Signifiant et significatif. Réflexions épistémologiques sur la sémiotique et l'analyse des données

Dario Compagno

DARIO COMPAGNO
Centre de recherche sur les médiations
Université de Lorraine
F-57000
dario.compagno@univ-lorraine.fr

SIGNIFIANT ET SIGNIFICATIF
RÉFLEXIONS ÉPISTÉMOLOGIQUES SUR LA SÉMIOTIQUE ET L'ANALYSE DES DONNÉES

Résumé. — L'article porte sur l'analyse de données en sciences humaines et sociales, en particulier selon une perspective sémiotique. Nous montrerons les affinités épistémologiques entre méthodes numériques et analyse sémiotique immanente. Ensuite, nous défendrons l'idée selon laquelle la discussion actuelle sur les limites des méthodes numériques reprend certaines des critiques qui ont suscité la constitution d'une approche pragmatique en sémiotique : l'analyse ne peut être détachée de l'interprétation. Cela peut conduire à formuler des méthodes quali-quantitatives d'enquête, qui sont plus qu'une juxtaposition de méthodes quantitatives et qualitatives traditionnelles. En effet, il s'agit d'identifier des procédures pour la transformation des expressions, capables de retrouver leur sens et leur dimension sociale sans aller au-delà des données. À cette fin, il est nécessaire de dépasser l'opposition entre qualitatif et quantitatif en couplant compétences statistiques et sémiotiques.

Mots clés. — sémiotique, statistiques, signification, significativité, méthodes quali-quantitatives

Il existe une affinité entre les tournants structuraliste et numérique dans les sciences humaines. Cette affinité profonde va bien au-delà de ce que l'on pourrait voir d'emblée. À sa naissance, la sémiotique structurale se voulait la plus qualitative de toutes les sciences humaines, s'ancrant au coup de force théorique de Louis Hjelmslev (1943 : 12), selon lequel il faut éliminer de l'analyse toute composante substantielle : « La linguistique doit chercher à saisir le langage non comme un conglomérat de faits non linguistiques (physiques, physiologiques, psychologiques, logiques, sociologiques), mais comme un tout qui se suffit à lui-même, une structure *sui generis* ». Cette proposition semble affranchir le sémiologue de toute quantification si, par définition, on ne peut mesurer que ce qui est en quelque sorte contingent. Les analyses sémiotiques sont effectivement libres des considérations statistiques pour se concentrer sur la saisie formelle des *fonctions sémiotiques* : les règles de corrélation entre des systèmes d'expressions (ou signifiants) et des systèmes de contenus (ou signifiés). Les fonctions sémiotiques sont *immanentes* aux pratiques sociales : bien qu'implémentées dans des *apparatus* cognitifs et socialisées par des normes, en soi, elles peuvent être étudiées de manière autonome. Noam Chomsky (1957) soutenait l'autonomie de la syntaxe et la possibilité de l'étudier avec des modèles formels ; de manière analogue, Louis Hjelmslev défendait l'autonomie de la sémantique, son indépendance de toute condition empirique de réalisation. Comme l'énonce Patrizia Violi (1997), parmi les trois approches théoriques majeures du sens (philosophie analytique, psychologie cognitive et sémiotique structurale), la sémiotique structurale est la plus caractérisée par une attention au langage *sui generis*, au sens comme dimension autonome de la réalité humaine et sociale, à ne pas confondre avec les référents physiques ou les représentations cognitives.

De son côté, l'analyse des données est issue de la statistique appliquée. Son but est de mieux comprendre des données pertinentes pour un certain domaine. Cela peut paraître complètement étranger à la sémiotique. L'analyse de données ne travaille que le contingent, les productions attestées de données. Les textes qui passent par une machine semblent destinés à perdre leur sens et à se détacher de la « vie sociale ». Faisons cependant attention.

En réalité, l'analyse sémiotique et l'analyse des données partagent un certain nombre de traits importants. En regardant mieux, on constate premièrement que les régularités visées par cette analyse (règles d'association, facteurs latents, corrélations, regroupements…) sont toujours *immanentes* à un ensemble de données. C'est-à-dire que les résultats de l'analyse sont intrinsèques à celles-ci. Ils dépendent strictement d'elles et sont indépendants de tout ce qui se trouve au-delà. Résultat d'une collecte empirique, les données deviennent la seule réalité d'intérêt pour l'analyste (« il n'y a pas de hors données »). Le rapport avec la réalité empirique de départ est alors pour le moins ambigu. On est évidemment censé trouver des résultats en accord avec les connaissances scientifiques acquises et consolidées – la vraisemblance reste bien un critère de scientificité. En même temps, on travaille sur des traces « élégantes » et autonomes qui décrivent des mondes possibles, dont la vérité factuelle est souvent l'objet de vérifications *a posteriori* et, en ce sens,

« transcendantes » à l'analyse des données elles-mêmes. La force de l'analyse de données réside dans la possibilité d'identifier des régularités jusque-là inconnues, de mettre au jour des rapports qui, sans elle, resteraient ignorés. Conséquence de son approche immanente, sa principale faiblesse peut être ainsi décrite : *correlation is not causation*, ce qui signifie pratiquement que ce que l'on trouve dans des données est souvent suffisant pour formuler des hypothèses, mais pas pour les vérifier.

La deuxième ressemblance fondamentale entre l'analyse sémiotique et l'analyse de données réside dans le fait que le deux ne travaillent que sur des *différences* : tout traitement algorithmique est fondé sur la manipulation d'éléments qui n'ont, pour la machine, qu'une valeur différentielle. Du fait que les données soumises à l'analyse sont coupées par la machine de la réalité transcendante, elles perdent toute substantialité pour devenir de pures identités structurales, des croisements de formes. De fait, pour l'ordinateur, les « observations empiriques » ne deviennent que des séries de valeurs qui peuvent se transformer en ne gardant qu'une relation très indirecte avec les arrangements du départ. « L'observation » est donc bien une opération structurale qui décompose un phénomène réel (une substance) en une série finie de dimensions formelles (variables), chacune décrivant un aspect isolé du phénomène[1]. Finalement, chaque « individu » est un point défini par sa position dans un système de représentation, par des coordonnées sur un espace multidimensionnel où chaque dimension est une variable[2]. À la fin du processus, la substance de départ (la personne, le mot) n'est connue que par sa forme, *par les écarts* qui permettent de différencier une observation d'une autre.

Si Google et Amazon, entreprises deleuziennes, arrivent si bien à nous connaître, c'est parce qu'elles ont renoncé à le faire « vraiment », à saisir notre « vraie nature », un vrai au-delà de toute vraisemblance, et se limitent à construire des simulacres des acteurs sociaux. Si l'on regarde les choses selon cette perspective, l'analyse de données n'est que la continuation de l'analyse structurale par d'autres moyens. On y trouve la même curiosité pour la trace, les mêmes soucis méthodologiques sur la lecture attentive de ce qu'on a sous les yeux, plutôt que de ce qui est toujours au-delà du regard (le réel qui précède et fonde toute trace). On vise la représentation plutôt que la chose, la dissolution pragmatique des concepts dans leurs effets compréhensibles. On refuse donc de se rendre à quelque chose qui serait toujours plus vrai que l'apparence, un « objet dynamique » (Peirce, 1993) qui nous échapperait nécessairement, une « présence phénoménologique » (Derrida, 1967) antérieure à toute constitution ancrée sur les signes. Dans cette perspective, l'ordre de la donnée n'est que l'ordre de la trace, assisté par ordinateur. Dans l'analyse des données, la réduction sémiotique de l'homme à ses signes externes trouve plus une réincarnation armée qu'un simple allié.

---

[1] On pourrait renvoyer aux travaux de B. Latour et de F. Bastide (2001) sur les conditions structurales du travail expérimental, et plus généralement aux réflexions de C. G. Hempel, M. Hesse et T. Kuhn sur le rapport entre description et théorie.

[2] Ce que l'on appelle aujourd'hui *tidy data* (« bonne représentation de données ») prescrit que chaque variable doit être mise sur une colonne, chaque observation sur une ligne, chaque type d'unité observationnelle sur une table (Wickham, 2014).

À l'aube des *big data*, on se découvre spécialement fragiles, nus face à l'observation systématique : l'essence cachée de l'esprit montre bien sa chair sémiotique (Doueihi, 2008). On est finalement obligés de voir les crochets qui peuvent saisir l'esprit et de comprendre la prise que les statistiques peuvent avoir sur la liberté, la personnalité, la pensée. La prévision sociale est aujourd'hui plus à portée que l'explication, à partir de corrélations de traces numériques ; au contraire, l'explication traditionnelle – qui cherche des causes isolées comparables aux causes physiques – est noyée dans la complexité des niveaux imbriqués dans des sociétés de plus en plus globalisées et hybridées. On est bien entré dans le quatrième paradigme (Hey, Tansley, Tolle, 2009), c'est-à-dire qu'on prévoit sans expliquer, situation inédite pour les sciences humaines et sociales (SHS). Toute collecte de données personnelles menace d'arriver, tôt ou tard, à constituer un simulacre aussi bon que la vraie personne, capable de se substituer à elle et de prendre sa place et son identité. Et cela sans avoir isolé une seule cause sociale.

