De la fouille de données à la fabrique de l'opinion. Enjeux épistémologiques et propositions

Egle Eensoo-Ramdani, Evelyne Bourion, Monique Slodzian, Mathieu Valette

DE LA FOUILLE DE DONNÉES À LA FABRIQUE DE L'OPINION

Enjeux épistémologiques et propositions

http://lcn.revuesonline.com

EGLE EENSOO-RAMDANI

EVELYNE BOURION

MONIQUE SLODZIAN

MATHIEU VALETTE

Dans cet article nous proposons une réflexion sur les enjeux épistémologiques et l'histoire de la fouille d'opinion. À l'aide d'exemples pris dans des recherches en cours, nous illustrons la situation de cette pratique dans le cadre de la théorie de l'information. Nous nous intéressons particulièrement à la question de l'identification et de la restitution de l'émetteur porteur d'opinion, de ses valeurs et de son mode d'expression à partir de l'analyse des textes. Des pistes issues de théories énonciatives et textuelles sont proposées pour une amélioration des performances de l'analyse de la subjectivité.

In this paper we offer our views on the epistemological background and the history of opinion mining. Using examples taken from current research, we

VERSION SOUMISE AUX *CAHIERS DU NUMERIQUE* LE 18 VI 2011. MERCI DE SE RÉFÉRER A LA VERSION PUBLIÉE : Eensoo-Ramdani, Egle, Evelyne Bourion, Monique Slodzian, Mathieu Valette (2011) « De la fouille de données à la fabrique de l'opinion. Enjeux épistémologiques et propositions », *Analyse d'opinions sur internet*, Luc Grivel, éd., *Les Cahiers du Numérique*, Volume 7, n°2, pp. 15-39.

VERSION SOUMISE AUX *CAHIERS DU NUMERIQUE* LE 18 VI 2011. MERCI DE SE RÉFÉRER A LA VERSION PUBLIÉE : Eensoo-Ramdani, Egle, Evelyne Bourion, Monique Slodzian, Mathieu Valette (2011) « De la fouille de données à la fabrique de l'opinion. Enjeux épistémologiques et propositions », *Analyse d'opinions sur internet*, Luc Grivel, éd., *Les Cahiers du Numérique*, Volume 7, n°2, pp. 15-39.

show the situation of the practice in the context of information theory. We focus on the issue of identification and restitution of the opinion sender and his/her values based on text analysis. Enunciative and text theories offer some clues to enhance performance of subjectivity analysis efficiency.

VERSION SOUMISE AUX *CAHIERS DU NUMERIQUE* LE 18 VI 2011. MERCI DE SE RÉFÉRER A LA VERSION PUBLIÉE : Eensoo-Ramdani, Egle, Evelyne Bourion, Monique Slodzian, Mathieu Valette (2011) « De la fouille de données à la fabrique de l'opinion. Enjeux épistémologiques et propositions », *Analyse d'opinions sur internet*, Luc Grivel, éd., *Les Cahiers du Numérique*, Volume 7, n°2, pp. 15-39.

Introduction

L'intérêt massif pour la fouille d'opinion est directement lié à une demande sociale forte, à savoir l'expansion fulgurante du commerce en ligne. Comment accéder aux jugements privés relatifs à tel ou tel produit afin de mieux anticiper les besoins et mieux évaluer l'impact du marketing visant tel ou tel segment de consommateurs ? On a vu ainsi émerger un nouveau champ de recherches dont la vigueur est attestée par la multitude d'articles et d'événements s'en réclamant. Pour résumer, la détection d'opinion vise à déterminer les caractéristiques des opinions, positives ou négatives, relatives à un objet particulier. Cependant, l'avalanche de travaux suscités par cet enjeu économique majeur a diffracté le champ « détection d'opinion » en lignes enchevêtrées, relevant d'une anthropologie en ligne à la frontière de plusieurs disciplines, de sorte que des notions distinctes issues de la philosophie analytique comme celle de « private states » (Quirk, 1985), de la psychologie comme *analyse des sentiments* ou encore l'opposition entre émotif et cognitif se trouvent recyclées de façon équivoque. Ainsi, l'équation entre *opinion*, *sentiment* ou *jugement d'évaluation* est-elle manifestement abusive. Parle-t-on des opinions comme des représentations subjectives et non raisonnées, auquel cas la notion est antithétique avec celle d'opinion publique, supposée refléter une rationalité pratique et une exigence de contrat social ? On s'interrogera sur le changement de paradigme opéré entre la notion d'opinion publique au sens de croyance conventionnelle et d'opinion individuelle dans le cadre d'un individualisme atomistique, tel qu'il est façonné par l'économisme libéral (Rawls, 1971). De quels individus parle-t-on ?

Par ailleurs, l'impact technologique et économique de la problématique ne signifie pas qu'elle soit nouvelle. N'oublions pas que la production d'opinions par sondage a précédé la fouille d'opinion sur le Web. Les critiques dont la première a fait l'objet sont-elles susceptibles d'éclairer certains questionnements relatifs à la seconde ?

L'instabilité épistémologique qui affecte le domaine suggère que l'on procède d'abord une clarification des concepts rencontrés. Ce travail d'élucidation présuppose que l'on aborde succinctement l'arrière-plan historique d'où ont émergé certains des concepts philosophiques en filigrane, comme l'opposition fait/valeur, développée par la tradition positiviste. Ce sera le point de départ de la partie consacrée aux enjeux épistémologiques (parties 2.1 et 2.2). À la suite de ces clarifications, nous nous interrogerons sur l'objet d'étude du champ fouille d'opinion (partie 2.3). Cela nous conduira à examiner les continuités et discontinuités entre sondage et fouille d'opinion par rapport

VERSION SOUMISE AUX *CAHIERS DU NUMERIQUE* LE 18 VI 2011. MERCI DE SE RÉFÉRER A LA VERSION PUBLIÉE : Eensoo-Ramdani, Egle, Evelyne Bourion, Monique Slodzian, Mathieu Valette (2011) « De la fouille de données à la fabrique de l'opinion. Enjeux épistémologiques et propositions », *Analyse d'opinions sur internet*, Luc Grivel, éd., *Les Cahiers du Numérique*, Volume 7, n°2, pp. 15-39.

au statut de l'individu (partie 2.4). La partie 3 sera consacrée aux études de cas et propositions méthodologiques. On s'appliquera à montrer la nécessité d'intégrer le cadre énonciatif dans l'analyse préalable à la détection d'opinion (partie 3.1). Plus précisément, on s'intéressera à la caractérisation de l'émetteur, pièce centrale de l'énonciation (partie 3.2.), ainsi qu'aux normes d'élaboration de textes pour lesquels il a opté (parties 3.3 et 3.4). Pour finir, on s'arrêtera sur la question de la subjectivité des concepts et de leur lexicalisation (partie 3.4).

La discussion qui sera poursuivie tout au long de l'article puisera dans l'état de l'art sans prétendre à l'exhaustivité. Les propositions des différents auteurs qui traitent du domaine depuis une dizaine d'années indiquent des pistes que nous essaierons d'évaluer en liaison avec les impératifs qu'il nous semble fondé de mettre en avant. Par ailleurs, nous illustrerons notre propos avec des exemples tirés des projets relevant de la détection d'opinion déjà réalisés, ou en cours de réalisation par notre équipe.

Enjeux épistémologiques

Nous proposons une esquisse de la préhistoire de l'anthropologie en ligne qui servira de cadre à la discussion des principaux concepts fondateurs du champ de la fouille d'opinion, tels que fait/valeur, opinion privée/publique, individualité, etc., informés par l'état de l'art. Nous serons amenés ainsi à examiner l'impact de deux courants épistémologiques majeurs sur la problématique de l'opinion : le positivisme logique d'une part, et le pragmatisme anglo-saxon de l'autre.

Fait-valeur : une dualité centrale pour le positivisme logique

L'idéal d'un langage pour la science tel qu'il a été longuement élaboré par le positivisme logique[1], partait du principe d'économie qui tend à éliminer tout ce qui n'est pas représentable et à ne représenter par des signes que les seuls objets. Ceci explique la centralité persistante de la dénomination et la

1. Le Cercle de Vienne représenté par Carnap, Schlick, Wittgenstein notamment, du début des années 20 à la guerre, entend imposer la primauté de la logique et de l'ontologie sur la théorie de la connaissance. Le positivisme logique, appelé encore empirisme logique, connaîtra un grand rayonnement, en particulier dans les pays anglo-saxons où les pères fondateurs trouvèrent refuge à la venue du nazisme. Russell, Quine, Tarski en sont des membres éminents. Les questions de philosophie du langage sont au cœur de leur problématique (définition sémantique du Vrai).

VERSION SOUMISE AUX *CAHIERS DU NUMERIQUE* LE 18 VI 2011. MERCI DE SE RÉFÉRER A LA VERSION PUBLIÉE : Eensoo-Ramdani, Egle, Evelyne Bourion, Monique Slodzian, Mathieu Valette (2011) « De la fouille de données à la fabrique de l'opinion. Enjeux épistémologiques et propositions », *Analyse d'opinions sur internet*, Luc Grivel, éd., *Les Cahiers du Numérique*, Volume 7, n°2, pp. 15-39.

prééminence de la terminologie dans l'ensemble des applications relevant de l'ingénierie des connaissances.

