Vers une méthodologie d'analyse des discours sur Internet fondée sur de principes sémantiques. Application à l'analyse de discours de marques, de journalistes et de clients.

Luc GRIVEL, (*, **), Olivier BOUSQUET (, ***)**
luc.grivel@univ-paris1.fr, olivier.obousquet@gmail.com

(*) Université Paris 1, 17 rue de la Sorbonne, Paris (France)
(**) Laboratoire Paragraphe, Université Paris 8, 2 rue de la Liberté, Saint-Denis (France)
(***) Chargé d'études, Harris Interactive, 5-7 rue du Sahel, Paris (France)

Mots clefs :
E-reputation, analyse sémantique, étude d'opinions, retour d'expérience, méthodologie, analyse de contenus, Internet, questions ouvertes, textes journalistiques

Keywords:
E-reputation, semantic analysis, opinion survey, experience learnings, methodology, content analysis, Internet, open end questions, journalistic articles

Stich Wörter
E-reputation, semantische analyse, Opinionstudie , Erfahrung Lehren, Methodologie, Informationsgehalt Analyse, Internet, journalistische Texte

Résumé

Cette contribution s'inscrit dans le cadre de l'axe « expérience de mise en place de systèmes de veille ».

Le travail décrit dans cet article est issu d'un projet de recherche concernant la mesure et la veille de l'opinion sur Internet, au sein du laboratoire Paragraphe de l'université de Paris8, en collaboration avec la société Harris Interactive dans le cadre d'une bourse CIFRE qui a démarré en 2010. L'expérimentation qui est décrite dans cet article concerne une mission de la société Harris Interactive pour un de ses clients.

L'objectif de l'étude commandée est de définir de manière précise les leviers de la relation client à l'égard des professionnels. La phase d'analyse de discours est liée à une phase d'étude qualitative. Elle a porté sur trois types de discours : des discours de marques, des discours journalistiques et des discours de clients.

Dans le premier type, il s'agit d'effectuer un benchmark de sites Internet de marques appartenant au même secteur que le commanditaire à des secteurs très variés. L'objectif de cette étape est d'identifier les clefs les plus exploitées par ces marques dans leurs discours à adressés à la cible. Ce benchmark s'est fait par l'aspiration des sections "Professionnels" des sites des marques. L'analyse des discours clients est fondée sur les réponses à des questions ouvertes issues de questionnaires de satisfaction. Celui qui nous intéressera ici est un questionnaire post-incident, faisant suite à un appel téléphonique à l'assistance technique ou au déplacement d'un technicien. L'étude des discours journalistiques se fonde sur l'analyse d'articles publiés sur des sites grand public spécialisés dans le secteur du commanditaire. Ces sites ont été choisis car ils représentaient le type de sources sur lesquelles la cible recherche de l'information.

Le choix de l'outil pour cette analyse de discours est TROPES. Après avoir justifié ce choix et décrit le fonctionnement de l'outil, la démarche et les résultats obtenus, nous entamons une discussion sur les apports économiques, scientifiques, méthodologiques d'une méthode d'analyse d'opinions basée sur l'analyse sémantique ainsi que sur ses limites techniques et méthodologiques.

Abstract

This article describes an analysis coming from a research project about opinion measurement and monitoring on the Internet. This research is realized within "Paragraphe" laboratory, in partnership with the market research institute Harris Interactive (CIFRE grant beginning in 2010). The experimentation we describe is taken from a Harris Interactive mission.

The purpose of that study was to define precisely CRM levers. The targets of the study were self-employed workers and very small businesses. The discourses analysis is linked to a qualitative study. It turns on three types of discourses : brand, journalists and clients discourses.

In the brand discourses analysis, we benchmarked brand website belonging to several businesses. In this first step, we tried to identify the brands most used words and promises to the target we were studying. To realize that benchmark, we downloaded "Professionals" sections of the websites. Clients discourses analysis is based on opened questions answers coming from satisfaction questionnaires. The questions we are studying have been asked after a call to hot line or after a technician intervention. Journalists discourses analysis is based on articles published on information websites specialized in Harris Interactive's client sector. These websites have been chosen because we considered they were representative of information sources that the target could consult.

The tool that has been chosen for these discourses analysis is TROPES. First, we will justify this choice and describe how the tool works and our approach of the mission. Then we will expose the results of the study. Finally, we will discuss economical, scientific and methodological contribution of an opinion analysis method based on semantic analysis and technical and methodological limits of such a method.

1) Introduction

Constatant la profonde transformation du rapport de notre[1] société à la communication, le laboratoire Paragraphe de l'université de Paris 8 et la société Harris Interactive sont impliqués actuellement dans un projet visant à développer une méthodologie d'analyse automatisée relative à la veille et à la mesure de l'opinion sur Internet. Pour mieux définir le périmètre des recherches à effectuer dans le cadre de ce projet financé en partie par une bourse CIFRE qui a démarré en 2010, nous avons mené au préalable une étude des processus des études marketing et d'opinions (cf. mémoire de master 1 d'Olivier Bousquet) puis une expérimentation d'analyse de discours assistée par un outil d'analyse sémantique (cf. mémoire de master 2 d'Olivier Bousquet). C'est cette expérimentation qui est décrite dans cet article, avec comme fil conducteur la question suivante :

Quels peuvent être les apports et les limites d'une démarche d'analyse sémantique assistée par ordinateur dans le cadre de l'analyse de discours issus du Web, dans une perspective de veille et d'analyse d'opinions ?

En premier lieu, nous définirons le contexte et le processus global de cette expérimentation. Nous décrirons ensuite sur quel type de sémantique nous avons choisi de nous appuyer, justifiant ainsi le choix de l'outil utilisé. Après avoir décrit le fonctionnement de l'outil, la démarche et les résultats obtenus, nous entamerons une discussion sur les apports économiques, scientifiques, méthodologiques d'une méthode d'analyse d'opinions basée sur l'analyse sémantique ainsi que sur ses limites techniques et méthodologiques.

2) Une expérimentation de la sémantique dans un cadre de gestion de la relation client

Nous avons mené cette expérimentation dans le cadre d'une mission de la société Harris Interactive pour un de ses clients. Il s'agissait d'améliorer la performance relationnelle et commerciale de sa marque auprès d'une cible particulière, les TPE. En effet, alors que les particuliers et les entreprises plus importantes (PME et grands groupes) sont des cibles plutôt bien définies, les populations envisagées ici sont parfois difficiles à cerner et la communication à leur égard reste difficile à positionner. L'objectif de l'étude commandée est donc de cerner les leviers permettant à de la marque d'améliorer ses performances.

2.1 Le processus d'étude.

L'étude a été réalisée en trois phases : une phase qualitative d'entretiens individuels avec des professionnels, une phase quantitative de validation et entre les deux, ce qui constitue à la fois la nouveauté et le centre de notre réflexion, une phase d'analyse des discours.

La phase qualitative exploratoire consiste à effectuer une exploration de l'ensemble des dimensions possibles de l'engagement des TPE à l'égard des marques du secteur du commanditaire. Il s'agit aussi de recueillir les attentes non satisfaites, d'identifier les évènements (ou non évènements) qui poussent les clients à basculer vers une autre marque.

