types d'homicide, de véritables publics émergent qui partagent une interprétation commune de ces occurrences distribuées (Parasie et Cointet, 2013).

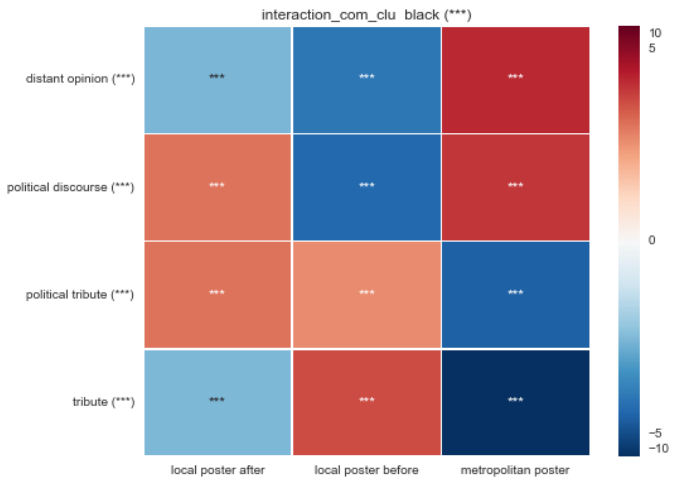


FIGURE 3.4: Matrice de contingence montrant les différences de registre dans la prise de parole entre commentateurs locaux avant et après l'arrivée d'un commentateur métropolitain sur la page d'un homicide.

Une autre des caractéristiques originales des corpus venant du web est qu'ils résultent souvent d'interactions entre locuteurs. En agrégeant tous les contenus au sein d'un corpus unique, on oublie trop souvent que l'on brise une conversation, ou en tout cas qu'on découpe une séquence d'interactions avec un contexte et une temporalité propre. Sans essayer de saisir la singularité de chaque fil de conversation, nous avons capturé figure 3.4, l'influence dans la dynamique de discussion de l'arrivée d'utilisateurs métropolitains sur la nature des discours portés sur la victime d'un homicide. La matrice de contingence contraste ainsi entre des situations de prise de parole de contributeurs « locaux » avant qu'un « métropolitain » ne fasse irruption dans la discussion d'un homicide avec des situations où ces mêmes contributeurs locaux répondent aux messages voire aux provocations des contributeurs métropolitains. Chaque commentaire a préalablement été catégorisé en quatre classes allant du registre d'expression le plus personnel et proche de la victime (*tribute*) au registre le plus politique et détaché de la victime (*distant opinion*) [16]. On voit clairement que les métropolitains dont le discours est fortement politique ont tendance, lorsqu'ils arrivent sur la page d'un homicide, à infléchir le discours des locaux (généralement dans le registre de l'hommage) dans une

16. Sans rentrer dans les détails, ces grandes classes résultent d'une catégorisation automatique des messages réalisée grâce à une classification hiérarchique ascendante sur les messages déjà indexés selon les modalités des deux axes (taille des êtres mobilisés, relation entre locuteur et victime) décrits précédemment.

direction beaucoup plus publique et politique.

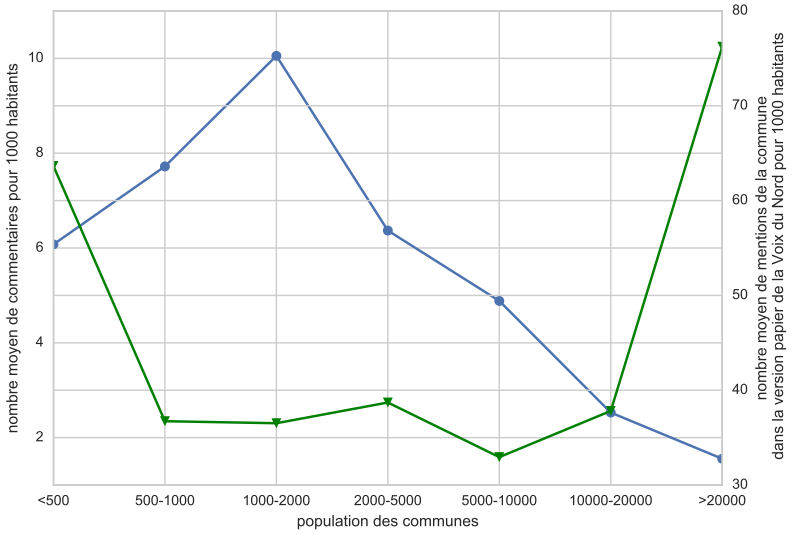


FIGURE 3.5: Pour chaque catégorie de tailles de commune, on calcule et représente [en bleu] le nombre total de commentaires que l'on divise par la population totale des communes. Le ratio permet d'estimer un taux de participation en ligne en fonction de la taille des communes (la figure diffère légèrement du diagramme publié dans la RFSP car nous avions alors choisi de nous appuyer sur le nombre de commentateurs uniques). La courbe verte mesure la couverture par la Voix du Nord des communes d'une taille donnée rapportée à leur population : si les plus grandes agglomérations reçoivent la plus forte attention par habitant, les très petites communes bénéficient également d'une forte attention relative.

Dans notre travail sur la Voix du Nord (Parasie et Cointet, 2012), nous nous sommes également efforcés de décrire les intervenants dans un forum de discussion aussi finement que possible. Ce projet visait à étudier une ensemble de près de 1 500 forums mis en place par la Voix du Nord à quelques mois des élections municipales de 2008 pour offrir aux habitants de la région Nord-Pas-de-Calais un espace de discussion du bilan du maire sortant, bilan qui était introduit par un court article rédigé par les journalistes du quotidien. En quelques mois, le dispositif a généré plus de 17 000 commentaires répartis de façon très inhomogène sur l'ensemble des communes.

Nous avons donc supposé que les interventions sur la page de chaque maire étaient rédigées localement, c'est à dire par des habitants des communes concernées. Aussi, les fortes disparités du taux de participation en ligne (reportées sur la figure 3.5) observées selon la taille des communes peuvent bien être imputées à la forme que prend le débat public localement. En se fondant uniquement sur l'intensité de la participation aux forums on constate en effet que (relativement à la taille de la population) le dispositif est beaucoup plus mobilisé dans les petites et moyennes communes (jusqu'à 10 commentaires pour 1000 habitants dans les communes de 1000 à 2000 habitants). le taux de participation en ligne semble inversement proportionnel à la place que le quotidien régional accorde aux communes, comme si l'ouverture de ces forums permettait de combler une lacune démocratique en offrant un nouvel espace de débat.

Pour mieux saisir la nature de cette participation en ligne, nous avons codé le contenu des commentaires, mais aussi l'alias utilisé par les commentateurs

(tout commentaire devant être signé). Cette opération manuelle a permis d'enrichir très significativement notre modèle. En effet, l'un des points les plus saillants de notre analyse a trait à l'usage variable qui est fait de l'anonymat en fonction de la morphologie des communes. Le codage des commentateurs a ainsi permis de montrer que les alias garantissant l'anonymat (« un citoyen », « une habitante de Libercourt » voire l'usage d'un personnage fictif « Danny Boon ») sont beaucoup plus fréquemment choisis dans les communes de petites tailles qui sont structurellement des espaces où règne une très forte inter-connaissance susceptible de limiter certaines prises de parole publiques [17]. En dépit d'une couverture presse aussi forte que pour les grandes villes, les forums ouverts par la Voix du Nord, en permettant à tout un chacun de s'exprimer sous couvert de l'anonymat, font office de voile dans les petites communes, ce qui facilite la prise de parole en assouplissant les contraintes liées à l'inter-connaissance généralisée. Avec ce modèle, on parvient à interpréter le pic observé figure 3.5 : les communes comportant entre 1000 et 2000 habitants souffrent doublement d'une sous-couverture médiatique et du poids de l'inter-connaissance. Les forums de la Voix du Nord y ont donc suscité une très forte participation.

17. C'est ainsi que nous expliquons le taux de participation très important dans les très petites communes.

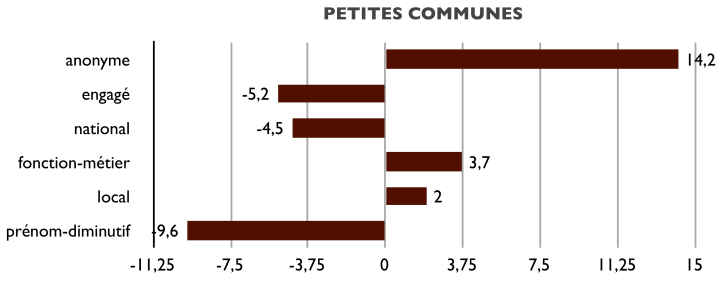


FIGURE 3.6: Spécificité (calculée avec une mesure de Cramer) des catégories d'alias pour les commentaires postés dans les communes de petite tailles (moins de 2000 habitants). En plus de la sur-représentation massive de l'anonymat (dont l'usage est d'ailleurs discuté et critiqué au sein même des forums), on notera la sous-représentation des alias renvoyant à des figures politiques (« engagé », « national »), ce qui signale en creux que les débats sont avant tout centrés sur la personnalité du maire sortant, de ses adjoints ou de ses opposants.

3.2 *Vibrations en milieux inconnus*

Dans cette partie, on tâchera plus précisément de saisir les changements induits par ces nouveaux espaces en terme de circulation de l'information. Les modèles traditionnels de formation de l'opinion s'expriment traditionnellement en sociologie des médias comme une flèche descendante partant des médias vers l'ensemble de la population après avoir éventuellement transité par les leaders d'opinion (Lazarsfeld et al., 1968). Si ce modèle est en crise, il n'est pas totalement caduque. Le web est en fait rempli d'ordres nouveaux et le village mondial est loin de constituer un tout homogène. Plutôt que de parler d'un espace public numérique, il serait plus juste de mettre l'expression au pluriel tant la nature des publics et des contenus échangés varie entre un forum de discussion sur les jeux vidéos, les messages privés échangés sur Facebook Messenger ou les tweets de commentateurs politiques. On illustrera

cette pluralité dans la première section (3.2.1). On s'intéressera par la suite aux variations de visibilité et de vitesse de l'information dans les espaces numériques, est-ce que le « pouvoir d'agir » des « vibrations » rend entièrement secondaire le choix des acteurs (Boullier, 2015) ? Peut-on expliquer les processus de concentration de l'information par leurs seules caractéristiques internes (section 3.2.2) ? Nous introduirons ensuite la notion de milieu et défendrons la nécessité d'enquêter sur les plateformes web avant même que d'interroger les processus politiques qu'elles supportent (section 3.2.3).

3.2.1 Un espace public pluriel, ou de la vertu des silos

Ouvert à une pluralité d'opinions inédite, le web ne peut pourtant pas être réduit, comme certains en ont émis l'hypothèse à ses débuts, à un nouvel espace mondial où les principes de délibération démocratique de l'espace public habermassien seraient enfin réunis. La première raison est très simple : le web est un espace pluriel, mosaïque. Les usagers de Twitter ne sont pas ceux de Youtube. Partager un lien sur Facebook n'est pas comparable avec la publication d'un commentaire sur Doctissimo. Les internautes se réunissent au sein de communautés d'intérêts partagés (Flichy, 2008) qui construisent autant d'espace de discussion et de modes de régulation (parfois très complexe, Wikipedia en étant l'archétype (Bryant et al., 2005)). Dès lors, comment suivre un débat de société qui traverse l'ensemble de ces arènes ? La question est d'autant plus cruciale que l'accessibilité des données est souvent régie par des infrastructures propres à quelques plateformes avec le risque d'une recherche elle aussi découpée en silos (Tufekci, 2014)...

S'il est difficile d'apporter une réponse définitive à la question, une rapide exploration menée sur les discussions en ligne autour de la loi Taubira sur le mariage pour tous illustre parfaitement le problème. C'est dans le cadre du projet Algopol que Linkfluence avait partagé les données que l'entreprise de surveillance du web social capture au sein de « son périmètre de veille ». Les données que nous avons analysées ont donc été capturées au sein de ce périmètre à compter du 1er janvier 2012 pour une durée d'un peu plus de deux ans en utilisant une requête minimaliste (« mariage gay » OU « mariage pour tous »). La méthodologie présentée au chapitre précédent a été déployée pour identifier les termes les plus pertinents du débat, puis cartographier les grandes thématiques du corpus. Le résultat est présenté ci-dessous (fig. 3.7)

Sans rentrer dans les détails, on retrouve les « cadres de discussion » suivants : le cluster jaune en haut à gauche intitulé « Manifestations » rassemble des termes qui documentent les manifestations et grands rassemblements de rue majoritairement anti-mariage pour tous qui ont précédé et suivi l'adoption

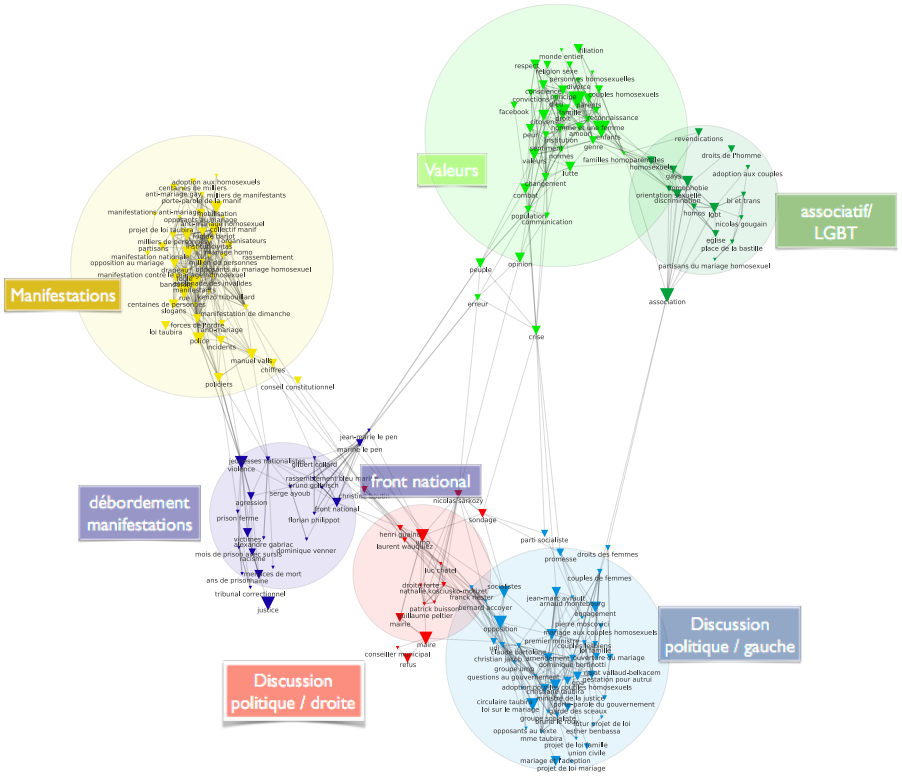


FIGURE 3.7: Carte sémantique des contenus échangés en ligne (dans le périmètre de veille de Linkfluence) sur le mariage pour tous. Les étiquettes attachées à chaque cluster ont été ajoutées manuellement.

de la loi. A proximité immédiate le cluster violet réunit des termes portant sur les « débordements » (*violence*, *agression*) autour des manifestations et leurs conséquences (*tribunal correctionnel*, *prison ferme*), auxquels étaient souvent associées des personnalités du Front National. Pour rappel *Jean-Marie Le Pen* et sa fille *Marine Le Pen* avaient choisi de prendre du recul par rapport au débat et de ne pas participer aux manifestations (il sont d'ailleurs assez périphériques au sein du cluster), au contraire de *Marion Maréchal Le Pen*, *Gilbert Collard* ou *Bruno Gollnisch*. Autres figures dans le même cluster, *Dominique Venner*, historien d'extrême droite, dont le suicide dans la cathédrale Notre Dame avait galvanisé les groupes nationalistes jugés responsables des violences (*Jeunesses Nationalistes* dont le leader est *Serge Ayoub*, *Civitas*, *Printemps Français*, etc.). Le cluster rouge adjacent (*Christine Boutin* servant de charnière avec le précédent) est essentiellement composé de personnalités politiques de droite. La présence des termes comme *maire/mairie*, *refus* s'explique par la médiatisation de déclarations de plusieurs maires UMP qui avaient déclaré à l'époque leur refus de célébrer des mariages entre personnes de même sexe. Le cluster bleu en bas à droite renvoie au travail parlementaire autour de la loi. On y retrouve essentiellement des personnalités de gauche. Deux clusters se font face au-dessus. Le cluster vert foncé réunit les éléments de discours et les acteurs de la société civile défendant la cause LGBT pro-mariage pour tous

(*droits de l'homme*, *homophobie* ou encore *discrimination*). Il fait écho au cluster vert clair, qui réunit les mots-clés qui résument les valeurs défendues par les opposants au mariage pour tous (théorie du *genre*, un *homme et une femme*, *famille*, etc.).

Mais comme on l'a déjà souligné ci-dessus, le web est loin d'être un espace homogène. Un des grands avantages lorsque l'on travaille à partir de données collectées par des compagnies comme Linkfluence [18], c'est justement de profiter de leur capacité à collecter tout un ensemble de traces numériques auprès d'un large éventail de sources. Le périmètre de veille comprend en effet des données venant de sources aussi variées que les médias sociaux (Facebook, Instagram, Twitter), les sites web de la presse, les blogs, etc. C'est en utilisant cette étiquette que l'on a tenté de révéler la forme que prend la discussion publique selon l'espace d'où l'on se place. Ainsi, on réalise en projetant les contenus propres à chaque type de source (voir les quatre « heatmaps » figure 3.8) que les médias traditionnels (en haut à gauche dont sont exclus les « pure-players ») ont essentiellement couvert les débats parlementaires et les manifestations en faisant peu de cas des arguments échangés par les deux camps (clusters verts supérieurs). Sans que les choses soient aussi tranchées, on retrouve au contraire une sur-représentation de ces deux clusters dans les forums. La blogosphère semble accorder un place beaucoup plus large aux discussions autour du cluster « Valeurs » et aux personnalités d'extrême droite.

18. Les leaders mondiaux de l'industrie de la distribution de données web sont américains et s'appellent Datasift ou Spinn3r.

Il s'agit d'une étude modeste. Néanmoins ce simple exemple montre combien certains choix cruciaux quant aux sources sélectionnées pour rendre compte d'un débat public peuvent avoir une influence majeure sur la représentation finale. Les médias traditionnels commentent essentiellement les événements et peuvent de ce fait passer sous silence les dimensions plus idéologiques des discussions qui prennent place dans d'autres arènes comme la blogosphère. Dès lors, on peut même s'interroger sur la nature de la carte d'ensemble que nous avons construite (fig. 3.7). Relève-t-elle d'une construction entièrement artificielle qui, par le truchement de quelques mots, rassemble des situations de prise de parole des acteurs et des milieux entièrement différents ? On aura l'occasion de revenir plus tard (section 3.3.3) sur cette question lorsqu'on interrogera la façon dont les big data déplacent la question de l'échantillonage.

3.2.2 *Concentrations*

Malgré (ou peut-être à cause) des idéaux de circulation ouverte et égalitaire de l'information largement documentés historiquement (Turner, 2010; Flichy,

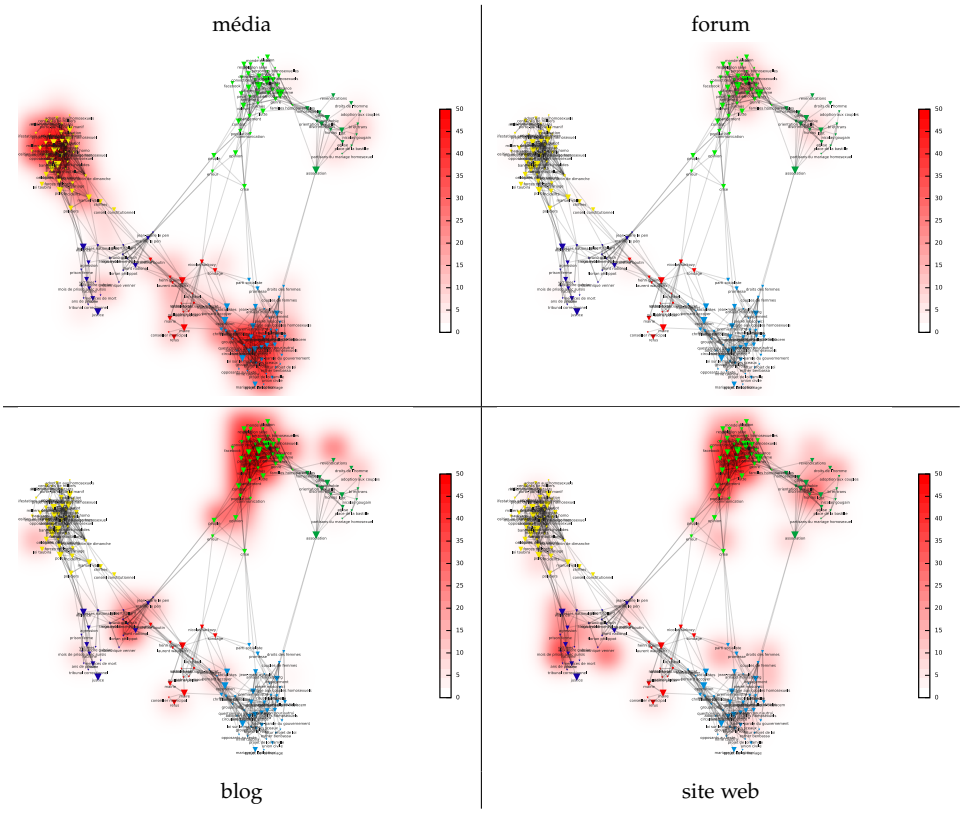


FIGURE 3.8: Les quatre "heatmaps" calculées en fonction des différentes sources de données : sites de média, forums de discussion, blogs, et autres sites web. Les zones rouges recouvrent le vocabulaire plus spécifiquement attaché à la source de données indiquée au-dessus ou en dessous de la carte.

2001), force est de constater que certains contenus ont beaucoup plus de visibilité que d'autres sur le web. Si les « gate-keepers » (White, 1950) traditionnels ont perdu de leur superbe [19] d'autres processus qu'il soient de nature technique, sociale ou mixte les ont remplacés pour opérer une sélection de l'information qui reconstruit une certaine forme de hiérarchie (au moins *a posteriori* puisque l'expression est rarement contrôlée à la source) dans un monde qui paraissait plat. L'exemple prototypique est le pagerank de Google qui vise à classer le web. En calculant le score de pagerank d'un site à l'aune du nombre de liens hypertextes qu'il reçoit et en fonction du pagerank de ces sites citant, Page et al. (1999) s'appuient sur un principe méritocratique (Cardon, 2013) qui remet de l'ordre dans la recherche d'information.

19. Selon Cardon (2015a), il serait faux d'affirmer que les grands médias ont perdu toute centralité avec l'émergence d'internet. Par contre ils ont bel et bien abandonné le monopole de la mise en circulation de l'information comme le montre très bien l'étude récente de Cagé et al. (2017).

De nouvelles formes de hiérarchisation de la visibilité des contenus en ligne (et de leurs auteurs) ont donc rapidement vu le jour à la faveur de nouveaux gate-keepers tels que les sites portails dans les années 90 (DiMaggio et al., 2001) ou les algorithmes de classement des moteurs de recherche (Hindman et al., 2003). La multiplication des métriques sur le web, qu'elles mesurent des audiences, de l'autorité, de la réactivité, de l'affinité (Cardon, 2015b) mais aussi la personnalisation des résultats (Bozdag, 2013) ou de façon plus procédurale les nouvelles formes de « bureaucraties » qui administrent les contributions

en ligne (Butler et al., 2008) sont autant de manières de remettre de l'ordre au sein du web. Mais plutôt que d'analyser la façon dont les algorithmes génèrent de la hiérarchie en distribuant de façon non équitable la visibilité des ressources en ligne (Hindman et al., 2003; Bucher, 2012), on préfère dans cette section mettre en lumière la manière dont, pour certains dispositifs sur le web, la liberté offerte aux individus permet de (re-)définir des fonctions de choix qui étaient jadis pré-déterminés par les professionnels de l'information. Finalement c'est la même stratégie que celle adoptée par des chercheurs de Facebook dans leur étude sur les bulles de filtre dans le newsfeed de Facebook (Bakshy et al., 2015). Ils ont ainsi montré que le filtrage de contenus politiques « non-alignés » n'est pas seulement la conséquence de l'algorithme de recommandation mais résulte également, et ce dans des proportions au moins comparables, de choix individuels.