Se réfugier dans la croyance en un supplément insaisissable dans l'homme, sa « vraie nature », âme ou intention, qui ne sera jamais collectée et simulée parce que jamais réellement exprimée, est alors un pari risqué. Heureusement, le XX^e^ siècle a déjà donné à penser l'extériorisation de l'intériorité et la dimension sémiotique des signifiés phénoménologiques (Peirce, 1993 ; Wittgenstein, 1933-1935 ; Derrida, 1967). Peut-être n'est-il pas nécessaire de formuler une épistémologie des *big data*, parce qu'elle existait déjà, un peu en avance sur ses applications. En effet, les assertions telles « mon langage est la somme totale de moi-même » (Peirce) ou « il n'y a rien de caché » (Wittgenstein) signifient qu'il n'y a pas de limite que l'analyse ne saurait franchir, pas de limite *a priori* à ce que l'on peut dire et prédire d'un individu sur la base de la mémoire de ses actes. La limite n'apparaît qu'*a posteriori*, elle ne dépend que des outils, de la qualité et de la quantité des données. L'analyse de données offre à la réflexion sémiotique des armes pour mieux combattre deux *idola* : le « mythe de l'intériorité », selon lequel l'homme cache quelque chose qui ne sortira jamais de sa bouche (Bouveresse, 1987) et celui de la « donnée complète », soit l'idée que le monde social cache quelque chose qui ne pourra jamais être converti en données, *rosa pristina* qui meurt si elle est cueillie (c'est-à-dire nommée), référent qui jamais ne deviendra signe. Sauf à parier sur ce cœur asémiotique de l'homme, situé au-delà de toute saisie objectivable, il est préférable de commencer à réfléchir avec l'analyse de données (plutôt que contre elle) pour ne pas se trouver dans la désagréable situation d'être objet et non plus sujet de réflexion[3].

Pas d'analyse sans interprétation

L'analyse des données a trouvé dans l'internet son parfait complément. Archive vivante de données déjà numérisées, formes déjà sans substance. L'euphorie pour l'objet a conduit à la constitution de *digital methods* pour la gestion, le traitement et l'analyse de données numériques (voir Rogers, 2013). Cependant, face aux données de l'internet,

[3] Comme l'écrit M. Doueihi (2008), on ne peut pas faire rentrer les technologies numériques de manière instrumentale dans les grilles traditionnelles des sciences de l'homme. Il est nécessaire de repenser ces dernières et même l'humain, à partir de l'apport scientifique et quotidien des technologies.

l'analyse de données montre aussi ses limites. Franck Rebillard (2011) évalue de manière détaillée les avantages et manques des *digital methods*. Pour le chercheur, certaines de ces limitations sont temporaires et pourront être dépassées avec l'évolution des technologies et des pratiques de recherche[4] ; en revanche, d'autres sont constitutives des méthodes numériques. Ces manques majeurs peuvent se diviser en deux classes : l'impossibilité de bien interpréter des données avec les seuls instruments quantitatifs et automatisés ; l'impossibilité à saisir la réalité sociale à partir des seules traces numériques. Ces deux classes de limites sont imbriquées parce que, pour Franck Rebillard, la réponse à l'une réside dans un dépassement de l'autre. C'est-à-dire que, si l'on ajoute à l'analyse des données numériques une appréhension parallèle de la réalité sociale (à travers des méthodes traditionnelles comme les questionnaires et les entretiens), on peut atteindre une meilleure compréhension des données elles-mêmes et, en même temps, des phénomènes sociaux auxquels elles font référence. L'auteur fait en particulier le point sur certaines expériences de recherche auxquelles il a pris part directement. À propos de l'une d'elles (projet « Internet, pluralité et redondance de l'information » – IPRI – financé par l'Agence nationale de la recherche), il écrit :

> « Pour autant, à chacune de ces étapes, constitution du corpus comme analyse de contenu, le recours aux *digital methods* s'est avéré nécessaire mais non suffisant : continuellement, l'expertise des chercheurs spécialistes de l'étude des médias a été sollicitée pour statuer sur les différents choix en amont et interpréter les résultats des processus d'automatisation en aval. Les méthodes de constitution du corpus de sites sont tout à fait révélatrices d'une telle imbrication » (*ibid.* : 369).

Dans l'argument de Franck Rebillard, ce qui frappe le sémiologue est une affinité évidente avec les critiques épistémologiques portées contre les approches immanentistes du sens. De ce fait, une analyse sans interprétation est-elle possible ? Est-il possible d'analyser un texte en dehors d'un corpus de références et/ou de circonstances de production et d'interprétation attestées ? Ces questionnements ont conduit à la constitution de la sémiotique interprétative d'Umberto Eco (1979, 1973-1975, 1975), de la sémantique interprétative de François Rastier (1987, 1989), de la sémio-pragmatique de Roger Odin (1983, 2000) et François Jost (2001), de la sémiotique des pratiques de Jacques Fontanille (2007, 2008). Pour rappel, le projet scientifique et philosophique d'Umberto Eco visait le dépassement d'une analyse pure en faveur d'une prise en compte des conditions de l'interprétation. *Œuvres ouvertes*, c'est-à-dire où le sens est à compléter par les lecteurs ; *structures absentes*, donc outils conceptuels pour la compréhension plutôt que résultats de l'explication. La description immanente (Hjelmslev) n'est pas suffisante parce que les règles qu'elle décrit ne valent que pour des interprètes situés et non pour un sujet idéal et transcendantal ; il faut alors coupler les règles immanentes aux parcours inférentiels (Peirce) qui activent les fonctions sémiotiques dans des situations d'interprétation. En France, la sémio-pragmatique invite à élargir le texte pour inclure ses instructions d'appropriation. D'ailleurs, Franck Rebillard (2011 : § 25) propose l'approche sémio-pragmatique comme complément qualitatif pour dépasser les limites de l'analyse des données :

---

[4] Par exemple, D. Mayaffre (2010) montre que l'évolution des outils a permis de dépasser des limites qui semblaient condamner l'étude assistée par ordinateur du langage à sa surface.

> « Après identification des différentes classes de discours portés sur un sujet d'actualité dominant [...] un échantillon d'une trentaine d'articles, représentatif de ces classes de discours et des différentes catégories de sites, est constitué. Il est alors l'objet d'une analyse sémiopragmatique (modes d'énonciation éditoriale, relations texte-image, représentations du lecteur, registres discursifs) qui vise à cerner les différents cadrages (cadrages seconds), opinions et points de vue exprimés. Une telle démarche permet d'atteindre un stade d'évaluation du pluralisme (la disparité de traitement journalistique) plus fin que de seules mesures quantitatives (variété et distribution des sujets). Cette analyse qualitative et sémiopragmatique de discours apporte donc des compléments aux analyses quantitatives et souvent purement lexicales autorisées par les *digital methods* ».

L'analyse pure, immanente – qualitative ou quantitative – doit être accompagnée par l'interprétation, donc, par une relation au contexte de réception prévu. Or, il semble que l'argument pourrait être remis en question afin de montrer des conséquences épistémologiques intéressantes pour la sémiotique. Considérations qui vont vers une définition de sortes de « *digital methods 2.0* »[5] ou, pour être plus précis, vers une sémiotique interprétative quali-quantitative. Il ne s'agit pas de nier cette argumentation, mais de prendre la direction qu'elle indique et d'aller jusqu'au bout de ses possibles. Dans une perspective sémiotique, il est loisible d'identifier deux points de développement dans l'argument présenté par le chercheur en sciences de l'information et de la communication. Nous bâtirons notre propos à partir d'eux. Premièrement, Franck Rebillard s'intéresse aux *digital methods* focalisées sur l'internet et considère les données *offline*, produites par les chercheurs avec des méthodes traditionnelles, comme essentiellement différentes et mieux liées à la réalité sociale. Pourtant, en prenant en considération un sens plus général du concept de donnée numérique, cette limitation peut être dépassée. En effet, les chercheurs qui conduisent des recherches « traditionnelles » finissent aujourd'hui par numériser et diffuser leurs données, et la réalité sociale (production, consommation) rentre sur l'internet « par la fenêtre »[6]. Secondement, Franck Rebillard donne l'impression de défendre une différence essentielle entre méthodes quantitatives et qualitatives : chacune doit être suivie de manière autonome, avec ses critères spécifiques, et l'intégration des deux ne peut se faire qu'*a posteriori*.