[1] les citoyens et les consommateurs, mais aussi les journalistes, les leaders d'opinion, et même les entreprises et les marques…

La phase d'analyse de discours est liée à la phase qualitative. Elle est destinée à entrer dans le processus de définition des items. Elle a porté sur trois types de discours : les discours de marques, les discours journalistiques et les discours de clients.

Dans le premier type, il s'agit d'effectuer un benchmark de sites Internet de marques appartenant au même secteur que le commanditaire ainsi qu'à des secteurs très différents. L'objectif de cette étape est d'identifier les clefs les plus exploitées par ces marques dans leurs discours à adressés à la cible. Ce benchmark s'est fait par l'aspiration des sections "Professionnels" des sites des marques.

L'étude des discours journalistiques se fonde sur l'analyse d'articles parus sur des sites grand public spécialisés dans le secteur du commanditaire. Ces sites ont été choisis car ils représentaient le type de sources sur lesquelles la cible est susceptible de rechercher de l'information.

L'analyse des discours des clients est fondée sur des réponses à des questions ouvertes issues de questionnaires de satisfaction. Celui qui nous intéressera ici est un questionnaire post-incident, faisant suite à un appel téléphonique à l'assistance technique ou au déplacement d'un technicien.

2.2 Choix de l'outil

Le choix de l'outil s'est effectué sur un critère principal : s'inscrire dans une démarche d'analyse sémantique pragmatique, c'est à dire prenant en compte le fait que la signification d'un discours ne peut pas s'envisager sans référence au contexte d'énonciation.

Un deuxième critère important pour le choix de l'outil a été sa facilité de paramétrage : il devait pouvoir être utilisé par tous les chargés d'études.

L'outil d'assistance choisi est Tropes. Ce logiciel trouve son origine, entre autres, dans les travaux du psycholinguiste Rodolphe Ghiglione[2], eux-mêmes influencés par Erving Goffman et Jaakko Hintikka, sur l'analyse automatique des contenus et plus particulièrement sur l'analyse cognitivo-discursive. Pour lui, les enjeux de la communication sont définis par le fait que tout locuteur est inscrit dans un système de communication : il ne parle pas seul. Sa parole, qui n'est que l'expression d'un « monde possible » propre au locuteur, se trouve donc confronté à l'autre, qui possède ses propres « mondes possibles ». La communication est donc un affrontement permanent, son fondement ne peut être qu'argumentatif. Dans ce contexte, les opérateurs syntaxiques jouent un rôle fondamental, car ils sont les armes de l'affrontement, les éléments qui inscrivent la présence et la personnalité du locuteur dans le discours. Ils sont donc des éléments centraux de la théorie de Ghiglione.

Ainsi, ce dernier distingue trois types de mots : les référents noyaux (RN), qui « nomment les objets du monde » (le monde étant ici à entendre comme "monde possible"), les verbes qui inscrivent les RN dans l'univers proposé, et les autres catégories de mots, définies en négatif comme n'étant ni RN ni verbe. Il s'agit entre autre des adjectifs, des modalisateurs, des connecteurs..., tous les mots qui inscrivent le locuteur dans le discours et qui permettent de moduler le sens de ce qui est dit.

Le sens se construit donc par l'articulation des ces trois types de mots au sein de l'unité minimale du sens qu'est la proposition. En effet, Tropes est fondé sur le principe de l'analyse propositionnelle : le discours est découpé en propositions, considérées comme des micro univers qui concentrent un sens simple et autonome.

Il fonde son analyse sur un découpage du texte en propositions, qui s'appuie sur un examen de la ponctuation ainsi que sur l'analyse de la syntaxe (prise en compte des conjonctions, des opérateurs de liaison…). Une proposition contient au minimum un actant (qui fait l'action), un acté (qui la subit) et un verbe (qui l'accomplit). C'est plus ou moins le schéma de la phrase simple : sujet, verbe, complément ; que l'on peut traduire dans le langage de TROPES par : RN

[2] GHIGLIONE [8]

actant, verbe, RN acté. Ce modèle minimal s'étend avec l'ajout de compléments. Dans chaque proposition, on retrouve donc des référents-noyaux, mis en relation par des verbes, définis par des adjectifs, intégrés à l'argumentation par les modalisateurs, les connecteurs et les pronoms.

Le logiciel propose une organisation de ces RN qui est déjà un premier stade d'analyse sémantique. Pour effectuer ce traitement, il utilise un dictionnaire des équivalents sémantiques, qui est une sorte de thésaurus de la langue française. Les mots inconnus au dictionnaire sont présentés individuellement, c'est-à-dire qu'ils ne sont pas intégrés dans le schéma présenté en page suivante. Ce thésaurus est à la base du travail de TROPES, dans ce que les éditeurs du logiciel appellent le « moteur d'analyse linguistique ».

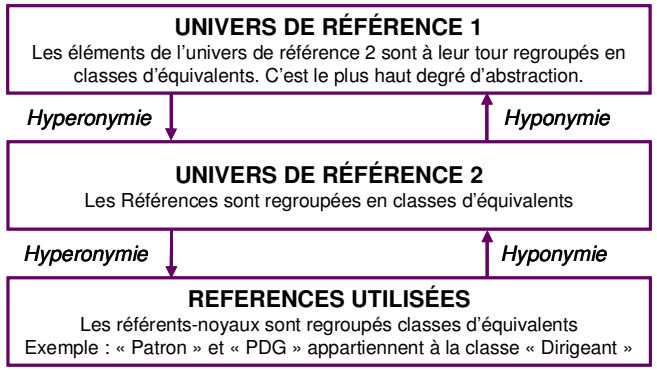

Figure 1 : L'organisation des références (Source : GHIGLIONE [8])

Tropes s'inscrit vraiment dans une démarche de sémantique pragmatique : en plus de proposer un thésaurus général de la langue française, il permet à l'analyste de construire ses propres thésaurus. Chaque analyse s'inscrivant dans un contexte particulier, la construction d'un thésaurus personnalisé permet d'assigner aux mots une signification unique, propre au contexte de l'analyse.

2.3 Paramétrage de l'outil

Ainsi, pour le cas qui nous intéresse ici, il a été nécessaire de définir un thésaurus adapté au contexte de cette étude. Ce thésaurus doit permettre de comparer, d'un côté les discours d'entreprises appartenant à des secteurs très différents entre eux, et de l'autre les discours de ces entreprises avec ceux des clients. Le point commun entre tous ces acteurs est donc bien la relation construite entre la marque et le consommateur. Or, le marketing propose un prisme d'analyse de cette relation : le mix marketing ou les 4 P (Produit, Prix, Distribution ou *Place*, Communication ou *Promotion*). En fait, c'est une libre adaptation de ce mix qui est utilisée à travers la définition de cinq entrées communes à tous les discours étudiés :

- « Matériel » inclut le vocabulaire lié à l'aspect matériel de l'offre de la marque (infrastructures et terminaux)
- « Services et relations » regroupe le lexique ayant trait aux services offerts par la marque, ainsi qu'à la relation client

- « Politique tarifaire » concerne le vocabulaire tournant autour des prix, des offres tarifaires, des promotions…
- « Marque » répertorie toutes les citations de la marque étudiée, mais aussi les marques internes
- « Client / professionnel » marque l'importance du vocabulaire désignant les clients, et en particulier les professionnels.