Le questionnement du rôle des médias dans ce nouvel espace informationnel n'a pas été sans interroger la profession elle-même. Et le journalisme de données a été l'une des manifestations de cette réflexion menée dans les rédactions (Parasie et Dagiral, 2012). Le « LA Times Homicide Report » relève de ce type d'initiative. Conscients des biais de sélection liés à la couverture de la criminalité dans l'agglomération de Los Angeles, des chiffres potentiellement manipulés par la police, les journalistes du « data desk » du LA Times ont mis en place ce blog à partir de 2007 (Young et Hermida, 2015). Il a pour ambition de couvrir l'intégralité des homicides, ceux-ci étant traités de façon totalement uniforme au travers de quelques variables descriptives et d'un article très standardisé [20]. En somme, les journalistes refusent le modèle traditionnel dans lequel ils sont supposés opérer une sélection des homicides remarquables pour leur lectorat. Plutôt que d'être prescripteurs de ce qui intéresse ou non le public, le dispositif du LA Times Homicide Report offre ainsi au lecteur un accès à l'intégralité des homicides commis dans le comté de Los Angeles. Chaque victime est traitée sur un pied d'égalité, les lecteurs du blog ont donc toute liberté pour naviguer sur le site et « sélectionner » les homicides à leur convenance.

20. Les articles étaient même rédigés automatiquement par un robot au début de l'expérimentation.

Traditionnellement, seuls 10% des homicides sont couverts par l'édition papier du LA Times. Dans ce nouvel espace, comme on vient de le voir tous les homicides bénéficient de la même exposition, mais les lecteurs ne portent pas la même attention à chacun des homicides, ou en tout cas ils ne réagissent pas de la même manière sur toutes les pages. C'est une caractéristique importante de ces pages que d'être ouvertes aux commentaires. Nous pouvons donc apprécier la variation, non pas de l'audience, mais de l'engagement que suscite [21] un homicide donné en fonction du nombre de commentaires publiés sur sa page. Pour reprendre la dénomination empruntée à Cardon (2015a), on peut dire que les homicides bénéficient tous de la même visibilité (la plateforme est même construite pour garantir cette égalité) mais qu'en fonction

21. Il faut souligner que le nombre de commentaires reçus ne constitue qu'une estimation très imparfaite de l'audience réelle d'une page web.

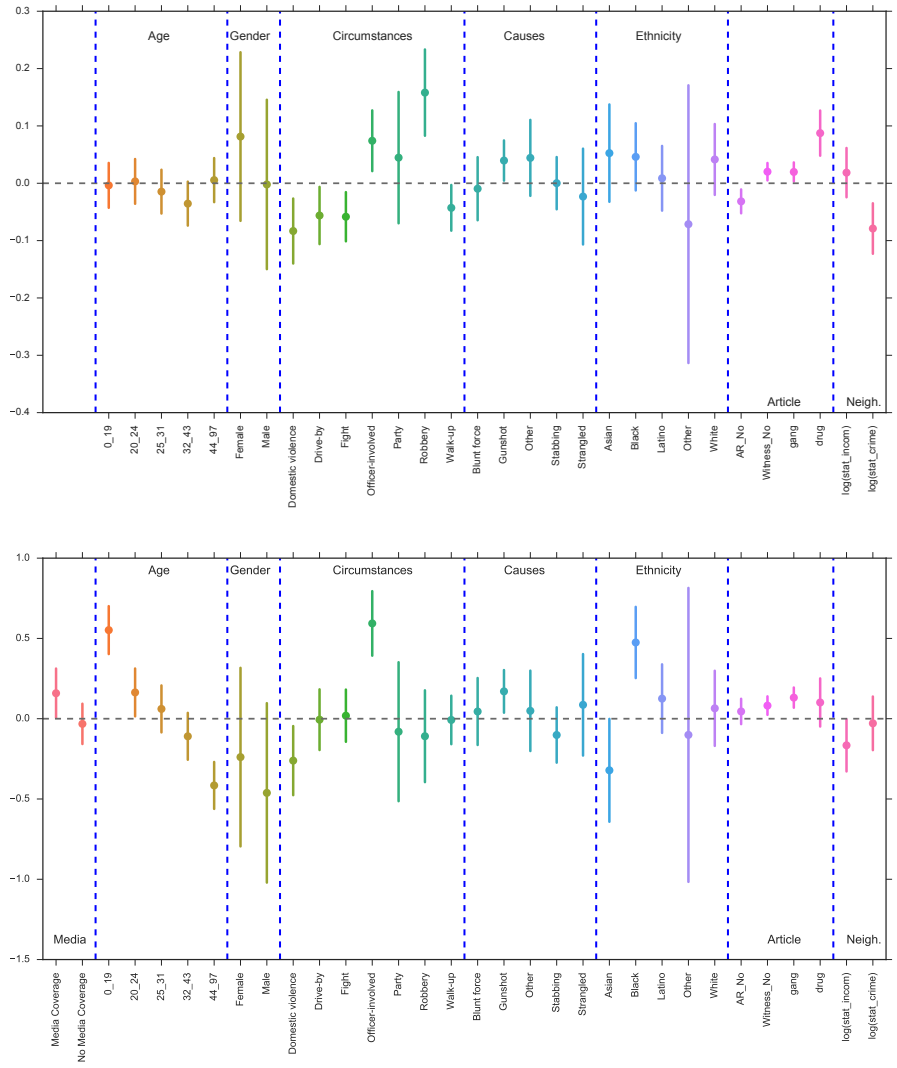


Figure 3.9: Coefficients résultant d'une régression sur la présence d'un article dans la version imprimée du Los Angeles Times [haut], le nombre de commentaires suscités par chaque homicide [bas]. Les barres correspondent à des intervalles de confiance à 95%. La couverture de l'homicide dans la version papier a été rajoutée comme variable dépendante supplémentaire pour prédire le nombre de commentaires sur la plateforme et son effet est clairement positif.

de leurs caractéristiques, ils peuvent être publicisés de manières très variables.

En premier lieu, soulignons, que classiquement les commentaires laissés sur les pages des victimes se distribuent de façon très inégale sur l'ensemble des homicides. 71% des commentaires portent ainsi sur seulement 20% des homicides. Dès lors on a voulu comprendre quelles étaient les caractéristiques principales susceptibles de provoquer une réaction des internautes. Nous avons donc soumis le nombre de commentaires à une régression linéaire par rapport à toute une série de variables qui décrivent l'homicide mais aussi par rapport à des caractéristiques propres au quartier où l'homicide a été commis. Au final, les variables que nous avons retenues ont trait à l'âge de la victime, son sexe, les circonstances et causes de l'homicide, l'ethnicité de la victime, le

contexte de l'homicide (informations mentionnées dans l'article tel que le lien potentiel avec un gang, la présence de drogue, l'existence de témoins, etc.), et des caractéristiques liées au quartier (revenu moyen des habitants et taux de criminalité).

Pour avoir un point de comparaison nous avons réalisé la même régression, en tâchant de prédire cette fois une variable binaire codant pour la couverture de l'homicide dans la version papier du LA Times. C'est en effectuant des recherches sur le web du nom de chaque victime que nous avons glaner cette information. Les résultats sont intéressants à double titre.

En premier lieu, on constate dans la première régression, qu'un homicide faisant suite à un vol (dont on imagine qu'il a dû mal tourner), impliquant si possible un policier, dans un quartier plutôt sûr (faible taux de criminalité) a de fortes chances d'être couvert dans l'édition papier du LA Times. La présence de drogue est également « souhaitable » mais la couleur de peau ou l'âge de la victime ne semblent pas être des critères de sélection particuliers. Dans le cas des commentaires, on observe quelques corrélations similaires comme la sur-sélection des homicides dans lesquels des membres des forces de l'ordre sont impliqués ou la faible couverture des crimes liés à des violences domestiques. En revanche, certaines propriétés semblent générer une réponse spécifique en ligne. Ainsi on commente très majoritairement des homicides de jeunes gens : moins de 25 ans, ou de façon encore plus forte, moins de 20 ans. Les homicides de personnes noires sont beaucoup plus commentés également. On voit bien émerger le portrait robot des victimes qui suscitent un engagement en ligne : des homicides de jeunes noirs par arme à feu commis dans les quartiers plutôt défavorisés de Los Angeles - ceux-là mêmes qui souffrent traditionnellement d'un déficit structurel de couverture par la presse.

En somme, on voit clairement apparaître des biais de sélection systématiques au sein de la plateforme qui semblent intimement liés à des processus politiques ancrés dans le territoire de la métropole : racisme, ségrégation sociale, problèmes de violence, (etc.) des déterminants classiques qui n'ont nul besoin de faire appel à une théorie mémétique pour expliquer leur prolifération en ligne.

Si la conclusion n'est pas très surprenante, il est utile d'insister sur l'enseignement méthodologique de ce travail : à savoir la comparaison entre processus en ligne et hors ligne. Notre comparaison de la couverture médiatique d'un homicide dans un journal et du nombre de commentaires qu'il suscite en ligne est discutable. Pour autant, cette comparaison permet d'estimer au moins à grand trait, en quoi l'espace numérique (un blog ouvert aux commentaires) diffère de son équivalent hors ligne (la rédaction d'un journal)

du point de vue de la sélection des homicides les plus « intéressants ».

3.2.3 *Sonder les espaces numériques*

L'un des principaux enseignements de l'école de la *Digital Method Initative*, dont on a introduit les principes dans le premier chapitre (section 1.2.3), consiste à interroger la politique des plateformes web en même temps que les processus politiques qu'elle peuvent porter. Marres (2015b) cite notamment le manifeste sur les médias tactiques des théoriciens et activistes des médias Garcia et Lovink (1997) : « (Tactical) Media are never impartial, they always participate » [22].

22. « Les médias (tactiques) ne sont jamais neutres, ils sont toujours partie prenante »

A titre d'exemple Marres mentionne la forte présence d'un hashtag #anonymous dans un corpus de tweets sur la World Conference on International telecommunications (ou plus précisément tiré d'un corpus de tweets mentionnant le hashtag #WCIT) dont la présence relève d'une stratégie d'occupation de l'espace médiatique sans rapport aucun avec les discussions de la conférence. Partant de ce constat classique, Marres et Moats (2015) suggèrent de traiter les deux questions (média et processus) en même temps et suggèrent deux chemins possibles pour faire de l'analyse de controverse en milieu numérique [23]. La première peut être qualifiée de « précautionneuse ». Elle hérite de la tradition de l'analyse de discours, elle répond aux biais naturels des médias numériques en tâchant d'éliminer toute forme de bruit. La position des acteurs dans un débat doit être cartographiée indépendamment de tout contexte technologique. L'autre voie, que les auteurs privilégient, est qualifiée d'empiriciste. Elle embarque ces biais numériques et propose d'adopter une attitude symétrique vis-à-vis des « dynamiques substantives » et des médias qui les supportent.

23. un projet déjà vieux de plus de 15 ans ! (Rogers et Marres, 2000)

On peut difficilement contester la nature politique des dispositifs numériques qui, comme on le discutait déjà dans la partie précédente, trient l'information, filtrent les amis, pré-définissent les préférences individuelles. Il serait également bien naïf de vouloir à tout prix évacuer l'analyse des plateformes des dynamiques socio-politiques quelles portent, elles sont déjà trop intriquées. Ce serait d'ailleurs faire injure à la réflexivité des acteurs que de tâcher de retirer un « biais » supposé dont ils sont parfaitement conscients et vis-à-vis duquel ils ont pour certains adopté des comportements stratégiques (Jin, 2014). Nous partageons donc la même posture symétrique. Reste à en tirer toutes les conséquences, car le défi méthodologique est de taille.

Autrement dit, nous nous devons d'être au moins aussi stratèges dans nos choix méthodologiques que les acteurs le sont dans les modes d'expression

qu'ils choisissent. Et ces choix sont décisifs dès la constitution du corpus. Imaginons que nous demandions, comme nous l'avons fait dans le passé, à des étudiants de rendre compte de la dynamique d'un événement tel que la COP21 sur Twitter. Quelles stratégies peuvent-ils mettre en œuvre pour délimiter le corpus ? Chercher tous les tweets qui mentionnent le hashtag #cop21 ou #copparis ? Mais quid des tweets sans hashtag ou avec d'autres hashtags comme #keepitintheground qui peuvent légitimement prétendre faire partie de la conversation. Quels types de biais introduit-on en privilégiant une requête par hashtag ou dans le contenu intégral des tweets ? Retrouve-t-on les mêmes sources dans les deux cas ? Les tweets avec hashtags sont-ils par nature plus viraux ou proviennent-ils d'acteurs représentant plus souvent des collectifs etc. ?

Assurément, le changement climatique n'est déjà qu'un lointain souvenir et nous sommes maintenant en train d'interroger le métabolisme des usages de la plateforme Twitter. Le point que nous défendons ici relève d'une double exigence. Il s'agit d'abord de suivre résolument les acteurs à la trace même lorsque leur expression est médiée par de nouvelles technologies numériques (et l'analyste ne se privera pas d'utiliser les API commerciales des plateformes pour construire son corpus). Mais il faut aussi, en voyageant à travers ces différents espaces, prendre acte de leur singularité : à quelles conventions obéissent-ils, quelles populations est-on susceptible d'y retrouver ? Bref, essayer, avant de relever ce qu'il y a de singulier dans un milieu donné, de le qualifier dans son régime normal de fonctionnement.

Après ce long détour, nous voici confortés dans l'idée que la connaissance des milieux importe autant que les processus particuliers qui les traversent. Une recherche trop portée par son objet empirique se condamnerait à ne découvrir que des évidences - un peu comme re-découvrir l'effet Matthieu à chaque fois qu'une étude scientométrique est entreprise pour analyser un nouveau champ scientifique. Dans le cas du suivi d'une controverse publique sur Twitter par exemple, on devrait pouvoir juger de la longévité d'un phénomène viral à l'aune de celle d'un hashtag saisi dans un tout autre contexte politique. En somme, il faut comprendre l'équilibre écologique du milieu dans lequel se déploient les problèmes publics avant de s'atteler à l'analyse de l'un d'eux.

Naturellement, analyser un milieu n'est pas chose aisée. La sociologie des médias s'attelle ainsi à décrypter les processus d'agenda setting ou de diffusion de l'influence depuis déjà quelque temps (Lippman, 1922; Katz et Lazarsfeld, 1966). Il est d'ailleurs intéressant de noter que l'analyse de contenu américaine trouve ses sources dans l'analyse de la presse. Si Evans et Aceves (2016) font remonter les premiers pas de l'analyse systématique de corpus textuels au projet d'étude à grande échelle de la presse allemande de Max

Weber en 1910, Krippendorff (2004), remonte encore plus loin dans le passé [24]. Il mentionne ainsi des travaux encore plus anciens dont ceux d'Eugene Löbl proposant un système de codage élaboré pour analyser la structure interne du contenu des journaux en 1903 ainsi qu'une étude, réputée être la première étude quantitative de contenu journalistique (Sumpter, 2001). Son titre « Do newspapers now give the news ? » résonne de façon frappante avec les débats actuels sur la presse et les « fake news ». Publiée en 1893, elle montre que les sujets couverts par les journaux new yorkais à la fin du XIX^{ème} s'orientaient de plus en plus vers les sports et les nouvelles à scandales au détriment des sujets plus sérieux tels que la littérature, les sciences ou la religion (Speed, 1893). Mais si presse, radio et télévision, sont déjà largement étudiés, Internet a chamboulé les anciens modèles, et généré de nouveaux usages. Dès lors, la question de l'analyse des traces textuelles, si ces traces ont été élaborées au sein des plateformes du web, ne peut faire l'impasse d'un modèle minimal de ces espaces. La sociologie des usages pourrait être d'une aide précieuse à ce titre. Mais elle doit également adapter ses méthodes à l'heure des big data (Proulx, 2015) ? C'est en tout cas la voie que nous avons expérimentée dans le projet Algopol.

24. Pour être précis, il va jusqu'à citer des thèses de théologie de la fin du XVII^{ème} siècle qui interrogeaient le caractère moral des contenus publiés dans la presse et qui inquiétaient alors l'Église.

Le long travail de codage des actions individuelles auquel mes collègues se sont livrés pour caractériser les comportements individuels des enquêtés de l'application Alogopol a déjà été longuement documenté dans la partie précédente (section 3.1.1). La suite de ce travail illustre parfaitement le type d'observations qu'une sociologie des usages équipée d'outils d'analyse quantitative peut espérer tirer des traces du web.

Une fois les comptes des enquêtés réduits à leur volume d'activité et aux 8 indicateurs compositionnels déjà décrits (correspondant aux possibilités de publier un contenu, partager, commenter, chez soi, chez autrui ou dans un groupe, etc.), nous avons cherché à les regrouper au sein de grandes classes d'usagers. La méthode des k-means a ainsi été appliquée pour décomposer la population d'enquêtés en 6 configurations d'activité. La diversité de ces classes est frappante. Si parmi ces groupes on retrouve des classes de comportement (intitulés « égovisibles » et « égocentrés ») qui renvoient en effet à l'image habituelle que l'on peut se faire de Facebook comme d'un lieu où prédominent les comportements égo-centrés d'utilisateurs se souciant essentiellement de leur réputation, elles sont loin d'être majoritaires. On trouve bien d'autres familles d'activité : les « partageurs » ou des adeptes de la « conversation distribuée » ou de la « conservation de groupe », mais aussi des inactifs.

Mais quelles conclusions peut-on tirer de cette caractérisation ? Certes la mise en évidence d'une très forte hétérogénéité des usages sur Facebook est utile pour tordre le cou à une certain nombre de préjugés notamment quant au caractère nécessairement nombriliste des pratiques dans les réseaux

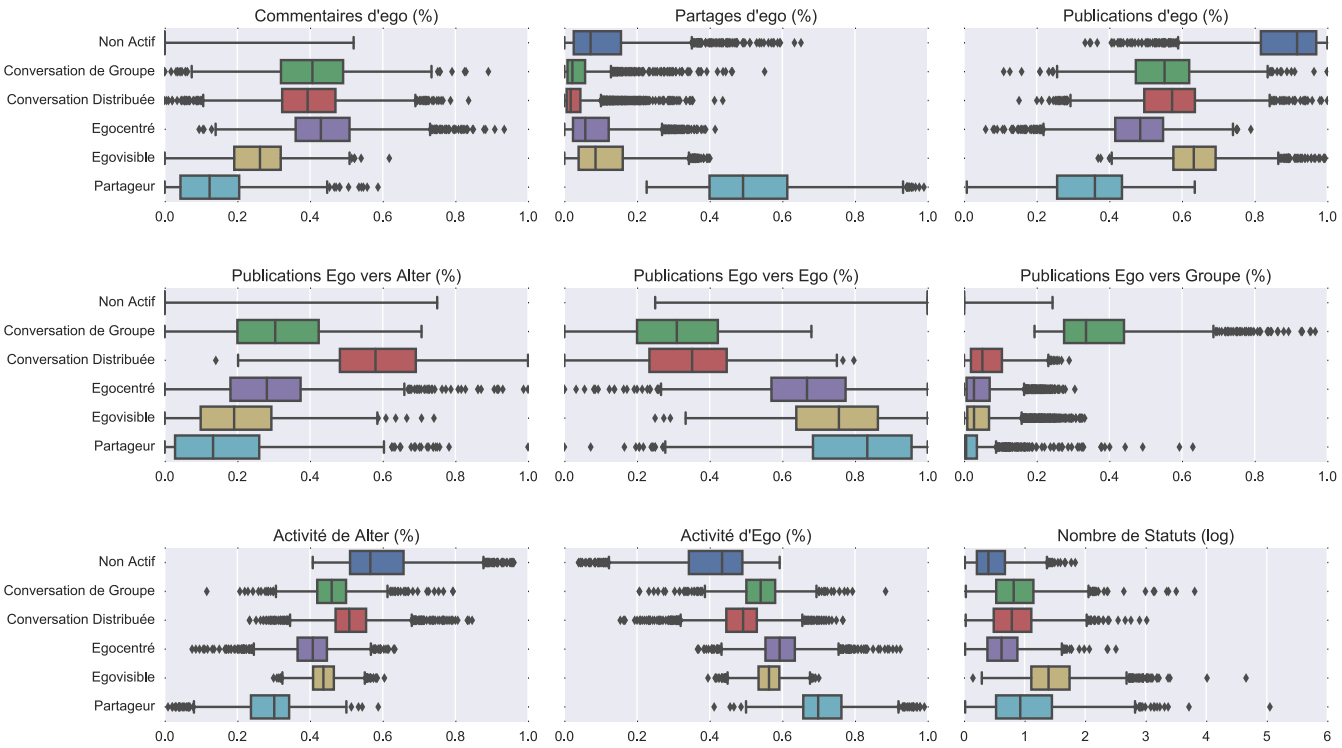


FIGURE 3.10: Typologie en 6 classes d'activité : non actif, conversation de groupe, conversation distribuée, égocentré, égovisible et partageur en fonction des 9 variables retenues. Chaque diagramme à moustache représente la distribution d'une variable sur les 6 classes d'enquêtés.

sociaux. Mais cette hétérogénéité constitutive doit également nous inciter à redoubler de prudence dans l'interprétation de résultats liés à des corpus construits sur Facebook. S'intéresser aux structures de partage de liens comme on l'a déjà fait au premier chapitre (section 1.1.3), c'est faire la part belle aux partageurs qui se sont fait une spécialité de cette pratique. Non seulement la classe des partageurs réunit des utilisateurs dont l'activité se compose d'actions de partage à plus de 50 % (contre moins de 10% en moyenne), mais ce sont également des utilisateurs très actifs (seconds en nombre de statuts). Or ils ne consistent qu'en une sous-population de taille réduite (à peine un millier de personnes) et assez singulière de notre échantillon global : les partageurs sont beaucoup plus âgés que la moyenne (cf. histogramme des âges figure 3.11) et par voie de conséquence sont très rarement des étudiants mais souvent des cadres (voir figure 3.12). Dire de la carte de la figure 1.5 qu'elle représente la structure du web selon les internautes relèverait alors d'un double mensonge faux. D'une part, elle est construite *via* des actions de partage et ne recouvre évidemment pas l'entièreté du web réellement visité par les internautes. D'autre part, elle se fonde sur un échantillon assez spécial d'internautes qui ont participé à l'expérience Algopol et qui ont un usage de Facebook qui privilégie le partage de liens à d'autres formes d'activité. Conscients de ce biais potentiel, nous avions contourné ce dernier biais en

nous assurant que chaque enquêté contribuait de la même façon à la carte finale. Enumérer les cooccurrences brutes de domaines cités chez les enquêtés aurait en effet largement sur-représenté les utilisateurs hyper-actifs et très partageurs. Aussi, nous avons fait en sorte que chacun contribue avec le même poids à la matrice de co-occurrence finale. Nous aurons l'occasion de discuter plus en détail ces choix méthodologiques dans la dernière section du chapitre (3.3.3).

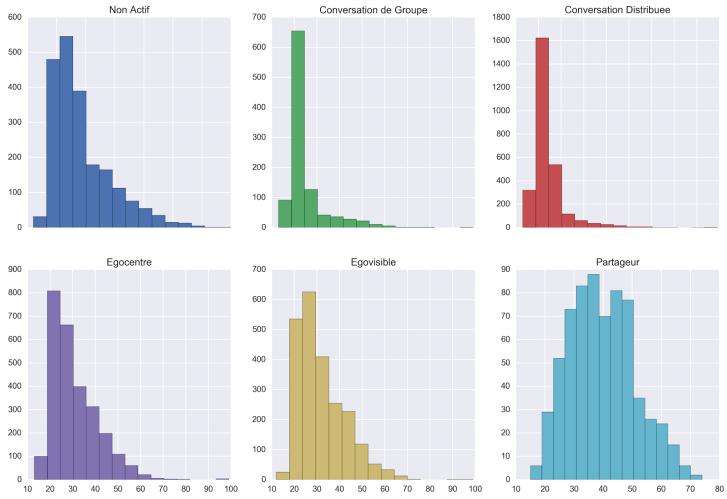


FIGURE 3.11: Distribution de l'âge des enquêtés en fonction de leur classe d'activité.

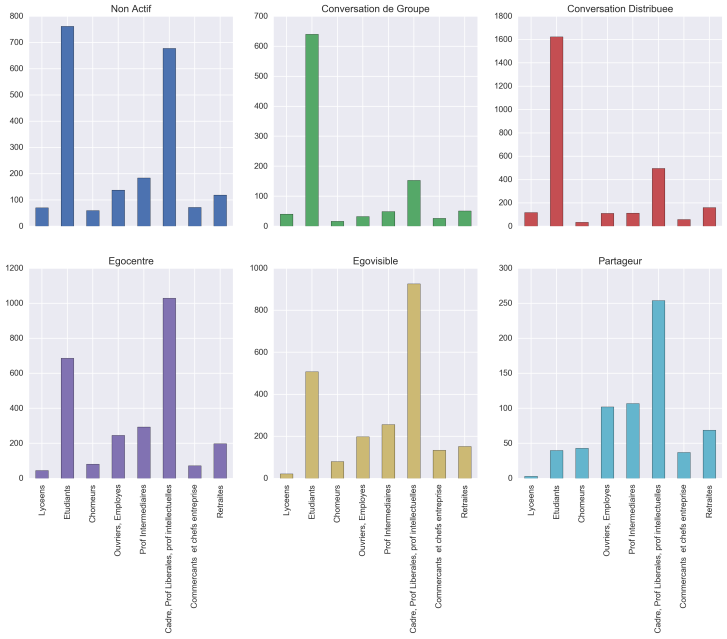


FIGURE 3.12: Distribution des catégories socio-professionnelles des enquêtés en fonction de leur classe d'activité.

