Available for free online at https://ojs.hh.se/

Journal of Intelligence Studies in Business 1 (2011) 76-86

A discourse analysis methodology based on semantic principles - an application to brands, journalists and consumers discourses

Luc Grivel* and Olivier Bousquet**

*INDEX-PARAGRAPHE, Université de Paris 8, 2, rue de La LIBERTE - Saint Denis, France, Université de Paris 1, (Panthéon-Sorbonne), Paris, France

**INDEX-PARAGRAPHE, Université de Paris 8, 2 rue de La LIBERTE - Saint Denis, France Harris Interactive, Paris, France

luc.grivel@univ-paris1.fr, olivier.obousquet@gmail.com

Received 20 July 2011; received in revised form 10 September 2011; accepted 29 December 2011

ABSTRACT: This is a R&D Paper. It 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 July 2010). The purpose of the study was to define CRM possibilities. The targets of the study were self-employed workers and very small businesses. The discourses analysis is linked to a qualitative study. It turns around three types of discourses: brands, journalists and clients' discourses. In the brand discourses analysis we benchmarked brand websites belonging to several businesses. In this first step, we tried to identify the most used words and promises by brands to the target we were studying. For that benchmark, we downloaded "Professionals" sections of the websites. Clients' discourses analysis is based on opened answers coming from satisfaction questionnaires. The questions we are studying have been asked after a call to a 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 were chosen because we considered them to be representative of information sources, which the target could consult.

Keywords: Discourse analysis, Brand management, Market research

1. Introduction

Regarding the deep change in our relation to communication, University Paris 8's Paragraphe laboratory and the market research institute Harris Interactive started a common project in 2010, aiming at developing a methodology to monitor and measure opinions on the Internet. To define our research area, we first analyzed various opinion and market research processes (Master 1 essay). Then

we did discourses analysis with computer tools assistance (Master 2 essay) for answering the question:

What are the contributions and challenges of a computer assisted semantic analysis, within the analysis of Web-coming discourses?

This article describes that experiment. The tool that has been chosen for this discourse analysis is TROPES. First, we will justify this choice and describe how the tool works, our approach of the mission and then present the results of the study. Finally, we will discuss economical, scientific and methodological contributions of an opinion analysis method based on semantic analysis, and identify the technical and methodological limits of this method.

1.1. A semantic experiment in a CRM context

The experiment takes place as part of a Harris Interactive mission for a client. The goal of that mission was to improve the brand's performances in terms of relations with specific clients: very small firms. These ones are particular targets and it can be difficult for a great brand to define the way to communicate with them. The research was aimed to identify the levers that could be pulled to improve the brand's performances.

2. Methodology

The research is divided in three phases: a qualitative phase (individual interviews of professionals), a quantitative one (validation and establishment of a decision model based on ideas defined in the first phase) and a phase of discourse analysis. This step is the one which we are focusing on in this article.

The qualitative phase is an exploration phase. Its goal is to explore every possible dimensions of very small firms' engagement to a brand, specifically in the brand's sector. We also needed to understand the expectations of the targets and the way they are satisfied or dissatisfied, and to identify experiences that can make clients leave the brand for another one.

The discourses analysis is linked to the qualitative phase. It is about three types of discourses:

- Brand discourses: this analysis is based on a benchmark of twenty websites of brands belonging to varied business sectors. We searched the "Professionals" sections of the websites. This discourse analysis enabled us to identify words, expressions and different types of discourses that these brands use when they are communicating to professionals.
- Journalistic discourses: this is based on analysis of articles. These articles are taken from mass specialized website, chosen because they represent the type of sources that are used by the targets.
- Consumer discourses: this analysis is based on answers to open ended questions in satisfaction surveys. The one that we used is a survey that had been sent after a call to a hotline or after an intervention by a technician.

2.1 Tool choice

A main criterion that led to the choice of the tool is that it should be based on semantic-pragmatic principles. That means that the tool had to allow an analysis, taking into account a specific conception of meaning: the meaning of a discourse can't be understood without a reference to the enunciation context.

A second criterion that has been important for the choice is ease of use: it should be as easy as possible since all research executives should be able to use it

The third criterion was linked to market research structures. The experiment research was an adhoc research and it was not certain that it would be followed by other similar studies. Thus the tool had to be adapted to this specific logic. For instance, a global monitoring solution, that most often implies a yearly subscription or additional software development, was not adapted.