Ces cinq entrées constituent la matrice de l'analyse des thèmes de CRM abordés aussi bien par les clients que par les marques. Elles sont communes à tous les secteurs et permettent une comparaison. Ensuite, leur contenu est adapté aux spécificités de chaque secteur. On est donc bien dans une situation de sémantique pragmatique : dans les thésaurus, les mots prennent un sens fixe et unique lié au contexte de leur utilisation.

Ces cinq entrées, qui sont elles-mêmes divisées en plusieurs branches, proposent cinq modalités de mise en valeur de l'offre d'une entreprise, cinq profils de marques non exclusifs. L'intérêt d'une telle classification du vocabulaire est de proposer des entrées suffisamment larges pour qu'elles puissent être adaptées à tous les secteurs. Ainsi, les cinq entrées restent toujours les mêmes, mais les notions présentes à l'intérieur sont différentes selon les secteurs, ce qui nécessite la création d'un thésaurus spécifique par secteur dont les entrées sont toujours les mêmes. Par exemple, et il s'agit ici d'un parti-pris, dans le secteur des Télécoms, le terme « Internet » appartient à l'entrée « Matériel », car il s'agit d'une infrastructure (et pas du service en lui-même) ; pour la bancassurance ou l'énergie, il appartiendra à l'entrée « Services et relation », car il devient un moyen de communication, un outil au service de la relation client.

La construction d'un thésaurus par l'analyste prend du temps. Le thesaurus initial a nécessité deux jours de travail pour être défini. Les suivants, qui ne sont que des adaptations du premier, ont nécessité chacun une demi journée de travail. Cette construction a été effectuée par extraction terminologique. Avant d'être analysés, les sites ont tous subi un premier traitement dans TROPES dont le seul but était d'en extraire le vocabulaire pour l'organiser. Dans un cadre aussi précis et bien défini, cette méthode est la plus rapide et la plus sûre pour constituer un thesaurus qui soit à la fois suffisamment exhaustif et ne contienne pas de mots superflus ou inutiles. Pour le dire avec Alexander Geyben, « si une expression fait partie de la langue, elle doit apparaître dans le corpus, et inversement, la fréquence d'une expression dans le corpus est le reflet de sa fréquence réelle dans la langue »[3]. Elle permet aussi, grâce à l'affichage du contexte des mots, de définir de manière sûre le sens précis des mots dans le contexte analysé, et donc de les classer correctement.

TROPES n'effectue pas un simple comptage des occurrences des mots du thesaurus dans les textes. Le logiciel effectue en amont tout un travail de reconnaissance et de catégorisation grammaticale des mots (noms, verbes, adjectifs…), et le fait de les classer dans un thésaurus constitue déjà un acte de sémantique pragmatique, puisque les mots intégrés au thesaurus y prennent un sens fixe lié au contexte de leur utilisation. Le thesaurus est donc l'outil central de l'analyse des discours, le moyen de comparaison des sites étudiés dans le benchmark et le lien entre les trois volets de l'analyse.

3) Résultats obtenus

L'utilisation de TROPES a permis d'obtenir des analyses variées et assez différentes de celles présentées usuellement dans les études. Voici donc une présentation de quelques analyses possibles, qu'il s'agisse de données documentaires, issues de questionnaires quantitatifs ou d'entretiens qualitatifs.

[3] GEYBEN Alexander, [7]

3.1 L'analyse de données secondaires (Benchmark documentaire)

Le benchmark des sites Internet et l'étude des articles journalistiques donnent un exemple de représentations et d'analyses de données secondaires. En ce qui concerne les sites Internet, chacun est synthétisé par une fiche d'identité personnelle. Celle-ci, présentée en figure 3, est divisée en deux grandes parties :

- A gauche, les informations de base sur le site : les statistiques générales du site (nombre de pages, de mots, et de notions utilisées), le Top 10 des notions les plus utilisées (ce terme remplace « classes d'équivalents » dans un souci de compréhension par le client), les pronoms les plus fréquents, les seuils de couverture du discours et la répartition en pourcentage des cinq entrées du thesaurus.
- La partie droite est réservée aux analyses et commentaires sur le site.

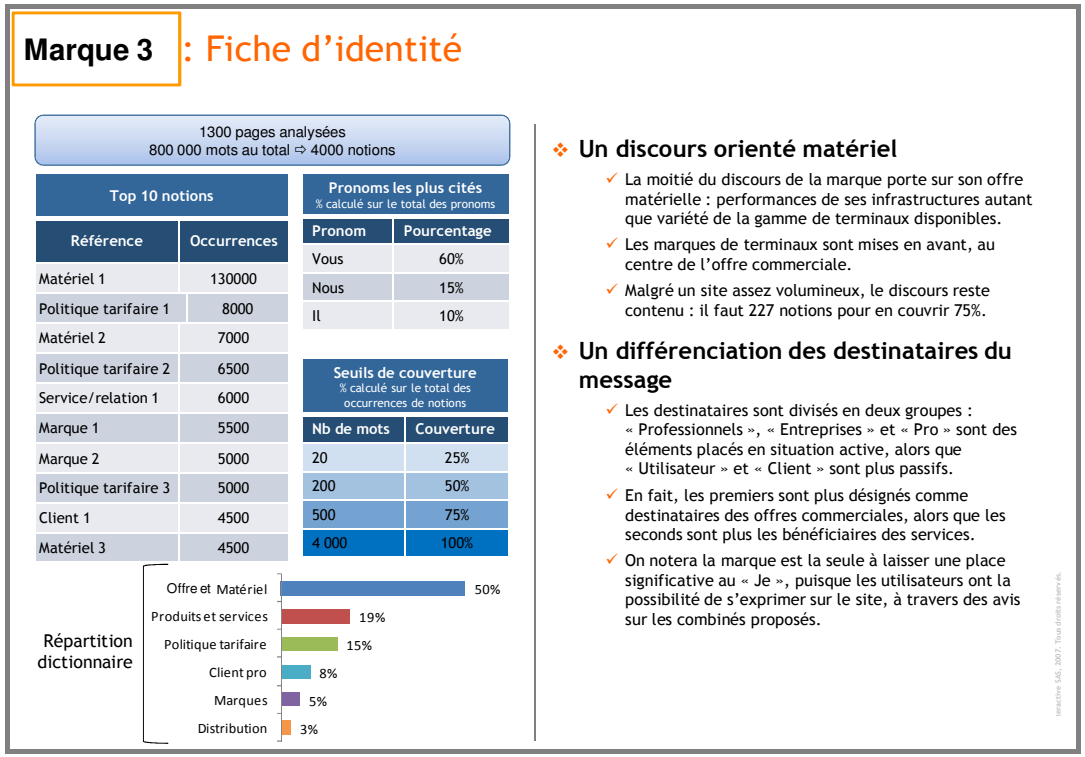

Figure 3 : Fiche d'identité d'une marque

La Figure 4, graphique de la répartition détaillée des notions (classes d'équivalents) définies dans le thésaurus, est fondamental dans la compréhension du discours de la marque puisque c'est lui qui expose l'organisation sémantique, déterminée par les objectifs de l'étude, de tout le vocabulaire utilisé par le site.