L'autre piste possible envisageable pour sonder les milieux que nous explorerons est celle de l'analyse des processus de « diffusion » dans ces nouveaux espaces. C'est une approche qui a été très investie par les sciences informatiques et la science des données (et qui peuvent répondre à différentes

dénominations : cascades informationnelles (Cheng et al., 2014), mémétique (Leskovec et al., 2009), etc.). Plutôt que de s'intéresser aux espaces numériques en dénombrant patiemment les activités individuelles, on s'intéresse aux processus de réplication et de circulation qu'ils portent.

Memetracker[25], large collection de familles[26] de « citations » publiées sur Internet (par citation on entend en réalité simplement des chaînes de caractères entre guillemets) offre la possibilité de sonder les différents espaces du web et leur écologie selon le principe déjà énoncé par Tarde que « toute existence va différant ». À la manière d'un radar dont l'écho revient plus ou moins rapidement et déformé en fonction des milieux que le signal a traversé, l'analyse des transformations des citations et de leur profil temporel de diffusion a permis de mieux comprendre le cycle de l'information. Ainsi les citations diffusent dans leur très grande majorité des sites de presse vers les blogs (Leskovec et al., 2009), les altérations dans les citations sont plus souvent le fait des grands sites de presse même si les blogs peuvent également introduire des altérations non intentionnelles (Simmons et al., 2011). Dans un travail plus tardif, nous avions avec Thierry Poibeau et Elisa Omodei dont nous encadrions la thèse à l'époque, essayé de caractériser plus finement les critères de stabilité des citations. On avait ainsi pu identifier la forte instabilité des noms propres et des déterminants, l'influence négative de la longueur des citations sur leur stabilité, mais surtout la façon dont des processus dynamiques endogènes permettent de stabiliser les différentes variantes d'une citation. Nous avions notamment montré que la multiplication d'une citation avait un effet bénéfique sur sa stabilité (cf figure 3.13). S'intéressant plus spécifiquement aux événements de substitutions Lerique et Roth (2016) on montré combien l'évolution des citations n'était pas aléatoire mais dirigée vers des mots plus courts et plus « simples », biais bien connu en psycholinguistique mais empiriquement démontré ici à grande échelle.

Si les nouveaux espaces numériques sont porteurs de processus dynamiques inédits comme la diffusion de mèmes, il est assez illusoire de tenter de les comparer à des processus connus tant la nature des entités en circulation et surtout les milieux qu'elles traversent sont entièrement nouveaux. Dans les études de sociologie de l'innovation, les processus de diffusion concernait généralement la diffusion d'un pratique professionnelle, la prescription d'un médicament (Rogers, 2010; Coleman et al., 1957), etc. Il nous faut donc avancer sans préjugés et tâcher d'étudier les propriétés endogènes de ces milieux numériques. Mais cela ne signifie pas pour autant qu'il faille évacuer *a priori* les « catégories sociales classiques » de nos analyses. Certes les plateformes du web en font bien peu de cas, privilégiant la connaissance des individus par leurs actions passées plutôt que par leur appartenance à telle ou telle catégorie. Les données du web semblent même construites (Denis et Goëta, 2014) pour effacer toute inscription sociale et politique des individus. Mais est-ce que

25. Les données originales peuvent être téléchargées à cette adresse : http://www.memetracker.org

26. Une famille réunit l'ensemble des mentions d'une citation donnée, sous toutes les formes qu'elle a pu prendre au gré de sa diffusion.

cette absence aplanit tous les comportements ? L'exemple des classes d'activité sur Facebook semble bien indiquer le contraire (voir figure 3.12). L'étude des milieux comme substrat de diffusion ne nous dispense pas d'essayer de comprendre comment un « taux de viralité » observé localement pourrait être imputé à une composition d'acteurs singulière, un changement d'espace linguistique, etc. (Romero et al., 2011; Friggeri et al., 2011).

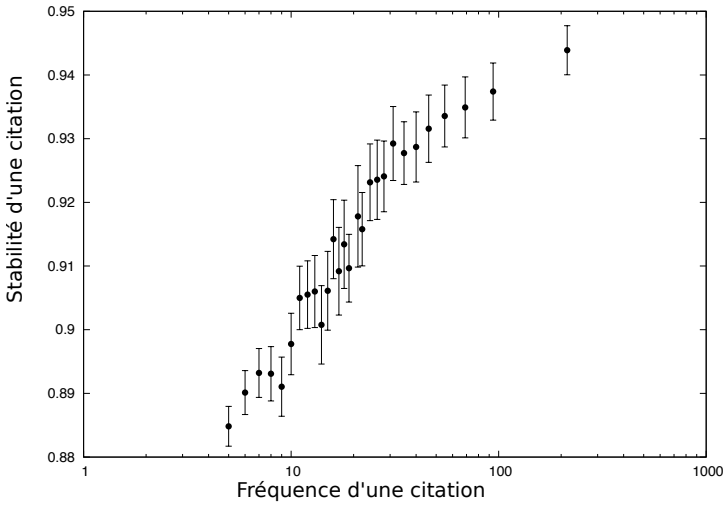


FIGURE 3.13: Stabilité d'une citation en fonction de sa popularité (Omodei et al., 2012).

Adopter une perspective tardienne qui s'attache plus volontiers aux dynamiques de réplication qu'à des ensembles stabilisés pose néanmoins un certain nombre de difficultés pratiques. La première question qui se pose est naturellement celle de la nature de ces entités circulantes. « Répliques », « processus d'aggrégation », « vibrations », Boullier (2015) préfère ne pas pré-définir la forme que pourrait prendre ces entités. Nous partageons son constat que si certaines traces semblent se plier de façon quasiment naturelle à ce type de questionnement (c'est le cas du memetracker, mais aussi des retweets sur Twitter), il faut encore développer les outils de traçabilité ad-hoc pour rendre compte des processus de transformations dans d'autres milieux. Les modèles de reconstruction de la dynamique sémantique des termes pourrait postuler à un tel rôle.

Par définition, les termes se répliquent de texte en texte de manière stable. Pour autant, comme on l'a déjà longuement explicité au chapitre précédent, le sens des termes est largement dépendant de leur contexte d'apparition. À titre d'illustration, la figure 3.14 montre le réseau égocentré du terme *constitution* saisi à travers la topologie du réseau sémantique des discours de l'État de l'Union construit à différentes périodes. On voit immédiatement combien la topologie est changeante et ce qu'elle trahit des variations de sens du terme au fil de l'histoire. Ainsi, on réalise que *consituents* n'est présent qu'à la première période et le terme *slavery* à la seconde. Si la dernière période

introduit la notion d'*ideals*, la troisième insiste sur les *land laws*. Durant les premières décennies de la démocratie américaine, l'environnement sémantique de *constitution* fait apparaître un sous-cluster décrivant spécifiquement le peuple américain. Il disparaît au moment de la Guerre de Sécession et de la Reconstruction. Durant cette période les thèmes associés à *constitution* se densifient, avant de se raréfier à nouveau dans la troisième période pour se limiter à des questions de jurisprudence.

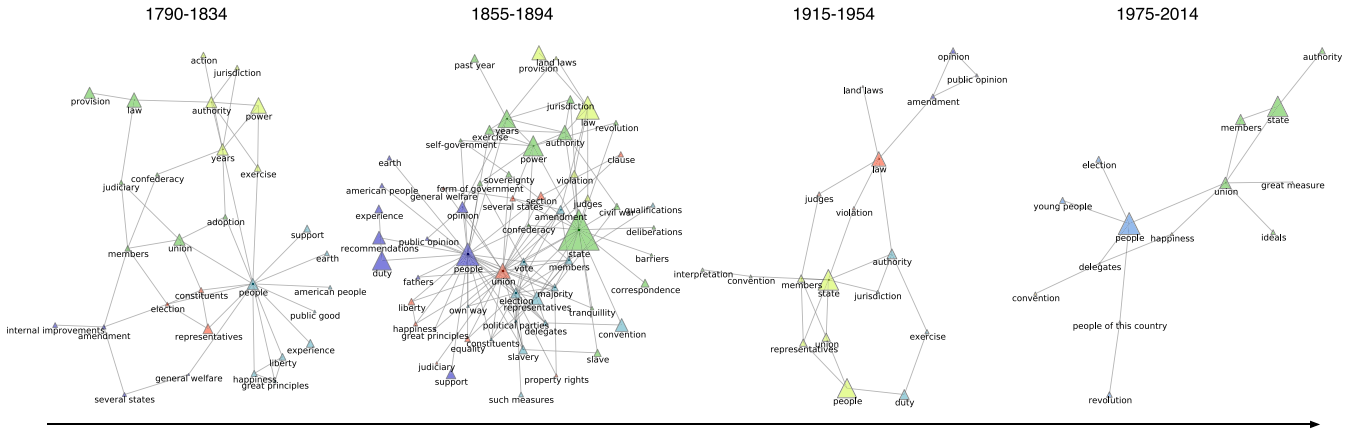


FIGURE 3.14: Réseau sémantique égocentré autour du terme *constitution* dans les discours de l'État de l'Union calculé durant quatre périodes différentes. (extrait de (Rule et al., 2015)).

Les modèles de plongement de mots permettent également de quantifier et formaliser ces transformations de façon rigoureuse. Hamilton et al. (2016) ont ainsi proposé une méthodologie pour mesurer le glissement sémantique de termes individuels à différentes périodes. Contrairement à la méthode que nous avions introduite au chapitre précédent pour mesurer les transformations globales des discours de l'État de l'Union (figure 2.7) en calculant un modèle mixte de plongement de mots et de paragraphes pour l'ensemble de la période d'observation, la méthode introduite par Hamilton et al. (2016) suppose de calculer un modèle sémantique différent à chaque période puis d'aligner les modèles à l'aide d'une transformation de l'espace qui rende comparable les positions des termes à différentes périodes. D'autres méthodes alternatives ont également été proposées, sans qu'une évaluation précise n'ait encore été menée pour les comparer (Szymanski, 2017), notamment la réinitialisation des positions des termes par les positions finales apprises durant la période précédente (Kim et al., 2014) ou en opérant une régression linéaire locale autour du voisinage d'un mot donné (Kulkarni et al., 2015).

L'avantage de la méthode que nous avons déjà présentée dans la section 2.1.4 est qu'elle permet de travailler sur des corpus de taille réduite. il serait en effet impossible de construire des modèles sémantiques calculés sur quelques discours seulement et de suivre la trajectoire sémantique d'un terme avec les

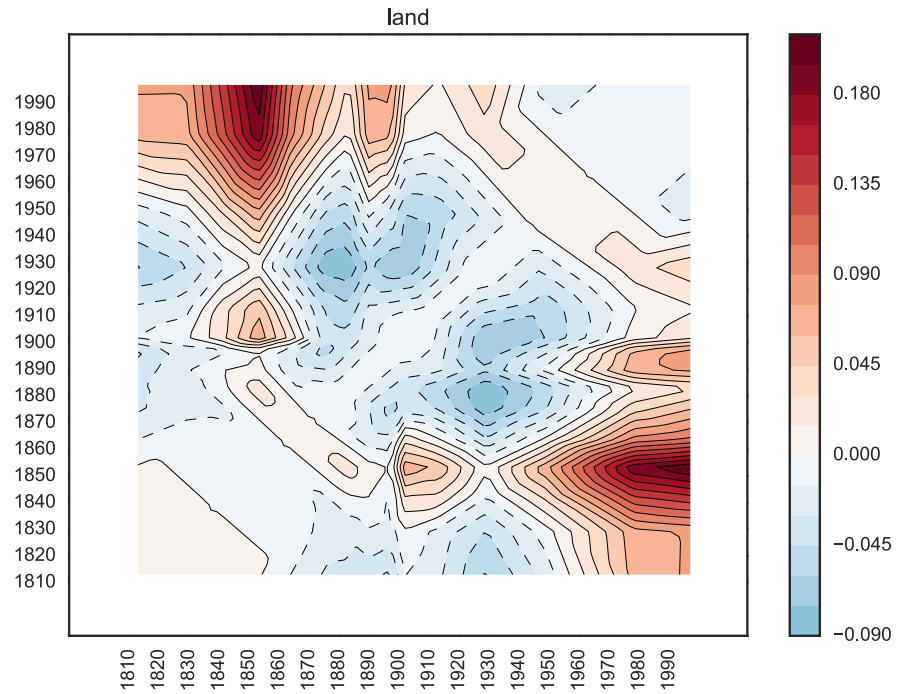


FIGURE 3.15: Diagramme de transformation de *land* : la distance sémantique normalisée (la distance brute est en effet normalisée par un profil de transformation moyen propre au couple de périodes considéré) entre les contextes de *land* saisis à deux périodes distinctes (la taille des périodes est ajustée en fonction des variations du volume d'apparition du terme, ce qui explique l'extension temporelle bornée du diagramme). On observe qu'entre la période contemporaine (depuis la seconde partie du XX^{ème} siècle) et le milieu du XIX^{ème} siècle le sens de *land* a significativement changé. Du fait de la normalisation, la table des couleurs utilisée s'étend du bleu au rouge, une zone bleue signifie que le terme est plus stable que la moyenne, tandis qu'une zone rouge montre que les contextes d'apparition du terme ont changé plus vite que la moyenne.

méthodes qu'on vient d'énumérer. Techniquement parlant, notre approche est plus parcimonieuse. Un seul modèle sémantique de base est calculé qui s'appuie sur l'ensemble des discours et de leurs paragraphes plongés dans le même espace, quelque soit leur année de publication. Par suite, on choisit un terme (*land* à titre d'exemple) et on infère sa position à une période donnée en soumettant au réseau de neurones déjà entraîné l'ensemble de ses contextes d'apparition durant un intervalle de temps donné. On calcule ensuite la similarité sémantique entre chaque position du terme inférée à différentes périodes. C'est cette matrice de départ que l'on interpole pour construire le diagramme de transformations (fig 3.15) qui permet de visualiser l'intensité des transformations subies par notre terme. On remarque ainsi la très forte différence de sens entre *land* tel qu'utilisé durant la période 1841-1865 et la période contemporaine 1978-2016.

Pour caractériser plus finement la nature de ces changements, la stratégie la plus directe consiste simplement à comparer le voisinage sémantique de *land* à ces deux périodes. C'est ce que la figure 3.16 vise à représenter. En premier lieu, on reprend la carte obtenue en appliquant l'algorithme de réduction de dimensionnalité t-SNE que l'on avait décrit au premier chapitre (section 1.2.2). Les 100 termes les plus proches de la position de *land* (non pas inférée pour une période mais apprise sur l'ensemble du corpus) permettent d'interpréter grossièrement la structuration de l'espace. Par la suite, on superpose à cette

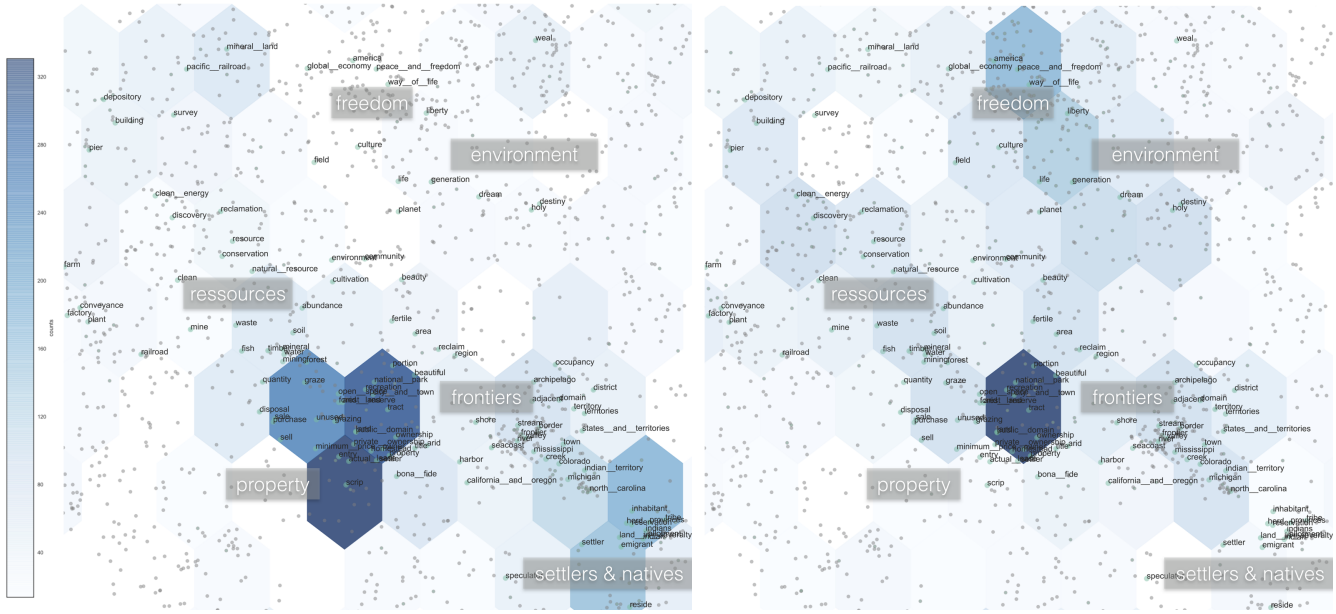


FIGURE 3.16: L'espace sémantique déjà décrit au premier chapitre et représenté dans la figure 1.15 sert de fond de carte pour représenter (sous la forme d'un histogramme hexagonal) la distribution des 500 termes les plus proches de la position du terme *land* telle qu'inférée en fonction de ses contextes d'apparition durant les périodes allant de 1841-1865 [à gauche] et courant de 1978 à 2016 [à droite]. Seuls les 100 termes les plus proches sont étiquetés.

carte un histogramme hexagonal représentant la distribution spatiale des termes les plus proches de la position de *land* calculé à chaque période [27].

La distribution spatiale des voisins les plus proches de *land* indique les différentes valences de sens que notre terme peut prendre à différentes époques. On note ainsi que si *land* était encore connoté par des questions relevant de l'occupation du territoire par des colons (*settlers*) et les problèmes de cohabitation posés par la présence de tribus indiennes, il est associé à des valeurs de liberté typiques de la culture américaine dans la période la plus récente. Si les notions de propriété et de frontière sont toujours convoquées, c'est de moins en moins pour traiter de questions d'exploitation agro-forestière, mais plutôt en lien avec la gestion des parc nationaux.

Sans permettre la même souplesse que la méthode précédente qui consistait à inférer des positions virtuelles de termes en considérant des sous-ensembles spécifiques du corpus, les méthodes d'alignement vectoriel sont également riches de promesses pour saisir les points de stabilité et de transformation dans un corpus. Dans la lignée d'Hamilton et al. (2016), j'ai notamment fait appel à la procédure d'alignement par une transformation Procrustes dans deux projets récents : l'analyse des abstracts de publications scientifiques sur le cancer (et ici plus spécifiquement sur le cancer du sein [28]) et l'analyse des résumés d'articles publiés par le New York Times dont j'ai constitué l'archive exhaustive avec l'aide d'Alix Rule.

Dans les deux cas, le corpus a été découpé en tranches temporelles (annuelle

27. Techniquement, pour chaque terme (parmi les 500 plus proches voisins de *land*), on incrémente la cellule hexagonale dans laquelle il se trouve d'un score proportionnel à sa similarité avec *land*.

28. Ce projet (*Oncology Metaknowledge Network*) réunit Alberto Cambrosio et James Evans. Le financement du CIHR permet de financer le post-doc d'Alexandre Hannud Abdo avec qui je collabore à Paris.

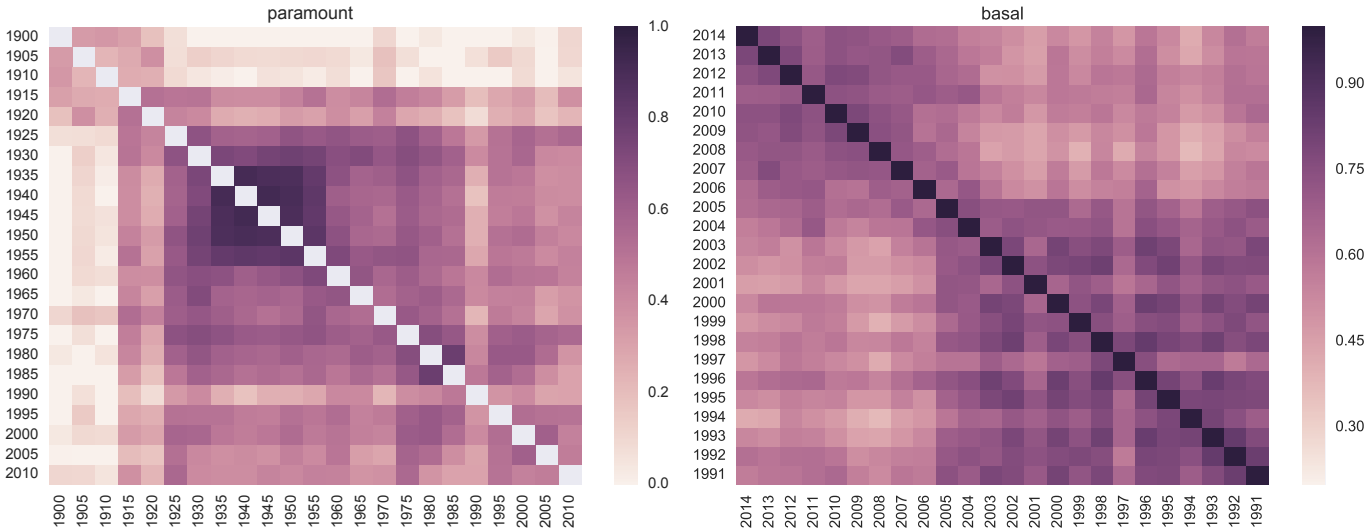


FIGURE 3.17: Diagrammes de similarité des termes *paramount* et *basal* calculés respectivement sur le corpus des résumés des articles du New York Times (regroupés par sous-corpus de 5 ans) et sur le corpus des abstracts Pubmed sur la recherche sur le cancer du sein. (Les cellules diagonales n'ont pas été représentées dans le cas de *paramount* mais valent 1 par construction).

dans le premier cas, tous les cinq ans dans le second). Un modèle sémantique a été entraîné pour chacun de ces sous-corpus. Les modèles successifs ont ensuite été alignés via une procédure Procrustes si bien que la position d'un mot à un pas de temps donné peut-être comparée à sa position à un autre pas de temps. La figure 3.17 représente pour chacun des jeux de données, l'un des termes dont la position a changé de la façon la plus dramatique (on a simplement calculé pour l'ensemble du vocabulaire, les termes dont la position variait le plus à travers l'intervalle de temps étudié, *paramount* et *basal* font ainsi partie des dix termes les plus fluctuants de chacun des jeux de données). Chaque cellule de la matrice correspond à un couple de périodes et représente la similarité entre la position du terme à chacune de ces périodes. Une zone sombre équivaut donc à une période de stabilité, tandis que les cellules blanches signalent de fortes disparités.

D'une certaine façon, le résultat peut paraître dérisoire, les experts du domaine auront tôt fait de remarquer que *basal* est essentiellement employé dans les années 90 dans son sens premier c'est à dire comme un adjectif signifiant basique. A compter du milieu des année 2000, de nombreuses recherches sont menées sur le *TNBC* (*Triple Negative Breast Cancer*) également appelé (même si ce n'est pas si simple (Keating et al., 2016)) *basal-like breast cancer*. La transition que l'on observe indexe bien ce changement, mais il s'agit d'une simple mutation linguistique. L'exemple de *paramount* au sein du corpus d'articles de presse (ou en tout cas de leur résumé) est plus intéressant. *Paramount Pictures Corporation* est l'un des plus anciens studios de cinéma aux États-Unis. Il a été créé en 1912, période à partir de laquelle on repère un changement de contexte dans les usages du terme *paramount* [29]. Naturellement, le terme était déjà présent dans notre archive mais il était exclusivement

29. Classiquement, dans les modèles de plongement de mots, le pré-traitement linguistique du texte est minimal. Une des opérations qui est habituellement effectuée pour normaliser le texte est de le passer en minuscule. Il n'est donc pas possible de distinguer les studios de l'adjectif.

utilisé avant les années 10' comme un adjectif. Il est intéressant de noter que, passé l'âge d'or du studio dans les années 30' où la mention de *paramount* était quasiment sysématiquement liée à la sortie d'une grande production hollywoodienne, le terme semble subir une « perturbation » durant la période 1990-1995. Cette ligne diparate s'explique par la nature des articles du New York Times couvrant l'actualité de la *Paramount* à l'époque. En effet, le studio faisait alors les grands titres car il était la cible de spéculations financières. Une fois la fusion avec le géant des médias *Viacom* actée en 1993, Titanic et Star Trek reprennent leur droit sur Wall Street.