Following these three fundamental criteria, the chosen tool has been Tropes. This software is based on the work of Ghiglione (1998), a psychological linguist. Inspired by Goffman and Hintikka, Ghiglione (1998) worked on automated content analysis, and more particularly on cognitive and discursive analysis. His idea is that communication issues are defined by the fact that every speaker takes place in a communication system: he never speaks alone. His speech is the expression of a "possible world", that is personal to the speaker, but which is in dispute with other people, who have their own "possible worlds". Communication is like a permanent clash between subjectivities, and it has to be based on argumentation. In that context, syntactical operators play a fundamental role; they are weapons used in the fight, the discourses elements that impose the speaker's personality. They are central elements in Ghiglione's theory.

Thus, this study distinguishes between three types of words: references that "name the objects of the world", verbs that place RN in the proposed universe, and the other categories of words, negatively defined as all words that are neither a reference nor a verb. These words are, among others, adjectives, modalities, connectors, all words that show the speaker through the discourse and adjust the meaning of what is said.

Therefore, meaning is built by the articulation of these three categories of words, inside phrases, considered as the smallest meaning unity. Tropes is

based on propositional analysis principles: discourses are cut in propositions (simple phrases), considered as micro-universes concentrating a simple and self-sufficient meaning.

The analysis is based on the text cutting in propositions, based on a punctuation and syntax analysis (conjunctions, syntactic links and so on). A proposition is at least made of an "Actant" (from French, which acts), an "Acted" (that is subjected to the action) and a verb (that makes the action). This minimalistic model can be extended, adding complements. In each proposition, we can find Referent Nucleus linked by verbs, defined by adjectives and integrated to argumentation thanks to modalities, connectors and pronouns.

The software allows a first step of meaning analysis through an organization of the references. This organization is based on an internal dictionary, like a generalist thesaurus of French language. When a text contains a word that is missing in the dictionary, this is individually underlined. This means that the word is not integrated in the following diagram. This thesaurus is the foundation of Tropes' work, in what the developers call a "linguistic analyze engine".

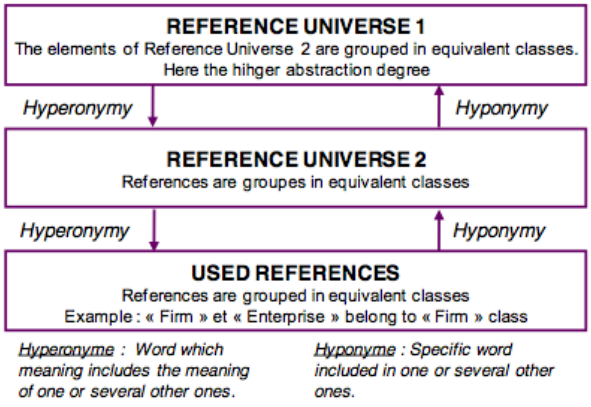

Figure 1 : References organization

Thus, Tropes allows a semantic-pragmatic approach. As well as proposing a generalist thesaurus of French language, it enables the analyst to build his own thesauruses. As every analyzes is inscribed in a specific context, the building of a particular dictionary allows the analyst to give a specific and unique meaning to every word, linked to the context.

2.2. Tool configuration

In this article, we have developed a specific thesaurus, suitable for the research context. The role of the thesaurus is to enable comparison between the discourses of firms from varied economic sectors, but also to compare the discourses of these firms with the client's. Their common point is the relation built between brand and consumer. Thus, Marketing has developed an angle of analysis for that relation: the Marketing Mix, or the 4P's (Product, Price, Place, Promotion). We use an adaptation of that tool and we have defined five common entries for all studied discourses:

- "Material": notions linked to material aspects of the firm offer (Infrastructures and terminals)
- "Relations and Services": vocabulary linked to the services offered by the firm, and client relationship
- "Pricing policy": vocabulary linked to prices, pricing offers, sales and so on
- "Brand": quotes of brands and sub-brands
- "Client or Professional": shows the consistency of the vocabulary naming the clients, particularly professional clients.

The five entries are the base in the analysis of the CRM themes quoted by the firms and the clients. These are common to all sectors and they enable comparison. In a context of topic-centered analysis, this choice seems to be problematic. As written by Pang and Lee (2008) comparing topic-centered analysis to sentiment-centered analysis templates in "traditional information extraction can differ greatly from one domain to another". This is why each entry's content was specific to each sector. In thesauruses, words have a unique meaning, linked to the context in which they are used, called pragmatic-semantic.