Dans l'exemple ici présenté, il montre que la Marque 2 présente son offre en mettant en avant sa dimension matérielle, en particulier ses infrastructures, et les services et relations proposés, à travers entre autres le terme « Solution », véritable mot magique de la relation client pour cette marque. Ces deux entrées représentent plus de la moitié du discours de la marque. On remarque aussi que les citations de la marque occupent près d'un cinquième du discours, avant même les termes liés à la politique tarifaire. On est donc face à une marque qui se met très en avant en tant que marque. Quant au client à qui on s'adresse, il s'agit avant tout d'une entreprise disposant d'une flotte de téléphones, bien plus que d'une TPE.

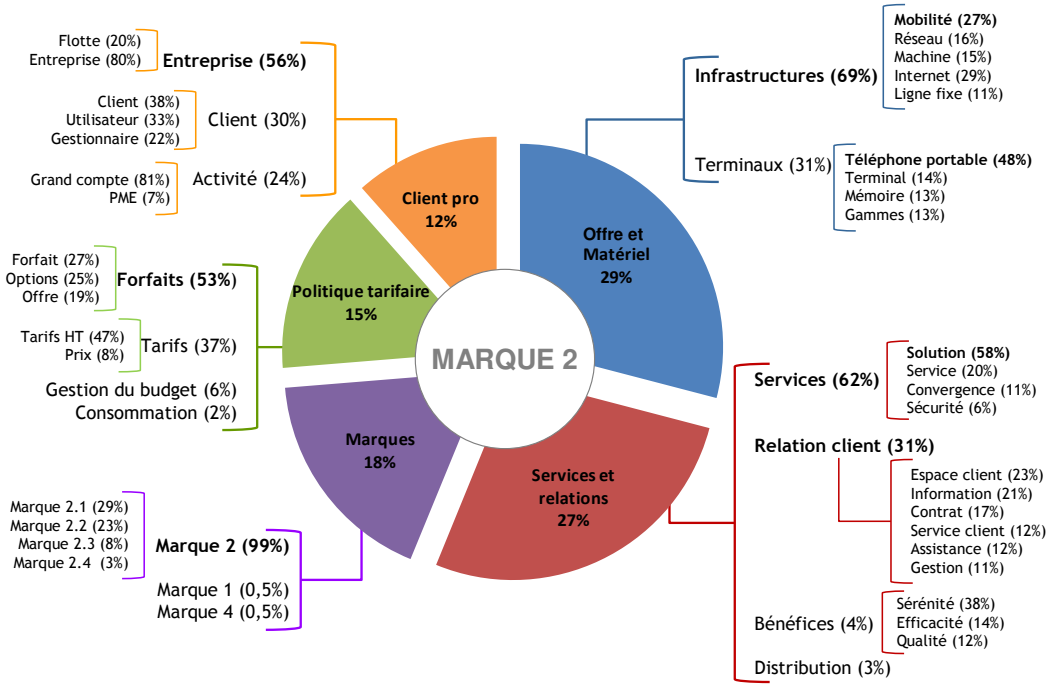

Figure 4 : répartition détaillée des notions (classes d'équivalence) définies dans le thésaurus

Dans le cas qui nous occupe, un autre type de données secondaires a été étudié : des articles journalistiques. Une double approche a été appliquée à ces articles : thématique et sémantique. L'approche thématique a permis, entre autres, de dégager la place des marques dans les articles (principale ou secondaire) et la tonalité. L'analyse sémantique a pu être poussée plus loin que sur les sites Internet, en particulier par l'utilisation du « graphe des acteurs » de TROPES. Ce graphique représente les relations entre les mots. Il est difficile à utiliser car il se fonde sur la structure syntaxique des phrases pour définir si les références sont actantes (agissent sur le verbe), ou actées (objets de l'action). Or les sites Internet présentent souvent des formulations non verbales, surtout les sites

commerciaux sans réel rédactionnel. Cette analyse, présentée en Figure 5, montre la place centrale des opérateurs et Fournisseurs d'Accès à Internet (FAI) dans le discours des journalistes.