La métamorphose temporaire d'un studio de cinéma en produit financier ne relève pas simplement d'un phénomène de renouvellement linguistique mais a directement trait à des dynamiques sociales et politiques environnantes. Sur ce dernier exemple, on perçoit combien cette modalité d'observation des vibrations sémantiques est susceptible de renseigner directement l'analyste sur des transitions majeures du corpus et du monde qu'il indexe.

3.3 *Epistémologie numérique*

On reviendra dans un premier temps (section 3.3.1) sur les débats toujours d'actualité sur l'avènement annoncé d'une science des données entièrement faite de corrélations, ringardisant les questions d'interprétation voire de causalité. Naturellement, passé les effets d'annonce de la dernière décennie, certaines voix plus nuancées se sont exprimées, faisant valoir un modèle où co-existent big data et approche qualitative et où les notions d'interprétation ou d'explication ont encore droit de cité. Suite à cette discussion, on cherchera à en tirer quelques conclusions quant aux propriétés souhaitables pour garantir l'harmonie d'une pensée sociologique équipée (et non pas entravée) de méthodes numériques (section 3.3.2). Dans un troisième temps, on inversera la perspective et plutôt que de considérer les méthodes et les traces numériques comme une menace pour les sciences sociales (Savage et Burrows, 2007), on interrogera leur capacité à remettre en question les principes de ce que Dominique Boullier appelle les sciences sociales de première et deuxième génération. On s'attardera ainsi plus particulièrement à la notion d'échantillonnage à l'heure des traces numériques en nous demandant notamment comment les « big data » peuvent changer les pratiques de recherche dans des situations aussi prosaïques que la construction d'un corpus thématique (section 3.3.3).

3.3.1 *Corrélations, prédictibilité et interprétation*

La crise des sciences sociales empiriques, déjà prophétisée (Savage et Burrows, 2007) il y a 10 ans, n'a fait que se confirmer avec l'avènement des big data (Burrows et Savage, 2014). Ce nouveau modèle a soulevé un feu nourri de critiques que l'on a déjà largement commentées dans les deux parties précédentes. Les sciences sociales sont ainsi accusées d'abriter des théories incohérentes les unes avec les autres, et de manquer d'assise empirique (Watts, 2017b). L'argument principal est de nature épistémologique et vise l'incapacité de la sociologie (et des sciences sociales en général) à produire des modèles causaux satisfaisants. Duncan Watts, ancien sociologue des réseaux à Columbia University et maintenant chercheur à Microsoft anime la discussion tambour battant et affirme dans l'AJS (American Journal Of Sociology) :

> « *if sociologists want their explanations to be causal, they must place less emphasis on understandability (i.e., sense making) and more on their ability to make predictions.* » (Watts, 2014) [30]

30. « si les sociologues veulent que les explications puissent être causales, alors, ils doivent moins se préoccuper de compréhension (*i.e.* faire sens) et travailler à leur capacité de prédictions. »

Le modèle que défend Watts est très clairement sous influence du workflow conceptuel de l'apprentissage automatique qui fait invariablement se succéder choix des données, définition d'une tâche, évaluation de la tâche. Dans son plaidoyer, Watts souhaite soumettre toutes les recherches en sciences sociales à ce même schéma, conditions nécessaires à rendre les sciences sociales plus « utiles » (Hofman et al., 2017). Sans être entièrement rejetées, les études dites exploratoires ne semblent plus guère bonnes qu'à générer des hypothèses avant que bagging et boosting ne prennent le relais pour induire les véritables modèles prédictifs.

La charge initiale n'est pas restée sans réponse. Deux chercheurs du MIT (de la Sloan Management School) : Turco et Zuckerman (2017) ont ainsi répondu à Watts en défendant la pratique interprétative comme un moyen privilégié pour identifier les mécanismes causaux : « To the contrary, we argue that theories that account for actors' situated intentions, beliefs, and opportunities aid in the identification of generalizable causal mechanisms ». Argument balayé d'un revers de main quelques mois plus tard dans le droit de réponse de Watts (2017a) : interprétation et causalité sont totalement disjointes, la recherche d'explications embarque nécessairement les catégories de sens commun contre lequelles le chercheur ne peut pas lutter. Trop verbeuses, les histoires que produisent la recherche interprétative ont tendance à alourdir inutilement les modèles (« overfitting »). Elles ne rendent plus compte que d'un cas singulier et ne permettent pas de tirer des conclusions plus générales.

Naturellement, résumé en si peu de mots, le compte rendu de ces débats peut paraître caricatural. Il n'empêche qu'ils permettent de se rendre compte de la distance qui nous sépare ici par exemple de la « compréhension causale »

de Weber pour qui les deux notions sont intimement liées (Gonthier, 2004). La multiplication des données numériques et la prolifération des méthodes d'analyse, notamment en apprentissage automatique, semblent plaider pour une causalité de type beaucoup plus mécanistique en lien avec une pensée positiviste qui se plairait sans doute à voir les épistémologies des sciences naturelles et des sciences sociales s'aligner quitte à les caricaturer toutes les deux.

Ce qui qui frappe également dans les méthodes employées en science des données, c'est la nature extrêmement inductive des procédures mathématiques[31] à tel point qu'Anderson (2008) écrit :

> « *Petabytes allow us to say : "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.* »[32]

Anderson exprime ici avec un enthousiasme non feint une croyance assez répandue dans la toute puissance de l'inférence[33] des big data. Pourtant certains auteurs n'ont pas manqué de signaler les limites de ces approches enitèrement athéoriques. Par exemple, Lazer et al. (2014) montrent que l'algorithme de prédiction des épidémies de grippe Google Flu après avoir été célébré(Ginsberg et al., 2009) a parfois de sérieux ratés : manquant certains épisodes viraux (en particulier lorsqu'ils sortent du cycle saisonnier habituel, ce qui est le cas de la pandémie de grippe A-H1N1) ou surestimant à d'autres moments largement la prévalence du virus par rapport aux statistiques du Centre pour le contrôle et la prévention des maladies américain (CDC). L'ensemble plutôt opaque de termes de recherche qui étaient utilisés pour estimer la propagation de la grippe est une source possible de l'erreur. Cette liste a été produite en identifiant le profil temporel des requêtes les mieux corrélées avec les profils temporels de l'épidémie tels qu'avérés par le CDC. Parmi les termes de recherche retenus, les chercheurs de Google avaient ainsi inclus des éléments comme « high school basket-ball » dont la corrélation avec les épisodes de grippe passés était visiblement fortuite (Ginsberg et al., 2009). Ce dernier a été exclu de la dernière version de l'algorithme, mais il illustre parfaitement le risque de corrélations parasites lors du traitement d'une énorme quantité de données sur un ensemble finalement réduit de points de données. Et cette « crise » n'est pas exclusive aux sciences sociales, bien au contraire ! La légitimité des méthodes employées dans bien d'autres disciplines est régulièrement remise en cause, et l'irruption du big data menace à nouveau de bouleverser les pratiques et l'épistémologie d'autres disciplines. Il est d'ailleurs cocasse d'observer que ce sont souvent les procédures statistiques comme la sacro-sainte *p-value* qui soulèvent le scepticisme dans la recherche biomédicale ou en psychologie (Goodman, 1999; Moher et al., 2009; Ioannidis, 2005; Chavalarias et al., 2016).

31. Antoine Mazières a fortement insisté dans sa thèse (Mazieres, 2016), que j'ai co-encadrée avec Christophe Prieur, sur la nature inductive du raisonnement en apprentissage automatique. Il y raconte la façon dont des approches aux origines parfois anciennes s'étaient concrétisées plus récemment en des algorithmes prêts à l'usage au sein de librairies à la diffusion très large à la faveur de données disponibles en masse, de moyens de calcul adaptés, et d'une demande de prédiction des comportement individuels toujours plus forte.

32. « Avec les petabytes on peut dire : "les corrélations nous suffisent". Nous n'avons plus besoin de modèles. Nous pouvons analyser les données sans avoir à formuler d'hypothèses sur ce qu'elles devraient montrer. Il nous suffit de faire ingurgiter les chiffres aux plus grands clusters de calcul que le monde n'a jamais connu et laisser les algorithmes statistiques trouver les motifs que la science a échoué à trouver jusque là »

33. Anderson la qualifie de statistique, mais l'adjectif est assez impropre pour des procédures d'apprentissage automatique qui viennent de l'intelligence artificielle.

Quand Watts souhaite tester les capacités de prédiction des modèles d'action sociale, Anderson simplifie encore l'équation et nous dispense même de la recherche d'un modèle pourvu que l'on sache prédire. Il est possible que l'article d'Anderson relève de la pure provocation mais il a néanmoins marqué un tournant dans les débats obligeant les chercheurs à prendre position dans différentes disciplines.

Mazzocchi (2015) souligne ainsi que dans l'épistémologie poppérienne, toute entreprise scientifique s'appuie nécessairement sur certaines hypothèses. L'induction pure n'existe pas. Même si elles sont traitées par des boîtes noires, les données « ingérées » par les algorithmes d'apprentissage doivent toujours être sélectionnés par l'expérimentateur. De plus, un résultat ne saurait être reconnu comme tel qu'en comparaison avec une certaine attente qui prouve l'existence d'une théorie sous-jacente.

Kitchin (2014) analyse ce type de posture comme une renaissance empiriste défendue par des intérêts commerciaux qu'il a tôt fait de renvoyer face à ses contradictions. Il identifie néanmoins parallèlement l'émergence d'une science des données (« data-driven science ») qui selon lui est susceptible de dépasser les limites de la science de la connaissance (« knowledge-driven science ») condamnée dans un environnement où la collecte des données était coûteuse à une logique purement déductive.

C'est un argument connexe que l'on retrouve chez Goldberg (2015) qui prend la défense des approches inductives comme un moyen d'émancipation de l'imagination sociologique. Plutôt que de défendre *mordicus* la méthode sociologique comme la recherche des causes contre les corrélations, il voit dans ces nouvelles méthodes l'occasion de soigner la « myopie catégorielle » des sciences sociales et de redonner ses lettres de noblesse à la méthode abductive (Timmermans et Tavory, 2012) qui permet de recalibrer sans cesse la théorie en fonction des découvertes empiriques. C'est le même sentiment libératoire que l'on retrouve chez Latour et al. (2012), pour qui la navigation dans les traces numériques nous dispense de postuler *a priori* l'existence de deux niveaux d'agrégation séparant les individus et la société. DiMaggio et al. (2001) font du caractère inductif de ces méthodes l'une des trois propriétés[34] fondamentales d'une procédure d'analyse textuelle « saine » :

> « *it must be inductive to permit researchers to discover the structure of the corpus before imposing their priors on the analysis, and to enable different researchers to use the same corpus to pursue different research questions.* »[35]

34. Les deux autres propriétés sont assez simples. Ils s'agit du caractère explicite de celles-ci (qui doivent permettre à d'autres chercheurs de tester leur propre interprétation des données ultérieurement) et de leur caractère automatique.

35. « elle doit être inductive pour permettre au chercheurs de découvrir la structure d'un corpus avant que ses idées préconçues n'orientent l'analyse, et pour permettre à différents chercheurs d'utiliser le même corpus pour répondre à des questions de recherche différentes. »

3.3.2 *Équiper sans entraver*

Partant du constat partagé de l'importance des processus inductifs, on réalise combien les conséquences pour les sciences sociales peuvent diverger selon les auteurs. Dans un cas, les algorithmes d'apprentissage identifient les corrélations, formulent les modèles d'eux mêmes ou, pour les meilleurs d'entre eux, prédisent le comportement des systèmes sociaux. Libéré de la chappe du « common sense » (Watts, 2014) des sociologues, ils peuvent enfin prédire optimalement. Dans l'autre cas, c'est l'imagination des sociologues qui se trouve libérée (DiMaggio et al., 2001). Ces derniers se saisissent alors des nouvelles catégories induites par l'analyste automatique pour recomposer une interprétation sociologique.

À ce stade du manuscrit, notre affinité pour le second modèle ne devrait pas surprendre le lecteur. Le choix n'est pas très difficile car après tout, si l'on prend les recommandations de Duncan Watts au sérieux, quel type de prédiction pourrait générer notre travail sur les discours de l'État de l'Union ? Si certains travaux (et une bonne part des recherches socio-historiques risquent d'être concernées) sont exclus par construction, c'est aussi parce que les problèmes que Watts jugent scientifiques sont finalement en nombre limité et renvoient à une vision relativement réductrice de la discipline. Dès lors, un paradoxe réside, comment les mêmes outils peuvent-ils être utilisés tantôt de façon instrumentale pour administrer la preuve, tantôt comme un outil d'exploration au cœur d'une démarche abductive. Autrement dit, à quelles conditions, l'utilisation de ces méthodes numériques permettent-elles d'échapper à un assèchement du raisonnement sociologique ?

On essaiera de répondre à la question de deux manières différentes. D'une part nous reviendrons brièvement sur notre projet d'analyse du processus d'émergence de la biologie de synthèse, qui nous servira à illustrer le concept de lecture multi-niveau des corpus. D'autre part, on présentera rapidement le design de l'infrastructure de recherche CorText .

Notre étude scientométrique sur la biologie de synthèse, même si elle n'a pas grand chose à voir avec la sociologie du web (encore que les réseaux de citations scientifiques sont célèbres pour avoir inspiré le principe du Page Rank), illustre à merveille le concept de lecture multi-niveau des corpus textuels. Pour rappel, ce projet questionne la dynamique d'émergence de la communauté des biologistes de synthèse depuis le début des années 2000 à partir d'un corpus de publications académiques. Pour l'étudier, nous avons largement fait varier les échelles : que l'on s'intéresse aux propriétés macroscopiques de la communauté (mesure de cohésion décrite ci-dessous), à l'hétérérogénéité thématique constitutive du champ (déjà décrite au chapitre précédent et illustrée par la

figure 2.22), ou aux trajectoires individuelles des chercheurs les plus centraux du domaine (dont on reproduit l'analyse ci-dessous). Fort de ces regards à différentes échelles, et des aller-retour réguliers entre connaissance de terrain (obtenue *via* travail d'analyse biographique et entretiens, etc.) et cartographies hétérogènes, nous avons pu, au-delà du seul cas le la biologie de synthèse, proposer un modèle original d'émergence des champs techno-scientifiques qui articule stratégies d'ouverture et de clôture progressive de la communauté et rôle des « boundary spanners » dans la construction de la légitimité du champ. Outre le changement de focale (macro, meso, micro déjà décrit ci-dessous), la richesse du modèle vient essentiellement des déplacements de point de vue qui ont permis de comprendre, entre autres choses, comment la centralité dans l'espace épistémique des publications académiques s'explique par la capacité des acteurs à mobiliser des ressources depuis d'autres espaces (industriels ou institutionels).

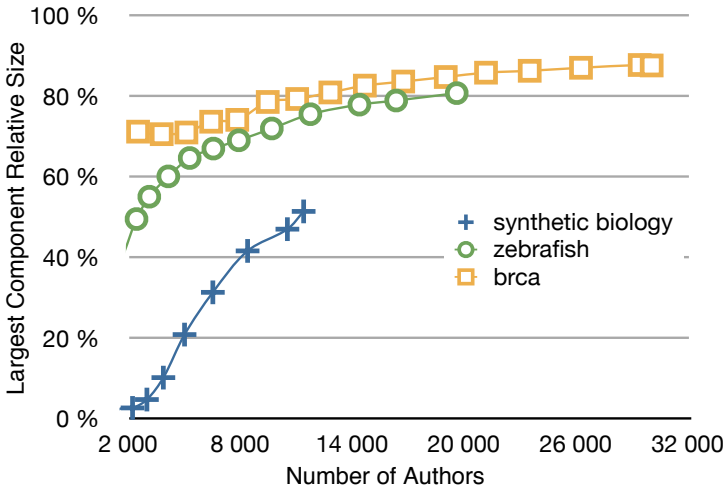


FIGURE 3.18: Evolution de la taille de la composante connexe principale (mesurée en proportion de la taille totale du réseau de collaboration) de trois communautés de recherche. Chaque point correspond à une année mais c'est bien la taille du réseau pour une année donnée qui figure en abscisse.

Pour détailler plus avant, notre analyse s'est déployée à trois échelles différentes. D'abord au niveau macroscopique, nous avons mesuré la taille de la plus grande composante connexe du réseau de collaboration extrait de l'ensemble des articles du corpus. En l'interprétant comme une mesure de cohésion de la communauté (Bettencourt et Kaiser, 2015), sa croissance nous a permis de conclure que la communauté de biologie de synthèse était bien en cours de cristallisation. Pour autant nous avons également pu observer en appliquant la même méthode à d'autre champs[36] que cette structuration était, à nombre total d'auteurs équivalent (et sans que le nombre moyen de co-auteurs par papier soit non plus significativement différent dans les différents domaines), beaucoup plus lente en biologie de synthèse que pour les recherches sur le zebrafish ou BRCA qui nous ont servi de points de référence (fig. 3.19). On interprète ce retard relatif de la communauté des biologistes de synthèse par rapport à des champs plus classiques comme une

36. C'est là une propriété des méthodes numériques dont il faudrait expliciter les conséquences plus longuement. Parce qu'elles sont explicites, l'ensemble des opérations de traitement du corpus original jusqu'à sa forme finale peut être appliquée indistinctement à tout domaine et toute source de données. Plutôt que de parler de reproductibilité (la machine générera a priori toujours le même résultat de toute façon, c'est plutôt de la robustesse des résultats qu'il faudrait s'inquiéter) il vaudrait sans doute mieux insister sur la capacité de ces méthodes à produire des comparaisons.

manifestation du caractère pluri-disciplinaire du domaine qui vise à fédérer des recherches assez hétérogènes. On met cette hétérogénéité en évidence en cartographiant la structure du réseau de co-citation (voir figure 2.22 au chapitre précédent) dont l'interprétation en quatre écoles est bien compatible avec les dires d'experts. Enfin, nous nous sommes intéressés à une échelle plus microscopique aux membres remarquables de la commuanuté. La représentation que nous avons construite et reproduite ci-dessous figure 3.19 est ainsi remarquable de ces changements de focale et de point de vue. L'histogramme hexagonal représente en premier lieu la distribution de centralité (dans le réseau de collaboration) et d'impact (en citations reçues) des chercheurs de la communauté. L'immense majorité des individus se concentrent dans la partie inférieure gauche. Quelques individus remarquables dont le nom est rajouté à la carte s'en éloignent néanmoins. Les écoles épistémiques (de nature co-citationnelle) auxquelles ces chercheurs contribuent sont re-projetés dans cette représentation [37]. On voit ainsi au sein d'un même diagramme s'aggréger des informations à différentes échelles (macroscopique pour la distribution de la population, mésoscopique pour les écoles, microscopiques pour les 20 chercheurs principaux) et relevant de différentes inscriptions (collaboration, impact, et citations émises). Mais ce diagramme serait purement descriptif s'il ne permettait de justifier l'argument des « boundary spanners ». Une recherche biographique ultérieure a ainsi permis de montrer que ces mêmes individus jouissaient également de positions extrêmement centrales dans d'autres espaces sociaux connexes. ils sont ainsi très présents dans l'industrie, via la participation à des start-ups, ce qui garantit l'afflux de ressources financières. Mais ils jouent également un rôle central dans les grandes institutions de la biologie de synthèse et dans les instances de régulation.

37. L'une de nos hypothèses de départ, qui est restée à l'état d'hypothèse, était que les principaux représentants de chaque école pouvaient développer des stratégies différentes pour construire leur légitimité.

Garantir une fluidité de circulation entre échelles (micro/meso/macro), dimensions (citations (émises et reçues)/collaborations) et points de vue (arène épistémique/industrielle/institutionnelle/etc.) est une opération coûteuse. Dans l'exemple qui nous occupe, par exemple, la question des homonymes est (relativement) secondaire jusqu'à ce qu'on cherche à identifier un groupe restreint d'individus, et il faut alors prendre en charge le problème. Mais ce n'est pas une difficulté propre aux recherches quantitatives. Ainsi, alimenter le tableau 3.2 n'a pas été une synécure pour Benjamin Raimbault. Les méthodes numériques que nous avons employées ont simplement permis de fluidifier le travail d'articulation et de déplacement des niveaux d'analyse.

C'est cette liberté de circulation de l'analyste que le design de la plateforme CorText tâche de préserver. Née en 2010, la plateforme CorText s'est construite autour d'un noyau de trois personnes engagés dans une entreprise commune de « créolisation » [38] qui mélangeait alors linguistique (avec Audrey Baneyx), ingénierie (représentée par Philippe Breucker) et sociologie (personnifiée par Marc Barbier). Le point de départ était donc résolument pluri-disciplinaire avec

38. Un créole qui s'oppose à l'émergence de pidgins tant redoutés par McFarland et al. (2015).

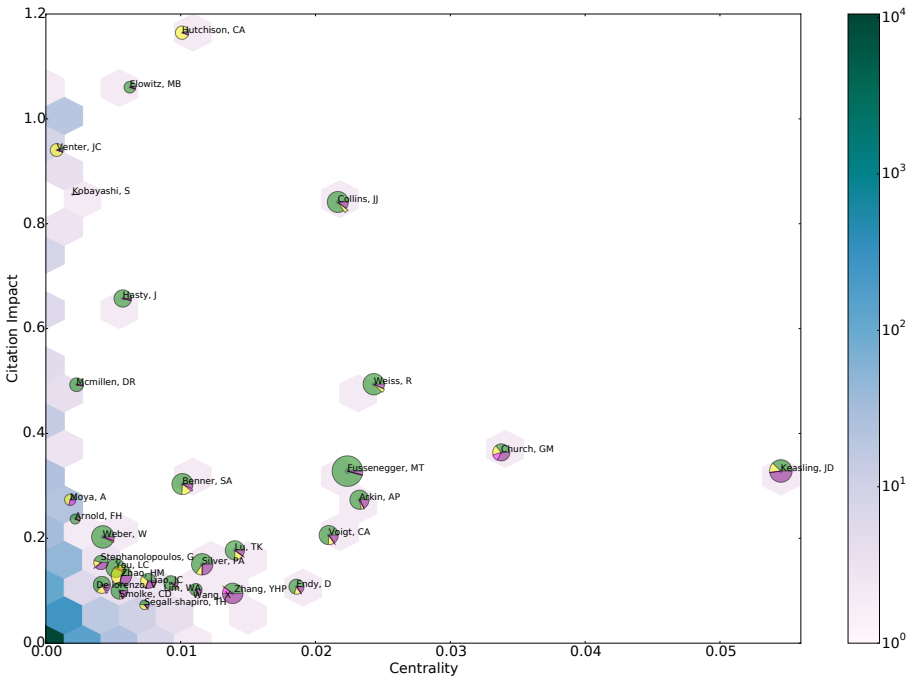


FIGURE 3.19: Diagramme stratégique des chercheurs en biologie de synthèse.

l'ambition originale de construire une véritable infrastructure de recherche au cœur d'un laboratoire de sociologie des sciences. D'autres compétences en traitement de la langue, physique, analyse de réseaux, développement web, design, visualisation de données, ont progressivement enrichi l'équipe de départ.

Le projet a finalement convergé vers la réalisation d'une plateforme d'analyse de données textuelles entièrement en ligne [39] intitulée : CorText Manager. Elle est opérationnelle dans sa nouvelle mouture (v2) depuis novembre 2016.CorText s'est avant tout construit au gré de projets de recherche variés (en histoire, sciences politiques, économie, sociologie et même en neurosciences (Mesmoudi et al., 2015)) avec des corpus de tout horizon (venus du web mais aussi issus des médias et des sciences). C'est sans doute de cette construction itérative au fil des collaborations qu'est née sa structure extrêmement modulaire. Tous les outils sont réunis dans un espace unique où les analyses peuvent être combinées et configurées librement pour construire des scénarios d'utilisation originaux.

39. https://managerv2.cortext.net

Les scripts d'analyse actuellement disponibles sont répertoriés dans 5 grandes familles : exploration de données, analyse du texte, traitement des données catégorielles, cartographie de données et analyse temporelle. Pour ne citer que quelques uns de ceux sur lesquelles les travaux présentés dans ce mémoire s'appuient, on retrouve notamment des capacités d'extraction

TABLE 3.2: Membres du « core-set » ordonnées en fonction de leurs écoles épistémiques d'appartenance (Appro.) (BE=Biological Engineering, ME=Metabolic Engineering, GE=Genome Engineering and PC=Protocell Creation) et de leur participation à des activités non-académiques estimées à travers trois indicateurs : secteur privé (Priv.), institutions (Inst.) et gouvernance (Gouv.).