Voici une présentation synthétique de notre propos alternatif :

1. il n'est pas possible de « sortir des données », parce que n'importe quelle information supplémentaire, si objectivable soit-elle, redevient donnée (c'est-à-dire que l'immanence peut s'agrandir, mais l'analyse reste immanente à des données[7]) ;

---

[5] « Loin d'être mécanique et purement quantitative, la deuxième génération des travaux des humanités numériques est "qualitative, interprétative, émotive, générative", comme l'écrivent les auteurs du *Digital Humanities Manifesto* 2.0 » (Gefen 2015 : 73).

[6] F. Rebillard (2011) remarque ultérieurement que les *digital methods* ne se focalisent que sur l'internet en dépit des autres médias, et prétendent voir en lui le seul objet d'intérêt. Ce n'est pas la perspective que nous défendons ; tout au plus, il serait possible de produire et numériser des données sur la télévision ou la presse.

[7] J. Fontanille (2007) insère les textes dans une succession de niveaux qui arrivent jusqu'aux pratiques sociales : pour comprendre chaque niveau *n*, il est toujours nécessaire d'atteindre le niveau *n+1*. L'immanence est donc constitutivement insuffisante, manquante, mais la constitution de niveaux ultérieurs est susceptible d'analyse.

2. il n'est donc pas nécessaire, ni possible, d'aller au-delà des données pour retrouver le social – en réalité, il faut des données de qualité, « denses » ;

3. avec les instruments technologiques actuels, il n'y a plus d'opposition entre méthodes qualitatives et quantitatives, mais continuité et complémentarité (c'est-à-dire que la sémiotique ne peut plus se passer des statistiques et inversement).

À bien y regarder, les solutions pratiques proposées par Franck Rebillard vont déjà dans cette direction. Pour interpréter les données du projet IPRI, son équipe a mis en place un échantillonnage des données et leur analyse sémiopragmatique, et ne cherche pas un « au-delà » des données. Au contraire, les chercheurs procèdent à un « zoom » qualitatif (*less data*) sur des résultats quantitatifs. Prenons un autre exemple. En étudiant la médiatisation des tueries de Montauban et Toulouse[8] avec Franck Rebillard, nous n'avons pas cherché *a posteriori* des moyens pour intégrer le qualitatif au quantitatif[9] ; au contraire, nous avons tenté de *retrouver le qualitatif (le sens, le social) avec des algorithmes* (Compagno, Rebillard, 2016). Comment peut-on réaliser cela ? Telle est la question qui fonde l'analyse quali-quantitative, qui n'est pas une simple juxtaposition d'analyse qualitative et quantitative.

Qualitatifs et quantitatifs

Résumons : sémiotique immanentiste et analyse de données semblent partager la même épistémologie, sauf à dire que la première suit une méthodologie qualitative et la deuxième quantitative. La sémiotique immanentiste est sortie d'une impasse grâce à l'intégration de considérations pragmatiques. Que faire alors pour l'analyse de données ? Peut-elle sortir de son impasse et intégrer la pragmatique ? C'est ce que propose Franck Rebillard : en couplant sémiopragmatique (qualitative) et analyse de données (quantitative), nous pouvons obtenir de bons résultats. Dans la distinction entre sémiotique et analyse de données et, en conséquence, dans la solution de Franck Rebillard, ce qui semble poser problème est l'appui sur un seuil net entre les concepts de qualitatif et de quantitatif. Quel sens donner aux deux termes ? Y a-t-il un seuil infranchissable entre *res extensa* quantifiable et esprit qualifiable ? Cela est important parce que, si nous arrivons à dépasser cet obstacle cartésien, une nouvelle voie s'ouvre pour les SHS : une troisième méthodologie qui est plus que la somme des deux précédentes.

---

[8] Entre le 11 et le 19 mars 2012, à Montauban et Toulouse, Mohammed Merah exécute 7 personnes, dont deux militaires de confession musulmane et trois enfants d'une école juive. Ces tueries sont revendiquées par des organisations terroristes islamistes. Le 22 mars 2012, l'assassin est mis hors d'état nuire par le Raid, l'unité d'élite de la Police nationale.

[9] Par exemple, nous n'avons pas conduit des entretiens. Dans le sens contraire, des résultats d'autres études sur la consommation des médias (enquête Mediapolis, Le Hay, Vedel, Chanvril, 2011) ont été utilisés pour enrichir les données dont nous disposions.

Il se trouve que l'on peut présenter l'opposition entre qualitatif et quantitatif de plusieurs manières, sur la base : (1) de la taille des données ; (2) du type de variables analysées ; (3) de la représentation mathématique ou discursive des données ; (4) des procédures de recherche employées ; (5) de la spécificité de l'homme ; (6) de la nature des différences pertinentes ou encore – position que nous défendons ici – sur la base de la dimension sémantique des données. Nous allons présenter rapidement ces options (pour une discussion approfondie, limitée aux travaux en sémiotique, voir Compagno, à paraître b) pour montrer que le seuil séparant qualitatif et quantitatif est, dans une bonne mesure, idéologique. Cela servira ensuite à souder le « significatif », propre à la statistique et vrai noyau des méthodes quantitatives, et le « signifiant », propre à la sémiotique et essence des méthodes qualitatives.

Dans les sciences sociales, on semble parfois assimiler qualitatif et quantitatif à la *taille* des données (1) : des statistiques sur de grands nombres sont quantitatives, tandis que des entretiens ciblés et conduits en profondeur sont qualitatifs. Mais c'est justement le fait d'être *ciblées* et créées *en profondeur* par des chercheurs *compétents* qui rend les données pertinentes pour l'analyse qualitative. Le travail humain, référence qui permettait d'établir un seuil entre *small* et *big*, est désormais affecté par les instruments à disposition du chercheur ; dès que l'on peut analyser avec des outils informatisés des milliers d'interactions spontanées (« situées » et *ecologically valid*) sur les réseaux socionumériques, la taille des données est le premier critère à perdre de sa valeur pour déterminer ce qui est quantitatif ou qualitatif. Comme l'écrit François Rastier (2011 : 51), « pas plus que le fréquent n'est assimilable au quantitatif, le rare ne se confond avec le qualitatif. Il n'y a pas d'opposition entre quantitatif (positiviste) et qualitatif (élitiste), mais une complémentarité ». Il semblerait plus intéressant de parler de données soignées, *dense data*, descriptions denses (*thick*, à la Geertz, 1973a, 1973b) des phénomènes étudiées. Mais comment précisément caractériser cette « densité » qualitative ?

L'analyse de données (2) considère comme qualitative toute *variable* discrète (*A*, *B*, *C* sont des valeurs d'une variable qualitative, pendant que *1*, *2* et *3* sont des valeurs d'une variable quantitative). Il s'agit d'une distinction importante, premièrement, d'un point de vue pratique, car nous ne pouvons pas appliquer les mêmes procédures de traitement et d'analyse aux deux types de variables[10]. Deuxièmement, ce point de vue reconnaît à certaines variables une vraie résistance à la représentation métrique. Il ne s'agit pas d'un problème de taille : on peut travailler avec peu de nombres et beaucoup de textes. Mais les textes ne peuvent pas être complètement quantifiés, sans résidus. Troisièmement, ce seuil est quand même perméable et le passage entre qualitatif et quantitatif est tout à fait possible sous conditions. Par exemple, on peut donner une représentation quantitative de textes à travers des matrices lexicales[11]. Les variables qualitatives sont donc

[10] Par exemple, l'analyse des correspondances ne fonctionne que sur des variables quantitatives ; pour les qualitatives, une analyse des correspondances multiples est nécessaire.

[11] Pour une présentation classique des méthodes en statistiques textuelles, voir L. Lebart et A. Salem (1994).

spéciales mais, en même temps, on peut les convertir partiellement en variables quantitatives. Dès lors, comment caractériser ce que l'on perd au passage, le bon grain qui n'est pas soluble dans la mesure ?

Cette question demande de s'interroger (3) sur le langage de description que nous choisissons d'utiliser. L'idée ébauchée est que les *nombres* ne permettent pas de saisir le « qualitatif ». Par exemple, Marion Carel (2013) invite à penser que les nombres n'ont pas de valeur argumentative propre (seul, l'énoncé « cet appartement coûte 100 000 € » ne dit rien d'intéressant, parce qu'il ne laisse pas savoir si l'appartement est donc cher ou non). Les *mots*, de leur côté, ne permettent pas des descriptions réellement maîtrisées. Mais alors où situer les langages formels et les mathématiques discrètes ? Les analyses de réseaux sont-elles quantitatives ? Les grammaires de Noam Chomsky sont-elles qualitatives[12] ? Plutôt que de faire des nombres un laissez-passer facile, on pourrait plutôt se demander quel est le rôle des nombres dans la recherche quantitative. Enfin, quelle est la raison pour laquelle on ne les utilise pas dans la recherche qualitative.