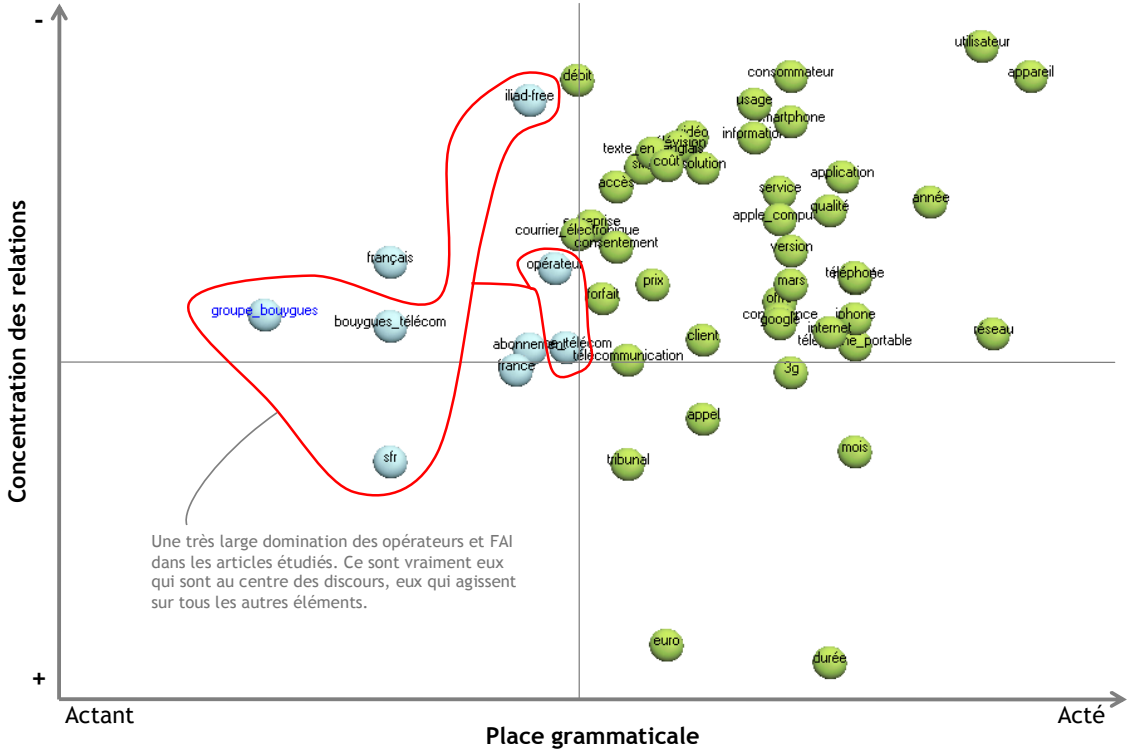

Figure 5 : Graphe de répartition des références actantes et actées

Tous les opérateurs sont présents en position d'actant. L'usager est plutôt de tendance actée, jamais présenté comme un professionnel, ce qui laisse montre que les sites généralistes ne s'adressent pas particulièrement aux professionnels. L'« utilisateur » et le « consommateur » subissent les actions des opérateurs, de tous les opérateurs puisqu'ils sont employés dans des contextes plutôt variés (leurs relations avec les autres mots sont peu concentrées). De manière générale, tout ce qui appartient à la base matérielle de l'offre a tendance à se trouver plus à droite du graphique, c'est-à-dire en position très passive, alors que les vocabulaires de la politique tarifaire et des services sont présents au milieu du graphique. On remarquera la présence des termes « appel » et «tribunal », en position actée et avec des relations faiblement concentrées : ceci s'explique par les articles traitant des déboires d'Orange avec la justice (d'où une concentration des relations, puisque le contexte est toujours le même). En fait, cette analyse ne possède pas un intérêt évident, en grande partie parce que le

corpus n'a pas été défini de manière assez précise. La plupart des informations intéressantes a été tiré de l'analyse thématique manuelle, ce qui pose la question de la pertinence d'une analyse sémantique dans ce cas précis.

3.2 L'analyse d'un discours client issu de questions ouvertes

Le dernier volet de l'étude concernait le discours des clients Professionnels. Il s'agissait là d'une source quelque peu différente, puisque ces données avaient été recueillies par un questionnaire quantitatif. Cet exemple peut donc représenter une manière nouvelle de traiter les réponses aux questions ouvertes.

L'approche de ces données est en effet double. Elles ont été analysées par une cartographie des notions, fondée sur le graphe Actant/Acté, et par une analyse sémantique alliant l'aspect lexical et syntaxique. L'approche par la cartographie des notions offre la possibilité d'une double analyse des réponses, à la fois thématique et discursive. La figure 6 présente la carte des principales notions concernant l'intervention d'un technicien sur place.

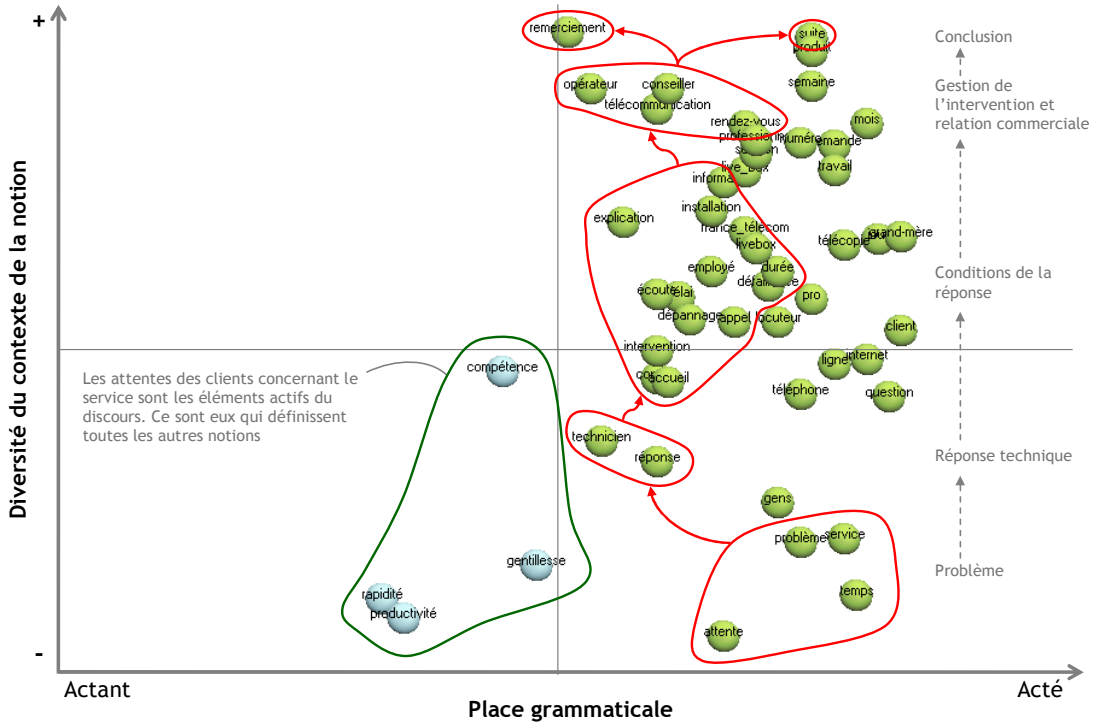

Figure 6 : Cartographie des notions concernant l'intervention d'un technicien sur site

Ce graphique apporte plusieurs types d'informations. Tout d'abord, on remarque que les termes placés en position d'actants désignent tous les attentes des clients à l'égard des techniciens. Il apparaît aussi que toutes ces notions se trouvent dans le bas du graphique, c'est-à-dire qu'elles sont utilisées dans des contextes peu variés. Ces attentes apparaissent donc comme un point de départ des réponses : beaucoup de phrases commencent par ces notions, qui sont une évidence pour les répondants. Une lecture des éléments placés en position d'actés permet de dégager une autre logique de discours. Au début, il y a un problème : les notions clefs (« problème », « attentes »), sont toujours associées entre elles. La perception de la réponse technique et des conditions de cette réponse appartient à un contexte plus diversifié car il existe une variété de problèmes, de réponses et de perceptions. Ceci est encore plus marqué en ce qui concerne la dimension commerciale de cette intervention (prise de rendez-vous, contact avec un conseiller). Enfin, les conclusions, qu'il s'agisse de remerciements ou d'attente d'une suite, s'inscrivent dans des contextes très variés, elles se nourrissent de tout ce qui a été présenté avant.

Cette analyse, complétée par une analyse grammaticale et sémantique permet de comprendre la manière dont les répondants s'impliquent dans leur réponse. On remarque ainsi que les adjectifs (« compétent », « rapide », « aimable », « bon », « efficace », « professionnel », « clair »…) et les verbes (« dépanner », « résoudre », « répondre », « satisfaire »…) utilisés montrent des attentes très fortes à l'égard de l'assistance technique. Pourtant ces attentes sont souvent déçues, ce qui est souligné par l'utilisation de connecteurs d'opposition (qui marquent un jugement contrasté : « mais », « au lieu de », « par contre »…), et de modalisations d'intensité et de négation (qui représentent à elles deux près de 60% des modalisateurs). Des verbes d'injonction (« devoir », « améliorer », « pouvoir », « falloir »…) associés aux adjectifs et verbes évoquant les attentes soulignent bien que les clients attendent du changement de la part de l'entreprise qui ne satisfait leurs besoins. Les opérateurs langagiers montrent donc un véritable engagement personnel (pris en charge par les modalisateurs notamment, mais aussi par les adjectifs et les connecteurs) des clients dans la relation avec la marque et l'insatisfaction face à l'impression de ne pas être entendu (par des injonctions au changement). Il transparaît donc que la relation technique avec la marque semble être le lieu d'un investissement personnel, voire émotionnel, pour les clients, et par là un lieu sensible de la relation client.