PI	École	Priv.	Inst.	Gouv.
CA Voigt	BE	3	4	1
P Silver	BE	2	4	2
D Endy	BE	3	3	6
J Keasling	BE ME	3	3	2
G Church	BE ME PE	3	3	2
J Collins	BE	3	2	1
R Weiss	BE	2	2	2
A Arkin	BE	2	2	0
W Lim	BE	1	2	1
M Fussenegger	BE	3	2	0
F Arnold	BE	3	2	0
C Smolke	BE	3	2	1
J Hasty	BE	2	1	0
T Lu	BE	3	1	0
X Wang	BE	1	1	0
D Mcmillen	BE	1	1	0
W Weber	BE	3	1	0
L You	BE	0	1	0
G Stephanopoulos	BE	3	1	0
T Segall-Shapiro	BE GE	1	0	0
J Liao	BE GE ME	3	0	0
M Elowitz	BE	2	0	1
S Benner	BE	3	0	0
V De Lorenzo	BE	3	0	2
Y Zhang	BE ME	2	0	0
C Hutchison	GE	2	0	0
C Venter	GE	3	0	1
H Zhao	GE ME	2	0	1
C Moya	PC GE	0	0	0

terminologique à partir de texte en langage naturel (section 2.1), des solutions de cartographie hétérogène (section 2.11), la possibilité de construire et comparer les profils d'évolution de termes sous la forme de séries temporelles (section 1.2.3), la détection automatique de périodes (section 2.1.4), des outils d'exploration de corpus qui permette aisément de lire les contenus individuels, et de segmenter les corpus, etc. [40]

40. La documentation en ligne : http//-docs.cortext.net liste l'ensemble des « scripts » à disposition.

Longtemps local, le développement de la plateforme ne dépend maintenant plus uniquement des formations organisées à Marne-La-Vallée, et des usages originaux émergent comme ce travail (Rykov et al., 2016) portant sur l'analyse de données Instagram ou cet autre sur la cartographie historique du corpus Bentham (Tieberghien et al., 2016). Largement fondé sur des logiques de partage et de collaboration héritées du monde du logiciel libre, CorText se diffuse à son tour et réunit une communauté de chercheurs qui de par leur pratique expérimentale de l'analyse de corpus font subir aux méthodes et aux données les mêmes épreuves que des bosons dans un accélérateur à particules.

Le design de la plateforme et de son interface (illustrée par une capture d'écran figure 3.20) laisse la maximum de liberté aux utilisateurs. Un point crucial tient au fait que la plateforme soit accessible en ligne. Un simple navigateur suffit, il permet à tout un chacun d'utiliser les outils d'analyse

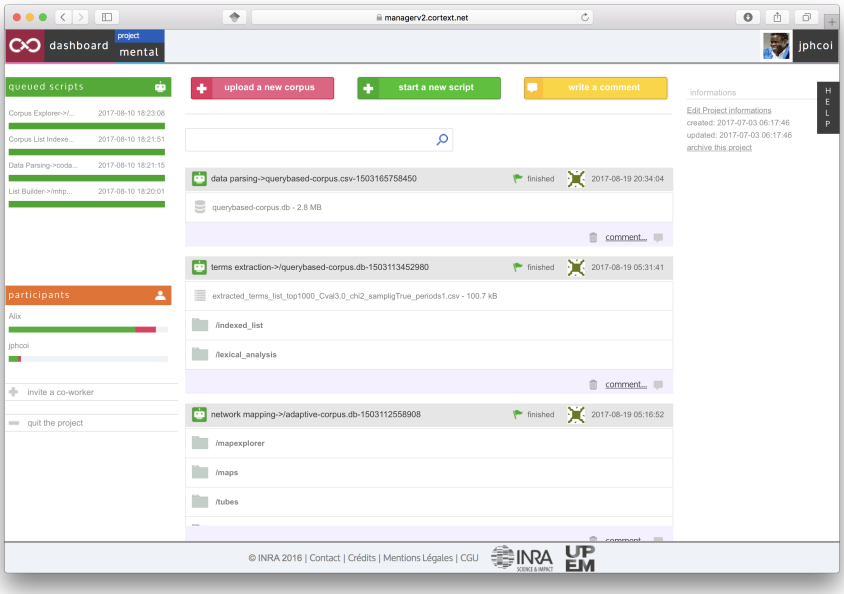


FIGURE 3.20: Vue projet de l'interface de CorText Manager. Dans cet exemple deux utilisateurs contribuent au même projet. Trois actions de base sont possibles : charger de nouvelles données, analyser les corpus existants et commenter les résultats.

quelque soit ses compétences en informatique. Pour autant, la facilité d'accès technique n'équivaut pas à facilité de prise en main. Si l'interface est simple et entièrement en ligne, dès que le premier menu de paramètre s'ouvre, elle place l'utilisateur face à ses responsabilités. En corollaire, les visualisations sont, pour la grande majorité, interactives et directement partageables en ligne. Autre conséquence importante, les mises à jour du code peuvent être apportées côté serveur de façon régulière sans que l'utilisateur ne soit importuné. Une très grande gamme de sources et de formats de fichiers peuvent être lus de sorte que la plateforme n'est pas l'apanage des seuls scientomètres ou exclusivement dédiée à l'étude des corpus médiatiques. Les utilisateurs sont donc libres de faire varier les points de vue et de mélanger au sein d'un même projet des données provenant de sources variées. L'interface du CorText Manager est également minimaliste au sens où trois actions de base sont possibles : ajouter de nouvelles données (*upload a new corpus*), commenter des résultats (*write a comment*), lancer de nouvelles analyses (*start a new script*). La structure de données utilisée est telle que la partie analytique de la plateforme repose à l'heure actuelle sur une petite vingtaine de scripts. Libre aux développeurs désireux de s'investir dans le projet d'ajouter un nouveau module de détection d'entités nommés ou de détection de classes sémantiques via un topic model. Aucune contrainte de langage ou de format ne s'y oppose. Les scripts sont donc indépendants, ils analysent les données à différentes échelles, en construisant des réseaux, extrayant les termes pertinents, représentant les dynamiques, etc. Libre à tout un chacun d'imaginer un scénario d'analyse adéquat pour répondre à une question de recherche donnée. Dernier point

important, les résultats générés par ces scripts laissent une trace au sein de la base de données utilisées au sein de le plateforme [41] mais aussi sous la forme de fichiers qui peuvent être lus par d'autres outils et logiciels (ainsi les cartes produites peuvent être téléchargées dans le format gexf du logiciel Gephi).

41. En travaillant sur la même base, on peut ainsi enchaîner les traitements suivants : détecter les mots clés dans un corpus d'articles de presse (script *term extractor*), éventuellement nettoyer la liste manuellement (*corpus term indexer*), cartographier le réseau sémantique structurant ces termes en grands thèmes (script *analysis*), avant de représenter l'évolution du nombre d'articles mobilisant ces thèmes (script *demography*).

3.3.3 *Echantillonnage de corpus*

Appliquée au web, la notion même de corpus et de délinéation semble particulièrement poussiéreuse. De nombreuses études analysent pourtant la façon dont les données brutes sont avant tout construites supposant des choix techniques et conceptuels forts tout le long d'une chaîne de traduction qui va des actions individuelles à un flux « json » délivré par le « firehose » des API des grandes plateformes du web (Denis et Goëta, 2013). Mais dès lors que les données numériques sont utilisées pour construire un corpus censé refléter quelque phénomène social, ces considérations sont bien vite évacuées et les questions d'échantillonnage, ou de biais de sélection des données sont rapidement englouties sous les superlatifs quant à leur taille ou leur instantanéité.

L'adage dit d'ailleurs qu'il est devenu inutile d'interroger un ensemble représentatif de la population alors même que les plateformes numériques enregistrent déjà l'intégralité des comportements individuels. Anderson (2008) affirme ainsi « Big data approaches are often presented as the final solution to sampling issues » et dans le même article :

> « *Forget taxonomy, ontology, and psychology. Who knows why people do what they do ? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves.* » [42]

42. « Oubliez taxonomies, ontologies ou toute forme de psychologie. Qui pourrait bien savoir pourquoi les gens font ce qu'ils font ? L'essentiel est qu'ils le fassent et que l'on puisse suivre et mesurer ces actions avec une fidélité sans précédent. Avec suffisamment de données, les nombres parlent d'eux mêmes »

En somme, nul besoin de construire un protocole expérimental complexe pour sonder la population quand les plateformes du web enregistrent tout. Le travail mené autour de l'application Algopol offre pourtant un contre-champ édifiant à la promesse d'un regard panoptique sur l'ensemble des actions individuelles que portent les traces du web. On a déjà largement commenté combien l'indexation des activités individuelles par l'API de Facebook pouvait être lacunaire et mal ajustée à la compréhension réelle des comportements des internautes. Mais on s'intéressera plutôt dans ces section aux biais de sélection liés aux stratégies de recrutement des enquêtés.

L'application a circulé en empruntant trois modes de diffusion différents. Une première série d'utilisateurs (735) ont été directement recrutés par l'institut de sondage CSA. Ils sont représentatifs (selon les critères classiques d'un institut de sondage) de la population des internautes français. Un petit nombre d'étudiants ont été directement recrutés à l'université Paris V (147 personnes).

Enfin la grande majorité des utilisateurs (dont 14263 ont contribué au classement ci-dessous) ont simplement été recrutés en ligne, de façon virale ou après avoir lu des articles mentionnant l'expérience dans les quotidiens nationaux comme Le Monde. La figure 3.23 montre comment les profils d'activité que nous avons décrits dans la section précédente se distribuent sur l'ensemble de ces populations. Naturellement les distributions divergent largement.

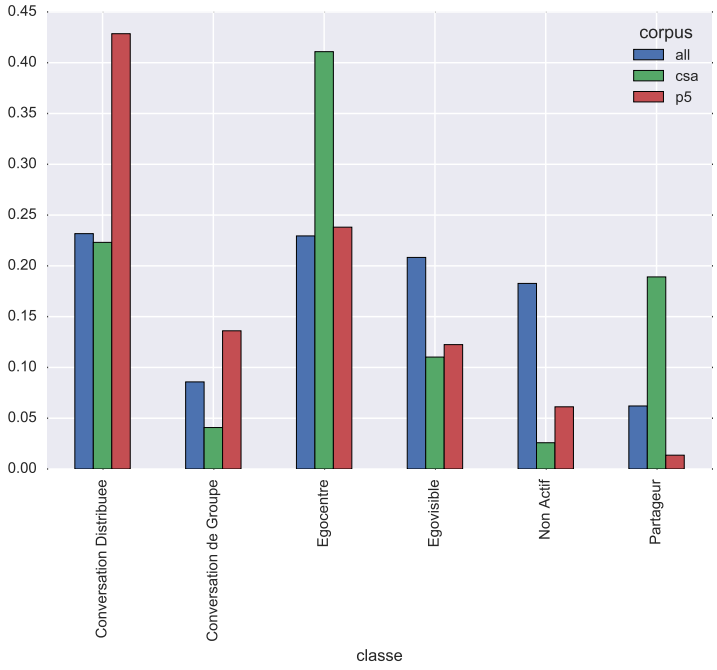


FIGURE 3.21: Distribution des classes de comportement des enquêtés en fonction de leur origine (Institut CSA en vert, Paris V en rouge, ou distribution moyenne en bleu)

De prime abord, le résultat n'est pas surprenant, l'échantillon « viral » d'enquêtés (qui ont installé l'application spontanément) est caractérisé, probablement de par les canaux de diffusion empruntés, par une sur-représentation de jeunes hommes, plutôt urbains et plutôt plus diplômés que la moyenne. Dès lors, si l'on se réfère aux histogrammes des figures 3.11 et 3.12, on ne sera pas surpris de constater que les partageurs, dont on a vu qu'ils étaient composés d'individus en moyenne plus âgés sont sur-représentés dans l'échantillon CSA. De la même façon, c'est au sein des enquêtés de Paris V que l'on retrouve la plus grande proportion d'individus relevant de la classe conversation distribuée dont on avait déjà repéré qu'elle était plus répandue chez les étudiants. Notre échantillon de départ est biaisé mais, grâce au formulaire que les enquêtés devaient remplir, nous avons toutes les données nécessaires pour comprendre la nature de ce biais.

Pour autant, une question demeure, à quel degré les classes que nous avons construites sont-elles sensibles à notre échantillon d'enquêtés. Nous avons fait appel à un simple kmeans pour construire ces classes. *A priori* l'algorithme identifie les k classes permettant de minimiser la distance de

chaque point au centroïde de sa classe. Mais on sait combien cet algorithme, comme bien d'autres (Ertöz et al., 2003) est sensible à l'hétérogénéité des tailles des catégories existantes. Ainsi en présence de bruit, des clusters de petites tailles peuvent être aisément agrégés à des clusters plus denses. Un échantillon d'usagers trop particulier et le risque est grand que certaines classes d'activité disparaissent de l'analyse [43].

43. C'est un cas un peu extrême car l'échantillon est de toute façon de taille très réduite. Mais il est par exemple très peu probable que la classe des partageurs puisse être retrouvée en analysant les seules données des étudiants de Paris V.

Nous n'avons pas mené l'expérience, mais il faudrait s'assurer qu'une clusterisation obtenue sur une sous-partie du corpus, même sélectionnée de façon biaisée (par exemple en considérant exclusivement des enquêtés CSA [44] ou en « redressant » notre échantillon) résulterait en une même partition. Il ne s'agit pas ici simplement de contrôler la sensibilité de la classification à la présence de quelques comptes exotiques à l'aide d'une procédure de bootstrap, mais bien d'évaluer la robustesse de nos résultats à l'aune d'autres logiques d'échantillonnage.

44. Naturellement, c'est plutôt l'échantillon viral qui est la version biaisée de l'échantillon CSA.

On le voit, même avec des données massives (plusieurs milliers de comptes) et avec un protocole très maîtrisé, les procédures statistiques sont plus indispensables que jamais. Les nombres parlent d'eux-mêmes, raison de plus pour les garder sous étroite surveillance si l'on veut pouvoir caractériser les milieux dans toute leur généralité au-delà des contingences d'une expérience. On réalise à travers cet exemple que les méthodes numériques sont toujours tributaires de questions classiques d'échantillonnage.

Pour donner une dernière illustration de ces difficultés, revenons sur la carte des partages déjà introduite au premier chapitre (figure 1.5). Compte tenu de la distribution singulière de l'échantillon d'enquêtés (individus plus jeunes que la moyenne, plus masculins, urbains et diplômés aussi), est-ce que notre tentative de cartographie des domaines du web à travers les partages de domaines serait entièrement à revoir ?

Nous avons simplement rajouté à l'aide d'une heatmap les sites que les individus de l'échantillon CSA mentionnaient préférentiellement sur la figure 3.22. Première remarque liminaire : on remarque sur cette carte les différents clusters périphériques qui correspondent à autant d'espaces francophones (l'application était uniquement disponible en français) étrangers : Belgique en bas à droite, Tunisie et Afrique en haut à gauche, Québec en haut à droite. Et pour chacun de ces clusters, on retrouve une grappe de sites régionaux (par exemple : *lapresse.ca*, *ledevoir.com*, *quebec.huffingtonpost.ca*). Même périphériques, la présence de ces sites nous interpelle quant à la véritable structure des ressources numériques françaises sur laquelle nous pensions enquêter. Du fait de ces « débordement » La division ne peut être purement thématique, elle est aussi en partie géographique.

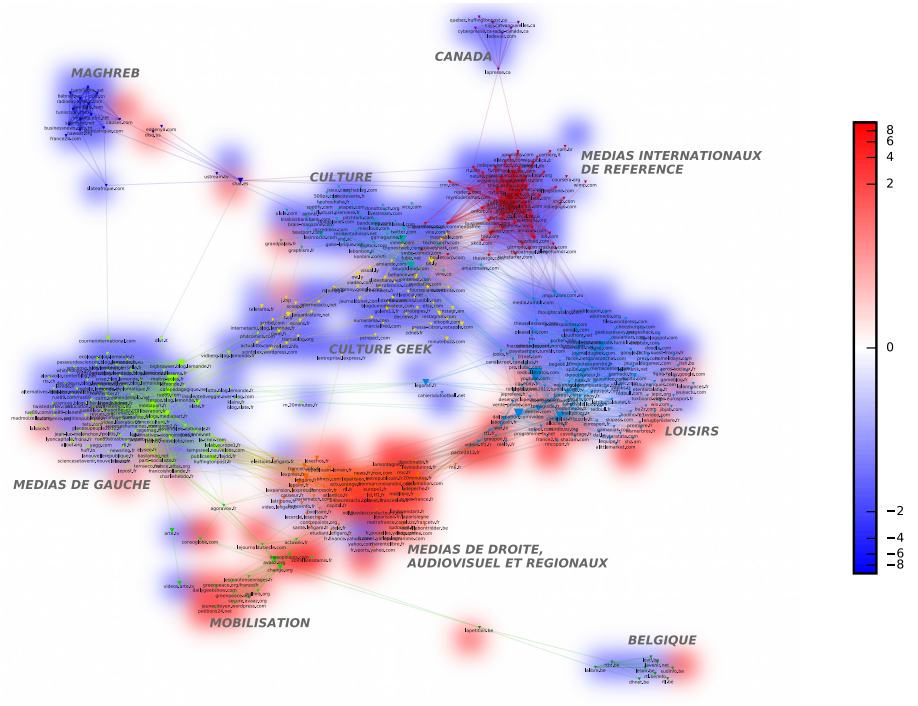


FIGURE 3.22: Heatmap des enquêtés CSA sur la carte des partages. Les sites recouverts par un halo rouge sont surreprésentés dans les comptes de l'échantillon CSA. La couleur bleue signale au contraire une situation de sous-représentation.

Par ailleurs, il apparaît très clairement que les enquêtés CSA publient de façon préférentielle à propos de ressources relevant des sites situés dans la partie inférieure de la carte qui se compose des univers des loisirs, des médias régionaux ou audiovisuels et de la mobilisation. Si le contraste est impressionnant visuellement, il faut préciser que tout est question d'échelle, le biais est bel et bien significatif, mais à cette échelle, tout signal est significatif, cela ne veut pas dire pour autant que son intensité soit importante.

On pourrait ainsi craindre que le fond de carte soit biaisé, ne rendant compte que des pratiques d'une portion congrue de la population. Si l'on se contente de comparer la distribution des fréquences des sites, cela ne semble pas vraiment être le cas. Ainsi, parmi les 600 domaines les plus populaires dans l'échantillon du CSA, les deux tiers figurent déjà sur la carte des partages. La figure 3.23 illustre cette relative stabilité. Elle montre la proportion de mentions reçues par les 20 domaines les plus cités dans chaque échantillon. Au total on retrouve seulement 25 domaines différents. À quelques détails près (*allocine* est apparemment plus souvent partagé par les étudiants), les distributions se ressemblent énormément. Dès lors on peut imaginer que la carte des partages des sites les plus cités par les enquêtés CSA aurait sans doute beaucoup ressemblé à la carte globale[45]. On pourrait imaginer que l'interprétation relationnelle du web (la façon dont les sites sont reliés) diffère en fonction des échantillons. C'est tout à fait possible mais assez improbable

45. La présence d'un blog dans le top 20 des sites les plus cités par les étudiant de Paris V est la conséquence de la taille très réduite de ce corpus, et ne peut pas vraiment être apparentée à un biais systématique.

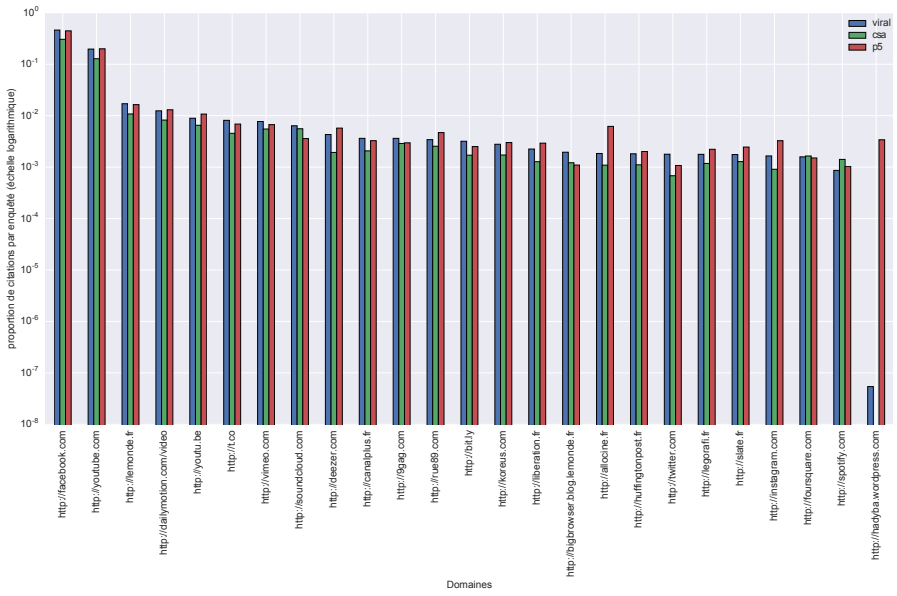


FIGURE 3.23: Distribution de la proportion moyenne de mentions de chaque domaine en fonction de l'origine des enquêtés (Institut CSA en vert, Paris V en rouge, ou échantillon viral en bleu). Un profil de citation est calculé pour chaque individu. Il est normalisé avant qu'un profil moyen soit calculé à l'échelle de la population.

vu la nature très claire des agrégats observés. Mais à nouveau, il s'agit de pures spéculations qui pour être entièrement clarifiées nécessiteraient sans doute un travail approfondi de définition d'un protocole de comparaison statistique entre cartes qui permette de dire de deux cartes qu'elles sont semblables au bruit près.

On peut également penser les questions d'échantillonnage de corpus en raisonnant leur contenu même. Les archives numériques qui se multiplient (on pense ici plus précisément aux archives d'articles de presse non nativement numériques qui peuvent couvrir des longues périodes historiques) ouvrent de très riches perspectives pour étudier les dynamiques culturelles ou l'évolution des problèmes publics avec la garantie de travailler sur un matériau d'enquête contemporain du problème dont il est question.

Souhaitant mener un tel travail sur le processus de désinstitutionalisation avec Alix Rule, nous nous sommes rapidement heurtés à un problème de taille. Pratiquement, Alix Rule s'intéresse dans sa thèse au mouvement de désinstitutionalisation psychiatrique aux Etats-Unis. Impulsé après la guerre suite aux protestations contre les traitements jugés inhumains à l'encontre des internés, il a entraîné une diminution drastique de la population des hôpitaux d'État en quelques décennies. Mais alors que la responsabilité de la prise en charge des malades mentaux avait déjà été transférée à l'échelle fédérale, ceux-là mêmes qui avaient plaidé pour la fermeture des hôpitaux d'État fustigeaient maintenant l'abandon des malades mentaux à leur propre sort. On le voit la chronologie des débats est particulièrement riche de renversements de

perspectives.

Pour suivre une telle discussion dans la presse, la pratique courante consiste à chercher les articles mentionnant un certain nombre de termes clés : *mental institution*, *involuntary commitment*, *psychosurgery*, *state department of mental hygiene*, etc. Mais on voit rapidement combien des références aussi explicites risquent de nous faire passer à côté de la nature éminemment dynamique de la discussion. Comment attraper un flux avec de simples requêtes ? A ce problème principal vient naturellement s'ajouter le fait que la référence explicite à un objet du monde dans un article de presse n'est pas nécessairement un bon indicateur de la position que cet objet occupe. À titre d'exemple, la figure 3.24 contraste l'évolution du nombre de lits disponibles dans les institutions psychiatrique non fédérales depuis 1940 (données tirées de (Dowdall, 1996)) et le nombre de mention du terme *mental hospital* dans le New York Times durant cette même période. La résurgence d'articles dans les années 80 semble pouvoir être expliquée par un double phénomène. D'une part, la diminution du financement des dispositifs d'aide social rend la vulnérabilité des malades mentaux sortis du système encore plus criante. D'autre part, de nombreux articles sont aussi plus « contingents », liés qu'ils sont à la tentative d'assassinat de Ronald Reagan le 30 mars 1981 par John Hinckley qui souhaitait ainsi impressionner Jodie Forster.