Une version plus précise du point précédent (4) présente l'analyse qualitative comme le *résidu* de la quantitative : l'analyse quantitative peut être définie précisément – les données sont représentées numériquement et sont représentatives de la population étudiée, les hypothèses sont définies préalablement et clairement réfutables, les expériences sont dessinées et reproductibles (par exemple, voir Tashakkori, Teddlie, 2010) et ce qui échappe à cette définition est de l'ordre du qualitatif. Jacques Derrida verrait sans doute dans cette opposition hiérarchisée le poids social du quantocentrisme. Cependant (ou alors justement), cette prise de distance finit par protéger le « mystère » du qualitatif de manière extrêmement efficace. En effet, les chercheurs qualitativistes peuvent tenter de montrer que leurs résultats sont représentatifs, réfutables, etc., pour attester du fait qu'ils font du quantitatif (« de la science »). Mais il est beaucoup plus compliqué, pour les chercheurs quantitativistes, de montrer que leurs résultats sont qualitatifs, vu qu'il n'y a pas de critère décisif pour repérer le qualitatif. Le qualitatif n'est défini que comme ce qui n'est pas quantitatif. Comparer le nombre d'adjectifs chez Victor Hugo et chez Émile Zola est-elle une étude qualitative ? Celle-ci dit-elle quelque chose sur ces auteurs ? Avons-nous besoin de mesurer les textes ? Pour répondre, il est nécessaire de donner une définition positive (non résiduelle) du qualitatif.

Apparemment, si c'est vers l'herméneutique qu'il faudrait se tourner pour une définition positive du qualitatif, en définitive, cette discipline n'a fait que supposer (5) une *spécificité irréductible* de l'humain. Pour Wilhelm Dilthey (1883), les raisons de l'interprétation ne peuvent pas être objectivées, la compréhension de l'esprit échappe

[12] « Le *logos*, le discours, c'est toujours du discret ; ce sont des mots entrant dans une certaine succession, mais des mots discrets. Et le discret appelle immédiatement le quantitatif. Il y a des points : on les compte ; il y a des mots dans une phrase : on peut les classer qualitativement par la fonction grammaticale qu'ils occupent dans la phrase, mais il n'empêche qu'il y a une incontestable multiplicité » (Thom, 1991 : 80-81).

à l'explication scientifique. Dans cette perspective, il semble correct d'affirmer qu'un texte d'Émile Zola ne peut ni ne doit être quantifié. Mais on peine à voir le propre du qualitatif : la compréhension reste le *résidu* de l'explication, résidu hypostatisé et glorifié en esprit. Le seuil entre sciences de la nature et sciences de l'homme semble tenir seulement par habitude.

Le mathématicien structuraliste René Thom (1972, 1980) suggérait plutôt (6) d'opposer des *différences* qualitatives et quantitatives. Les différences qualitatives sont des discontinuités observables dans un système, tandis que les différences quantitatives sont des « petites » variations qui n'influencent pas le passage d'un état à l'autre. La morphogenèse est le passage (souvent abrupt) d'une forme qualitative à une autre, sur la base de l'accumulation de différences quantitatives. On pourrait dire que les études qualitatives voient le tas là où les quantitativistes comptent les grains de sable. La définition formelle de René Thom relève de la topologie : nous pouvons faire varier les mesures d'une figure (la longueur des côtés d'un triangle) sans en changer de type (le fait qu'il s'agisse d'un triangle). Or, cette définition de *qualitatif* prend en compte trop de choses, incluant même les phénomènes physiques – une avalanche serait alors un parfait exemple de phénomène qualitatif. Il faudrait limiter les formes que nous voulons appeler « qualitatives » à ce qui relève de l'homme et du social[13]. Avec un exemple linguistique, il y a un seuil qualitatif (*emic*) entre *p* et *b*, qui différencie ces deux phonèmes, en-deçà de toute description quantitative des petites nuances dans leur prononciation par des « parlants » différents (formants *etic*[14]). Cet exemple classique rappelle que toute description qualitative en linguistique demande la prise en compte du *sens* de ce qui est dit. Pour différencier *p* et *b*, on doit saisir qu'il y a une différence de sens qui sépare *poulet* et *boulet*, à travers une épreuve de commutation. Avec ce que nous appelons « analyse qualitative », nous voulons en réalité prendre en compte les seules *formes signifiantes*, douées de sens pour les humains.

Signifiant et significatif

Enfin, si nous cherchons à donner une définition positive de l'analyse qualitative, nous sommes obligé de faire référence à la dimension sémiotique de l'expérience humaine. Les définitions résiduelles de l'analyse qualitative ne saisissent pas le fait que toute recherche qualitative porte sur le *sens*. La densité dont parle Clifford Geertz (1973b) est une *densité signifiante*, qui prend en compte la richesse sémantique des pratiques humaines. C'est cette richesse que nous visons en SHS. Hypostatiser

[13] R. Thom (1991 : 132) reprend la « carte de Tendre » pour dessiner une « carte du sens ». Sur cette carte, les disciplines sont situées dans la continuité, organisées par une quête générale de connaissance qui est le sens : « Ce qui limite le vrai, ce n'est pas le faux, c'est l'insignifiant ».

[14] « Les formants sont les éléments minimaux composant les suites transformationnelles terminales (obtenues par l'application de toutes les transformations aux suites syntagmatiques terminales) ». Accès : http://www.universalis.fr/encyclopedie/formant-linguistique/.

la dimension sémantique, comme le fait l'herméneutique, serait une erreur : nous risquons de présupposer ce que nous voulons expliquer en introduisant un « mystère de l'homme ». Si, au contraire, nous ne voulons pas introduire l'hypothèse selon laquelle l'homme aurait une spécificité irréductible (âme, esprit), on doit accepter l'idée qu'il y a du qualitatif – du sens – tout simplement *là où des signes sont maniés à la manière des humains*. Pas de seuil essentiel entre quantitatif et qualitatif donc, mais continuité et pragmatisme. Justement, la sémiotique définit ce qui est *signifiant* par rapport aux signes publics et aux règles qui les régissent (pas de langage privé, pas d'introspection, pas de monologue intérieur, « rien de caché »). En définitive, pour la sémiotique, l'étude du sens passe par *une étude des expressions*, de leurs règles de combinaison et de transformation dans d'autres expressions : « pure extériorité » (Foucault), « pure surface » (Deleuze). L'homme peut receler des mystères sans que, pour autant, il y ait un mystère de l'homme.

Si qualitatif veut vraiment dire « à la manière des humains », toute analyse sur des données issues de pratiques signifiantes est *déjà* potentiellement qualitative. Si l'analyse qualitative est une description soignée de données, comme le suggère Franck Rebillard, alors tout, dans ses résultats, est potentiellement objectivable et peut devenir donnée. Il n'y a donc aucune discontinuité entre qualitatif et quantitatif, juste le besoin d'identifier et de codifier des procédés pour passer d'un côté à l'autre. Et l'ordinateur est l'instrument idéal pour travailler avec des descriptions denses[15]. Que se passe-t-il si l'on s'approprie les instruments de l'analyse des données pour traiter les signes dans leur dimension signifiante ? L'interprétation a-t-elle besoin de travailler sur les textes originaux, fussent-ils nombreux, pour saisir des contenus, ou peut-elle se limiter à étudier des diagrammes ? Ce qui semble possible aujourd'hui est *une extension des moyens de l'analyse qualitative*, qui passe par une « opérationnalisation » (Moretti, 2015) de ses concepts et par l'utilisation d'instruments statistiques. Cette extension intéresse toutes les SHS, mais, pour la sémiotique, elle est cruciale : il s'agit de se demander *si les données peuvent rester signifiantes après leur traitement statistique*.

La connaissance des statistiques permet de cerner ce qui est significatif. *Significatif* veut dire « pas dû au hasard » : un écart de la norme est significatif si, à travers des calculs, nous pouvons être confiants du fait qu'il est le résultat d'un vrai phénomène à l'œuvre – telle est la définition de la « significativité » donnée par Ronald Fisher (1935 : 15-18) au fondement des statistiques modernes. Notre cerveau n'est normalement pas capable de dire d'emblée si des écarts sont significatifs. De nombreuses études en psychologie montrent que notre raisonnement est limité par des biais. Ceux-ci font suivre des parcours heuristiques, mais ils ne sont pas corrects d'un point de vue logique ou statistique (voir par exemple Guimellli, 1999). C'est la raison pour laquelle nous profitons de l'étude des statistiques dans

[15] Nous ne voulons pas équiparer *big data* et *thick descriptions*, mais indiquer que l'analyse de données permet d'explorer les corrélations entre plusieurs niveaux de description d'un fait social d'une manière impossible autrement. L'analyse de données permet même de repérer des « nouveaux observables » qu'il ne serait pas possible de remarquer autrement (Rastier, 2011).

la pratique scientifique où nous avons besoin de corriger ces biais : parce qu'il n'est pas spontanément visible à nos yeux, il est nécessaire d'apprendre à identifier le significatif. Le monde humain n'est pas « statistiquement solide », mais fondé sur des « réseaux de signifiés » (Geertz, 1973a : 5). La sémiotique étudie justement la tessiture de renvois qui rend le monde humain *signifiant*. Et, tout comme le (statistiquement) significatif n'est pas visible à l'œil nu, pareillement, nous devons apprendre à bien voir et décrire ce qui est (sémiotiquement) signifiant. Nous pouvons profiter de la sémiotique pour apprendre à identifier les variations expressives qui sont corrélées à des variations de contenu. Beaucoup de choses se passent dans notre quotidien qui sont *signifiantes sans être statistiquement significatives*, des signes auxquels nous donnons une importance dans nos interprétations sans que ces interprétations soient des validations correctes d'hypothèses bien définies ou des explications causales. En parallèle, des choses peuvent être *statistiquement significatives sans être signifiantes*, phénomènes qui ne passent pas par le biais de la culture. L'opposition épistémologique entre méthodes qualitatives et quantitatives en SHS semble exacerber la distance entre ces deux perspectives.