Par voie de conséquence, le projet d'une synthèse de la science confié à la logique formelle excluait les questions de métaphysique – à commencer par l'éthique et la psychologie –, au nom du principe de vérifiabilité (Carnap, 1928). Il s'agissait de s'affranchir de l'individualité pour garantir la vérité. C'est ainsi que les jugements de valeurs, affirmations non factuelles, subjectives et non représentables ont été délibérément écartés, la forme logique devant être un pur reflet de la structure des faits. Dans l'énoncé « Jean aime Marie » l'enjeu n'est pas de détecter le sentiment de Jean pour Marie. Ici, il ne s'agit pas d'investigation d'un « private state ».

En quoi la problématique fait/valeur trouve-t-elle un prolongement dans le débat actuel sur l'opinion ? Considérer la tâche de détection de la subjectivité comme autonome par rapport à la détection d'opinion elle-même revient à assumer une option logiciste puisqu'il s'agit de séparer les faits des opinions. Autrement dit, dans maints travaux, on suppose que les segments linguistiques de longueur et de structure différentes peuvent être classés *a priori* dans l'une ou l'autre des catégories – fait ou valeur – et que seuls les segments véhiculant des valeurs sont pertinents pour l'analyse. En effet, différentes expériences (Mihalcea *et al.,* 2007) ont montré un gain de performance sur la détection de polarité d'opinion si une détection des passages subjectifs a été effectuée préalablement. Cependant, un tel mode de catégorisation nous paraît problématique. Considérons un exemple tiré du domaine des critiques d'ordinateurs portables (Pang et Lee, 2008) : le syntagme « long battery life », considéré *a priori* comme factuel, assigne en fait une valeur positive à l'ordinateur, valeur contenue contenue dans la dénomination puisque la durée d'autonomie de la batterie est parmi les critères positifs importants pour cet outil technique. Il serait illusoire de croire que l'on aurait mieux répondu à l'exigence de séparation fait-valeur en proposant la dénomination « batterie d'une durée de dix heures », d'une factualité renforcée. En effet, cette assertion sera interprétée comme positive dans le contexte actuel de notre société, eu égard à son développement technologique, même si l'interprétation est dépendante du degré d'expertise du lecteur du message.

Héritage conceptuel du logicisme

L'effacement progressif de l'empirisme logique et l'émergence de nouveaux paradigmes venus du monde anglo-saxon dans les années 70 (philosophie de l'esprit et renouveau du pragmatisme en premier lieu) critiquant le dogmatisme

VERSION SOUMISE AUX *CAHIERS DU NUMERIQUE* LE 18 VI 2011. MERCI DE SE RÉFÉRER A LA VERSION PUBLIÉE : Eensoo-Ramdani, Egle, Evelyne Bourion, Monique Slodzian, Mathieu Valette (2011) « De la fouille de données à la fabrique de l'opinion. Enjeux épistémologiques et propositions », *Analyse d'opinions sur internet*, Luc Grivel, éd., *Les Cahiers du Numérique*, Volume 7, n°2, pp. 15-39.

logiciste ont imposé un retour de la subjectivité, réhabilitant, entre autres, les jugements de valeur issus de l'éthique, de la psychologie et de l'esthétique. Il en est résulté une greffe inattendue, consistant à plaquer le réductionnisme réservé aux concepts élaborés par la langue de la science sur les concepts relatifs à des jugements de valeur subjectifs.

La compositionnalité constitue un des concepts fondamentaux du logicisme qui a été critiqué à la fois par des philosophes du langage (Putnam, 2004) et des linguistes (Rastier, 1987). À l'origine appliquée à l'analyse conceptuelle et largement mise en œuvre par les techniques informatiques, elle est mise à profit par les spécialistes de la fouille d'opinion, même si certains travaux émettent des réserves.[2] Ainsi, (Turney, 2002) fait remarquer que la stratégie consistant à repérer les morceaux de texte supposés positifs ou négatifs et à les additionner pour déterminer la valeur finale du signe ne donne pas les mêmes résultats selon le domaine (artefacts ou objets culturels). Par exemple, lorsque l'on traite les opinions relatives à des voitures par une approche compositionnelle, la valeur de l'entier est proche de la somme des parties, mais ce n'est pas le cas pour les films. Dans ce dernier cas, le jugement porté a un caractère holistique (le jugement global est plus complexe que la somme de ses parties). On voit bien là le risque qu'il y a à considérer l'opinion indépendamment de l'objet sur lequel elle porte, ainsi que de l'émetteur et du contexte d'émission qui inclut les pratiques individuelles et collectives, comme on le verra dans la deuxième partie de l'article. Même si les pratiques en cours dans d'autres domaines de la recherche d'information comme la recherche documentaire tendent à contextualiser les sources textuelles à interroger (modélisation de l'utilisateur et de la communauté d'utilisateurs, localisation géographique et linguistique, etc.), la détection d'opinion ne s'est pas vraiment engagée dans cette voie.

Plus généralement, l'extension aux jugements de valeur de cette approche héritée du logicisme a également conduit à considérer la variation des jugements éthiques comme des entités indifférenciées par rapport au contexte prédicatif (« Néron est cruel » vs « la guerre est cruelle ») et à proposer des ontologies du jugement moral, esthétique, voire des sentiments à l'instar de SentiWordNet (Esuli et Sebastiani, 2006). Bien que cette approche continue à susciter des travaux (Hatzivassiloglou *et al.*, 1997 ; Kamps et Marx, 2002 ; Mullen et Collier, 2004), elle cède du terrain aux approches contextuelles. Cela va du calcul des simples n-uplets (Pang et Lee, 2002 ; Dave *et al.*, 2003) à l'extraction des

2. La compositionnalité évoquée ici réfère à la conception de l'opinion comme somme des expressions d'opinion d'un texte. Cette problématique diffère de celle d'agrégation d'opinions en vue de dégager l'opinion de la majorité.

VERSION SOUMISE AUX *CAHIERS DU NUMERIQUE* LE 18 VI 2011. MERCI DE SE RÉFÉRER A LA VERSION PUBLIÉE : Eensoo-Ramdani, Egle, Evelyne Bourion, Monique Slodzian, Mathieu Valette (2011) « De la fouille de données à la fabrique de l'opinion. Enjeux épistémologiques et propositions », *Analyse d'opinions sur internet*, Luc Grivel, éd., *Les Cahiers du Numérique*, Volume 7, n°2, pp. 15-39.

syntagmes (Turney, 2002 ; Riloff et Wiebe, 2003) jusqu'à l'identification d'éléments contextuels multiniveaux influant sur la valeur des unités linguistiques (Polanyi et Zaenen 2004 ; Vernier *et al.*, 2009). Cependant, la contextualisation souvent mise en avant par les auteurs semble illusoire puisque l'idée de l'invariabilité de « l'orientation sémantique » est toujours présupposée. Par exemple, (Turney, 2002 ; Gamon et Aue, 2005) fondent leur travaux sur les syntagmes adjectivaux plutôt que sur les adjectifs isolés mais le calcul de leur sémantisme se fait toujours en mesurant la distance (information mutuelle) entre ces syntagmes et les mots « poor » et « excellent », dont le sens est supposé immuable. Même si des mesures statistiques comme l'information mutuelle rendent compte des proximités textuelles entre les unités lexicales, d'autres variables textuelles sont négligées. Sans même aborder les phénomènes complexes comme l'ironie ou la négation (« Nice hotel but beach is not excellent »[3]), nous observons que l'objet de l'opinion, le domaine du texte ou encore le but communicatif ne sont pas pris en compte, comme si les qualifications utilisées pour s'exprimer dans nos univers personnels (ou communautaires) existaient indépendamment de ces univers eux-mêmes. Ainsi on suppose à tort que les « palpitations du cœur » dans le discours d'un cardiologue en exercice sont de même nature que les « palpitations » dans le discours d'un romancier.

(Putnam, 2004) critique explicitement le réductionnisme ontologique: « there are a host of ethical judgements which are not happily formulated using the moral philosopher's favorite words, ought, must, musn't, good, bad, right, wrong, duty, and obligation–the idea that all ethical issues can be expressed in the meager vocabulary is a form of philosophical blindness ». Il serait dérisoire de croire, ironise-t-il, qu'à l'aide d'une poignée de mots tirés des champs lexicaux « expression des sentiments » ou « jugement moral », on produira un fragment de la Théorie Scientifique Unifiée telle qu'elle a été imaginée par le logicisme viennois (Carnap, 1928). De plus, si on envisage le langage ordinaire – au sens donné par (Wittgenstein, 1961) – et non plus celui de la science, l'enchevêtrement des faits et valeurs relatifs à l'éthique ou à l'esthétique, par exemple, est si profond et si général qu'il paraît vain d'essayer de caractériser *a priori* le sémantisme des adjectifs d'évaluation (d'attribuer une valeur positive ou négative à tel adjectif sans tenir compte des conditions d'énonciation, de

3. Source : http://www.epinions.com/review/Pelangi_Beach_Resort_Langkawi___Malaysia/content_247211789956 (Consulté le 23/12/2010).

VERSION SOUMISE AUX *CAHIERS DU NUMERIQUE* LE 18 VI 2011. MERCI DE SE RÉFÉRER A LA VERSION PUBLIÉE : Eensoo-Ramdani, Egle, Evelyne Bourion, Monique Slodzian, Mathieu Valette (2011) « De la fouille de données à la fabrique de l'opinion. Enjeux épistémologiques et propositions », *Analyse d'opinions sur internet*, Luc Grivel, éd., *Les Cahiers du Numérique*, Volume 7, n°2, pp. 15-39.

l'intention normative ou descriptive du locuteur dans une situation donnée, etc.). Cette tendance était pourtant prédominante dans les premiers travaux de fouille d'opinion jusqu'à ce que son inadéquation soit démontrée par (Pang et Lee, 2002) entre autres. Un des résultats de cette étude a été de constater que la classification sur la base des mots simples sans aucun traitement donnait de meilleurs résultats que la classification sur des adjectifs supposés exprimer une opinion. En effet, en prenant compte la totalité du matériau textuel, on vise à modéliser des faits textuels complexes (et éventuellement porteurs d'opinions) alors qu'en réduisant le texte à un ensemble d'attributs homogènes (comme les adjectifs), la richesse de l'expression nous échappe.