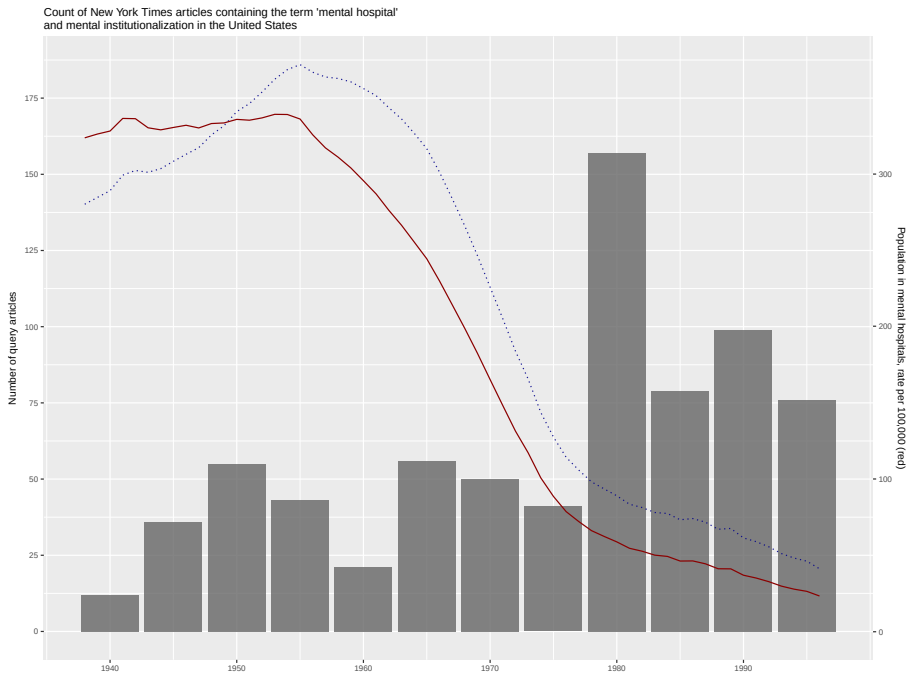


FIGURE 3.24: Évolution du nombre d'articles mentionnant le terme *mental hospital* dans notre corpus de résumés d'articles (ledes) du New York Times [barres]. Les deux courbes rouges (continue) et bleu (en pointillés) mesurent la diminution massive du nombre d'hôpitaux et de lits occupées à partir des années 50.

La construction de corpus *via* une opération de sélection (par opposition

aux corpus naturels, dont les discours de l'État de l'Union sont un exemple) est une pratique évidemment extrêmement répandue qu'il s'agisse de corpus médiatiques comme dans le travail de DiMaggio et al. (2013) sur le financement public des Arts, d'études scientométriques (dont les nôtres qu'elles concernent la biologie de synthèse (Raimbault et al., 2016), les services écosystémiques (Tancoigne et al., 2014), ou les réseaux d'expression génétiques (Cointet et al., 2012b)) ou même de corpus de tweets censés indexer un mouvement politique (Occupy) (Conover et al., 2013). Pour autant la question de la délinéation du corpus reste toujours secondaire. Certains chercheurs ont néanmoins identifié ce problème. King et al. (2017) commentent de la manière suivante la pratique traditionnelle de la construction de corpus par mots-clés :

> « *Although all substantive results depend on this choice, researchers usually pick keywords in ad-hoc ways that are far from optimal and usually biased.* »[46]

46. « Bien que tous les résultats substantifs en dépendent, les chercheurs choisissent généralement leur mot-clé en suivant des intuitions ad-hoc qui sont loin d'être optimales et sont probablement biaisées. »

Pour autant la solution des auteurs consiste à accompagner l'analyste dans la construction d'une requête de plus en plus complexe grâce à un outil automatique de recommandation de mots-clés. Cette solution ne fait, selon nous, que déplacer le problème.

Notre travail actuel propose de construire des corpus en utilisant un critère purement géométrique. L'idée est simple. Plutôt que de s'évertuer à sélectionner un texte en fonction de la présence (voire l'absence pour les requêtes les plus évoluées) de tel ou tel mot, le corpus est défini par sa position dans un espace sémantique pré-calculé. Pratiquement, nous faisons naturellement appel aux modèles de plongements mixtes de mots et de paragraphes (Le et Mikolov, 2014). Une première requête simple est proposée, on est déduit un centre en calculant la position moyenne de l'ensemble des articles mentionnant le terme. Le corpus final est obtenu en agrégeant tous les articles se trouvant à proximité de ce point.

Il n'est pas ici question de montrer l'intégralité des résultats ni même de démontrer la validité de notre méthode mais d'en illustrer les propriétés à travers un exemple. Imaginons que l'on souhaite définir le corpus correspondant au traitement médiatique d'Internet en 2000. À partir d'une base de données composée de plus de 190 000 paragraphes publiés dans le quotidien cette année, on entraîne un modèle sémantique dans lequel les paragraphes et les mots sont conjointement plongés. 3 079 paragraphes mentionnent explicitement le terme *Internet*. Une fois le centre de tous ces paragraphes calculé, la mesure de la similarité entre un article et ce point de l'espace permet aisément de se définir un corpus qui indexe sémantiquement le concept qu'*Internet* recouvrait à l'époque. En comparaison avec la méthode classique, on parvient ainsi très facilement à identifier les faux-positifs et des vrais-négatifs. Lorsque le terme *Internet* est mentionné de façon secondaire au sein d'un paragraphe parlant en réalité d'un tout autre sujet comme de football, le paragraphe se retrouve à très longue distance du centre du corpus :

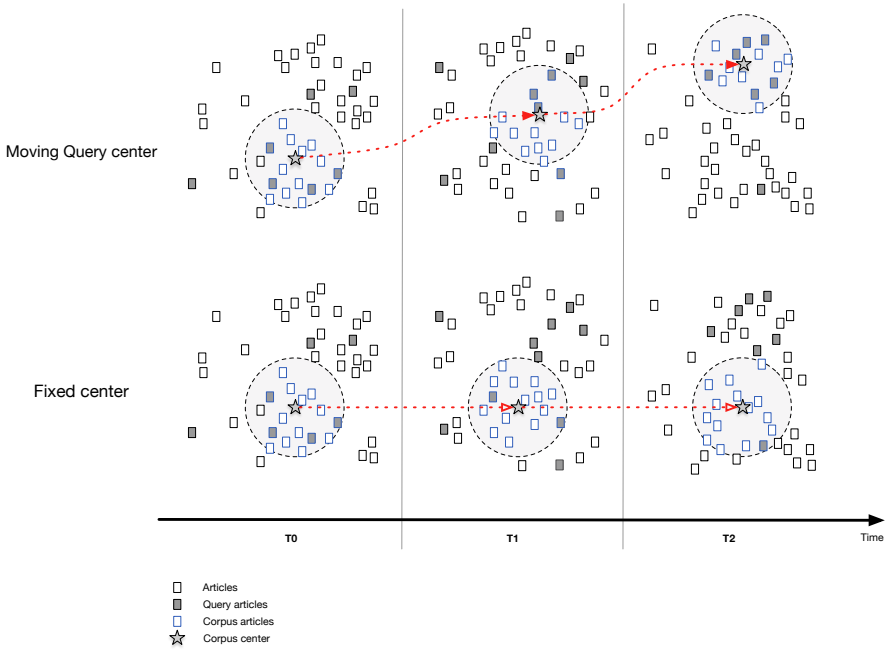


FIGURE 3.25: Les deux stratégies à centre mobile (haut) ou à centre fixe (bas) sont schématisés. Chaque rectangle correspond à un article dans l'espace sémantique. Les corpus sont construits en agrégeant les articles inclus dans la sphère qui entoure le centre à chaque période.

> « *With Nicolas Anelka now injured for two months (video footage of his operation available on the internet at http ://www.marca.es/ he's found plenty of time to maintain his own personal web-site, www.nicolasanelka.net.Free from the 'interfering medium' of press interpretation, Nicolas speaks directly to his many fans about why he left Arsenal. 'The Arsenal players really drove me crazy. They would never let me play my French rap music, which they ought to have done because it is so good.' Nicolas also explains why he took and missed the penalty that cost Real Madrid victory over Corinthians - it was his agent-brother Didier's idea : 'Besides, Hierro had missed a couple and Savio, next in line, was hesitating. So I had a go. Anyway, we drew 2-2 and both goals were mine'* » David Hills, 16 Janvier 2000, *Said & Done* [47]

En revanche, un paragraphe dont le sujet porte effectivement sur le « cyberespace » sans mentionner une seule fois le mot *Internet* et qui aurait été exclu dans un corpus « classique » à base de requête est manifestement très proche du centre calculé :

> « *Shopping phones Many of the world's big electronic retailers believe that 3G phones will lead to a massive increase in online shopping.* » Ashley Norris, 20 Avril 2000, *This is the future calling* [48]

L'autre avantage décisif d'une approche par plongement est la possibilité qui nous est offerte de construire des corpus dynamiques qui prennent en compte les évolutions sémantiques structurelles de l'espace tout en garantissant la continuité d'un point de vue. C'est ce que la figure 3.25 illustre. Ainsi dans notre modèle dit à « centre fixe », le centre du corpus est estimé à la période où la conversation atteint son pic d'intensité (au moment aussi où, on en fait l'hypothèse, la densité d'articles autour du centre est la plus forte), puis il est fixé et est transporté de période en période par le truchement

47. Similarité au centre de −0.18 (la mesure de cosine s'étend de -1 à 1), ce qui fait de ce paragraphe le faux positif le plus lointain parmi les paragraphes mentionnant *Internet*.

48. Similarité au centre de 0.72, ce qui fait de ce paragraphe le 45ème paragraphe le plus proche du centre du corpus, soit le vrai négatif le plus « vrai ».

des opérations d'alignement inter-temporel déjà décrites. Cette stratégie de délimitation permet de se poser des questions du type : quels discussions actuelles occupent aujourd'hui la place prise jadis par les discussion sur les asiles psychiatrique d'État ? La méthode dite à centre mobile suppose de déplacer le centre au gré des nouveaux usages des termes de la requête : suivre la dérive des discussions qui l'entourent jusqu'à épuisement ou transformation.

Selon les hypothèses retenues et en faisant varier les stratégies (à centre fixe ou mobile) à partir du point de départ qui consiste en la discussion sur les *mental hospital* dans les années 1955-1960, la dynamique discursive peut prendre différentes directions que l'on décrit qualitativement mais qui pourrait naturellement faire l'objet d'une analyse automatique de corpus. Dans le cas de la stratégie à centre fixe, on observe que la discussion concerne dans un premier temps (1965-1980) des scandales sanitaires à l'encontre de malades mentaux (négligence, abus) avant de se concentrer plus spécifiquement sur le rôle que jouent les tribunaux pour déterminer et défendre leurs droits. La dimension juridique persiste dans les années 90, portant de plus en plus sur la justice criminelle (en lien ou non avec les questions de santé mentale). La stratégie à centre mobile regroupe les interventions publiques en matière de santé. Le corpus se tarit après la conversation sur les actions publiques contre le SIDA dans les années 80. Autre exemple, appliqué à la requête *atomic* ou *nuclear* dans les années 40, le corpus finit par « découvrir » de lui même la course à l'Espace.

Il est encore un peu tôt pour détailler plus avant les résultats et poursuivre l'exposé de ces méthodes encore très expérimentales. Ce dernier exemple montre cependant très bien combien les formalismes les plus abstraits et en apparence lointains (comme les réseaux de neurones des plongements de mots) peuvent tout à fait se mettre au service de questions classiques (même si peu investies) mais essentielles pour l'enquête sociologique.

Conclusion

La division traditionnelle entre les deux cultures de Snow (1959) - si elle a jamais été justifiée - n'a jamais semblé si inappropriée pour décrire la situation présente. En effet, le schéma classique qui sépare les « humanités » supposées explorer le monde social avec des outils purement qualitatifs d'une part et les « scientifiques » essayant de valider des hypothèses hautement formalisées à l'aide d'outils quantitatifs d'autre part, offrre une description bien réductrice des pratiques de recherche actuelle. Pour autant il ne s'agit pas de céder à l'angélisme car de nombreuses tensions traversent le champ de l'analyse textuelle en sciences sociales.

On peut ainsi entendre l'inquiétude de McFarland, Lewis, et Goldberg (2015) voyant dans la collision des champs de l'intelligence artificielle, des sciences sociales et de l'industrie des médias sociaux une source d'incertitudes pour la sociologie dont l'épistémologie est menacée par l'obsession prédictive des big data. Les auteurs développent la métaphore coloniale[49] pour illustrer la situation actuelle craignant que les collaborations entre disciplines ne prennent un tournant purement instrumental. Il est certain que les propos parfois outranciers de figures comme Watts, ou Anderson, peuvent alimenter ces craintes. Mais ces prises de position sont aussi l'occasion pour les sciences sociales de clarifier leur position par rapport aux nouvelles formes que prend le social et la nécessaire mise à jour de leur méthodologie d'enquête. Sans doute parce qu'elles avaient suscité un très forte attente (Lazer et al., 2009), les promesses ouvertes par l'étude des traces numériques ont déçu (Boyd et Crawford, 2011). Est-ce une raison pour les renvoyer aux oubliettes ? Il est probable qu'informaticiens et physiciens continueront d'occuper le terrain des études numériques tant que les sciences sociales feront preuve de frilosité.

49. « the colonization of sociology (and social science more generally) by computer science perspectives and practices »

Naturellement, sur ces questions, la sociologie est traversée de courants pluriels et éminemment contradictoires. Quand certains crient au loup (Savage et Burrows, 2007) d'autres conçoivent les traces numériques comme une formidable opportunité pour nous mouvoir dans des espaces à la dimensionnalité étrange (Latour et al., 2012), et enfin débarrasser la sociologie de

concepts encombrants (Latour, 2014). Alors que certains défendent une ligne interprétative dure et décrète impossible toute activité de codage automatique d'un texte (Biernacki, 2012), d'autres sont près à faire table rase du passé et voudraient ranger l'entreprise sociologique sous la courbe ROC de l'apprentissage automatique (Hofman et al., 2017), ou réduire la culture à des séries temporelles (Michel et al., 2011).

« Numbers, numbers, numbers » Latour (2012) avec de nombreux auteurs (Desrosières, 1985; Porter, 1995; Hacking, 2006) s'accordent sur le rôle croissant des approches quantitatives en sciences sociales, qui déplacent sans cesse l'équilibre entre approches hypothético-déductives et approches exploratoires plus inductives. Pour autant, il n'est pas si clair que l'usage de méthodes issues de l'informatique comme les topic models ou même les plongements de mots condamne la sociologie à aligner son épistémologie avec celle des sciences naturelles. Ainsi de nombreux sociologues témoignent au contraire de la nature libératrice de ces méthodes dans leur pratique (DiMaggio, 2015; Goldberg, 2015) pourvu que le raisonnement sociologique puisse être construit main dans la main avec les méthodes quantitatives dans un mouvement abductif (Timmermans et Tavory, 2012).

Après ce long panorama des méthodes d'analyse automatiques de texte courant sur les 50 dernières années, on ne peut être que surpris de la très grande diversité de ces approches : modélisation plus ou moins complexe de l'énonciation, modèle géométrique ou topologique, à liste ou graphique, logique hypothético déductive ou fortement interprétative, etc. Il semble exister autant d'approches que de façons de faire de la sociologie et les différentes ères des sciences sociales que décrit Boullier (2015) semblent toutes co-exister. Mieux, de leur hybridation naissent de nouvelles questions et des dispositifs d'enquête originaux. Ainsi, comme on l'a vu au chapitre précédent, les questions d'échantillonnage (caractéristiques de l'ère des sondage), que certains avaient déjà voué aux gémonies, semblent plus d'actualité que jamais dans un espace public numérique où les identités sont si fluides et insaisissables. Les statistiques, que l'on avait sans doute enterrées un peu vite, s'avéreront sans aucun doute précieuses pour apprécier la robustesse et la sensibilité des modèles. Symétriquement, on a aussi montré que les approches les plus récentes en apprentissage automatique sont aussi susceptibles de nous aider à formaliser sous un jour nouveau des problèmes de sciences sociales extrêmement classiques comme la question de la délimitation d'un corpus.

Projets principaux

Dans cette annexe, on rappelle brièvement, la nature des six corpus de données les plus discutés dans ce mémoire de synthèse. Certains s'appuient sur des données historiques (les discours sur l'État de l'Union, articles du New York Times) ou contemporaines (Biologie de Synthèse, données Facebook). Certains sont nativement numériques pour reprendre l'expression chère à Richard Rogers (interactions sur Facebook, ou commentaires en ligne sur les plateformes du LA Times ou de la Voix du Nord) ou ont bénéficié d'un processus de numérisation (discours politiques, compte-rendus de négociation climatiques). Enfin, ces travaux renvoient à des champs de connaissance assez divers : les sciences politiques (et la sociologie des médias) lorsque la nature des publics dans les espace en ligne est interrogée (Voix du Nord, LA Times), les science studies quand on s'interroge sur le travail de construction d'un nouveau champ scientifique (Biologie de Synthèse), la sociologie des usages lorsqu'on décrit à partir de traces égo-centrées l'activité d'un internaute sur une plateforme web (Algopol), la socio-histoire quand des processus historiques dynamiques sont décortiqués au travers d'une série de discours qui traverse les siècles (État de l'Union, etc.

Le compte rendu factuel des traces récoltées montre que, dans quasiment l'intégralité des projets, les questions de recherche naissent de la confrontation des données d'une source avec d'autres données complémentaires obtenues auprès d'autres sources. Le tableau récapitulatif ci-dessous (3) tente de rendre compte de ces opérations essentielles. Pour chaque projet, la nature du corpus est indiquée, mais aussi la stratégie de collecte des données adoptée. Les stratégies de modélisation des actes d'énonciation propres à chaque projet et ajustées au matériau et à la question de recherche sont également indiquées. Enfin, on essaye de remobiliser les grandes catégories d'analyse définies à la fin du premier chapitre pour décrire les résultats dans chaque projet.

TABLE 3: Typologie des méthodes d'analyse de corpus en sciences sociales

Projet	**Nature du corpus**	**Stratégie de collecte**	**Informations Complémentaires**	**Modélisation des énoncés**	**Type d'analyse**	**Réalisations**
LA Times Homicide Report	commentaires publics (web)	Scraping site web	Morphologie du conté de Los Angeles, couverture de la criminalité dans la version papier du journal	Pré-codage suivant un schéma actantiel qui articule locuteur / victime / reponsable / juge	contrastive, corrélation entre type d'homicides, quartiers et type de locuteurs	article en cours de rédaction
Voix du Nord	forum (web)	Scraping site web	Morphologie du Nord-Pas-de-Calais	Catégories pré-codées inspirée de la sociologie de la critique, codage du locuteur *via* son alias	contrastive, corrélation entre taille de commune et rôle des forums	(Parasie et Cointet, 2012)
Biologie de Synthèse	Articles scientifiques	Corpus numérisé (plateforme scientifique)	Enquête biographique sur membres du core-set	extraction automatique de groupes nominaux pertinents	Cartographie sémantique et co-citationnelle, et le leur corrélation	(Raimbault et al., 2016)
État de l'Union	Discours politiques	corpus déjà numérisée	—	extraction lexicale automatique	Cartographie sémantique temporelle	(Rule et al., 2015)
Algopol	Traces des comportements d'utilisateurs sur Facebook	Application dédiée, API Facebook	Questionnaire pour obtenir données socio-démographiques	traces composées d'actions de partages, commentaires, clics, etc.	Carte des partages, détection de sous-populations d'usagers	(Bastard et al., 2017)
COP	Compte-rendus de discussions diplomatiques	Scraping depuisCorpus numérisé	Guidelines de rédaction des rapports	extraction lexicale automatique	Cartographie sémantique, calcul de modalités	(Venturini et al., 2014a; Baya-Laffite et Cointet, 2014)

Bibliographie

Abello, J., Resende, M. G., Sudarsky, S., 2002. Massive quasi-clique detection. In : Latin American Symposium on Theoretical Informatics. Springer, pp. 598–612.

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R., 2011. Sentiment analysis of twitter data. In : Proceedings of the workshop on languages in social media. Association for Computational Linguistics, pp. 30–38.

Anderson, C., 2008. The end of theory : The data deluge makes the scientific method obsolete. Wired 16 (07).

Armatte, M., 2008. Histoire et préhistoire de l'analyse des données par jp benzecri : un cas de généalogie rétrospective1. Journal Electronique des Probabilités et de la Statistique 4 (2).

Arora, S., Li, Y., Liang, Y., Ma, T., Risteski, A., 2016. Linear algebraic structure of word senses, with applications to polysemy. arXiv preprint arXiv :1601.03764.

Bakshy, E., Messing, S., Adamic, L. A., 2015. Exposure to ideologically diverse news and opinion on facebook. Science 348 (6239), 1130–1132.

Barabási, A.-L., 2016. Network science. Cambridge University Press.

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., Bonneau, R., 2015. Tweeting from left to right : Is online political communication more than an echo chamber ? Psychological science 26 (10), 1531–1542.

Bastard, I., Cardon, D., Charbey, R., Cointet, J.-P., Christophe, P., 2017. Facebook, pour quoi faire ? configurations d'activités et structures relationnelles. Sociologie.

Baya-Laffite, N., Cointet, J.-P., 2014. Cartographier la trajectoire de l'adaptation dans l'espace des négociations sur le climat. Réseaux (6), 159–198.

Bearman, P., 2015. Big data and historical social science. Big Data & Society 2 (2), 2053951715612497.

Bearman, P. S., Stovel, K., 2000. Becoming a nazi : A model for narrative networks. Poetics 27 (2-3), 69–90.

Beaudoin, V., 2016. Statistical analysis of textual data : Benzécri and the french school of data analysis. In : Léon, J., Loiseau, S. (Eds.), History of Quantitative Linguistics in France. RAM-Verlag, Stüttinghauser Ringstr. 44 D-58515 Lüdenscheid Germany, pp. 173–193.

Behrisch, M., Bach, B., Henry Riche, N., Schreck, T., Fekete, J.-D., 2016. Matrix reordering methods for table and network visualization. In : Computer Graphics Forum. Vol. 35. Wiley Online Library, pp. 693–716.

Benford, R. D., Snow, D. A., 2000. Framing processes and social movements : An overview and assessment. Annual review of sociology 26 (1), 611–639.

Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C., 2003. A neural probabilistic language model. Journal of machine learning research 3 (Feb), 1137–1155.

Benzécri, J.-P., 1973. L'analyse des données. Vol. 2. Dunod Paris.

Benzécri, J.-P., 1976. Histoire et préhistoire de l'analyse des données. Les cahiers de l'analyse des données 1 (1), 9–32.

Benzécri, J.-P., 1981. Pratique de l'analyse des données. Dunod.

Berelson, B., 1952. Content analysis in communications research.

Berelson, B., Lazarsfeld, P. F., 1948. The analysis of communication content. np.

Bergström, M., 2011. La toile des sites de rencontres en france. Réseaux (2), 225–260.

Bettencourt, L., Kaiser, D. I., 2015. Formation of scientific fields as a universal topological transition. arXiv preprint arXiv :1504.00319.

Beuscart, J.-S., 2014. Des données du web pour faire de la sociologie… du web ? In : Big Data, entreprises et sciences sociales, chaire de sociologie du travail créateur.

Biernacki, R., 2012. Reinventing evidence in social inquiry : Decoding facts and variables. Springer.

Blanco, R., Lioma, C., 2012. Graph-based term weighting for information retrieval. Information retrieval 15 (1), 54–92.

Blei, D. M., 2012. Probabilistic topic models. Communications of the ACM 55 (4), 77–84.

Blei, D. M., Lafferty, J. D., 2006. Dynamic topic models. In : Proceedings of the 23rd international conference on Machine learning. ACM, pp. 113–120.

Blei, D. M., Lafferty, J. D., 2007. A correlated topic model of science. The Annals of Applied Statistics, 17–35.

Blei, D. M., Ng, A. Y., Jordan, M. I., 2003. Latent dirichlet allocation. Journal of machine Learning research 3 (Jan), 993–1022.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. Journal of statistical mechanics : theory and experiment 2008 (10), P10008.

Boellstorff, T., 2013. Making big data, in theory. First Monday 18 (10).

Bollen, J., Mao, H., Zeng, X., 2011. Twitter mood predicts the stock market. Journal of computational science 2 (1), 1–8.

Boltanski, L., Schiltz, M.-A., Darré, Y., 1984. La dénonciation. Actes de la recherche en sciences sociales 51 (1).

Bonin, E., Dallo, A., 2003. Hyperbase et lexico 3, outils lexicométriques pour l'historien. Histoire & mesure 18 (3), 389–402.

Bonnafous, S., 1983. Le congrès de metz (1979) du parti socialiste : Processus discursifs et structures lexicales à travers les motions mitterrand, rocard et ceres. Langages (71), 3–123.

Bonnafous, S., Tournier, M., 1995. Analyse du discours, lexicométrie, communication et politique. Langages, 67–81.

Bosco, C., Patti, V., Bolioli, A., 2013. Developing corpora for sentiment analysis : The case of irony and senti-tut. IEEE Intelligent Systems 28 (2), 55–63.

Boudin, F., 2013. A comparison of centrality measures for graph-based keyphrase extraction. In : International Joint Conference on Natural Language Processing (IJCNLP). pp. 834–838.

Boullier, D., 2015. Les sciences sociales face aux traces du big data. Revue française de science politique 65 (5), 805–828.

Boullier, D., Lohard, A., 2012. Opinion mining et Sentiment analysis : Méthodes et outils. OpenEdition Press.

Bourdieu, P., 1979. La distinction. critique sociale du jugement. Paris, éditions de Minuit.

Bourdieu, P., 2000. Les structures sociales de l'économie. Paris, Seuil.

Bowker, G. C., 2014. Big data, big questions| the theory/data thing. International Journal of Communication 8, 5.

Boyack, K. W., Klavans, R., 2010. Co-citation analysis, bibliographic coupling, and direct citation : Which citation approach represents the research front most accurately ? Journal of the American Society for Information Science and Technology 61 (12), 2389–2404.