Ce type d'analyse peut être compatible avec une codification plus traditionnelle des questions ouvertes, qui dégage des thèmes généraux, mais sans s'attarder sur les enjeux linguistiques. D'ailleurs, l'analyse sémantique peut être utilisée dans le cas d'un traitement par codification. Elle permet de gagner du temps et d'éviter une trop forte influence des premières réponses sur la suite de l'analyse.

4) Discussion et conclusion

L'adoption de cette méthode nous a donc permis d'analyser de manière assez approfondie les discours d'une vingtaine de marques de secteurs différents, de confronter les leviers de leur relation client, afin de mieux envisager celle du commanditaire.

D'un autre côté, l'analyse des discours de clients a permis de confronter ces discours au ressenti des clients lors de leurs expériences de contact avec la marque. Les réponses aux questions ouvertes ont été analysées de manière plus profonde que les discours de marques, car elles s'y prêtaient mieux. En effet, sur les sites Internet la langue n'est généralement pas très construite : par exemple, les phrases sont souvent non verbales, ce qui pose des problèmes pour une analyse syntaxique ou une détermination des actants (les mots qui agissent sur le verbe) et des actés (les mots qui sont objets de l'action). Les réponses aux

questions ouvertes sont très différentes : il s'agit de vraies phrases, construites selon une syntaxe déterminée qu'il est possible d'analyser de manière plus poussée. L'approche de ces données a donc été double : nous avons réalisé une cartographie des notions (par une représentation graphique des notions actantes et actées), ainsi qu'une analyse sémantique alliant l'aspect lexical et syntaxique.

Cette double approche a permis de pénétrer au cœur du discours des clients, d'analyser réellement la manière dont ceux-ci s'impliquent dans leurs discours à l'égard (et à l'intention) de la marque. En plus de dégager des attentes (qu'on aurait pu définir grâce à une analyse thématique plus classique), il a été possible de comprendre la manière dont ces attentes sont présentées, et surtout la manière dont leur résolution (ou non) est vécue par les clients. Une approche graphique a permis de reconstituer la dynamique générale des discours, tandis que l'analyse des formes syntaxiques a permis de comprendre comment, à travers l'utilisation de verbes, d'adjectifs, de modalisateurs ou de connecteurs particuliers, les clients s'impliquent personnellement dans le discours.

4.1 Apports d'une sémantique automatisée dans la compréhension des opinions

D'un point de vue scientifique et méthodologique, elle permet de réaliser des analyses plus fines et plus exhaustives. Pour reprendre l'exemple des questions ouvertes, les outils d'analyse sémantique permettent de construire des plans de code en exploitant l'ensemble d'un corpus de réponses et plus seulement sur une extraction aléatoire d'un échantillon de réponses. Le plan de code obtenu est donc plus précis car porté par un regard exhaustif sur l'information.

La sémantique automatique permet aussi d'aborder les discours d'un point de vue différent des analyses de contenu traditionnelles (bien que, on l'a vu, elle soit aussi un outil qui les soutient). Elle permet notamment d'entrer dans la peau de l'émetteur du message. En effet, la sémantique, qui se concentre sur la construction du sens, dépasse l'analyse de contenu. Elle prend à la fois en compte le contenu (ce qui est dit, le *dictum*) et le contenant (comment c'est dit, le *modus*). En s'attardant de manière systématique sur les constructions syntaxiques, sur l'utilisation des modalisateurs, des adjectifs et de tous les mots qui marquent la présence du locuteur et du contexte d'énonciation dans le discours, la sémantique offre une vision plus complète des discours. Elle souligne surtout la manière dont les locuteurs s'impliquent dans le discours, que cela relève de l'émotion ou de l'argumentation. Elle permet aussi d'inscrire les discours et ceux qui les prononcent dans des groupes plus larges. Pour le dire avec Roland Barthes, « toute parole est fatalement inscrite dans un sociolecte »[4], c'est-à-dire que les discours ne sont jamais les discours d'un seul individu : chaque individu partage une intersubjectivité avec les membres de son (ou plutôt de « ses ») groupe(s). L'analyse sémantique pragmatique devrait permettre de pénétrer cette intersubjectivité du discours[5], c'est-à-dire de prendre en compte l'inscription du locuteur dans un contexte, dans une interaction, dans une histoire. Elle doit permettre de dégager les spécificités discursives propres à des groupes définis *a priori* (selon des critères « objectifs » comme le sexe ou l'âge) ou *a posteriori* (par la prise en compte des régularités et des invariants du discours).

Enfin, l'analyse sémantique automatique permet de traiter de manière qualitative des échantillons beaucoup plus importants. Jusqu'à présent, les études qualitatives se limitent à l'interrogation d'échantillons relativement restreints. Cette limite est avant tout pratique : interroger qualitativement un grand nombre

[4] BARTHES [1]
[5] LARSSON [10]

de personnes revient très cher, à la fois en temps de recueil et en temps d'analyse et d'interprétation. Avec les outils informatiques d'assistance, et notamment grâce à l'analyse sémantique automatisée, il est possible d'analyser les réponses qualitatives d'échantillons plus importants. Il devient ainsi envisageable d'atteindre le seuil de saturation de l'expérience, c'est-à-dire le moment où toutes les expériences possibles sur un sujet auront pu être envisagées.

4.2 Limites techniques

D'un point de vue technique, la principale limite de l'approche sémantique en elle-même, celle qui semble constituer le frein le plus important, est la difficulté de s'adapter au langage utilisé dans les discours spontanés. Prenons le forum, comme exemple de lieu de discours spontanés. Sur un forum dédié à une entreprise, le nom de l'entreprise en question va de soi et les participants en parlent à la troisième personne. Comment alors repérer automatiquement les moments où l'on en parle, puisqu'on ne peut évidemment pas décréter que tout référent d'un pronom de troisième personne est l'entreprise ? Comment aussi gérer les tours de parole, puisque ce qui caractérise ces discussions ce sont les interruptions, les prises de parole inopportunes ? Et puis, comment gérer les décalages dans les débats ?

En effet, les tours de parole ne se suivent pas logiquement mais chronologiquement. Ce qui décide de l'ordre de parution des participations, ce n'est pas la logique des réponses mais le temps de rédaction d'une contribution. Les contributions ne sont pas forcément des réponses au dernier mail publié, mais à une question qui a pu être posée, cinq ou six mails auparavant. Le traitement des anaphores est déjà un réel problème au sein d'une phrase. Que dire, dans le cas des tours de parole sur des forums de discussions, puisque souvent le référent d'une anaphore ne figure même pas dans la proposition précédente ! Sur le plan du style, les auteurs « écrivent comme ils parlent », utilisent beaucoup d'abréviations, omettent les signes diacritiques, les majuscules, ou les marques de ponctuation, et commettent des coquilles, des fautes d'orthographe, d'accord ou de syntaxe. En effet, il reste très difficile pour des outils configurés selon les modèles de la langue canonique d'analyser des constructions langagières aussi informelles et concises que celles employées dans les forums, sur Twitter, etc.