Transformer plus pour comprendre mieux

Nous avons plaidé pour une approche de l'analyse de données en SHS capable de combiner compétences statistiques et sémiotiques pour saisir le sens des phénomènes culturels. Il reste à voir quelle est la spécificité méthodologique de cette approche. Les outils informatiques aident l'interprétation quand ils servent à opérer une *transformation des expressions*. Le contenu des expressions est (jusqu'aujourd'hui) inaccessible à la machine ; cependant, les logiciels peuvent travailler le versant expressif des mots et des autres signes. Les textes se décomposent et recomposent selon des traits (*features*) accessibles à la machine. En transformant les expressions, on produit des *diagrammes* interprétables à partir de corpus que l'analyste ne peut souvent pas saisir dans sa globalité pour des raisons de taille ou de complexité. Comment lire des milliers d'articles, ou même de *tweets*, en se faisant une idée exacte de leurs ressemblances et différences, en identifiant des structures narratives, en réalisant des regroupements motivés ? Il est nécessaire d'utiliser des « instruments optiques » qui permettent le *distant reading* (« lecture à distance ») des données (Moretti, 2008). Les calculs significatifs opérés par la machine se trouvent donc au milieu du processus, mais les diagrammes résultants sont signifiants comme les textes de départ. Il s'agit des « mêmes données » transformées. Le défi méthodologique de l'analyse quali-quantitative consiste alors à *repérer des familles de transformations des expressions capables de garder, dans les diagrammes résultants, le rapport originaire que ces expressions entretiennent avec leurs contenus, c'est-à-dire leur sens*[16]. En d'autres termes, il faut apprendre à

[16] D. Mayaffre (2008 : 61) écrit que « le traitement lexicométrique s'appuie toujours sur l'expression, sur des formes bien définies, et ne saurait compter autre chose que des réalités linguistiques matérielles ou

transformer les expressions sans affecter la fonction sémiotique. Et si une partie de la fonction sémiotique doit forcément varier (c'est-à-dire perdre une partie du sens), au moins les contenus pertinents doivent rester constants. L'analyse quali-quantitative est donc un « art » de la transformation topologique, qui déforme les métriques sans toucher aux relations qualitatives. À travers des isomorphismes, on vise à rendre visible l'organisation de contenus responsable des discontinuités et des régularités expressives dans les corpus (voir Valette, 2008 ; Garric, Longhi, 2012). Cette compétence n'a rien à voir avec le fait de savoir conduire des entretiens ou le savoir-faire du terrain[17] ; elle est peut-être plus proche des analyses structurales des mythes et des récits, avec leurs schèmes et formules.

Pour Umberto Eco (1975 : 166), la production de diagrammes signifiants opère par *ratio difficilis*, car ceux-ci sont « motivés par l'organisation sémantique de leur contenu ». Sans prendre en compte l'organisation sémantique, les statistiques ne font qu'une description de la seule expression (comme si l'on se demandait si les mots anglais sont moyennement plus courts que les mots français). Au contraire, les diagrammes qui prennent en compte la dimension signifiante de leur objet « sont mis en corrélation avec certains aspects du sémème qui leur correspond, devenant par là même des expressions, dont les traits sont également des traits sémantiques, à la fois marques sémantiques transformées et projetées sur le plan syntactique » (*ibid.*). Le sémioticien précise qu'« un ordinateur qui aurait reçu des instructions pour produire ces objets [les diagrammes] devrait aussi avoir reçu des instructions d'ordre sémantique » (*ibid.*). On ne peut pas (encore) donner directement des instructions sémantiques : il faut utiliser l'ordinateur comme un « instrument optique » qui transforme les expressions et c'est au chercheur de s'assurer que la fonction sémiotique ne soit pas perdue au passage.

Un exemple d'analyse : retour sur expérience

Faisons maintenant référence à la dimension méthodologique d'une recherche sur le discours politique sur Twitter (Compagno, 2016, à paraître a). L'objectif de ce retour sur expérience n'est pas de donner une interprétation détaillée des données, pour laquelle nous renvoyons à aux deux contributions mentionnées ci-dessus.

Le corpus analysé contient la totalité des *tweets* partagés par les candidats français, italiens et britanniques pendant la campagne aux élections européennes de 2014. Nous visions la différenciation des « styles » de communication adoptés par les candidats. L'intérêt du corpus réside dans sa dimension internationale et donc sur la possibilité de repérer des tendances à travers les pays. Le corpus a

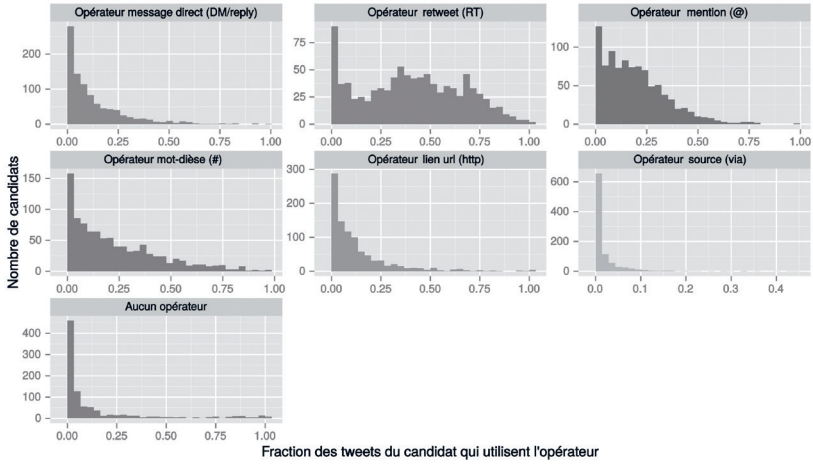
implémentables du texte ». Dans le traitement d'un corpus, l'accès au contenu « est médiatisé par la statistique qui est là pour déceler les associations qui font sens *puisqu'on ne saurait les attribuer au hasard* ».

[17] Les observations ethnologiques et les connaissances implicites qui en dérivent deviennent données et métadonnées quand elles passent à l'analyse et à l'interprétation collectives.

été constitué par l'équipe internationale du projet TEE2014[18], porté par la Maison des sciences de l'homme de Dijon. Un premier travail important a été réalisé pour obtenir des données de qualité en repérant le plus grand nombre de comptes possible à partir des listes de candidats à l'élection, en considérant les différents comptes que pouvaient utiliser certains candidats et ceux créés pour l'occasion. L'extraction et le prétraitement des *tweets* ont été réalisés par un outil conçu à cet effet (Leclercq *et al*., 2016).

Un premier défi consistait à traiter un corpus multilingue. Nous devions identifier des traits pertinents dont l'usage était indépendant de la langue. Une analyse exploratoire a donc été consacrée à la *feature selection* (« sélection de traits »), qui a isolé comme variables les plus prometteuses les opérateurs technolangagiers propres à Twitter : mots-dièse (#), *retweets* (RT), mentions (@), hyperliens (http). Cela entrait en résonance avec des résultats déjà obtenus par d'autres chercheurs de l'équipe (Einspänner, Dang-Anh, Thimm, 2013). Nous avons donc utilisé la présence d'opérateurs pour décrire les *tweets*. La figure 1 montre les distributions d'usage de chaque opérateur par les candidats.