L'opinion comme objet d'étude

On s'interrogera légitimement sur le contenu de termes comme « opinion », « sentiment », « subjectivité », etc., qui sont souvent donnés comme interchangeables. Dans de nombreux travaux, une opinion est, souvent implicitement, conçue comme localisable dans un énoncé, car se réduisant à une unité minimale (un mot, un syntagme, une proposition) (Bethard *et al.*, 2004 ; Turney, 2002, Vernier *et al.*, 2009 ; Hatzivassiloglou et Wiebe, 2000). Dans les travaux qui analysent les opinions relatives à des objets de consommation culturels ou matériels (review mining), l'opinion est l'unité à laquelle il est possible d'attribuer une polarité (négatif/positif). À l'inverse, un « point de vue » ou un « positionnement » reflète un sentiment général qui se dégage d'un texte entier et qui répond à plusieurs critères (Pang et Lee, 2008). Ce terme est souvent utilisé pour les travaux traitant des sujets politiques ou idéologiques. Un « point de vue » ne se caractérise pas toujours par une polarité mais discrétise un espace de valeurs plus complexe (par exemple, appartenir à un parti politique ou certain courant idéologique). Le traitement des émotions est lui-même souvent inclus dans ce vaste domaine ajoutant encore un risque de confusion d'autant que le cadre applicatif est très large allant de la fouille de données orientée vers la détection d'émotions (Mishne et Rijke, 2006) à la robotique (Picard, 1997). Au bilan, l'essentiel du débat nous paraît porter sur l'objet d'étude lui-même plus que sur la différenciation terminologique. Cette somme d'incertitude se répercute directement sur les tâches d'annotation. Ainsi, comment s'assurer de la validité de certaines mesures d'évaluation (précision, rappel, F-score, etc.) alors que l'objet soumis à la mesure reste insuffisamment clarifié ?

C'est dans le cadre des systèmes question-réponse que l'on a cherché le plus à qualifier ce qu'est la subjectivité et comment elle s'exprime. (Bethard *et al.*, 2004) définissent l'opinion comme « une phrase ou une partie de phrase qui pourrait répondre à la question « Qu'est-ce que X pense de Y ? ». De plus,

VERSION SOUMISE AUX *CAHIERS DU NUMERIQUE* LE 18 VI 2011. MERCI DE SE RÉFÉRER A LA VERSION PUBLIÉE : Eensoo-Ramdani, Egle, Evelyne Bourion, Monique Slodzian, Mathieu Valette (2011) « De la fouille de données à la fabrique de l'opinion. Enjeux épistémologiques et propositions », *Analyse d'opinions sur internet*, Luc Grivel, éd., *Les Cahiers du Numérique*, Volume 7, n°2, pp. 15-39.

seules les opinions clairement exprimées sont prises en considération. Il s'agit en effet d'une façon de délimiter le flou définitoire qui règne dans le champ de la « détection d'opinion » et de l'« analyse des sentiments ». Ces auteurs ont aussi le mérite d'évoquer l'interprétation (sous forme d'« inférence qu'on peut faire en se basant sur les choix lexicaux de l'auteur »), même si ce n'est que pour l'exclure de leur travail. On constate ainsi que la notion de jugement de valeur est devenue progressivement synonyme de préférence subjective (par exemple, bon/mauvais, recommandable/non recommandable) donnant à la subjectivité une interprétation restreinte à l'univers de l'individu consommateur. Cette question sera abordée plus en détail dans la partie 2.2 qui tente de d'établir un pont entre sondage d'opinion et fouille d'opinion.

Certains travaux récents (Somasundaran *et al.*, 2007) mettent également en cause une conception simplificatrice du « langage subjectif ». Ils font observer qu'il s'agit d'un ensemble d'expressions répondant à des objectifs très variés (opinions, croyances, émotions, jugements, etc.). Il s'ensuit que la pertinence des réponses apportées à une question dépend aussi de l'exactitude avec laquelle le type d'expression subjective recherché est détecté. Pour illustrer ce point, ils donnent les exemples suivants :

Q1 Are you worried about climate change?

Q2 What will be the effect of reporting Iran to the Security Council ?

Il est clair qu'en réponse, ce n'est pas le même type de « discours subjectif » que l'on attend (dans le premier cas on vise un sentiment et dans le second un argument). Conscients des limites de la division binaire objectif-subjectif, (Somasundaran *et al.*, 2007) proposent de qualifier plus finement les expressions subjectives et de distinguer avant tout les sentiments (opinions, émotions, évaluations) des arguments (croyances, arguments pour/contre). Partant de ces constats et propositions nous proposons d'étendre l'approche méthodologique au cadre de l'énonciation comprenant l'émetteur, la réception et le contexte d'énonciation (voir partie 3).

Individualité et opinion

Pour mieux cerner la problématique de l'individualité dans les travaux actuels, il est utile de voir son évolution dans le passage du sondage à la fouille d'opinion. Les controverses sur la dualité objectivité/subjectivité renvoient à un débat plus général sur la notion d'individualité telle qu'elle a évolué dans le cadre du pragmatisme anglo-saxon, de plus en plus radicalisé par l'utilitarisme. L'une des critiques les plus fortes de cette évolution est à mettre au crédit de

VERSION SOUMISE AUX *CAHIERS DU NUMERIQUE* LE 18 VI 2011. MERCI DE SE RÉFÉRER A LA VERSION PUBLIÉE : Eensoo-Ramdani, Egle, Evelyne Bourion, Monique Slodzian, Mathieu Valette (2011) « De la fouille de données à la fabrique de l'opinion. Enjeux épistémologiques et propositions », *Analyse d'opinions sur internet*, Luc Grivel, éd., *Les Cahiers du Numérique*, Volume 7, n°2, pp. 15-39.

Richard Rorty pour qui le pragmatisme dans sa forme triviale actuelle a conduit à dévaloriser les notions d'objectivité, de vérité, voire de morale, et à promouvoir dans le champ intellectuel l'idée-force de « rentabilité ». L'aboutissement en serait le phénomène intimement lié à la finalité initiale du Web : « l'individualité marchande », autrement dit l'utilisateur-consommateur. Il s'ensuit que l'intérêt économique est présenté comme une justification scientifique de la fouille d'opinion (Vernier *et al.*, 2009 ; Pang et Lee, 2008).

Le rétrécissement de la notion d'individualité à celle de souci de soi, note pour sa part (Putnam, 2002), fait qu'un comportement rationnel est supposé être identique à un comportement guidé par l'intérêt personnel. Cette nouvelle vision de l'individu est antagonique avec celle des Lumières qui l'associait à des valeurs éthiques et des exigences de rationalité. La rupture entre individualité et contrat social rencontre l'opposition public/privé. Ce débat sur le comportement individuel vs comportement collectif (dualité privé/public) concerne directement la problématique de l'opinion. Qui sont ces individus dont on parle et dont on cherche à détecter l'opinion? Quelle est la part de rationalité contenue dans l'opinion majoritaire ? La subjectivité de l'opinion correspond-elle à une maximalisation de l'intérêt personnel comme le prétend le courant pragmatiste issu de l'utilitarisme ou a-t-elle une signification intersubjective, voire sociale ? Rarement posée, la question est particulièrement pertinente si l'on s'avise de problématiser le cadre épistémique du Web 2.0 et ses nouvelles pratiques sociales. Une conception atomistique de l'individualité marchande aurait-elle invalidé subrepticement celle d'une individualité sociale, telle qu'elle était imaginée, non sans idéalisme, par les sociologues français d'après-guerre, tel J. Stoetzel, fondateur de l'IFOP, lorsqu'ils entendaient fonder une « science de l'opinion publique » ?

Il n'est pas inintéressant de considérer de ce point de vue les deux valeurs d'*opinion* telles qu'elles sont actualisées dans le *sondage d'opinion* d'une part et la *fouille d'opinion* de l'autre.

Même si elle souffre d'une connotation négative (suspicion sur les procédures de sondage) depuis son éclosion à la fin des années 30, l'opinion publique, dans sa dimension « publique » précisément, présuppose l'émanation d'un public éclairé, déployé dans l'espace « public » (Habermas, 1988). À cet égard, elle comporte une part de rationalité qui la rend légitime comme base de connaissance objective des faits sociaux, au nom de quoi on lui appliquera des techniques quantitatives raffinées. Si les sondages d'opinion se soumettront toujours davantage aux contraintes du marché, ils demeureront rattachés au débat scientifique sous la forme d'une controverse entre humanisme (analyse

VERSION SOUMISE AUX *CAHIERS DU NUMERIQUE* LE 18 VI 2011. MERCI DE SE RÉFÉRER A LA VERSION PUBLIÉE : Eensoo-Ramdani, Egle, Evelyne Bourion, Monique Slodzian, Mathieu Valette (2011) « De la fouille de données à la fabrique de l'opinion. Enjeux épistémologiques et propositions », *Analyse d'opinions sur internet*, Luc Grivel, éd., *Les Cahiers du Numérique*, Volume 7, n°2, pp. 15-39.

des phénomènes sociaux par les sciences sociales) et empirisme (mesure des phénomènes sociaux avec des modèles mathématiques).