Boyd, D., 2006. Friends, friendsters, and top 8 : Writing community into being on social network sites [electronic version]. First Monday 11.

Boyd, D., Crawford, K., 2011. Six provocations for big data. In : A decade in internet time : Symposium on the dynamics of the internet and society. Vol. 21. Oxford Internet Institute Oxford.

Boyd-Graber, J., Mimno, D., Newman, D., 2014. Care and feeding of topic models. In : Airoldi, E. M., Blei, D., Erosheva, E. A., Fienberg, S. E. (Eds.), Handbook of Mixed Membership Models and Their Applications. Chapman and Hall/CRC, Ch. 12, pp. 225–255.

Bozdag, E., 2013. Bias in algorithmic filtering and personalization. Ethics and Information Technology 15 (3), 209–227.

Breiger, R. L., Boorman, S. A., Arabie, P., 1975. An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. Journal of mathematical psychology 12 (3), 328–383.

Brinton, L. J., 2000. The structure of modern English : A linguistic introduction. John Benjamins Publishing.

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., Lai, J. C., 1992. Class-based n-gram models of natural language. Computational linguistics 18 (4), 467–479.

Brunet, É., 2006. Navigation dans les rafales. les Actes des 8èmes JADT, 15–29.

Brunet, E., 2009. Comptes d'auteurs. Champion.

Bruns, A., 2007. Methodologies for mapping the political blogosphere : An exploration using the issuecrawler research tool. First Monday 12 (5).

Bruns, A., Burgess, J. E., 2011. The use of twitter hashtags in the formation of ad hoc publics. In : Proceedings of the 6th European Consortium for Political Research (ECPR) General Conference 2011.

Bryant, S. L., Forte, A., Bruckman, A., 2005. Becoming wikipedian : transformation of participation in a collaborative online encyclopedia. In : Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work. ACM, pp. 1–10.

Bucher, T., 2012. Want to be on the top ? algorithmic power and the threat of invisibility on facebook. new media & society 14 (7), 1164–1180.

Burke, K., 1969. A grammar of motives. Univ of California Press.

Burrows, R., Savage, M., 2014. After the crisis ? big data and the methodological challenges of empirical sociology. Big Data & Society 1 (1).

Butler, B., Joyce, E., Pike, J., 2008. Don't look now, but we've created a bureaucracy : the nature and roles of policies and rules in wikipedia. In : Proceedings of the SIGCHI conference on human factors in computing systems. ACM, pp. 1101–1110.

Cagé, J., Hervé, N., Viaud, M.-L., et al., 2017. L'information à tout prix. INA.

Caliskan-Islam, A., Bryson, J. J., Narayanan, A., 2016. Semantics derived automatically from language corpora necessarily contain human biases. arXiv preprint arXiv :1608.07187.

Callon, M., Courtial, J. P., Laville, F., 1991. Co-word analysis as a tool for describing the network of interactions between basic and technological research : The case of polymer chemistry. Scientometrics 22 (1), 155–205.

Callon, M., Courtial, J.-P., Turner, W. A., Bauin, S., 1983. From translations to problematic networks : An introduction to co-word analysis. Information (International Social Science Council) 22 (2), 191–235.

Callon, M., Latour, B., 2006. Le grand léviathan s' apprivoise-t-il. Sociologie de la traduction. Textes fondateurs, 11–33.

Callon, M., Rip, A., Law, J., 1986. Mapping the dynamics of science and technology : Sociology of science in the real world. Springer.

Calude, C. S., Longo, G., 2016. The deluge of spurious correlations in big data. Foundations of science, 1–18.

Campello, R. J., Moulavi, D., Sander, J., 2013. Density-based clustering based on hierarchical density estimates. In : Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, pp. 160–172.

Cardon, D., 2008. Le design de la visibilité. Réseaux (6), 93–137.

Cardon, D., 2013. Dans l'esprit du pagerank. Réseaux (1), 63–95.

Cardon, D., 2015a. L'espace public élargi. opinion, critique et expressivité à l'ère d'internet. Ph.D. thesis, Paris Est.

Cardon, D., 2015b. A quoi rêvent les algorithmes. Nos vies à l'heure : Nos vies à l'heure des big data. Le Seuil.

Chateauraynaud, F., 2003a. Marlowe. vers un générateur d'expériences de pensée sur des dossiers complexes. Bulletin de méthodologie sociologique. Bulletin of sociological methodology (79), 6–32.

Chateauraynaud, F., 2003b. Prospéro : une technologie littéraire pour les sciences humaines. CNRS.

Chateauraynaud, F., 2011. Argumenter dans un champ de forces. Essai de balistique.

Chateauraynaud, F., 2013. La radicalité est-elle soluble dans l'argumentation ? la sociologie des controverses et l'endogénéisation de la critique sociale. In : Pourquoi la controverse ? Définitions, enjeux et méthodes. Université de Liège.

Chateauraynaud, F., Debaz, J., 2011. Processus d'alerte et dispositifs d'expertise dans les dossiers sanitaires et environnementaux.

Chateauraynaud, F., Debaz, J., Charriau, J., Marlowe, C., 2013. Une pragmatique des alertes et des controverses en appui à l'évaluation publique des risques. Paris : Anses.

Chavalarias, D., 2004. Métadynamiques en cognition sociale. Ph.D. thesis, Université Libre de Bruxelles.

Chavalarias, D., Cointet, J.-P., 2013. Phylomemetic patterns in science evolution—the rise and fall of scientific fields. PloS one 8 (2), e54847.

Chavalarias, D., Cointet, J.-P., Cornilleau, L., Duong, T. K., Mogoutov, A., Villard, L., Roth, C., Savy, T., 2011. Streams of media issues, monitoring world food security. Tech. rep., Presented at the United Nations.

Chavalarias, D., Wallach, J. D., Li, A. H. T., Ioannidis, J. P., 2016. Evolution of reporting p values in the biomedical literature, 1990-2015. Jama 315 (11), 1141–1148.

Chen, C., 2006. Citespace ii : Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for information Science and Technology 57 (3), 359–377.

Chen, C., Ibekwe-SanJuan, F., Hou, J., 2010. The structure and dynamics of cocitation clusters : A multiple-perspective cocitation analysis. Journal of the American Society for Information Science and Technology 61 (7), 1386–1409.

Chen, D., Manning, C. D., 2014. A fast and accurate dependency parser using neural networks. In : Emnlp. pp. 740–750.

Chen, Y., Keogh, E., Hu, B., Begum, N., Bagnall, A., Mueen, A., Batista, G., July 2015. The ucr time series classification archive. www.cs.ucr.edu/~eamonn/time_series_data/.

Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J. M., Leskovec, J., 2014. Can cascades be predicted ? In : Proceedings of the 23rd international conference on World wide web. ACM, pp. 925–936.

Cho, J. Y., Lee, E.-H., 2014. Reducing confusion about grounded theory and qualitative content analysis : Similarities and differences. The Qualitative Report 19 (32), 1.

Chuang, J., Ramage, D., Manning, C., Heer, J., 2012. Interpretation and trust : Designing model-driven visualizations for text analysis. In : Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 443–452.

Church, K., Gale, W., Hanks, P., Hindle, D., Moon, R., 1994. Lexical substitutability. bts atkins and a. zampolli (eds.). computational approaches to the lexicon : 153-177.

Cilibrasi, R. L., Vitanyi, P. M., 2007. The google similarity distance. IEEE Transactions on knowledge and data engineering 19 (3).

Cointet, J.-P., Gallinari, P., 2017. Réseaux sociaux et émergence d'opinions et de mobilisation. In : Les Big Data à découvert. CNRS Editions, Ch. Web, réseaux sociaux et recherche d'information, pp. 140–141.

Cointet, J.-P., Mogoutov, A., Bourret, P., El Abed, R., Cambrosio, A., 2012a. Les réseaux de l'expression génique-émergence et développement d'un domaine clé de la génomique. médecine/sciences 28, 7–13.

Cointet, J.-P., Mogoutov, A., Bourret, P., R, E.-A., A, C., 2012b. Les réseaux de l'expression génique : émergence et développement d'un domaine clé de la génomique. Médecine/Sciences.

Coleman, J., Katz, E., Menzel, H., 1957. The diffusion of an innovation among physicians. Sociometry 20 (4), 253–270.

Conant, J., 1998. Wittgenstein on meaning and use. Philosophical Investigations 21 (3), 222–250.

Conein, B., 2004. Cognition distribuée, groupe social et technologie cognitive. Réseaux (2), 53–79.

Conover, M. D., Ferrara, E., Menczer, F., Flammini, A., 2013. The digital evolution of occupy wall street. PloS one 8 (5), e64679.

Coulter, N., Monarch, I., Konda, S., 1998. Software engineering as seen through its research literature : A study in co-word analysis. Journal of the Association for Information Science and Technology 49 (13), 1206–1223.

Dalud-Vincent, M., 2010. Les «choix» du sociologue avec alceste-du paramétrage des unités de contexte aux résultats obtenus. Bulletin de méthodologie sociologique 107 (1), 23–48.

Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., Kleinberg, J., 2012. Echoes of power : Language effects and power differences in social interaction. In : Proceedings of the 21st international conference on World Wide Web. ACM, pp. 699–708.

Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., Potts, C., 2013. No country for old members : User lifecycle and linguistic change in online communities. In : Proceedings of the 22nd international conference on World Wide Web. ACM, pp. 307–318.

Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., Livingston, B., et al., 2010. The youtube video recommendation system. In : Proceedings of the fourth ACM conference on Recommender systems. ACM, pp. 293–296.

De Domenico, M., Lancichinetti, A., Arenas, A., Rosvall, M., 2015. Identifying modular flows on multilayer networks reveals highly overlapping organization in interconnected systems. Physical Review X 5 (1), 011027.

de Nooy, W., 2015. Structure from interaction events. Big Data & Society 2 (2), 2053951715603732.

Debaz, J., 2013. Turbulences épistémiques et perturbateurs endocriniens# 2 les effets cocktails.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R., 1990. Indexing by latent semantic analysis. Journal of the American society for information science 41 (6), 391.

Demazière, D., Brossaud, C., Trabal, P., Van Meter, K. M., 2006. Analyses textuelles en sociologie (logiciels, méthodes, usages). Didact. Méthodes(Rennes).

Denis, J., Goëta, S., 2013. La fabrique des données brutes. le travail en coulisses de l'open data. In : Penser l'écosystème des données. Les enjeux scientifiques et politiques des données numériques.

Denis, J., Goëta, S., 2014. Exploration, extraction and 'rawification'. the shaping of transparency in the back rooms of open data. In : After The Reveal.

Open Questions on Closed Systems - Neil Postman Graduate Conference, Feb 2014, New York, United States.

Desrosières, A., 1985. Histoires de formes : statistiques et sciences sociales avant 1940. Revue française de sociologie, 277–310.

Desrosières, A., 2008. Analyse des données et sciences humaines : comment cartographier le monde social ? Journal électronique des probabilités et de la statistique 4 (2), 11–19.

Diesner, J., 2015. Small decisions with big impact on data analytics. Big Data & Society 2 (2), 2053951715617185.

DiMaggio, P., Jun. 2015. Adapting computational text analysis to social science (and vice versa). Big Data & Society 2 (2).

DiMaggio, P., Hargittai, E., et al., 2001. From the 'digital divide'to 'digital inequality' : Studying internet use as penetration increases. Princeton : Center for Arts and Cultural Policy Studies, Woodrow Wilson School, Princeton University 4 (1), 4–2.

DiMaggio, P., Nag, M., Blei, D., 2013. Exploiting affinities between topic modeling and the sociological perspective on culture : Application to newspaper coverage of us government arts funding. Poetics 41 (6), 570–606.

Dos Santos, C. N., Gatti, M., 2014. Deep convolutional neural networks for sentiment analysis of short texts. In : COLING. pp. 69–78.

Dowdall, G. W., 1996. The eclipse of the state mental hospital : Policy, stigma, and organization. SUNY Press.

Dumais, S. T., 1991. Improving the retrieval of information from external sources. Behavior Research Methods 23 (2), 229–236.

Dunning, T., 1993. Accurate methods for the statistics of surprise and coincidence. Computational linguistics 19 (1), 61–74.

Duric, A., Song, F., 2012. Feature selection for sentiment analysis based on content and syntax models. Decision support systems 53 (4), 704–711.

Ertöz, L., Steinbach, M., Kumar, V., 2003. Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In : Proceedings of the 2003 SIAM International Conference on Data Mining. SIAM, pp. 47–58.

Evans, J. A., Aceves, P., 2016. Machine translation : mining text for social theory. Annual Review of Sociology 42, 21–50.

Evert, S., 2005. The statistics of word cooccurrences : word pairs and collocations.

Fabo, P. R., Plancq, C., Poibeau, T., 2016. Climate negotiation analysis. In : Digital Humanities 2016. pp. 663–666.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., Strahan, E. J., 1999. Evaluating the use of exploratory factor analysis in psychological research. Psychological methods 4 (3), 272.

Firth, J. R., 1957. A synopsis of linguistic theory, 1930-1955. Studies in linguistic analysis.

Fisher, R. A., 1936. The use of multiple measurements in taxonomic problems. Annals of eugenics 7 (2), 179–188.

Flichy, P., 2001. L'imaginaire d'internet. Réseaux, Paris, La Découverte.

Flichy, P., 2008. Internet et le débat démocratique. Réseaux (4), 159–185.

Fligstein, N., Brundage, J. S., Schultz, M., 2014. Why the federal reserve failed to see the financial crisis of 2008 : the role of "macroeconomics" as a sense making and cultural frame.

Frantzi, K., Ananiadou, S., Mima, H., 2000. Automatic recognition of multi-word terms :. the c-value/nc-value method. International Journal on Digital Libraries 3 (2), 115–130.

Franzosi, R., 1989. From words to numbers : A generalized and linguistics-based coding procedure for collecting textual data. Sociological methodology, 263–298.

Franzosi, R., De Fazio, G., Vicari, S., 2012. Ways of measuring agency : an application of quantitative narrative analysis to lynchings in georgia (1875–1930). Sociological Methodology 42 (1), 1–42.

Friggeri, A., Cointet, J.-P., Latapy, M., et al., 2011. A real-world spreading experiment in the blogosphere. Complex Systems 19 (3), 235.

Fung, P., McKeown, K., 1997. A technical word-and term-translation aid using noisy parallel corpora across language groups. Machine translation 12 (1-2), 53–87.

Gao, Z. J., Song, Y., Liu, S., Wang, H., Wei, H., Chen, Y., Cui, W., 2011. Tracking and connecting topics via incremental hierarchical dirichlet processes. In : Data Mining (ICDM), 2011 IEEE 11th International Conference on. IEEE, pp. 1056–1061.

Garcia, D., Lovink, G., 1997. The abc of tactical media. first distributed via the nettime listserv.

Garfield, E., et al., 1972. Citation analysis as a tool in journal evaluation. Science 178 (4060).

Gelbukh, A., Sidorov, G., Lavin-Villa, E., Chanona-Hernandez, L., 2010. Automatic term extraction using log-likelihood based comparison with general reference corpus. In : International Conference on Application of Natural Language to Information Systems. Springer, pp. 248–255.

Gentzkow, M., Shapiro, J. M., 2006. Media bias and reputation. Journal of political Economy 114 (2), 280–316.

Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J., Reyes, A., 2015. Semeval-2015 task 11 : Sentiment analysis of figurative language in twitter. In : Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). pp. 470–478.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., Brilliant, L., 2009. Detecting influenza epidemics using search engine query data. Nature 457 (7232), 1012–1014.

Glaser, B. G., Strauss, A. L., 1967. The discovery of grounded theory : strategies for qualitative theory.

Goldberg, A., 2015. In defense of forensic social science. Big Data & Society 2 (2), 2053951715601145.

Gomaa, W. H., Fahmy, A. A., 2013. A survey of text similarity approaches. International Journal of Computer Applications 68 (13).

Gomez-Uribe, C. A., Hunt, N., 2016. The netflix recommender system : Algorithms, business value, and innovation. ACM Transactions on Management Information Systems (TMIS) 6 (4), 13.

Gonçalves, P., Benevenuto, F., Cha, M., 2013. Panas-t : A pychometric scale for measuring sentiments on twitter. arXiv preprint arXiv :1308.1857.

Gonthier, F., 2004. Weber et la notion de «compréhension». Cahiers internationaux de sociologie (1), 35–54.

Goodman, S. N., 1999. Toward evidence-based medical statistics. 1 : The p value fallacy. Annals of internal medicine 130 (12), 995–1004.

Graeff, E., Stempeck, M., Zuckerman, E., 2014. The battle for 'trayvon martin' : Mapping a media controversy online and off-line. First Monday 19 (2).

Grimmer, J., King, G., 2011. General purpose computer-assisted clustering and conceptualization. Proceedings of the National Academy of Sciences 108 (7), 2643–2650.

Grimmer, J., Stewart, B. M., 2013. Text as data : The promise and pitfalls of automatic content analysis methods for political texts. Political analysis, 267–297.

Güçlü, U., van Gerven, M. A., 2015. Semantic vector space models predict neural responses to complex visual stimuli. arXiv preprint arXiv :1510.04738.

Hacking, I., 2006. The emergence of probability : A philosophical study of early ideas about probability, induction and statistical inference. Cambridge University Press.

Hamilton, W. L., Leskovec, J., Jurafsky, D., 2016. Diachronic word embeddings reveal statistical laws of semantic change. arXiv preprint arXiv :1605.09096.

Harris, Z., 2012. Papers on syntax. Vol. 14. Springer Science & Business Media.

Harris, Z. S., 1954. Distributional structure. Word 10 (2-3), 146–162.

Hastie, T. J., Tibshirani, R. J., Friedman, J. H., 2011. The elements of statistical learning : data mining, inference, and prediction. Springer.

Healey, P., Rothman, H., Hoch, P. K., 1986. An experiment in science mapping for research planning. Research Policy 15 (5), 233–251.

Himelboim, I., McCreery, S., Smith, M., 2013. Birds of a feather tweet together : Integrating network and content analyses to examine cross-ideology exposure on twitter. Journal of Computer-Mediated Communication 18 (2), 40–60.

Hindman, M., Tsioutsiouliklis, K., Johnson, J. A., 2003. Googlearchy : How a few heavily-linked sites dominate politics on the web. In : annual meeting of the Midwest Political Science Association. Vol. 4. Citeseer, pp. 1–33.

Hofman, J. M., Sharma, A., Watts, D. J., 2017. Prediction and explanation in social systems. Science 355 (6324), 486–488.

Hofmann, T., 2000. Learning the similarity of documents : an information-geometric approach to document retrieval and categorization. Advances in Neural Information Processing Systems 12, 914–920.

Hu, B., Tang, B., Chen, Q., Kang, L., 2016. A novel word embedding learning model using the dissociation between nouns and verbs. Neurocomputing 171, 1108–1117.

Huang, E. H., Socher, R., Manning, C. D., Ng, A. Y., 2012. Improving word representations via global context and multiple word prototypes. In : Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics : Long Papers-Volume 1. Association for Computational Linguistics, pp. 873–882.

Ingold, T., 1996. Key debates in anthropology. Psychology Press.

Ioannidis, J. P., 2005. Why most published research findings are false. PLos med 2 (8), e124.

Jenny, J., 1997. Méthodes et pratiques formalisées d'analyse de contenu et de discours dans la recherche sociologique française contemporaine. etat des lieux et essai de classification. Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique 54 (1), 64–122.

Jensen, P., Morini, M., Karsai, M., Venturini, T., Vespignani, A., Jacomy, M., Cointet, J.-P., Mercklé, P., Fleury, E., 2015. Detecting global bridges in networks. Journal of Complex Networks, cnv022.

Jin, C., 2014. Entre discours critiques et usages a-critiques, sommes-nous sous l'emprise des algorithmes du web ? Mémoire de Master 2, Management de la communication Paris IV.

Joseph, K., Wei, W., Carley, K. M., 2017. Girls rule, boys drool : Extracting semantic and affective stereotypes from twitter. In : 2017 ACM Conference on Computer Supported Cooperative Work.(CSCW).

Joshi, M., Das, D., Gimpel, K., Smith, N. A., 2010. Movie reviews and revenues : An experiment in text regression. In : Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 293–296.

Jurafsky, D., Chahuneau, V., Routledge, B. R., Smith, N. A., 2014. Narrative framing of consumer sentiment in online restaurant reviews. First Monday 19 (4).

Kageura, K., Umino, B., 1996. Methods of automatic term recognition : A review. Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication 3 (2), 259–289.

Katz, E., Lazarsfeld, P. F., 1966. Personal Influence, The part played by people in the flow of mass communications. Transaction Publishers.

Keating, P., Cambrosio, A., Nelson, N. C., 2016. "triple negative breast cancer" : Translational research and the (re) assembling of diseases in post-genomic medicine. Studies in History and Philosophy of Science Part C : Studies in History and Philosophy of Biological and Biomedical Sciences 59, 20–34.

Kim, S., Li, F., Lebanon, G., Essa, I. A., 2013. Beyond sentiment : The manifold of human emotions. In : AISTATS. pp. 360–369.

Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D., Petrov, S., 2014. Temporal analysis of language through neural language models. arXiv preprint arXiv :1405.3515.

King, G., Lam, P., Roberts, M. E., 2017. Computer-assisted keyword and document set discovery from unstructured text. American Journal of Political Science.

Kitchin, R., 2014. Big data, new epistemologies and paradigm shifts. Big Data & Society 1 (1), 2053951714528481.

Klingenstein, S., Hitchcock, T., DeDeo, S., 2014. The civilizing process in london's old bailey. Proceedings of the National Academy of Sciences 111 (26), 9419–9424.

Kosinski, M., Stillwell, D., Graepel, T., 2013. Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences 110 (15), 5802–5805.

Krippendorff, K., 2004. Content analysis : An introduction to its methodology. Sage.

Kulkarni, V., Al-Rfou, R., Perozzi, B., Skiena, S., 2015. Statistically significant detection of linguistic change. In : Proceedings of the 24th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, pp. 625–635.

Kushner, S., 2016. Read only : The persistence of lurking in web 2.0. First Monday 21 (6).

Labbé, C., Labbé, D., 1994. Que mesure la spécificité du vocabulaire. Grenoble, CERAT. Repris dans : Lexicometrica 3, 2001.

Lafon, P., 1980. Sur la variabilité de la fréquence des formes dans un corpus. Mots 1 (1), 127–165.

Lancichinetti, A., Fortunato, S., 2009. Community detection algorithms : a comparative analysis. Physical review E 80 (5), 056117.

Lancichinetti, A., Fortunato, S., Kertész, J., 2009. Detecting the overlapping and hierarchical community structure in complex networks. New Journal of Physics 11 (3), 033015.

Lasswell, H. D., Lerner, D., de Sola Pool, I., 1952. The comparative study of symbols : An introduction. Vol. 1. Stanford University Press.

Latour, B., 1993. Le topofil de boa-vista. la référence scientifique : montage photophilosophique. Raisons pratiques 4, 187–216.

Latour, B., 2012. Tarde's idea of quantification. In : Candea, M. (Ed.), The Social After Gabriel Tarde : Debates and Assessments. Routledge, London, pp. 145–162.

Latour, B., 2014. Changer de société, refaire de la sociologie. La découverte.

Latour, B., Jensen, P., Venturini, T., Grauwin, S., Boullier, D., 2012. 'the whole is always smaller than its parts'–a digital test of gabriel tardes' monads. The British journal of sociology 63 (4), 590–615.

Latour, B., Woolgar, S., 2013. Laboratory life : The construction of scientific facts. Princeton University Press.

Launay, M., 1980. Le syndicalisme chrétien en france 1885-1940. Ph.D. thesis.

Lazarsfeld, P. F., Berelson, B., Gaudet, H., 1968. The peoples choice : how the voter makes up his mind in a presidential campaign. New York Columbia University Press 1948.

Lazer, D., Kennedy, R., King, G., Vespignani, A., 2014. The parable of google flu : traps in big data analysis. Science 343 (6176), 1203–1205.

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., et al., 2009. Life in the network : the coming age of computational social science. Science (New York, NY) 323 (5915), 721.

Le, Q. V., Mikolov, T., 2014. Distributed representations of sentences and documents. In : ICML. Vol. 14. pp. 1188–1196.

Lebart, L., Morineau, A., Fénelon, J.-P., 1979. Traitement des données statistiques(méthodes et programmes). Dunod.

Lee, M., Martin, J. L., 2015. Surfeit and surface. Big Data & Society 2 (2), 2053951715604334.

Lee, M. D., Navarro, D. J., Nikkerud, H., 2005. An empirical evaluation of models of text document similarity. In : Proceedings of the Cognitive Science Society. Vol. 27.

Lelu, A., 2011. Relevant eigen-subspace of a graph : A randomization test. In : CAP 2011. pp. pages–4.

Lemercier, C., Claire, Z., 2010. Méthodes quantitatives pour l'historien. La Découverte.

Lenci, A., Benotto, G., 2012. Identifying hypernyms in distributional semantic spaces. In : Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation. Association for Computational Linguistics, pp. 75–79.