Figure 1. Utilisation des opérateurs par les candidats



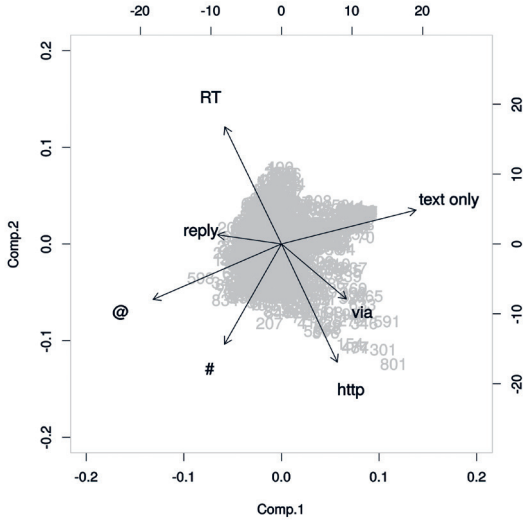
Nous nous sommes trouvés face à deux problèmes. D'abord, est-il possible de dire quelque chose sur le sens d'un *tweet* à partir de la simple présence d'un opérateur ? Des études montrent que l'on peut faire un usage diversifié d'un opérateur ; par

[18] Coordonné par A. Frame (Centre interlangues-texte, image, langage, Université de Bourgogne), le projet « Twitter aux élections européennes : une étude contrastive internationale des utilisations de Twitter par les candidats aux élections au Parlement Européen en mai 2014 » (TEE2014) a été financé entre 2013 et 2015 par l'Institut des sciences humaines et sociales et le Réseau national des maisons des sciences de l'homme.

exemple, la mention peut être utilisée pour citer quelqu'un, pour le viser dans une critique, pour lui adresser une information (Kondrashova, Frame, 2016). La fonction technique se sémiotise justement quand sa potentialité de signification est prise en charge par un sujet énonçant et dirigée pour des buts communicatifs (au sens austinien de « *twitter* c'est faire »). Il semble que si l'on isole le signe qui est l'opérateur, son sens doive se perdre parce qu'on le détache du « texte » qui le rend signifiant, et donc que seule une approche qualitative traditionnelle (avec ses limites) puisse prendre en charge les actes de langages réalisés par les candidats sur Twitter. Second problème, comment interpréter les distributions des usages des opérateurs ? Elles ne sont pas parlantes, il s'agit de diagrammes qui ne gardent pas la relation originaire avec leur contenu (il s'agit d'une transformation « plate » des expressions, voir *supra*, § 5). Même si les opérateurs étaient des signes encore signifiants une fois détachés des textes où ils apparaissaient, une description statistique trop simple n'aurait pas constitué un bon outil pour la recherche de leur sens. Voici donc que nous nous trouvions face à deux critiques très classiques faites aux *digital methods*.

La solution fut trouvée en acceptant ces critiques et en les retournant contre elles-mêmes. Si un signe isolé n'a pas de sens, parce que le sens n'est produit que dans son rapport avec d'autres signes (sur des séries syntagmatiques ou paradigmatiques), alors nous pouvions quantifier *ce rapport*. S'il n'y a de sens que dans la différence, nous pouvions utiliser l'ordinateur pour maîtriser la différence. Plutôt que de travailler avec les opérateurs un par un, nous avons constitué un espace multidimensionnel de représentations où chaque opérateur constituait une dimension. Nous avons ensuite situé chaque candidat dans cet espace, sur la base de son usage des opérateurs. Ce diagramme est plus compliqué à réaliser, mais plus simple à interpréter : il permet de voir le sens des pratiques énonciatives mieux que ne le font de simples distributions (figure 2).

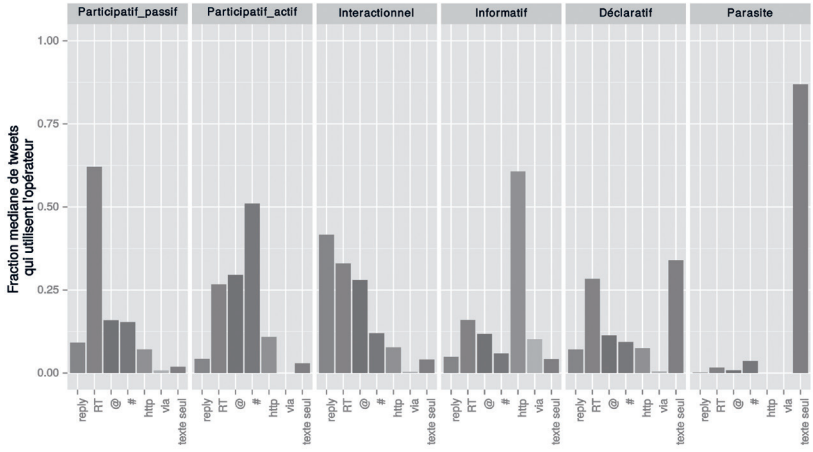
Figure 2. Candidats et opérateurs



Cette démarche commence à montrer que qualitatif et quantitatif ne s'opposent pas (statistiques *versus* sémiotique) ; au contraire, pour faire du qualitatif massif, il faut du quantitatif critique. Chaque candidat – simulacre verbal, fait de mots, rien de plus qu'un sous-corpus – a donc été décomposé à partir d'une sélection de traits pertinents et recomposé comme un point dans un espace complexe. On ne se demande plus si un opérateur a toujours le même sens, on se demande seulement s'il y a des tendances d'usage dans l'ensemble des opérateurs, des idiolectes qui sont des trajectoires à l'intérieur des possibilités offertes par la « grammaire » de Twitter. Il s'agit bien de *distant reading* : nous perdons du détail sans renoncer au sens. Le « qualitatif » n'est alors qu'une attribution de la description : il n'y a pas de qualitatif mais *plus* de qualitatif. Dans notre cas, nous aurions pu faire mieux en repérant dans un petit sous-corpus aléatoire des occurrences typiques de mots associés à certains usages des opérateurs (par exemple, des mots connotés négativement précédant une mention pourraient indiquer une critique envers quelqu'un[19]). Ensuite, en utilisant des algorithmes d'apprentissage automatique, on aurait pu classifier semi-automatiquement le corpus entier. Cela ne serait pas sorti de notre démarche « immanente » et aurait juste demandé des ressources supplémentaires. D'ailleurs, ne pas émettre des hypothèses précises sur l'usage des opérateurs, au moins dans le cadre d'une recherche exploratoire comme la nôtre, permet de repérer plus facilement des inattendus, des nouveaux observables indépendants de nos attentes.

L'étape suivante a consisté en la segmentation de l'espace de représentation en régions, c'est-à-dire des regroupements ou *clusters* de candidats. Chaque *cluster* inclut un nombre variable de candidats « proches », métaphoriquement et statistiquement, dans leur utilisation de Twitter. Nous pouvons décrire le profil médian de chaque *cluster* (figure 3).

Figure 3. Profils des regroupements

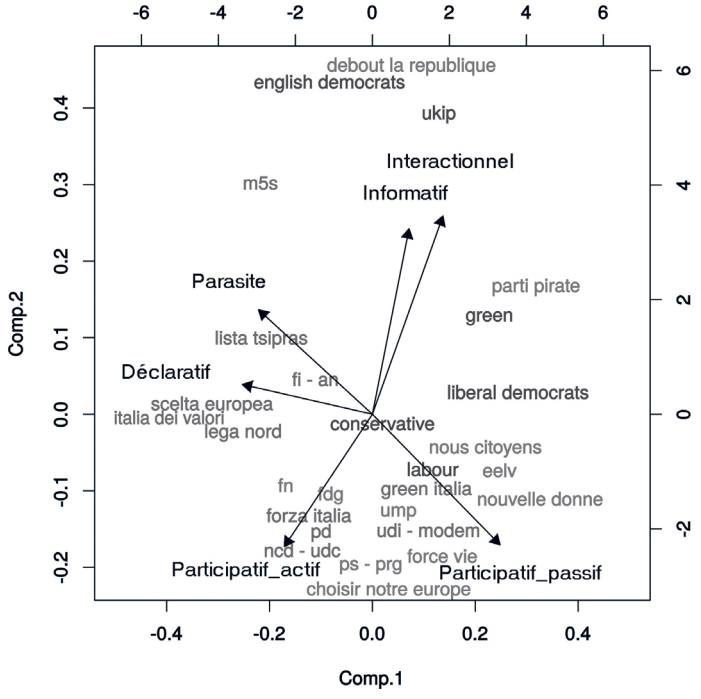


[19] Pour une analyse détaillée des fonctions des *hashtags*, voir A. Jackiewicz (2016, et sa contribution dans le présent dossier) et A. Mercier (à paraître).

Chaque *cluster* est caractérisé par l'usage prépondérant d'un opérateur. Cette information a constitué la base d'interprétation des *clusters*, c'est-à-dire en leur donnant un nom[20]. Cependant, seule, cette information n'est pas suffisante : pour bien interpréter, nous avons cherché d'autres informations (toujours « immanentes » à nos données), tels les logiciels pour *tweeter* utilisés de préférence par les candidats de chaque *cluster*. Nous avons même lu quelques exemples de *tweets*. Par exemple, c'est de cette manière que nous avons compris qu'un *cluster* (que nous avons nommé « Parasite ») était constitué de *tweets* produits de manière automatique et « frustre » par une application, n'affichant que des liens vers des contenus sur Facebook.