Il est ainsi indispensable d'inclure les abréviations dans les listes terminologiques. Il est bien évidemment inconcevable d'inclure dans les dictionnaires toutes les orthographes incorrectes. Il me semble donc indispensable d'inclure un correcteur automatique d'orthographe ou du moins un outil qui tenterait de rapprocher les formes inconnues avec les formes lemmatisées des dictionnaires terminologiques.

Pour traiter des textes en langage naturel de ce type, il est nécessaire d'effectuer en amont un énorme travail de nettoyage des textes qui reste très chronophage.

Dans sa chronophagie, le nettoyage des textes peut être rapproché d'une autre étape, celle de la constitution du corpus. En effet, en analysant sémantiquement et de manière informatisée des quantités conséquentes d'informations, on se place d'emblée dans une logique de linguistique de corpus. Or, dans ce domaine, la définition des sources est primordiale. Le regroupement de textes dans un corpus est en soi une première démarche sémantique. En effet, en choisissant les éléments de son corpus, l'analyste propose une première étape d'interprétation des sources à travers la problématique de son étude. Elles doivent présenter une cohérence, et même avoir une dimension de représentativité : si on étudie par exemple l'image d'une marque à travers des discours issus de forums, c'est que l'on considère que les forums choisis pour l'analyse sont représentatifs de ce qui peut se dire sur l'ensemble des forums de la toile, voire de ce que pensent l'ensemble des consommateurs de la marque (on est ici confronté au problème de la connaissance des émetteurs des discours sut Internet) .

C'est pourquoi le corpus doit être défini avec une infinie prudence, le plus souvent en relation avec le commanditaire. Cette nécessité peut constituer une limite à l'introduction de la sémantique automatisée dans une démarche économique, car elle implique d'accorder beaucoup de temps à la définition du corpus.

D'autres limites techniques s'ajoutent sitôt que l'on cherche à traiter de l'information issue d'Internet. En effet, pour analyser des pages Web, la solution la plus simple et la plus rapide est de les aspirer, de les enregistrer en local, pour effectuer des traitements dans un second temps. Cette étape de l'aspiration des sites est aujourd'hui soumise à des limites liées à la structure des pages Web et à la façon dont elles sont générées.

Comment reconnaître dans une page Web où se situe le message que l'on veut analyser ?

Bien sûr, il existe des heuristiques. Outre le contenu recherché, le plus souvent une page web contient une barre de navigation, des publicités, des liens vers autres articles, des informations légales, etc. La barre de navigation est riche en pointeurs HTML ; les publicités sont riches en pointeurs et en images; La publicité peut changer à chaque chargement de la page ; les liens vers les autres articles sont riche en texte en en pointeurs ; et les informations site et légales sont constants sur site. Le « message » à l'opposé, est riche en texte, pauvre en pointeurs.

L'automatisation de l'aspiration des pages web pose aussi le problème de la hiérarchisation des informations étudiées. Il est en effet impossible de connaître, lors de l'analyse, l'étendue du public d'une page, et donc de hiérarchiser les pages aspirées les unes en fonction des autres. Ainsi, on peut se demander si l'information qui se trouve dans une page très visitée (comme par exemple sur la page d'accueil d'un site) a la même valeur que celle qui se trouve sur une page peu visitée.

4.3 Renouveau et incertitudes méthodologiques

La première des remises en cause liées à la sémantique automatisée est celle de la séparation entre qualitatif et quantitatif. En effet, cette démarche emprunte aux deux types de méthodologies. Son objet est le discours brut, en cela elle peut être considérée comme une méthodologie qualitative. Cependant, elle applique à cette matière première des traitements quantitatifs : pour pouvoir être traitée informatiquement, la langue doit être modélisée et donc se voir appliquer une logique mathématique. De plus, les outils de traitement automatique du langage fournissent des données quantitatives : les occurrences des formes linguistiques sont exprimées sous forme de statistiques, de graphiques ou de tableaux, qui sont avant tout des représentations quantitatives. Pourtant, l'analyse sémantique automatisée reste ouverte sur le qualitatif car il est possible, et même nécessaire, d'effectuer des retours permanents au texte brut. Ce retour qualitatif permet de replacer les formes linguistiques soulignées par l'outil dans le contexte de leur contexte d'utilisation. Cette recontextualisation est nécessaire pour la compréhension des textes. Sans elle, le risque de contresens est très important. A aucun moment la quantification ne peut se suffire à elle-même. La sémantique automatisée propose donc véritablement une méthodologie et des outils à double tranchant : elle applique à un matériau qualitatif (le discours) une logique statistique et probabiliste, pour des résultats oscillant en permanence entre les deux logiques.

L'analyste qui utilise de tels outils doit donc être à l'aise avec les deux versants de la méthodologie. Celle-ci sort des cadres habituels des études de marché (où elle a été expérimentée) et des sciences humaines et sociales : elle peut donc se heurter à de fortes réticences (dans chacun des deux milieux). Le consensus n'est pas une évidence.

La méfiance potentiellement soulevée par les méthodes de statistique textuelle peut s'analyser sur deux plans. Le premier, le plus évident, est celui de la fiabilité et de la pertinence des résultats. Quelle confiance peut-on accorder à ces outils, et surtout quelle est la valeur des informations qui sont fournies par l'outil ? L'autre point de méfiance, peut-être moins avouable, est celui d'une crainte d'entrer en concurrence avec la machine, de voir la valeur de l'analyse humaine rabaissée par le recours à un outil. En fait, ces deux réticences sont liées. Car il est nécessaire de comprendre que l'outil n'est rien sans l'intelligence humaine, sans la faculté d'interprétation de l'analyste. L'outil ne peut être qu'une aide pour l'analyste, qui garde toute sa légitimité et est même le garant de la fiabilité et de la pertinence des résultats.

Pour bien comprendre la place de l'analyste dans le process d'une étude utilisant la sémantique automatisée, il peut être utile de faire appel à une distinction faite par François Rastier[6] et de l'adapter à notre propos. Dans son analyse des systèmes de compréhension, ce dernier distingue les trois termes *analyse*, *interprétation* et *compréhension*. Un système de compréhension est définit comme " tout système qui tente de passer d'un arbre syntaxique à un réseau sémantique, et de faire des inférences au sein de ce réseau." Rastier propose un cheminement en trois étapes. A chaque étape, il est possible de distinguer le rôle de la machine de celui de l'analyste.

La première étape, celle de l'arbre syntaxique, correspond à *l'analyse*. On peut la rapprocher des analyses morphologique et syntaxique évoquées plus haut dans la description des méthodes de l'analyse automatique des textes. Cette analyse est entièrement réalisée par le logiciel, qui définit les mots et leurs relations logiques.