Lerique, S., Roth, C., 2016. The semantic drift of quotations in blogspace : a case study in short-term cultural evolution. Cognitive Science.

Leskovec, J., Backstrom, L., Kleinberg, J., 2009. Meme-tracking and the dynamics of the news cycle. In : Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 497–506.

Levy, O., Goldberg, Y., 2014. Neural word embedding as implicit matrix factorization. In : Advances in neural information processing systems. pp. 2177–2185.

Levy, O., Goldberg, Y., Dagan, I., 2015. Improving distributional similarity with lessons learned from word embeddings. Transactions of the Association for Computational Linguistics 3, 211–225.

Levy, O., Goldberg, Y., Ramat-Gan, I., 2014. Linguistic regularities in sparse and explicit word representations. In : CoNLL. pp. 171–180.

Leydesdorff, L., 2008. On the normalization and visualization of author co-citation data : Salton's cosine versus the jaccard index. Journal of the Association for Information Science and Technology 59 (1), 77–85.

Leydesdorff, L., Rafols, I., 2009. A global map of science based on the isi subject categories. Journal of the American Society for Information Science and Technology 60 (2), 348–362.

Leydesdorff, L., Welbers, K., 2011. The semantic mapping of words and co-words in contexts. Journal of Informetrics 5 (3), 469–475.

Leydesdorff, L., Zaal, R., 1988. Co-words and citations relations between document sets and environments. Informetrics, 105–119.

Li, J., Jurafsky, D., 2015. Do multi-sense embeddings improve natural language understanding ? arXiv preprint arXiv :1506.01070.

Lin, Y., Michel, J.-B., Aiden, E. L., Orwant, J., Brockman, W., Petrov, S., 2012. Syntactic annotations for the google books ngram corpus. In : Proceedings of the ACL 2012 system demonstrations. Association for Computational Linguistics, pp. 169–174.

Lippman, W., 1922. Public opinion.

Loiseau, S., 2016. Lexicométrie : A linguistic school in france in the 1960s-1980s. history, theories and methods. In : Léon, J., Loiseau, S. (Eds.), History

of Quantitative Linguistics in France. RAM-Verlag, Stüttinghauser Ringstr. 44 D-58515 Lüdenscheid Germany, pp. 69–93.

Luong, T., Socher, R., Manning, C. D., 2013. Better word representations with recursive neural networks for morphology. In : CoNLL. pp. 104–113.

Maaten, L. v. d., Hinton, G., 2008. Visualizing data using t-sne. Journal of Machine Learning Research 9 (Nov), 2579–2605.

Manning, C. D., 2016. Computational linguistics and deep learning. Computational Linguistics.

Manning, C. D., Raghavan, P., Schütze, H., et al., 2008. Introduction to information retrieval. Vol. 1. Cambridge university press Cambridge.

Manning, C. D., Schütze, H., 1999. Foundations of statistical natural language processing. Vol. 999. MIT Press.

Manovich, L., 2011. Trending : The promises and the challenges of big social data. Debates in the digital humanities 2, 460–475.

Marchand, P., Ratinaud, P., 2012. L'analyse de similitude appliquée aux corpus textuels : les primaires socialistes pour l'élection présidentielle française (septembre-octobre 2011). Actes des 11eme Journées internationales d'Analyse statistique des Données Textuelles. JADT 2012, 687–699.

Marres, N., 2012. The redistribution of methods : on intervention in digital social research, broadly conceived. The sociological review 60 (S1), 139–165.

Marres, N., 2015a. Why map issues ? on controversy analysis as a digital method. Science, Technology, & Human Values 40 (5), 655–686.

Marres, N., 2015b. Why map issues ? on controversy analysis as a digital method. Science, Technology, & Human Values 40 (5), 655–686.

Marres, N., Moats, D., 2015. Mapping controversies with social media : The case for symmetry. Social Media+ Society 1 (2).

Mayaffre, D., 2005. De la lexicométrie à la logométrie. Astrolabe, 1–11.

Mazieres, A., 2016. Cartographie de l'apprentissage artificiel et de ses algorithmes.

Mazzocchi, F., 2015. Could big data be the end of theory in science ? EMBO reports, e201541001.

McCallum, A. K., 2002. Mallet : A machine learning for language toolkit, http ://mallet.cs.umass.edu.

McFarland, D. A., Lewis, K., Goldberg, A., Sep. 2015. Sociology in the Era of Big Data : The Ascent of Forensic Social Science. The American Sociologist.

McFarland, D. A., McFarland, H. R., 2015. Big data and the danger of being precisely inaccurate. Big Data & Society 2 (2).

McNally, R., 2005. Sociomics ! using the issuecrawler to map, monitor and engage with the global proteomics research network. Proteomics 5 (12), 3010–3016.

Mercklé, P., 2010. Le modèle de la distinction est-il (déjà) pertinent ? In : Premiers résultats de l'enquête longitudinale sur les pratiques culturelles des enfants et des adolescents, Trente ans après La Distinction, Colloque à Paris.

Mesmoudi, S., Rodic, M., Cioli, C., Cointet, J.-P., Yarkoni, T., Burnod, Y., 2015. Linkrbrain : Multi-scale data integrator of the brain. Journal of neuroscience methods 241, 44–52.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., Aiden, E. L., Jan. 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. Science (New York, NY) 331 (6014), 176–182.

Mihalcea, R., Tarau, P., 2004. Textrank : Bringing order into texts. In : Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004. Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013a. Efficient estimation of word representations in vector space. arXiv preprint arXiv :1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J., 2013b. Distributed representations of words and phrases and their compositionality. In : Advances in neural information processing systems. pp. 3111–3119.

Milios, E., Zhang, Y., He, B., Dong, L., 2003. Automatic term extraction and document similarity in special text corpora. In : Proceedings of the sixth conference of the pacific association for computational linguistics. Citeseer, pp. 275–284.

Miller, G. A., Charles, W. G., 1991. Contextual correlates of semantic similarity. Language and cognitive processes 6 (1), 1–28.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., Just, M. A., 2008. Predicting human brain activity associated with the meanings of nouns. science 320 (5880), 1191–1195.

Mogoutov, A., Cambrosio, A., Keating, P., Mustar, P., 2008. Biomedical innovation at the laboratory, clinical and commercial interface : A new method for mapping research projects, publications and patents in the field of microarrays. Journal of Informetrics 2 (4), 341–353.

Mogoutov, A., Vichnevskaia, T., 2006. Réseau-lu, outils d'analyse exploratoire des données hétérogènes en sciences sociales, réseaux, temps, paroles et textes. Les approches qualitatives dans les études de population. Théorie et pratique, AUF, Éditions des archives contemporaines, Coll. Manuels, Paris, 245–269.

Moher, D., Liberati, A., Tezlaff, J., Altman, D., Antes, G., Atkins, D., Barbour, V., Barrowman, N., Berlin, J., Clark, J., et al., 2009. Preferred reporting items for systematic reviews and meta-analyses : the prisma statement. Annals of internal medicine 151 (4), 264–269.

Mohr, J. W., Bogdanov, P., 2013. Introduction—topic models : What they are and why they matter. Poetics 41 (6), 545–569.

Mohr, J. W., Wagner-Pacifici, R., Breiger, R. L., Bogdanov, P., 2013. Graphing the grammar of motives in national security strategies : Cultural interpretation, automated text analysis and the drama of global politics. Poetics 41 (6), 670–700.

Moretti, F., 2004. Distant Reading, 2nd Edition. Addison–Wesley.

Moretti, F., 2005. Graphs, maps, trees : abstract models for a literary history. Verso.

Moretti, F., 2011. Network theory, plot analysis. New Left Review.

Moretti, F., Pestre, D., 2015. Bankspeak : the language of world bank reports. New Left Review 92, 75–99.

Newman, D., Noh, Y., Talley, E., Karimi, S., Baldwin, T., 2010. Evaluating topic models for digital libraries. In : Proceedings of the 10th annual joint conference on Digital libraries. ACM, pp. 215–224.

Newman, M. E., 2006. Modularity and community structure in networks. Proceedings of the national academy of sciences 103 (23), 8577–8582.

Newman, M. E., 2013. Spectral methods for community detection and graph partitioning. Physical Review E 88 (4), 042822.

Ollion, É., Boelaert, J., 2015. Au delà des big data. les sciences sociales et la multiplication des données numériques. Sociologie 6 (3), 295–310.

Ollivier, G., 2010. Vers une sociologie des usages des outils de la sociologie ? exploration du côté des logiciels d'analyse textuelle. Journée des Sociologues de l'INRA.

Omodei, E., Cointet, J.-P., Poibeau, T., 2014. Mapping the natural language processing domain : Experiments using the acl anthology. In : LREC 2014, the Ninth International Conference on Language Resources and Evaluation. ELRA, pp. 2972–2979.

Omodei, E., Poibeau, T., Cointet, J., 2012. Multi-level modeling of quotation families morphogenesis. In : Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom). IEEE, pp. 392–401.

Page, L., Brin, S., Motwani, R., Winograd, T., 1999. The pagerank citation ranking : Bringing order to the web. Tech. rep., Stanford InfoLab.

Palla, G., Derényi, I., Farkas, I., Vicsek, T., 2005. Uncovering the overlapping community structure of complex networks in nature and society. Nature 435 (7043), 814–818.

Parasie, S., Cointet, J.-P., 2012. La presse en ligne au service de la démocratie locale. Revue française de science politique 62 (1), 45–70.

Parasie, S., Cointet, J.-P., 2013. Connecting the dots. an ecological perspective on how publics gather around criminal data. In : Conférence de la Society for the Social Studies of Science (4S).

Parasie, S., Dagiral, E., 2012. Data-driven journalism and the public good :"computer-assisted-reporters" and "programmer-journalists" in chicago. New Media & Society, 20.

Pecina, P., Schlesinger, P., 2006. Combining association measures for collocation extraction. In : Proceedings of the COLING/ACL on Main conference poster sessions. Association for Computational Linguistics, pp. 651–658.

Peixoto, T. P., 2014. Hierarchical block structures and high-resolution model selection in large networks. Physical Review X 4 (1), 011047.

Pennington, J., Socher, R., Manning, C. D., 2014. Glove : Global vectors for word representation. In : EMNLP. Vol. 14. pp. 1532–1543.

Phillips, D., 1995. Correspondence analysis. Social research update 7, 1–8.

Plantin, J.-C., 2013. Chapitre 11-d'une carte à l'autre : Le potentiel heuristique de la comparaison entre graphe du web et carte géographique. U, 228–245.

Ploux, S., Victorri, B., 1998. Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes. Traitement automatique des langues (39), 161–182.

Poibeau, T., 2014. La linguistique est-elle soluble dans la statistique ? Revue Sciences/Lettres (2).

Porter, T. M., 1995. Trust in numbers : The pursuit of objectivity in science and public life. Princeton University Press.

Prost, A., 1974. Vocabulaire des proclamations électorales : de 1881, 1885 et 1889. Vol. 9. Presses universitaires de France.

Proulx, S., 2015. La sociologie des usages, et après ? Revue française des sciences de l'information et de la communication (6).

Prpić, J., Shukla, P. P., Kietzmann, J. H., McCarthy, I. P., 2015. How to work a crowd : Developing crowd capital through crowdsourcing. Business Horizons 58 (1), 77–85.

Raimbault, B., Cointet, J.-P., Joly, P.-B., 2016. Mapping the emergence of synthetic biology. PloS one 11 (9), e0161522.

Řehůřek, R., Sojka, P., May 2010. Software Framework for Topic Modelling with Large Corpora. In : Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. ELRA, Valletta, Malta, pp. 45–50.

Reinert, A., 1983. Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. Les cahiers de l'analyse des données 8 (2), 187–198.

Reinert, M., 1990. Alceste une méthodologie d'analyse des données textuelles et une application : Aurelia de gerard de nerval. Bulletin de méthodologie sociologique 26 (1), 24–54.

Reinert, M., 1993. Les «mondes lexicaux» et leur «logique» à travers l'analyse statistique d'un corpus de récits de cauchemars. Langage et société 66, 5–39.

Reinert, M., 1998. Alceste, un logiciel d'aide pour l'analyse de discours, notice simplifiée. CNRS : Université de Saint-Quentin-en-Yvelines.

Reinert, M., 1999. Quelques interrogations à propos de l'"objet" d'une analyse de discours de type statistique et de la réponse "alceste". Langage et société 90 (1), 57–70.

Reinert, M., 2001. Approche statistique et problème du sens dans une enquête ouverte. Journal de la société française de statistique 142 (4), 59–71.

Reinert, M., 2003. Le rôle de la répétition dans la représentation du sens et son approche statistique par la méthode" alceste". SEMIOTICA-LA HAYE THEN BERLIN- 147 (1/4), 389–420.

Reinert, M., 2007. Postures énonciatives et mondes lexicaux stabilisés en analyse statistique de discours. Langage et société (3), 189–202.

Reinert, M., 2008. Mondes lexicaux stabilisés et analyse statistique de discours. Actes de la JADT 2008, 981–993.

Rogers, E. M., 2010. Diffusion of innovations. Simon and Schuster.

Rogers, R., 2009. The end of the virtual : Digital methods. Vol. 339. Amsterdam University Press.

Rogers, R., 2013. Digital methods. MIT press.

Rogers, R., Marres, N., 2000. Landscaping climate change : A mapping technique for understanding science and technology debates on the world wide web. Public Understanding of Science 9, 1–23.

Romero, D. M., Meeder, B., Kleinberg, J., 2011. Differences in the mechanics of information diffusion across topics : idioms, political hashtags, and complex contagion on twitter. In : Proceedings of the 20th international conference on World wide web. ACM, pp. 695–704.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P., 2004. The author-topic model for authors and documents. In : Proceedings of the 20th conference on Uncertainty in artificial intelligence. AUAI Press, pp. 487–494.

Rosvall, M., Bergstrom, C. T., 2008. Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences 105 (4), 1118–1123.

Ruan, Y.-P., Ling, Z.-H., Hu, Y., 2016. Exploring semantic representation in brain activity using word embeddings. In : Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas. pp. 669–679.

Rubenstein, H., Goodenough, J. B., 1965. Contextual correlates of synonymy. Communications of the ACM 8 (10), 627–633.

Ruiz, P., Poibeau, T., 2015. Combining open source annotators for entity linking through weighted voting. In : Joint Conference on Lexical and Computational Semantics (* SEM 2015). pp. 211–215.

Rule, A., Cointet, J.-P., Bearman, P. S., 2015. Lexical shifts, substantive changes, and continuity in state of the union discourse, 1790–2014. Proceedings of the National Academy of Sciences 112 (35), 10837–10844.

Rykov, Y., Nagornyy, O., Koltsova, O., 2016. Semantic and geospatial mapping of instagram images in saint-petersburg. power 26, 75.

Sagi, E., Dehghani, M., 2014. Measuring moral rhetoric in text. Social Science Computer Review 32 (2), 132–144.

Sagi, E., Diermeier, D., Kaufmann, S., 2013. Identifying issue frames in text. PLoS one 8 (7), e69185.

Sahlgren, M., 2006. The word-space model : Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. Ph.D. thesis, Institutionen för lingvistik.

Salem, A., 1987. Pratique des segments répétés : essai de statistique textuelle. Klincksieck.

Salem, A., 1988. Approches du temps lexical [statistique textuelle et séries chronologiques]. Mots 17 (1), 105–143.

Saussure, F. d., 1967. Cours de linguistique générale, publié par charles bally. Paris (Payot).

Savage, M., Burrows, R., 2007. The coming crisis of empirical sociology. Sociology 41 (5), 885–899.

Schmidt, B. M., 2012. Words alone : Dismantling topic models in the humanities. Journal of Digital Humanities 2 (1), 49–65.

Schmidt, B. M., 2015. Plot arceology : A vector-space model of narrative structure. In : Big Data (Big Data), 2015 IEEE International Conference on. IEEE, pp. 1667–1672.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E., et al., 2013. Personality, gender, and age in the language of social media : The open-vocabulary approach. PloS one 8 (9), e73791.

Sennrich, R., Haddow, B., 2016. Linguistic input features improve neural machine translation. arXiv preprint arXiv :1606.02892.

Shahaf, D., Horvitz, E., Mankoff, R., 2015. Inside jokes : Identifying humorous cartoon captions. In : Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 1065–1074.

Shi, F., Foster, J. G., Evans, J. A., 2015. Weaving the fabric of science : Dynamic network models of science's unfolding structure. Social Networks 43, 73–85.

Shlens, J., 2014. A tutorial on principal component analysis. arXiv preprint arXiv :1404.1100.

Shwed, U., Bearman, P. S., 2010. The temporal structure of scientific consensus formation. American sociological review 75 (6), 817–840.

Simmons, M. P., Adamic, L. A., Adar, E., 2011. Memes online : Extracted, subtracted, injected, and recollected. ICWSM 11, 17–21.

Sloan, L., Morgan, J., Burnap, P., Williams, M., 2015. Who tweets ? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. PloS one 10 (3), e0115545.

Small, H., 1973. Co-citation in the scientific literature : A new measure of the relationship between two documents. Journal of the Association for Information Science and Technology 24 (4), 265–269.

Smyrnaios, N., Ratinaud, P., 2017. The charlie hebdo attacks on twitter : A comparative analysis of a political controversy in english and french. Social Media+ Society 3 (1), 2056305117693647.

Snow, C. P., 1959. Two cultures. Science 130 (3373), 419–419.

Snow, D. A., Benford, R. D., et al., 1988. Ideology, frame resonance, and participant mobilization. International social movement research 1 (1), 197–217.

Speed, J. G., 1893. Do newspapers now give the news ? In : Forum. Vol. 15. pp. 705–711.

Sternitzke, C., Bergmann, I., 2008. Similarity measures for document mapping : A comparative study on the level of an individual scientist. Scientometrics 78 (1), 113–130.

Sumpter, R. S., 2001. News about news : John g. speed and the first newspaper content analysis. Journalism History 27 (2), 64.

Szymanski, T., 2017. Temporal word analogies : Identifying lexical replacement with diachronic word embeddings. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistic, Vancouver.

Tan, P.-N., Kumar, V., Srivastava, J., 2002. Selecting the right interestingness measure for association patterns. In : Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 32–41.

Tancoigne, E., Barbier, M., Cointet, J.-P., Richard, G., 2014. The place of agricultural sciences in the literature on ecosystem services. Ecosystem Services 10, 35–48.

Tarde, G., 1904. La logique sociale. F. Alcan.

Teil, G., Latour, B., 1995. The hume machine : Can association networks do more than formal rules. Stanford Humanities Review 4 (2), 47–65.

Templeton, C., 2011. Topic modeling in the humanities : An overview. Maryland Institute for Technology in the Humanities Blog.

Teng, C.-Y., Lin, Y.-R., Adamic, L. A., 2012. Recipe recommendation using ingredient networks. In : Proceedings of the 4th Annual ACM Web Science Conference. ACM, pp. 298–307.

Teten, R. L., 2003. Evolution of the modern rhetorical presidency : Presidential presentation and development of the state of the union address. Presidential Studies Quarterly 33 (2), 333–346.

Tibshirani, R., Walther, G., Hastie, T., 2001. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society : Series B (Statistical Methodology) 63 (2), 411–423.

Tieberghien, E., Mélanie-Becquet, F., Fabo, P. R., Poibeau, T., Terras, M., Causer, T., 2016. Mapping the bentham corpus. In : Digital Humanities 2016.

Timmermans, S., Tavory, I., 2012. Theory construction in qualitative research : From grounded theory to abductive analysis. Sociological Theory 30 (3), 167–186.

Titov, I., McDonald, R., 2008. Modeling online reviews with multi-grain topic models. In : Proceedings of the 17th international conference on World Wide Web. ACM, pp. 111–120.

Tournier, M., 2007. Les mots de mai 68. Presses Univ. du Mirail.

Tufekci, Z., 2014. Big questions for social media big data : Representativeness, validity and other methodological pitfalls. Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media.

Turco, C. J., Zuckerman, E. W., 2017. Verstehen for sociology : Comment on watts. American Journal of Sociology 122 (4), 1272–1291.

Turner, F., 2010. From counterculture to cyberculture : Stewart Brand, the Whole Earth Network, and the rise of digital utopianism. University of Chicago Press.

Van Eck, N. J., Waltman, L., 2010. Software survey : Vosviewer, a computer program for bibliometric mapping. Scientometrics 84 (2), 523–538.

Van Eck, N. J., Waltman, L., Noyons, E. C., Buter, R. K., 2010. Automatic term identification for bibliometric mapping. Scientometrics 82 (3), 581–596.

Venant, F., 2008. Représentation géométrique et calcul dynamique du sens lexical : application à la polysémie de livre. Langages (4), 30–52.

Venturini, T., 2012. Building on faults : how to represent controversies with digital methods. Public understanding of science 21 (7), 796–812.

Venturini, T., Baya Laffite, N., Cointet, J.-P., Gray, I., Zabban, V., De Pryck, K., 2014a. Three maps and three misunderstandings : A digital mapping of climate diplomacy. Big Data & Society 1 (2), 2053951714543804.

Venturini, T., Cardon, D., Cointet, J.-P., 2014b. Méthodes digitales : approches quali/quanti des données numériques.

Venturini, T., Jacomy, M., Meunier, A., Latour, B., 2017. An unexpected journey : A few lessons from sciences po médialab's experience. Big Data & Society 4 (2).

Wagner, C. S., Leydesdorff, L., 2005. Network structure, self-organization, and the growth of international collaboration in science. Research policy 34 (10), 1608–1618.

Wang, C., Blei, D., Heckerman, D., 2012. Continuous time dynamic topic models. arXiv preprint arXiv :1206.3298.

Wang, X., McCallum, A., 2006. Topics over time : a non-markov continuous-time model of topical trends. In : Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 424–433.

Wasserman, S., Faust, K., 1994. Social network analysis : Methods and applications. Vol. 8. Cambridge university press.

Watts, D., 2017a. Response to turco and zuckerman's "verstehen for sociology". American Journal of Sociology 122 (4), 1292–1299.

Watts, D. J., 2014. Common sense and sociological explanations. American Journal of Sociology 120 (2), 313–351.

Watts, D. J., 2017b. Should social science be more solution-oriented ? Nature Human Behaviour 1, 0015.

Weeds, J., Weir, D., 2003. A general framework for distributional similarity. In : Proceedings of the 2003 conference on Empirical methods in natural language processing. Association for Computational Linguistics, pp. 81–88.

Weeds, J., Weir, D., 2005. Co-occurrence retrieval : A flexible framework for lexical distributional similarity. Computational Linguistics 31 (4), 439–475.

Weeds, J., Weir, D., McCarthy, D., 2004. Characterising measures of lexical distributional similarity. In : Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, p. 1015.

Wehbe, L., Vaswani, A., Knight, K., Mitchell, T. M., 2014. Aligning context-based statistical models of language with brain activity during reading. In : EMNLP. pp. 233–243.

Weisz, G., Cambrosio, A., Cointet, J.-P., 2017. Mapping global health : A network analysis of an heterogeneous publication domain. Biosocieties.

White, D. M., 1950. The "gate keeper" : A case study in the selection of news. Journalism & Mass Communication Quarterly 27 (4), 383–390.

Whittaker, J., 1989. Creativity and conformity in science : Titles, keywords and co-word analysis. Social Studies of Science 19 (3), 473–496.

Wilks, S., 1935. The likelihood test of independence in contingency tables. The annals of mathematical statistics 6 (4), 190–196.

Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., Nevill-Manning, C. G., 1999. Kea : Practical automatic keyphrase extraction. In : Proceedings of the fourth ACM conference on Digital libraries. ACM, pp. 254–255.

Wittgenstein, L., 1953. Philosophical investigations. John Wiley & Sons.

Wong, W., 2009. Determination of unithood and termhood for term recognition. In : Handbook of research on text and web mining technologies. IGI Global, pp. 500–529.

Xie, J., Kelley, S., Szymanski, B. K., 2013. Overlapping community detection in networks : The state-of-the-art and comparative study. Acm computing surveys (csur) 45 (4), 43.

Young, M. L., Hermida, A., 2015. From mr. and mrs. outlier to central tendencies : Computational journalism and crime reporting at the los angeles times. Digital Journalism 3 (3), 381–397.

Zalio, P.-P., 2007. Les entrepreneurs enquêtés par les récits de carrières : de l'étude des mondes patronaux à celle de la grammaire de l'activité entrepreneuriale. Sociétés contemporaines (4), 59–82.

Zhang, Z., Iria, J., Brewster, C., Ciravegna, F., 2008. A comparative evaluation of term recognition algorithms. In : LREC.

Zitt, M., Bassecoulard, E., 2006. Delineating complex scientific fields by an hybrid lexical-citation method : An application to nanosciences. Information processing & management 42 (6), 1513–1531.

Zitt, M., Lelu, A., Bassecoulard, E., 2011. Hybrid citation-word representations in science mapping : Portolan charts of research fields ? Journal of the American Society for Information Science and Technology 62 (1), 19–39.

Index exhaustif des projets