Une première différence qui sépare la problématique de la production d'opinion par sondage et celle de la détection d'opinion sur le Web concerne le statut épistémique de l'opinion de la majorité. En quoi participe-t-elle à la connaissance et donc au progrès général ? Les pionniers du sondage n'avaient de cesse d'invoquer le progrès scientifique même si l'usage social (domination des instituts de sondage et impératifs politiques et commerciaux) tend à l'emporter.

Les techniques de sondage d'opinion proposaient des méthodes quantitatives pour représenter l'état de l'opinion publique à un moment donné (par exemple, sondage d'opinion préélectoral) et prétendaient contribuer à une meilleure connaissance de la société (structuration de l'opinion selon les classes sociales, l'âge, le sexe, etc.). Les biais que présupposent les méthodes quantitatives ont été largement critiqués par des sociologues comme Bourdieu : « l'opinion publique est un artefact pur et simple dont la fonction est de dissimuler que l'état de l'opinion à un moment donné du temps est un système de forces, de tensions et qu'il n'est rien de plus inadéquat pour représenter l'état d'opinion qu'un pourcentage » (Bourdieu, 1980).

Est-ce que la fouille d'opinion est en mesure de surmonter les limitations du sondage d'opinion ? Introduit-elle de nouveaux biais et quels sont-ils ?

En proposant d'aller « attraper » les opinions directement dans la production textuelle, elle évite l'artifice de la stratégie de questionnement et elle proclame une ambition d'une autre envergure. Tout d'abord, elle exclut toute forme d'expertise et de médiation. Dans le sondage, en effet, ce rôle est joué en amont par le questionnaire qui oriente les réponses et les risques d'interprétation sont relativement maîtrisés par le sondeur. Mais surtout, dans la fouille d'opinion, l'analyste est confronté à la textualité. Les normes de production textuelle doivent être préalablement élucidées avant de prétendre accéder à la maîtrise des données. La taille et la variabilité de la production textuelle sur le Web impliquent que l'analyste se dote des moyens d'affronter la textualité dans toute sa complexité s'il veut garantir la validité de ses analyses.

S'il veut se rapprocher d'une exigence d'objectivité lui permettant d'étayer son interprétation, l'analyste doit, selon nous, élaborer des critères d'interprétation (émetteur, doxa de l'émetteur, genre, objectif rhétorique, etc.). On remarquera qu'il ne s'agit pas ici de se contenter de travailler sur des éléments contextuels limités (fenêtres de mots, propositions, phrases) mais de

VERSION SOUMISE AUX *CAHIERS DU NUMERIQUE* LE 18 VI 2011. MERCI DE SE RÉFÉRER A LA VERSION PUBLIÉE : Eensoo-Ramdani, Egle, Evelyne Bourion, Monique Slodzian, Mathieu Valette (2011) « De la fouille de données à la fabrique de l'opinion. Enjeux épistémologiques et propositions », *Analyse d'opinions sur internet*, Luc Grivel, éd., *Les Cahiers du Numérique*, Volume 7, n°2, pp. 15-39.

prendre en compte le texte dans son ensemble et en le situant dans son corpus de provenance. Ce programme de travail est justifié par la complexité de la tâche de fouille et l'exigence de vérifiabilité des interprétations produites.

La fabrique de l'opinion

Compte tenu des travaux existants qui pointent les insuffisances décrites plus haut, nous nous proposons d'échafauder ci-dessous une proposition susceptible d'améliorer la maîtrise des contenus textuels non factuels. Si l'on admet que le jugement de valeur ne se laisse pas réduire au mot-clé, on est amené à concevoir la fouille d'opinion comme une recherche d'indices de valeur actualisés dans des discours.

Le message et son cadre énonciatif

Pour situer la fouille d'opinion par rapport à la fouille de données dans le cadre du schéma de la communication hérité de la théorie de l'information, on pourrait dire qu'elle correspond à un étalement de la problématique du message vers l'émetteur et le récepteur. La fouille de données se focalise en effet, pour des raisons d'objectifs, sur le message seul. Comme nous l'avons vu précédemment, il s'agit d'identifier et d'extraire des contenus d'information du message (*i.e.* le texte ou le document) en dehors de toute question interprétative. En les affranchissant de leurs conditions d'émission, on les pare d'une objectivité et d'une neutralité nécessaires aux applications.

La fouille d'opinion, parente de la fouille de données, s'est peu démarquée de cette approche centrée sur le message. En cela, l'opinion demeure presque constitutive, comme fixée au message en dépit de la variété des récepteurs possibles. Pourtant, un texte est susceptible d'être interprété de différentes façons. Bien qu'il soit difficile de statuer sur la nature du récepteur, qui relève de l'extra-textuel, on peut identifier les différents parcours interprétatifs, c'est-à-dire les choix d'interprétation que peuvent effectuer les récepteurs à partir du même matériau textuel. C'est donc sur le producteur de ce matériau, l'émetteur, qu'il convient de porter notre attention. L'identification de l'émetteur est parfois effectuée dans le cadre d'études (Kobayashi *et al.,* 2007 ; Schulz *et al.*, 2010 ; Kim et Hovy, 2006) ou de campagnes d'évaluation (NTCIR 2007 ; Seki, 2008), mais son rôle reste souvent inexpliqué et, dans le meilleur des cas, il s'agit plus de s'assurer de la source du message que d'évaluer l'incidence de l'émetteur sur le message et ses interprétations possibles par les récepteurs.

Conçues comme un dépassement de la théorie de l'information, les théories linguistiques de la production, de l'énonciation ou de l'interprétation,

VERSION SOUMISE AUX *CAHIERS DU NUMERIQUE* LE 18 VI 2011. MERCI DE SE RÉFÉRER A LA VERSION PUBLIÉE : Eensoo-Ramdani, Egle, Evelyne Bourion, Monique Slodzian, Mathieu Valette (2011) « De la fouille de données à la fabrique de l'opinion. Enjeux épistémologiques et propositions », *Analyse d'opinions sur internet*, Luc Grivel, éd., *Les Cahiers du Numérique*, Volume 7, n°2, pp. 15-39.

empreintes de phénoménologie (Jakobson, 1973 ; Benveniste, 1966, Culioli, 1990, Charaudeau, 1992), ont pourtant souligné l'incidence du foyer énonciatif et de son univers mondain sur le message : par delà le simple sujet grammatical, l'énonciateur et le message sont co-construits, celui-ci portant les traces de la construction de celui-là. Dès lors, nous nous proposons de considérer l'opinion comme le fait sémantique qui résulte de cette co-construction.

La mise en texte de l'émetteur

La notion de polyphonie que (Ducrot, 1980), inspiré de (Bakhtine, 1977), élabore dans ses travaux sur l'argumentation nous semble particulièrement utile à notre propos. Elle permet de faire l'hypothèse qu'un émetteur peut avoir une multiplicité de voix, indice d'une diversité de points de vue et donc d'opinions exprimées. Cette diversité est souvent considérée comme un problème dans la détection de l'opinion. Ainsi, (Pang *et al.*, 2002) constatent que la présence de segments linguistiques de différentes polarités (en l'occurrence des mots) baisse significativement la performance de la détection et démontre l'inefficacité de l'approche « sac de mots » où un texte n'est qu'un ensemble de mots ayant tous la même valeur caractérisante quels qu'ils soient et ne se différenciant que par leur nombre d'occurrences. Toutefois, le seul fait que le message soit polyphonique donne des informations sur les intentions de l'émetteur. Différentes études tendent à montrer que les marqueurs de polyphonie tels que certains connecteurs argumentatifs d'opposition (*cependant, toutefois, mais*, etc.) sont sur-représentés dans les textes polémiques ou ceux qui soutiennent une position sujette à polémique. Par exemple, dans le cadre du projet PRINCIP (Valette, 2004) où étaient contrastés deux corpus issus du Web, de textes racistes et antiracistes, il a été observé que la conjonction de coordination « mais » était spécifique au sous-corpus raciste[4]. On remarque un phénomène tout à fait similaire dans un des corpus constitués[5] pour le projet C-Mantic[6]. En

4. Suivant le calcul de spécificités tel qu'implémenté dans Hyperbase (Brunet, 2001)

5. Le corpus est constitué principalement de pages Web portant sur le tabagisme. Pour pouvoir cerner les différences entre les propos sur le tabac issus de différents groupes émetteurs, nous avons constitué des sous-corpus combinant des ensembles de textes caractérisés suivant les axes suivants : « institutionnel vs personnel », « pro-tabac vs anti-tabac » et « militant vs non-militant ». Ainsi, le corpus « institutionnel » comprend les sites Web des industriels du tabac (Altadis, JTI, Philip Morris, BAT - 95 textes, 38293 mots) et des organismes de prévention (OFT, RHST, INPES, etc. - 540 textes, 986931 mots) tandis que le corpus « personnel » provient de blogs (comme Skyblog) et de forum (Atoute). La distinction « militant » (1160 textes, 1042674 mots) - « non-militant » (132 textes, 118000 mots) ne concerne que les textes du corpus « personnel ».

contrastant au moyen d'un calcul de spécificités[7] des textes issus des sites Web émanant d'institutions : d'une part les industriels du tabac et d'autre part les organismes de prévention, il apparaît que les connecteurs « toutefois » et « cependant » sont caractéristiques des textes des industriels du tabac (score de 8). Or, ces textes semblent alterner des considérations tantôt hostiles, tantôt valorisantes à propos du tabac. Par exemple :

> [À propos des risques liés au tabagisme, NdA] Nous ne pensons pas qu'il soit possible de les éliminer totalement ; il existe toutefois des moyens par lesquels nous nous efforçons de les réduire. Nous nous engageons à mettre au point des produits à risques réduits.
>
> Les gens fument par plaisir, cependant des risques réels sont associés à ce plaisir. (http://www.jti.com ; consulté le 22/02/2008)

Dans ces deux exemples, on repère des segments linguistiques correspondant dans ce corpus à des arguments opposés de part et d'autre des connecteurs identifiés (par exemple : « par plaisir » *vs* « risque réels »). Il ne s'agit pas ici d'allouer une valeur sémantique particulière à ces différents segments, mais de prendre en compte le double positionnement de l'émetteur qui nous renseigne sur les modalités de production du texte. L'émetteur, en élaborant son texte, choisit, consciemment ou non, un ensemble de contraintes de production destinées à donner au lecteur un cadre d'interprétation (dans le cas présent, il s'agit d'un argumentaire marketing obligé de composer avec une législation restreignant la promotion du tabac et un impératif commercial tout en revendiquant le statut d'entreprise responsable). L'ensemble des contraintes régissant la production et l'interprétation d'un texte constitue ce que l'on appelle un genre textuel, tel qu'il a été théorisé par (Rastier, 2001) notamment.