La seconde étape, celle du réseau sémantique, correspond à *l'interprétation* et peut être le fait du logiciel comme de l'analyste. Il s'agit ici de dégager une "signification" dans le sens que lui donne Rastier, c'est à dire "du sens appauvrit car coupé de son contexte". Certains logiciels permettent d'automatiser cette étape. C'est par exemple le cas de Tropes qui utilise deux méthodes pour dégager de la signification. Tout d'abord, il extrait les marqueurs syntaxiques, les modalisateurs, etc., qui organisent l'énoncé en marquant la présence du locuteur (voire des interlocuteurs) dans le discours. Parallèlement, il classe et hiérarchise les notions du texte grâce à son thésaurus de la langue française. Ainsi, pour chaque mot, le logiciel propose une signification en définissant des liens de synonymie, d'hyponymie et d'hypéronymie. Par exemple, des termes comme "TPE", "PME" et "société" sont considérés comme synonymes et appartiennent à la classe d'équivalents "Entreprise". Toutefois, cette signification est toujours très abstraite et le risque de polysémie reste fort, car l'interprétation ne tient pas compte du contexte.

C'est alors qu'on entre dans la dernière étape du système de compréhension, qui est la *compréhension* en elle-même. Cette étape se situe au niveau purement mental, donc humain, et permet de passer de la signification au sens, c'est à dire de "créer des inférences au sein du réseau sémantique". L'analyste utilise tous les éléments dégagés par le logiciel, il les relie, afin de dégager le sens final du texte. Ainsi, la construction d'un thésaurus personnalisé permet de définir pour chaque mot un sens précis lié au contexte particulier de l'analyse. C'est à ce niveau que se dégage la valeur ajoutée de l'analyse, car elle est nourrie par les connaissances de l'analyste, qui dispose de données externes au discours analysé, d'une mémoire et d'une réflexion critique lui permettant d'extraire l'information utile du texte.

Il apparait donc, à travers ce recours au système de compréhension, que la complémentarité est réelle entre le logiciel et l'analyste. En fait, il ne faut pas perdre de vue que le logiciel, aussi puissant soit-il, reste avant tout un outil d'assistance, une béquille pour l'analyste dont les capacités à repérer les informations stratégiques sont toujours au centre de la valeur de l'étude.

---

[6] RASTIER [12]

L'autre remise en cause impliquée par l'analyse sémantique automatisée concerne la notion de « données ». En plaçant le discours au cœur de ses préoccupations, celle-ci met en évidence le jeu de l'échange entre celui interroge et celui qui est interrogé. La réflexion sur la sémantique permet de replacer les paroles (prononcées ou écrites) dans le contexte de leur production, dans l'échange qui a été à l'origine de la formulation du discours (même dans le cas d'un discours extrait d'Internet, il existe un échange au moins implicite entre un émetteur et un récepteur). Elle permet aussi de réintroduire la personne interrogée dans le ou les groupes auxquels il appartient. Là où les analyses plus traditionnelles mettent en évidence les particularités de cibles (définies par des critères objectifs tels que l'âge, le sexe ou la consommation d'un produit), l'analyse sémantique s'attache désormais à des publics (appartenant à des groupes sociaux divers, ayant un passé, évoluant dans un contexte particulier…)[7]. C'est donc la question de la connaissance de l'émetteur du discours qui est placée au centre des réflexions par l'analyse sémantique. Ce qui implique d'autres interrogations, notamment par rapport à la collecte de discours de consommateurs sur le Web : on ignore le plus souvent qui sont les personnes qui s'expriment sur le Web et qui elles représentent. Il serait donc intéressant de s'interroger sur l'identité de cet internaute, de mener une réflexion sur les critères de connaissance de l'internaute, de se demander si être un internaute qui s'exprime sur un forum, ce n'est pas déjà une forme d'identité. Cette question de la validité d'une analyse menée sur de discours de personnes dont on ignore presque tout semble poser une limite de la méthode.

Enfin, la mise en place d'une solution d'analyse sémantique automatisée a un coût non négligeable. Il s'agit d'un investissement de R&D. L'achat d'un outil, le temps de découverte puis de formation des salariés, tout cela coûte cher et constitue un pari sur l'avenir. La mise en place d'une solution d'analyse sémantique automatisée est au minimum un investissement à moyen terme, pas toujours évidente dans un secteur comme les études de marché dont la visibilité peine à dépasser quelques mois.

Eu égard à toutes ces limites, l'utilisation de la sémantique automatique pour l'analyse d'opinions reste donc pour l'instant avant tout une méthodologie de complément, qui vient soutenir les méthodologies existantes. Elle complète ces méthodologies de deux manières différentes : en permettant un traitement plus rapide et plus facile de certaines étapes (questions ouvertes, comptages qualitatifs…), et en proposant un éclairage différent sur une problématique traitée par ailleurs selon des méthodologies traditionnelles d'analyse de contenu.

[7] cf. MARC, TCHERNIA [14]

Bibliographie

[1] BARTHES Roland, *Le bruissement de la langue. Essais critiques IV*, Seuil, Points Essais, 1984, 439 p.

[2] BEAUDOIN Jean-Pierre, *L'opinion,c'est combien ? Pour une économie de l'opinion*, Village Mondial, 2005, 237p.

[3] BOURDIEU Pierre, « L'opinion publique n'existe pas », in *Questions de sociologie*, Les Editions de Minuit, Reprise, 1984, p. 222 à 235

[4] CONDAMINES Anne, « L'interprétation sémantique de corpus : le cas de la structuration de terminologies », in *Revue française de linguistique appliquée*, XII-1, Juin 2007, p. 39 à 52

[5] DEMAZIERE Didier (dir), *Analyses textuelles en sociologie – Logiciels, méthodes, usages*, PUR, Méthodes, 2006, 219 p.

[6] FUCHS Catherine (dir), *Linguistique et traitement automatique des langues*, Hachette-Classiques, HU Linguistique, 1993, 303 p.

[7] GEYBEN Alexander, « Quelques problèmes observés dans l'élaboration de dictionnaires à partir de corpus », in *Langages*, 171, Septembre 2008

[8] GHIGLIONE Rodolphe (dir), L'analyse automatique des contenus, Dunod, Psycho Sup, 1998, 168 p.

[9] JENNY Jacques, « Quali / Quanti – Distinction artificielle, fallacieuse et stérile ! », *1er congrès de l'AFS*, Groupe RTF 20, Session n°4, 25 février 2004, consultable à l'adresse http://testconso.typepad.com/files/jenny-quanti-quali.pdf (le 22 août 2009)

[10] LARSSON Björn, « Le sens commun ou la sémantique comme science de l'intersubjectivité humaine », in *Langages*, 170, Juin 2008, p. 28 à 40

[11] MARTIN Robert, *Sémantique et automate*, PUF, Ecritures électroniques, 2001, 190 p.

[12] RASTIER François, CAVAZZA Marc, ABEILLE Anne, *Sémantique pour l'analyse. De la linguistique à l'informatique*, Masson, 1994, 240 p.

[13] TAMBA Irène, *La sémantique*, PUF, Que sais-je ?, 2005, 128 p.

[14] MARC Xaxier, TCHERNIA Jean-François (dir), *Etudier l'opinion*, PUG, 2007, 260 p.