Le regroupement statistique est une procédure tout à fait *structurale*. Les *clusters* n'ont aucune identité substantielle ; pures formes, ils sont le produit de l'identification de seuils, d'écarts le plus possible aptes à segmenter des groupes homogènes à l'intérieur d'eux-mêmes et hétérogènes l'un par rapport à l'autre. Nous pouvons utiliser une analyse en composantes principales pour simplifier l'espace de représentation et visualiser le système tensif (au sens de Fontanille, Zilberberg, 1998) produit par les opérateurs (figure 4).

Figure 4. Regroupements et partis



[20] Quand il énonce que la nomination est l'opération sémiotique la plus basale et nécessaire, R. Barthes (1970 : § 11) permet de sortir du domaine des nombres pour entrer dans celui du sens.

Voici un diagramme parlant, que nous pouvons tenter d'interpréter au sens donné à ce terme par Umberto Eco dans la continuité de Charles S. Peirce et de l'herméneutique. Il signifie donner un sens à des éléments et relations devant nous (« immanents ») sur la base de notre connaissance du monde. Il faut reconnaître l'apport du chercheur pour ne pas passer sous silence les présupposés de l'analyse, conditions aussi importantes que l'implémentation d'algorithmes. D'ailleurs, il n'y a pas d'algorithme neutre, et une stratégie de traitement est déjà, comme toute stratégie, une série de choix. Le diagramme existe parce qu'il arrive à la fin d'une série de productions sémiotiques s'appuyant l'une sur l'autre. Sa lisibilité réside dans l'histoire de sa production. Ce diagramme n'a de sens que pour un *lecteur modèle*, capable de comprendre les procédures qui ont conduit à sa constitution (il s'agit d'une image technique qui prévoit une lecture experte, Fontanille, Dondero, 2012), et capable de bien situer et « faire vivre » les termes affichés. Les noms des *clusters* ne sauraient être neutres ; les noms des partis se réfèrent à une société vivante qui est montrée et non masquée par les traces de son discours. Ce diagramme ne dit des choses qu'à partir de nos compétences et attentes.

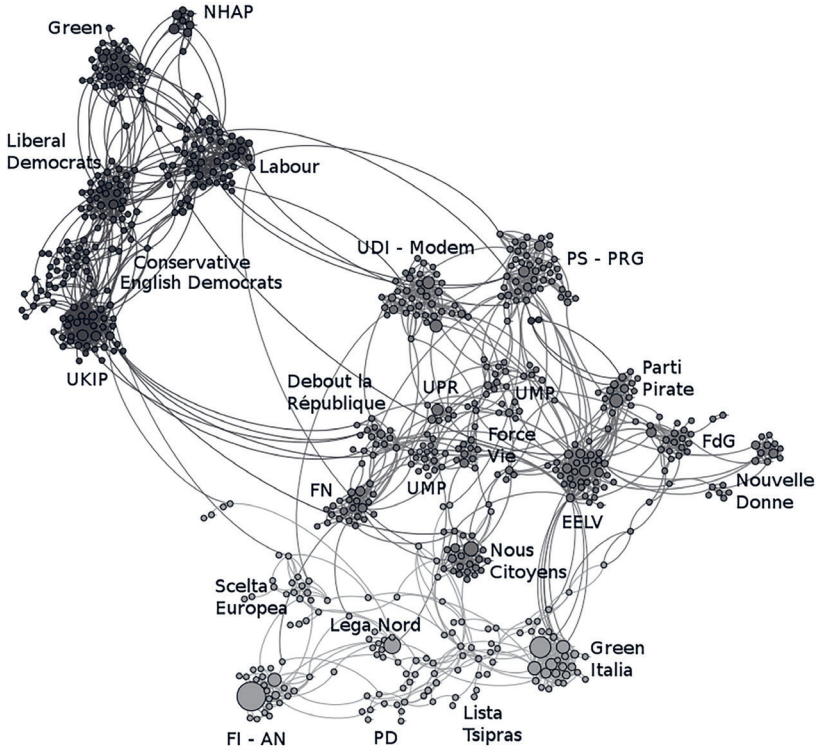
Par exemple, certains partis italiens (le Mouvement 5 étoiles, Lista Tsipras, Forza Italia – Alleanza Nazionale) font un usage très basique ou même parasite de Twitter, sans utiliser des opérateurs et en s'appuyant massivement et sans soins sur des outils automatisés d'éditorialisation. D'autres partis italiens et les partis français occupent une région différente du diagramme : ils acceptent l'usage majoritaire de Twitter en employant des mots-dièse et en partageant les messages d'autrui. Différence majeure transalpine : les Italiens utilisent plus de mots-dièse et les Français *retweettent* plus. Les Italiens visent des mini-publics (communautés hétérogènes faites de candidats et de citoyens) en cherchant à remonter dans la liste des *trending topics*[21] ; d'ailleurs, on ne peut pas nier que les *hashtags* sont un opérateur assez bruyant, petits cris (*gridolini*) lancés de derrière des cages-dièse. Les Français donnent davantage voix à d'autres usagers – souvent des membres du même parti – et, de cette manière, constituent d'autres types de communautés en se donnant plus d'importance l'un l'autre (la citation est toujours un peu narcissique). Le diagramme montre que chaque pays suit des normes d'usages relativement homogènes et que, en même temps, on peut repérer des tendances internationales (les partis les plus « innovants », tels le Parti pour l'indépendance du Royaume-Uni – UKIP – et le Mouvement 5 étoiles, occupent la partie supérieure du diagramme, éloignés de l'usage normatif du médium).

Conclusion

Notre analyse s'arrêta là. Cependant, n'aurait-il pas été possible d'aller plus loin à la recherche d'un sens encore plus dense ? Les limites d'une étude dépendent beaucoup de l'analyste et ne relèvent pas automatiquement des défauts de la méthode employée. Par exemple, nous pourrions étendre le travail déjà fait avec une nouvelle image, qui montrerait les liens de *retweets* et de mentions parmi les candidats des trois pays (figure 5).

[21] Les *trending topics* sont des termes les plus utilisés à un certain moment par les usagers.

Figure 5. Liens de *retweets* et mentions entre les candidats anglais, français et italiens



Dans cette nouvelle image, chaque point est un candidat, les liens sont des *retweets* ou des mentions, les gradations de gris délimitent les pays et la taille des nœuds représente le nombre de liens reçus par chaque candidat. On pourrait y voir de nombreuses choses. D'abord, que les partis « font *cluster* » et distribuent presque de manière égalitaire l'attention parmi tous leurs membres. Cela est moins vrai pour l'Italie où très peu de partis émergent et, souvent, autour d'un seul leader. On voit aussi qu'un espace de discussion européen existait. France et Italie sont rapprochées et on reconnaît le clivage gauche-droite. Paradoxalement, ce qui relie la France et la Grande-Bretagne, ce sont surtout les partis anti-européens. Parmi d'autres éléments remarquables, le centre droit français (l'Union pour un mouvement populaire – UMP) et le centre gauche italien (Partito democratico – PD) sont dispersés, le premier divisé mais central, le dernier presque sur le point de disparaître.

On pourrait aller encore plus loin dans l'analyse : différencier *retweets* et mentions, repérer les liens vers des classes de citoyens comme les journalistes, suivre la temporalité des partages, etc. Une interprétation plus profonde demanderait aussi de convoquer les compétences interprétatives d'experts du discours politique européen. Notre but était simplement d'exemplifier l'utilité des méthodes quali-quantitatives et de mettre en avant la spécificité de leur dimension épistémologique – combinant

sémiotique et statistiques. Les résultats de notre analyse sont provisoires et révisables. Mais quelle analyse ne l'est pas ? Pour autant, nous ne croyons pas que ses limites sont intrinsèques aux méthodes numériques ; au contraire, ces dernières permettent une exploration sémantique à grande échelle, impossible avec les méthodes qualitatives ou quantitatives traditionnelles.

Nous avons présenté des raisons pour prendre en compte la sémiotique comme complément nécessaire des statistiques pour l'analyse de données signifiantes. Nous nous sommes concentré sur la dimension épistémologique des nouvelles méthodes quali-quantitatives qui forment bien plus que la somme des méthodologies qualitatives et quantitatives traditionnelles. En particulier, nous avons mis l'accent sur la complémentarité entre significativité statistique et signification sémiotique, montrant que les deux doivent être prises en compte pour étudier le sens à grande échelle.

Références

Barthes R., 1970, *S/Z*, Paris, Éd. Le Seuil.

Bastide F., 2001, *Una notte con Saturno. Scritti semiotici sul discorso scientifico*, éd. par B. Latour, Rome, Meltemi.

Bouveresse J., 1987, *Le Mythe de l'intériorité*, Paris, Éd. de Minuit.

Carel M., 2013, « L'argumentation dans la langue », *in* : Carel M., Ducrot O., dirs, séminaire *Sémantique des langues naturelles*, Paris, École des hautes études en sciences sociales, 18 nov.

Chomsky N., 1957, *Structures syntaxiques*, trad. de l'anglais par M. Braudeau, Paris, Éd. Le Seuil, 1969.