Le genre comme cadre d'interprétation

L'idée que les genres textuels sont prégnants dans l'élaboration linguistique s'impose peu à peu dans le TAL (Beauvisage, 2001 ; Malrieu et Rastier, 2001 ; Jacques et Aussenac-Gilles, 2006 ; Finn et Kushmerick, 2006). La détection de genre a fait également l'objet de la campagne d'évaluation DEFT'08 où la distinction entre des normes journalistiques (corpus *Le Monde*) et encyclopédiques (corpus Wikipédia) constituait une des tâches du défi (Grouin *et al.*, 2008). Dans ce cadre, (Béchet *et al.*, 2008) par exemple, concluent que l'identification préalable du genre améliore la classification thématique. La question des normes d'élaboration des textes touche en conséquence la fouille

---

6. ANR-07-MDCO-002 C-Mantic
7. Calcul de spécificités tel qu'implémenté par Lexico3 (Salem et al., 2003).

VERSION SOUMISE AUX *CAHIERS DU NUMERIQUE* LE 18 VI 2011. MERCI DE SE RÉFÉRER A LA VERSION PUBLIÉE : Eensoo-Ramdani, Egle, Evelyne Bourion, Monique Slodzian, Mathieu Valette (2011) « De la fouille de données à la fabrique de l'opinion. Enjeux épistémologiques et propositions », *Analyse d'opinions sur internet*, Luc Grivel, éd., *Les Cahiers du Numérique*, Volume 7, n°2, pp. 15-39.

d'opinion. (Vernier *et al.*, 2009) dans la tâche de détection de la subjectivité de DEFT'09, définissent des « descripteurs empiriques » destinés à affiner la caractérisation des textes (passage rapporté, interview, erratum). Sans expliciter leur lien avec la notion de genre, ils observent toutefois qu'il s'agit de contraintes liées au corpus : « ces caractéristiques s'éloignent quelque peu des définitions théoriques sur la subjectivité pour se rapprocher, de façon *ad-hoc*, des contraintes liées au corpus du Monde ». Selon nous, ces caractéristiques répondent à des normes de production imposées par le but de la communication. Ces descripteurs empiriques sont donc des indices du genre.

Une fois un tel constat fait, nous remarquons que le genre n'est pas un phénomène se superposant au texte : le genre se construit et se reconnaît dans les éléments textuels qui peuvent être très variés. Dans le cadre du défi DEFT'08, (Béchet *et al.* 2008) par exemple s'intéressent au rôle de la ponctuation dans la détection du genre. Nous pouvons affirmer que l'emploi de tel ou tel signe de ponctuation dans une proportion relativement forte ou faible est à considérer sur le plan du genre comme tout déficit ou pic d'un autre type de signe. Un autre fait intéressant émerge de ces travaux : la fusion de plusieurs modèles de classification en augmente la robustesse. Plus qu'une simple combinaison mathématique, cette fusion correspond à la multiplication des points de vue sur un même phénomène et donc, modélise la complexité de la tâche d'interprétation.

Dans la même lignée, (Valette, 2004 ; Valette et Grabar, 2004), constatant que le vocabulaire est insuffisant pour distinguer les opinions racistes et antiracistes, s'appuient sur l'étude de modalités sémiotiques d'élaboration des textes. Insuffisantes pour identifier avec précision une opinion, ces modalités très contrastées sont indubitablement des indices de genre et renforcent la plausibilité de certains parcours interprétatifs : par exemple, un usage massif des capitales d'imprimerie et des points d'exclamation, une dominante rouge ou noire sont des marqueurs attestés de genres polémiques tels que le pamphlet ou la diatribe, genres d'opinion par excellence. Outre leur utilité pour la détection de la subjectivité dans les textes, ces indices de genres participent à la reconnaissance des intentions des émetteurs et aident ainsi à interpréter les contenus d'information. Autrement dit, ces indices du plan global qui confinent à l'extra-linguistique (présence ou absence de bannière, d'une image d'arrière-plan, etc.) peuvent orienter une interprétation locale.

L'approche adoptée dans le projet C-Mantic, dont un des objectifs est de caractériser les différents discours sur le tabac, va plus loin que l'opposition binaire pro *vs* anti (par exemple, raciste *vs* antiraciste). Ainsi, elle distingue, entre

les auteurs pro-tabac et les auteurs anti-tabac, ceux qui sont militants et ceux qui ne le sont pas. Cette distinction s'est imposée à l'analyse du corpus de blogs en raison de la diversité des modalités d'expression du positionnement pro- et anti-tabac (lire en annexe les textes 1, 2, 3, 4 extraits du site Skyblog).

Les textes militants pro- et anti-tabac se caractérisent par une virulence de ton s'exprimant dans des genres polémiques comme le pamphlet. Revendicatif, le texte 3 classé « militant pro-tabac » comporte plusieurs critères l'inscrivant dans une tradition générique bien établie : le tract politique. Parmi les indices de genre on repère les capitales d'imprimerie, la police grasse (marqueurs d'intensité), l'impératif (« exigeons que », etc.), l'agrammaticalité des énoncés et les liens vers des sites de la même communauté (*Fumeurs en colère* et *Fumeurs Electeurs*) indiquant une action collective. Ces indices de présentation matérielle sont cohérents par rapport aux thèmes récurrents dans le corpus avec des lexicalisations variées : la liberté des fumeurs[8] et le rejet du groupe antagoniste[9]. Les indices d'intensité sont similaires pour les textes militants anti-tabac (texte 4), ce qui justifie leur regroupement.

À la différence des textes militants, les textes non-militants choisissent un mode d'expression plus modéré avec des genres informatifs (en particulier pour les anti-tabac) et égocentrés (plutôt pour les pro-tabac). Le thème des « groupes antagonistes » est également présent dans les textes non-militants pro-tabac mais avec les lexicalisations différentes associant le groupe aux notions de plaisir et de convivialité plutôt qu'à la lutte politique. On trouve des énoncés comme, dans le texte 1, « cet article est réservé aux fumeurs et à leurs amis » et « Ici, on est chez nous ! ». Le groupe des non fumeurs est qualifié de « chiants », rabat-joie (« ils ne boivent presque pas ») et dévalorisé (« des tics grotesques »). À la place du ton polémique, on trouve chez les non-militants de l'ironie comme dans le paragraphe sur les qualités de la cigarette, où l'énonciateur prend le contrepied des reproches entonnés par les non-fumeurs. Le thème de la santé est présent dans les textes anti-tabac aussi bien militants que non-militants (*cf.* textes 2 et 4) en reprenant des informations dans les textes institutionnels anti-tabac. Alors que les textes militants présentent ce thème avec un ton dramatisant (employant par exemple des majuscules comme

8. « REAGISSONS CONTRE ce Décret discriminatoire "Anti Fumeur" qui veut nous exclure l'accès à tous les commerces et entreprises [...]!!! »

9. « Ainsi les intolérants au Tabac pourront rester qu'entre eux si ça leur chante »

marqueurs d'intensité[10]), les textes non-militants se contentent d'être informatifs[11].

Ainsi l'axe « militant-non-militant » s'est imposé pour les textes d'émetteurs individuels en particulier en raison de la confusion possible entre militant pro-tabac et militant anti-tabac, les indices du trait /militant/ pouvant être plus forts que ceux de l'axe « pro-anti ».

Le genre comme reflet de la pratique sociale

À l'instar du sondage d'opinion, la fouille d'opinion peut contribuer à l'étude de la société et des phénomènes sociétaux. Au lieu de considérer l'opinion d'un groupe comme la somme des opinions des individus atomistiques, on peut s'intéresser aux opinions comme des constructions collectives dont on retrouve des traces dans les matériaux textuels émis par et pour des membres des groupes sociaux visés.

Ainsi, l'étude des corpus de textes disponibles sur le Web et donc destinés à la communication publique, peut nous renseigner sur les pratiques sociales dans la mesure où le genre textuel les reflète. L'analyse textométrique et sémiotique du corpus pro- et anti-tabac (pour la constitution du corpus, voir note 3) nous a permis de mettre en évidence les caractéristiques du profil sociologique des deux protagonistes du champ « tabagisme » : l'industrie du tabac et les organisations de lutte contre le tabagisme. Elle fournit aussi des indicateurs pour la détection de la polarité des textes venant du même univers sociologique (l'univers des acteurs sociaux – individuels ou collectifs – situés en un lieu et une époque donnés et partageant des intérêts communs).

Nous avons contrasté deux corpus de polarité différente mais partageant un fond commun, à savoir le caractère institutionnel des émetteurs. Un calcul d'écart réduit (Lexico 3, Salem et al. 2003) permet de mettre en évidence la spécificité de mots comme *groupe*, *produits*, *intérêts*, *entreprise*, etc. dans le corpus des industriels du tabac (InddusT) et des mots désignant les parties du corps (*cœur, cardiaque, corps, poumons, cerveau, etc.*) et les pathologies (*cancer, tumeur, infarctus, etc.*) pour le corpus des organismes de prévention (OP).

Le caractère particulier de ces mots est lié aux thématiques abordées par les deux types de corpus : le discours commercial d'une part, le discours médical-santé publique de l'autre. Ainsi, nous constatons que les différences entre

10. « ARRETER DE FUMER, C'EST VOTRE VIE QUI EST EN JEU »
11. « saviez-vous que c'est [le tabac] aussi l'ennemi juré de votre beauté »

VERSION SOUMISE AUX *CAHIERS DU NUMERIQUE* LE 18 VI 2011. MERCI DE SE RÉFÉRER A LA VERSION PUBLIÉE : Eensoo-Ramdani, Egle, Evelyne Bourion, Monique Slodzian, Mathieu Valette (2011) « De la fouille de données à la fabrique de l'opinion. Enjeux épistémologiques et propositions », *Analyse d'opinions sur internet*, Luc Grivel, éd., *Les Cahiers du Numérique*, Volume 7, n°2, pp. 15-39.

pratiques sociales entraînent à première vue des différences thématiques. Pourtant, on peut remarquer dans les textes de ces deux protagonistes un fond thématique commun – la santé – qui témoigne de la constitution d'un espace social interconnecté. Néanmoins, la présentation de ce fond thématique, ou plutôt sa mise en texte, diffère selon l'émetteur. Pour le corpus IndduT, c'est la surreprésentation des verbes au présent de l'indicatif[12], donnant une dimension d'intemporalité (la vérité énoncée est vraie de tout temps), et introduits par le pronom « nous »[13], marqueur de revendication identitaire, qui constitue le discours de présentation de soi de ces entreprises (un des éléments de l'identité affichée des industriels est le statut revendiqué d'entreprise responsable). En outre, on a pu observer que les parties concernant les questions de santé tentent de minimiser l'impact des recherches scientifiques, en jetant le discrédit sur leur validité[14]. *A contrario*, le corpus OP fait état d'un savoir médical ou législatif daté et documenté[15].

Ainsi, du fait qu'il donne un cadre pour l'interprétation – sans lequel un bon nombre de caractéristiques textuelles seraient inutilisables – le genre doit être obligatoirement pris en compte dans toute analyse de matériau textuel hétérogène.

Subjectivité des concepts

S'il est possible d'identifier des marqueurs d'opinion diffus au niveau de la structure générique (ou contrainte de genre) d'un texte et non pas seulement au niveau de marqueurs discrets (tels que des adjectifs), l'opposition fait/valeur et son avatar objet/opinion méritent d'être questionnés. Davantage que l'allocation de valeurs subjectives à un objet, qu'il soit référentiel ou conceptuel, l'expression d'une opinion sur un objet consiste *à faire partager la représentation* que l'on en a. Si l'on s'affranchit de la perspective ontologique pour adopter un point de vue textuel, un concept n'est pas en effet détaché des conditions linguistiques de son élaboration et la représentation linguistique du concept ou

---

12. soutenons, pensons, sommes, voulons, reconnaissons, considérons, souhaitons, etc.
13. nous, nos, notre
14. La forte présence, statistiquement significative, dans leur corpus de toutefois (souvent en introducteur de phrase) s'explique par cette posture de réfutation ; le parallèle existe, bien sûr, dans le corpus anglo-américain où se retrouvent la prégnance de we et our des verbes comme to believe, to aim, to support, to be commited to, et aussi however. De même on y observe l'alternance, à propos de la cigarette, entre is a cause of disease (chez les IndduT) en face de is responsible dans le corpus OP.
15. On y trouve des segments comme Références bibliographiques ou entretien réalisé le.

VERSION SOUMISE AUX *CAHIERS DU NUMERIQUE* LE 18 VI 2011. MERCI DE SE RÉFÉRER A LA VERSION PUBLIÉE : Eensoo-Ramdani, Egle, Evelyne Bourion, Monique Slodzian, Mathieu Valette (2011) « De la fouille de données à la fabrique de l'opinion. Enjeux épistémologiques et propositions », *Analyse d'opinions sur internet*, Luc Grivel, éd., *Les Cahiers du Numérique*, Volume 7, n°2, pp. 15-39.

de l'objet inclut les valeurs de l'émetteur. Bien que l'objet soit stable, ce qui lui permet notamment d'intégrer une ontologie, ses représentations sémantiques sont non seulement subjectives mais composites. Concrètement, elles s'insèrent dans un réseau de co-occurrents constitutifs qu'il importe de prendre en compte. Par exemple, un film se décrit en termes de casting, mise en scène, montage, etc. ; un jeu vidéo en termes de graphisme, jouabilité, bande son, scénario, etc. (*cf.* DEFT 2007). Autrement dit, c'est davantage sur le réseau co-occurrentiel constitutif de l'objet que porte l'évaluation que sur l'objet lui-même. L'identification de ces réseaux n'est pas triviale. Par exemple, dans la tâche de détection de textes racistes mentionnée *supra*, ni *racisme* ni même *race*, pourtant pertinents d'un point de vue référentiel, ne sont des concepts pertinents, à la différence d'*étranger, jeune de banlieue* ou *immigration*. Dès lors, pour accéder aux opinions véhiculées par le texte, il peut être nécessaire de typer le réseau co-occurrentiel d'un mot-pôle. L'identification de ses variations permet de déterminer non pas l'opinion négative ou positive d'un émetteur relativement au mot-pôle mais l'opinion générale qui est la sienne sur une problématique donnée (raciste ou antiraciste, pour ou contre, partisan ou détracteur, etc.).

Considérons l'exemple d'*étranger* dans le corpus raciste/antiraciste déjà évoqué. Ses mesures de rappel antiraciste et raciste y sont relativement similaires de telle sorte que la seule présence du mot *étranger* ne suffit pas à déterminer si un texte est raciste ou non (comme par exemple, le seul mot *jouabilité* serait insuffisant pour déterminer la valeur allouée à un jeu vidéo). Mais un test d'écart réduit permet d'isoler les spécificités positives des contextes d'actualisation du mot-pôle dont on conserve les plus discriminantes. Il montre que, parmi les cooccurrents les plus saillants d'*étranger* dans le sous-corpus raciste, on trouve par exemple *illégalité, naturalisation, délinquants* tandis que dans le sous-corpus antiraciste, on peut identifier *régularisation, emplois, droit*. Ces items constituent des représentations subjectives du concept *étranger* qui ne sont pas elles-mêmes polarisées mais qui permettent d'attribuer une valeur raciste ou antiraciste au texte. Alors que le mot-pôle *étranger* a dans le corpus une distribution relativement homogène, il suffit d'une classification effectuée sur le réseau co-occurrentiel pondéré pour savoir si le texte où le mot *étranger* est actualisé est un texte antiraciste ou raciste, avec une précision de 90,41% dans le premier cas et de 64,71% dans le second cas[16].

16. Lire (Valette, 2004) pour plus de détails.

VERSION SOUMISE AUX *CAHIERS DU NUMERIQUE* LE 18 VI 2011. MERCI DE SE RÉFÉRER A LA VERSION PUBLIÉE : Eensoo-Ramdani, Egle, Evelyne Bourion, Monique Slodzian, Mathieu Valette (2011) « De la fouille de données à la fabrique de l'opinion. Enjeux épistémologiques et propositions », *Analyse d'opinions sur internet*, Luc Grivel, éd., *Les Cahiers du Numérique*, Volume 7, n°2, pp. 15-39.

Conclusion

Les cadres de la véric onditionnalité ou du principe de pertinence, propres à l'approche logico-grammaticale des mots et des énoncés, sont insuffisamment productifs pour guider vers le sens de ces textes qui sont à considérer comme des objets culturels, émis par des acteurs précis, historiquement et spatialement situés et qui mettent en mots leur *doxa*, au sein de champs de pratiques.

La fouille de texte s'est fixée comme objectif la découverte de connaissances à partir des textes non structurés, hétérogènes et surtout produits par et pour des humains. La fouille d'opinion se place dans la même lignée avec un impératif supplémentaire : prendre en compte les valeurs. Insuffisamment défini, cet objet d'étude mériterait lui-même une mise en perspective théorique et historique pour mieux comprendre ses enjeux. Cependant, nous pouvons déjà observer que pour pouvoir dépasser les limites des sondages d'opinion, la fouille d'opinion doit se doter d'un appareil théorique et d'un cadre descriptif afin de traiter le matériau que sont les textes. En effet, il serait illusoire de croire qu'on peut accéder directement à la subjectivité comme on accède au lexique, puisque la première est à construire à partir d'indices tissés dans le texte.

Dès lors qu'on admet que les textes ne se résument pas à une liste de mots agencés par des règles syntaxiques, mais qu'ils sont des objets complexes nécessitant un parcours interprétatif, on doit s'interroger sur les éléments à prendre en compte et les constructions nécessaires à partir de ces éléments. Dans le cadre de cet article, nous avons proposé une réflexion sur la nécessité de caractériser le cadre énonciatif du message, en particulier l'émetteur. Au lieu de penser l'émetteur comme une entité référentielle, nous avons proposé de le considérer comme un objet sémiotique construit à partir d'un ensemble de caractéristiques présentes dans le texte. L'émetteur se définit comme une construction complexe à partir d'éléments textuels simples (lexies, ponctuation, connecteurs) ou construits (co-occurrences de certains lexies et utilisation de signes sémiotiques tels que le gras ou les lettres capitales). L'identification de l'émetteur se fait en parallèle avec celle du genre, lequel est le cadre donné par l'émetteur pour permettre la bonne interprétation de son message. Au même titre que l'émetteur, le genre est une construction méta-textuelle effectuée à partir des indices textuels.