Compagno D., 2016, « Families of practices. A bottom-up approach to differentiate how French candidates made use of Twitter during the 2014 European Campaign », pp. 33-52, *in*: Frame A., Mercier A., Brachotte G., Thimm C., eds, *Tweets from the Campaign Trail. Researching Candidates' Use of Twitter During the European Parliamentary Elections*, Berne, P. Lang.

Compagno D., à paraître a, « Les candidats européens aux prises avec Twitter. Une comparaison des styles d'usage en France, Italie et Royaume-Uni », *in* : Brachotte G., Frame A., dirs, *Twitter aux Élections Européennes 2014 : regards internationaux sur un outil de communication politique*, Paris, Éd. L'Harmattan.

Compagno D., dir., à paraître b, *Quantitative Semiotic Analysis*, Berlin, Springer.

Compagno D., Rebillard F., 2016, « La médiatisation de Mohammed Merah en mars 2012. Jalons pour une analyse transmédiatique appuyée sur des données numériques », pp. 21-40, *in* : Touboul A., Hare I., Rampon J.-M., Tétu J.-F., dirs, *Informer avec Internet. Reprises et métamorphoses de l'information*, Besançon, Presses Universitaires de Franche-Comté.

Dilthey W., 1883, *Introduction aux sciences de l'esprit*, trad. de l'allemand par S. Mesure, Paris, Éd. du Cerf, 1992.

Derrida J., 1967, *La Voix et le phénomène*, Paris, Presses universitaires de France.

Dondero M. G., Fontanille J., 2012, *Des images à problèmes . Le sens du visuel à l'épreuve de l'image scientifique*, Limoges, Presses universitaire de Limoges.

Doueihi M., 2008, *La Grande Conversion numérique*, Paris, Éd. Le Seuil.

Eco U., 1973-1975, *Le Signe*, trad. de l'italien par J.-M. Klinkenberg, Paris, Éd. Le Livre de poche, 1992.

Eco U., 1975, *La Production des signes*, trad. de l'italien par M. Bouzaher, Paris, Éd. Le Livre de poche, 1992.

Eco U., 1979, *Lector in fabula*, trad. de l'italien par M. Bouzaher, Paris, Grasset, 1985.

Einspänner J., Dang-Anh M., Thimm C., 2013, « Computer-assisted content analysis of Twitter data », pp. 97-108, *in*: Bruns A., Weller K., Burgess J., Mahrt M., Puschmann C., eds, *Twitter and Society*, New York, P. Lang.

Fisher R. A., 1935, *The Design of Experiments*, Edinburgh, Oliver and Boyd.

Fontanille J., 2007, « Textes, objets, situations et formes de vie. Les niveaux de pertinence du plan de l'expression dans une sémiotique des cultures », pp. 213-240, *in* : Bertrand D., Costantini M., Dambrine S., Alonso J., dirs, *Transversalité du sens. Parcours sémiotiques*, Paris, Presses universitaires de Vincennes.

Fontanille J., 2008, *Pratiques sémiotiques*, Paris, Presses universitaires de France.

Fontanille J., Zilberberg C., 1998, *Tension et signification*, Liège, Éd. Mardaga.

Frame A., Mercier A., Brachotte G., Thimm C., eds, 2016, *Tweets from the Campaign Trail. Researching Candidates' Use of Twitter During the European Parliamentary Elections*, Berne, P. Lang.

Garric N., Longhi J., dirs, 2012, « L'analyse de corpus face à l'hétérogénéité des données », *Langages*, 187.

Geertz C., 1973a, *The Interpretation of Cultures*, New York, Basic.

Geertz C., 1973b, « La description dense. Vers une théorie interprétative de la culture », trad. de l'anglais par A. Mary, *Enquête*, 6, pp. 73-105, 1998.

Gefen A. 2015, « Les enjeux épistémologiques des humanités numériques », *Socio*, 4, pp. 61-74.

Guimelli C., 1999, *La Pensée sociale*, Paris, Presses universitaires de France.

Hey T., Tansley S., Tolle K., eds, 2009, *The Fourth Paradigm. Data Intensive Scientific Discovery*, Redmond, Microsoft.

Hjelmslev L., 1943, *Prolégomènes à une théorie du langage*, trad. du danois par A.-M. Léonard, Paris, Éd. de Minuit, 1968.

Jackiewicz A., 2016, « Les outils multimodaux de Twitter comme moyens d'expression des émotions et des prises de position », *Cahiers de praxématique*, 66. Accès : http://praxematique.revues.org/4247.

Jost F., 2001, *La Télévision du quotidien. Entre réalité et fiction*, Bruxelles/Paris, De Boeck/Institut national de l'audiovisuel.

Kondrashova T., Frame A., 2016, « Exploring the dialogical dimension of political tweets: A qualitative analysis of "Twitter styles" of UK candidates during the 2014 EU Parliamentary Elections », pp. 53-74, *in*: Frame A., Mercier A., Brachotte G., Thimm C., eds, *Tweets from the Campaign Trail. Researching Candidates' Use of Twitter During the European Parliamentary Elections*, Berne, P. Lang..

Lebart L., Salem A., 1994, *Statistiques textuelles*, Paris, Dunod.

Leclercq E., Savonnet M., Grison T., Kirgizov S., Basaille I., 2016, « SNFreezer: a Platform for Harvesting and Storing Tweets in a Big Data Context », pp. 19-32, *in*: Frame A., Mercier

A., Brachotte G., Thimm C., eds, *Tweets from the Campaign Trail. Researching Candidates' Use of Twitter During the European Parliamentary Elections*, Berne, P. Lang.

Le Hay V., Vedel T., Chanvril F., 2011, « Usages des médias et politique : une écologie des pratiques informationnelles », *Réseaux. Communication, technologie, société*, 170, pp. 45-73.

Mayaffre D., 2008, « De l'occurrence à l'isotopie. Les co-occurrences en lexicométrie », *Syntaxe et Sémantique*, 9, pp. 53-72.

Mayaffre D., 2010, *Vers une herméneutique matérielle numérique. Corpus textuels, logométrie et langage politique*, mémoire d'habilitation à diriger des recherches, Université Nice Sophia Antipolis.

Mercier A., à paraître, « Hashtags. Tactiques de partages et de commentaires d'informations », *in* : Mercier A., Pignard-Cheynel N., dirs, *Partager et commenter l'info sur les réseaux sociaux*, Paris, Éd. de la Maison des sciences de l'homme.

Moretti F., 2008, *Graphes, cartes et arbres*, trad. de l'anglais par É. Dobenesque, Paris, Éd. Les Prairies ordinaires.

Moretti F., 2015, « L'opérationnalisation ou, du rôle de la mesure dans la théorie littéraire moderne », *Critique*, 819-820, pp. 712-734.

Odin R., 1983, « Pour une sémio-pragmatique du cinéma », *Iris*, 1 (1), pp. 67-82.

Odin R., 2000, « La question du public. Approche sémiopragmatique », *Réseaux. Communication, technologie, société*, 99, pp. 49-72.

Peirce C. S., 1993, *À la recherche d'une méthode*, trad. de l'anglais par M. Balat et J. Deledalle-Rhodes, Perpignan, Presses universitaires de Perpignan.

Rastier F., 1987, *Sémantique interprétative*, Paris, Presses universitaires de France.

Rastier F., 1989, *Sens et textualité*, Paris, Hachette.

Rastier F., 2011, *La Mesure et le grain*, Paris, H. Champion.

Rebillard F., 2011, « L'étude des médias est-elle soluble dans l'informatique et la physique ? À propos du recours aux *digital methods* dans l'analyse de l'information en ligne », *Questions de communication*, 20, pp. 353-376.

Rogers R., 2013, *Digital Methods*, Cambridge, MIT Press.

Tashakkori A., Teddlie C., eds, 2010, *Sage handbook of mixed methods in social and behavioral research*, Los Angeles, Sage.

Thom R., 1972, *Stabilité structurelle et morphogenèse*, Paris, InterÉditions.

Thom R., 1980, « Qualità/Quantità », pp. 470-476, *in*: Romano R., dir., *Enciclopedia Einaudi*, vol. XI, Turin, G. Einaudi.

Thom R., 1991, *Prédire n'est pas expliquer*, Paris, Eshel.

Valette M., dir., 2008, « Textes, documents numériques, corpus. Pour une science des textes instrumentée », *Syntaxe et Sémantique*, 9.

Violi P., 1997, *Meaning and Experience*, trad. de l'italien par J. Carden, Bloomington, Indiana University Press, 2001.

Wickham H., 2014, « Tidy Data », *Journal of Statistical Software*, 59 (10). Accès : https://www.jstatsoft.org/article/view/v059i10.

Wittgenstein L., 1933-1935, *Le Cahier bleu et le Cahier brun*, trad. de l'anglais par G. Durand, Paris, Gallimard, 1988.