Ajoutant de la complexité dans la modélisation, émetteurs et genres sont pourtant nécessaires à la détection des valeurs véhiculées par le texte car on ne peut les isoler de leurs contextes d'apparition sans qu'elles perdent leur signification. Comme nous avons vu dans les analyses sur le tabagisme, être « pro-tabac » implique des attitudes sensiblement différentes selon que l'on a affaire à un émetteur personnel ou institutionnel, militant ou non. Cette modulation des constituants d'un thème ne peut

VERSION SOUMISE AUX *CAHIERS DU NUMERIQUE* LE 18 VI 2011. MERCI DE SE RÉFÉRER A LA VERSION PUBLIÉE : Eensoo-Ramdani, Egle, Evelyne Bourion, Monique Slodzian, Mathieu Valette (2011) « De la fouille de données à la fabrique de l'opinion. Enjeux épistémologiques et propositions », *Analyse d'opinions sur internet*, Luc Grivel, éd., *Les Cahiers du Numérique*, Volume 7, n°2, pp. 15-39.

être perçue que dans le cadre d'une approche différentielle et d'une analyse alternant plan local et plan global.

Bibliographie

Bakhtine, M., *Le Marxisme et la philosophie du langage : Essai d'application de la méthode sociologique en linguistique*, Paris, Editions de Minuit, 1977.

Béchet, F., El-Bèze, M., et Torres-Moreno, J.-M., « En finir avec la confusion des genres pour mieux séparer les thèmes », *Actes de l'atelier de clôture de la 4ème édition du DÉfi Fouille de Texte*, Avignon, 2008.

Bethard, S., Yu, H., Thornton, A., Hatzivassiloglou, V., Jurafsky, D., « Automatic Extraction of Opinion Propositions and their Holders », *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text*, 2004.

Benveniste, E., « De la subjectivité dans le langage », *Problèmes de linguistique générale* 1, 1966 [1958].

Biber, D., *Variation across speech and writing*, Cambridge, CUP, 1988.

Bourdieu, P., « Culture et politique », *Questions de sociologie*, Paris, Editions de Minuit, 1980.

Brunet, E. *Manuel d'utilisation d'Hyperbase*. Nice, 2001.

Carnap, R., Der logische Aufbau der Welt, *La Construction logique du Monde*, 1928.

Charaudeau P., *Grammaire du sens et de l'expression, Hachette Education* (Ed.), 1992.

Culioli A., « Pour une linguistique de l'énonciation ». *Opérations et représentations*, tome1, Paris, Ophrys, 1990.

Dave, K., Lawrence, S., Pennock, D.M., « Mining the peanut gallery: Opinion extraction and semantic classification of product reviews », Proceedings of The Twelfth International World Wide Web Conference, Budapest, 2003.

Ducrot, O., *Les mots du discours*, Éditions de Minuit, Paris, 1980.

Esuli, A. et Sebastiani, F., « SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining », *Proceedings of LREC-06, 5th Conference on Language Resources and Evaluation*, Genova, 2006.

Finn, A., Kushmerick, N., « Learning to classify documents according to genre », *Journal of the American Society for Information Science and Technology (JASIST)*, vol. 7, 2006.

Gamon, M., Aue, A., « Automatic identification of sentiment vocabulary: Exploiting low association with known sentiment terms », *Proceedings of the ACL Workshop on*

Feature Engineering for Machine Learning in Natural Language Processing, Ann Arbor, Michigan, USA, 2005

Grouin, C., Berthelin, J-B., El Ayari, S., Hurault-Plantet, M., Loiseau, S., « Présentation de DEFT'08 », *Actes de l'atelier de clôture du 4ème DÉfi Fouille de Texte*, Avignon, 2008.

Habermas, J., *L'espace public. Archéologie de la publicité comme dimension constitutive de la société bourgeoise*, Payot, 1988.

Hatzivassiloglou, V., McKeown, K. R., « Predicting the semantic orientation of adjectives », *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, Morristown, NJ, USA, 1997.

Hatzivassiloglou, V., Wiebe, J., « Effects of adjective orientation and gradability on sentence subjectivity ». *Proceedings of the International Conference on Computational Linguistics (COLING)*, Saarbrücken, Germany, 2000.

Jacques, M.-P. et Aussenac-Gilles, N., « Variabilité des performances des outils de TAL et genre textuel », *Traitement Automatique des langues*, vol 47, n° 1, 2007, p. 11-32.

Jakobson, R., *Essais de linguistique générale*, Paris, Minuit, 1973.

Kamps, J., Marx, M. « Words with Attitude », *1st International WordNet Conference*, 2002. p. 332-341.

Kim, S., Hovy, E., « Extracting opinions, opinion holders, and topics expressed in online news media text », *SST '06: Proceedings of the Workshop on Sentiment and Subjectivity in Text, Morristown*, NJ, USA, 2006.

Kobayashi, N., Inui, K., Matsumoto, Y., « Extracting Aspect-Evaluation and Aspect-Of Relations in Opinion Mining », *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic, 2007.

Malrieu, D., Rastier, F., « Genres et variations morphosyntaxiques », *Traitement Automatique des langues*, vol. 42, n°2, 2001, p. 548-577.

Mihalcea, R., Banea, C., Wiebe, J. « Learning multilingual subjective language via cross-lingual projections », *Proceedings of the Association for Computational Linguistics (ACL)*, Prague, Czech Republic, 2007.

Mishne, G., de Rijke, M., « Moodviews: Tools for blog mood analysis », *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, Stanford University, California, USA, 2006.

Mullen, T., Collier, N., « Sentiment analysis using support vector machines with diverse information sources », *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Barcelona, Spain, 2004.

Pang, P., Lee, L., « Thumbs up? Sentiment Classification using Machine Learning Techniques », *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, University of Pennsylvania, Philadelphia, USA, 2002.

Pang, P., Lee, L., *Opinion Mining and Sentiment Analysis*, Now Publishers Inc., 2008.

Picard, R., *Affective Computing*. MIT Press, 1997.

Polanyi, L., Zaenen, A., « Contextual Lexical Valence Shifters », *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, 2004.

Putnam, H., *Ethics without ontology*, Harvard, 2004.

Quirk, R., Greenbaum, S., Leech, G., Svartvik, J. A, *Comprehensive Grammar of the English Language*. Longman, New York, 1985.

Rastier, F., *Sémantique interprétative*, PUF, 1987.

Rastier, F., *Arts et sciences du texte*, PUF, 2001.

Rawls, J. A *Theory of Justice*, The Belknap Press of Harvard University Press, 1971.

Riloff, E. et Wiebe, J., « Learning extraction patterns for subjective expressions », *Proceedings of the 2003 Conference on Empirical methods in natural language processing*, Morristown, NJ, USA, 2003.

Rorty, R., *Consequences of Pragmatism*, Minneapolis , University of Minnesota Press, 1982.

Salem A. Lamalle C. Martinez W., Fleury S., Fracchiolla B., Kuncova A., Maisondieu A. *Lexico3 – Outils de statistique textuelle. Manuel d'utilisation*, Syled-CLA2T, Université de la Sorbonne nouvelle – Paris 3, 2003.

Schulz, J. M., Womser-Hacker, C., Mandl, T., « Multilingual Corpus Development for Opinion Mining », *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 2010.

Seki, Y., « A multilingual Polarity Classification Method using Mult-label Classification Technique Based on Corpus Analysis », *Proceedings of NTCIR-7 Workshop Meeting*, Japan, 2008.

Somasundaran, S., Wilson, T., Wiebe, J., Stoyanov, V., « QA with attitude: Exploiting opinion type analysis for improving question answering in on-line discussions and the news », *Conference on Weblogs and Social Media*, Boulder, Colorado, USA, 2007.

Stoetzel, J., *Les sondages d'opinion publique*, Paris, Ed. du Scarabée, 1948.

Turney, P., « Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews », *Proceedings of the Association for Computational Linguistics (ACL)*, 2002.

Valette, M., « Sémantique interprétative appliquée à la détection automatique de documents racistes et xénophobes sur Internet », *Approches Sémantiques du Document Numérique, Actes du 7e Colloque International sur le Document Electronique*, 2004, P. Enjalbert et M. Gaio, eds.

Valette, M. et Grabar, N., « Caractérisation de textes à contenu idéologique : statistique textuelle ou extraction de syntagmes ? l'exemple du projet PRINCIP », *Actes des 7èmes Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, 2004, Louvain-la-Neuve, Belgique., UCL-Presses Universitaires de Louvain.

Vernier, M., Monceaux, l., Daille, B., « DEFT'09 : détection de la subjectivité et catégorisation de textes subjectifs par une approche mixte symbolique et statistique », *Actes de l'atelier de clôture de la 5ème édition du DÉfi Fouille de Textes*, 2009.

Wilson, T., Wiebe, J., Hoffmann, P., « Recognizing contextual polarity in phrase-level sentiment analysis ». HLT'05: *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Morristown, NJ, USA, 2005.

Wittgenstein, L. *Tractatus logico-philosophicus, [Suivi de] Investigations philosophiques*. 1961, Paris, Editions Gallimard.

Annexe : textes analysés

Texte 1 : non-militant pro-tabac

Qu'est ce qu'ils sont chiants ces non-fumeurs!!!

[photo supprimée]

Attention! cet article est réservé aux fumeurs et à leurs amis!
Pas de conseil à la con sur notre santé! Ici, on est chez nous!
Cet article se doit de rassembler ceux qui en ont marre du "diktat non-fumeur" et de leurs pathétiques historiettes sur un épique combat contre le tabac!
Il faut le dire: les non-fumeurs sont chiants!!! Ils ne boivent presque pas, se plaignent tout le temps de la moindre odeur de tabac, nous expliquent combien on serait riches sans nos clopes, que Tata Micheline est morte du tabagisme passif, etc!
En bref, ras-le-bol de ces leçons données sous couvert d'humanisme. Ils ont gagné, alors qu'ils nous foutent la paix!!! Sans quoi, on pourrait devenir mauvais perdants...
Si comme moi vous trouvez que la cigarette:
- est indispensable avec un café,
- stimule vos neurones (c'est sans doute pour ça que les fumeurs sont plus intelligents!),
- occupe cette deuxième main que l'on ne sait jamais où fourrer à un cocktail,

- fait de vous le roi de la soirée à 4h, quand plus personne n'en a,
- ça sent bon,
- c'est enfin un produit de luxe, inaccessible aux masses laborieuses,
- n'a jamais troublé le bon fonctionnement de votre organisme
- évite de se ronger les ongles, de se gratter frénétiquement la tête, d'avoir des tics grotesques (regardez les non-fumeurs... la vérité sera flagrante),

Quand nous serons assez nombreux, nous pourrons demander à l'état d'inscrire (dans les mêmes proportions que celles des paquets de clopes) un CONDUIRE TUE sur le capot des voitures!
Vos idées et réactions sont les bienvenues... si tant est qu'elles soient fumeuses!!!
Tabacologiquement vôtre,

Deborah

Texte 2 : évaluation négative

la cigarette

Tout le monde sait que le tabac tue et que les maladies liees au tabac augmentent chaque annee ; mais saviez-vous que c'est aussi l'ennemi jure de votre beaute ?

Eh oui, les 4 000 agents chimiques presents dans une (seule) cigarette recouvrent la peau d'un voile gris (principalement au niveau des pomettes) et donnent mauvaise haleine. Ils creusent aussi les rides autour des yeux et de la bouche, deshydratent et accelerent le vieillissement de la peau. Et au cas ou ca ne suffirait pas, ils empuantissent aussi les cheveux et jaunissent les mains.

Il rend la peau grise

Pour etre saine, la peau doit etre regulierement alimentee par du sang frais. Malheureusement, la nicotine comprime les vaisseaux sanguins et prive, par consequent, la peau d'oxygene.

Il donne mauvaise haleine

La fumee asseche l'interieur de la bouche, et notamment les membranes responsables de la salivation. Sans salive pour tuer les bacteries, on se retrouve rapidement avec l'haleine d'un phoque. Si les fumeurs ont tous mauvaise haleine (sucer des pastilles a la menthe ne change rien), c'est parce que les produits chimiques deposes a chaque bouffee tapissent le fond de la bouche et produisent des odeurs particulierement desagreables.

Il gache votre beau sourire

VERSION SOUMISE AUX *CAHIERS DU NUMERIQUE* LE 18 VI 2011. MERCI DE SE RÉFÉRER A LA VERSION PUBLIÉE : Eensoo-Ramdani, Egle, Evelyne Bourion, Monique Slodzian, Mathieu Valette (2011) « De la fouille de données à la fabrique de l'opinion. Enjeux épistémologiques et propositions », *Analyse d'opinions sur internet*, Luc Grivel, éd., *Les Cahiers du Numérique*, Volume 7, n°2, pp. 15-39.

Des etudes ont montre que le tabac entraine certaines maladies buccales, comme les gingivites (inflammations des gencives) car la fumee detruit le mecanisme de defense de la bouche.

Il fait ronfler comme un cochon

Physiquement, ce n'est pas grave, mais ca peut quand meme porter un sacre coup a votre image.

A 30 ans, il vous donnera l'air d'en avoir 50

Le tabac accelere le vieillissement car il bourre le corps d'antioxydants (des molecules qui detruisent les cellules saines du corps), deshydrate et ride la peau, et provoque des maladies de poumons et de bronches. Mais rejouissez-vous : au bout de 8 heures d'abstention, le taux d'oxyde de carbone dans le sang redevient normal. 2 jours plus tard, vos chances d'avoir une crise cardiaque diminuent ; et 1 an plus tard, l'apparition de rides prematurees est ralentie.

(suit un paragraphe intitulé « la beauté des paresseuses »)

Arretez de fumer et voici ce que vous gagnerez en :

- 20 minutes : votre tension arterielle baisse

- 8 heures : votre taux de monoxyde de carbone redevient normal

- 2 jours : votre risque d'infarctus diminue ; vous retrouvez votre odorat.

- 3 jours : vos bronches commencent a se dilater

- 2 semaines : votre circulation sanguine s'ameliore

- 1 mois : l'inflammation de vos sinus diminue ; vous vous sentez moins fatiguee

- 2 moins : votre capacite pulmonaire augmente

- 6 mois : vous recuperez votre energie

- 1 ans : vos rides prematurees disparaissent ; vos risques de maladies cardiovasculaires diminuent de 50%

- 3 a 5 ans : vos risques de cancer du poumon redeviennent normaux

- 10 ans : votre sante et les risques de maladies liees au tabac sont les memes que pour un non-fumeur

Huit bonnes raisons pour vous arreter de fumer :

- vous depenserez beaucoup moins d'argent

- vous vous sentirez plus en forme

- vous ne trepasserz pas en montant un escalier

- vous sentirez bon

- vous aurez une haleine fraiche

VERSION SOUMISE AUX *CAHIERS DU NUMERIQUE* LE 18 VI 2011. MERCI DE SE RÉFÉRER A LA VERSION PUBLIÉE : Eensoo-Ramdani, Egle, Evelyne Bourion, Monique Slodzian, Mathieu Valette (2011) « De la fouille de données à la fabrique de l'opinion. Enjeux épistémologiques et propositions », *Analyse d'opinions sur internet*, Luc Grivel, éd., *Les Cahiers du Numérique*, Volume 7, n°2, pp. 15-39.

- des tas de gens auront envie de vous embrasser
- votre famille arretera de rouspeter
- vous ne mourrez pas prematurement

La sante des paresseuses

Et encore, je n'ai pas parle de l'effet nocif de la cigarette sur la vie sexuelle ...

Texte 3 : militant pro

blog du fumeur, sarko fumeur en public prouve que son décret est idiot

SARKOSY FUMEUR de Cigare information révélée au J.T. sur A2 le Mardi 19/06/07 à 20h par leur journaliste Darmon reçu lors de la présentation du nouveau gouvernement, <<le Président a reçu les journalistes en bras de chemise FUMANT UN CIGARE... >> (sic)

Information confirmée le lendemain 20/06/07 au J.T. sur M6.

Mr SARKOSY en fumant en public prouve que ce décret Anti Tabac est idiot et n'est pas démocratique...

FUMEURS (H/F) et Tolérants au Tabac REAGISSONS CONTRE ce Décret discriminatoire "Anti Fumeur" qui veut nous exclure l'accès à tous les commerces et entreprises,

Si on ne dit rien, la France ça va être RAMADAM toute l'année !!!

EXIGEONS QUE LA LIBERTE SOIT LAISSER AU COMMERCANT ou CHEF D'ENTREPRISE D'ACCEPTER ou non LES FUMEURS DANS LEUR ETABLISSEMENT (comme en Espagne).

Ainsi les intolérants au Tabac pourront rester qu'entre eux si ça leur chante...

Allez consulter un blog très bien documenté sur :http://stenograf.blog.lemonde.fr/

Le site des Fumeurs en colère : http://www.doktorglub.com/dotclear/index.php

Le Forum des Fumeurs Electeurs (version en Français) : http://www.electeurs-fumeurs.eu

Cordialement Ludo à Nîmes (fumeur et non mouton) fumeur-en-colere@hotmail.fr

Texte 4 : militant anti

VERSION SOUMISE AUX *CAHIERS DU NUMERIQUE* LE 18 VI 2011. MERCI DE SE RÉFÉRER A LA VERSION PUBLIÉE : Eensoo-Ramdani, Egle, Evelyne Bourion, Monique Slodzian, Mathieu Valette (2011) « De la fouille de données à la fabrique de l'opinion. Enjeux épistémologiques et propositions », *Analyse d'opinions sur internet*, Luc Grivel, éd., *Les Cahiers du Numérique*, Volume 7, n°2, pp. 15-39.

Arrêter la cigarette...

connaissez les avantages qu'offre le fait d'arrêter de fumer??

LES AVANTAGES SUR LA SANTE:

*Réduction du risque du cancer, d'infarctus et de maladies respiratoires pour nous et notre famille.

*Diminution de la toux, des grippes et des rhumes.

*Diminution des bronchites et des problèmes respiratoires pour nos enfants.

*Récupération rapide et complète des fonctions du coeur et des poumons.

*Augmentation de la capacité de concentration et du rendement intellectuel.

*Augmentation de la résistance pour les activités sportives.

L'AUTO ESTIME:

*Satisfaction personnelle d'avoir remporté une victoire importante.

* sensation de liberté liée au fait de ne plus être conditionné par une dépendance physique et psychologique. ARRETER DE FUMER, C'EST VOTRE VIE QUI EST EN JEU:

[tableau supprimé]