These five entries are themselves switched in several branches. They underline five ways a firm can showcase its offer with five brand profiles (different but not exclusive). The entries are large enough to be operative for all sectors. Thus, the five entries are always the same, but the notions that compose them are specific to each sector. This resulted in creating a specific thesaurus for each sector (without changing the five entries). For example, in the telecommunication sector, we classified the notion "Internet" in the "Material entry" because we considered that it is an infrastructure (and not a service itself). For energy or bank/insurance sectors, "Internet" is classified as "Relations and Services" because it becomes a communication device, a tool linked to client relations.

Building a thesaurus is quite time consuming. It took two days to build the first thesaurus. The following ones, which are just adaptations of that first one, have been built in half a day each. This work has been made possible by notions extraction. Before the discourses analysis, websites have been analyzed with Tropes to extract the vocabulary and notions that should be organized. In such a framework, this method appears as the safest to build an efficient thesaurus, which means a thesaurus that is exhaustive but without unnecessary notions and words. Geyken (2008)

states that if an expression is part of the language, it must appear in the corpus, and conversely the frequency of an expression in the corpus is the image of its frequency in language. The software shows the words in their context, which allows the analyst to define the meaning of words in the specific context of the text, and to classify them correctly.

In the end, we notice that the thesaurus, on the contrary to what it may seem, do not only make a lexical analysis. Even if that method is focused on the vocabulary used in the text, Tropes does not only count occurrences of lexical forms. (By lexical form this article refers to a series of characters between two spaces or punctuation signs). The software has previously created a word based on recognition and categorization of words (nouns, adjectives, verbs and so on) and the fact that the analyst classifies these forms in a thesaurus is a first step in pragmatic-semantic. Words included in the thesaurus have a unique meaning, linked to the usage context. Therefore the thesaurus appears both as the central tool of computer-assisted discourses analysis and as a way to compare the various websites of the benchmark, as well as the element that links the three steps of analysis.

3. Data, Analyze and Implications

Through Tropes, we obtained various analyses quite different from the ones we usually obtain in market research. In this part, we are presenting some possible analysis on different data: documentary data and open-ended questions.

3.1 Analyzing secondary data: websites benchmark and journalistic articles analysis

The website benchmark and the journalistic articles analysis are two examples of secondary data analysis. When it comes to websites, each one is synthesized in a personal identity card. This is described in Figure 2 and it is divided in two parts:

- On the left side, basic information about the website: general statistics (number of pages, words, used notions), Top ten most used notions (what we call "Notions" is actually the "equivalent classes", but translated into a more accessible word here), frequently used pronouns, discourses concentration and the distribution (in percents) of the five entries of the Thesaurus.
- The right part is dedicated to analysis and commentaries about the website.

Figure 3 represents the detailed distribution of equivalent classes defined in the thesaurus. It is fundamental to understand brand discourse. It describes the semantic organization of all the vocabulary on the website and that organization is partly determinate by the objectives of the study.

In this example, the discourse brand underlines the material dimension of its offer, particularly concerning infrastructures. The discourse brand also insists on relations and services. We notice the importance of the word "Solution", which appears as a central word in a client's relation to that brand. More than half of the brand discourse is contained in the two entries material, relations and services. We also notice that brand quotations are more than one notion out of five, which is more than the pricing policy. This brand seems to be self-centered, when it comes to highlighting its brand. Concerning the client, a firm belonging to a mobile fleet, is not a small firm.

In the case we are describing here, another type of secondary data has been studied: journalistic articles. A double approach has been necessary: thematic and semantic. The thematic approach defines the importance of brands in articles (principal or secondary place) and the tonality of the articles. Semantic analysis has been more precise and complete than the one on websites. We used Tropes' "Actant chart". This chart represents relations between words. It is based on syntactic structure of sentences. On the horizontal axe, references are defined as "Actant" (acts on the verb), or "Acted" (object of the action): The further to the right a notion is, the more passive it is in the text. Vertically, this chart represents concentration of relations between notions. The higher a notion is, the wider is its usage context. Thus, websites are often less redacted, with lots of non-verbal phrases (at least in commercial websites). The analysis presented in Figure 4 describes the central place of telecommunication companies in journalists' discourses.

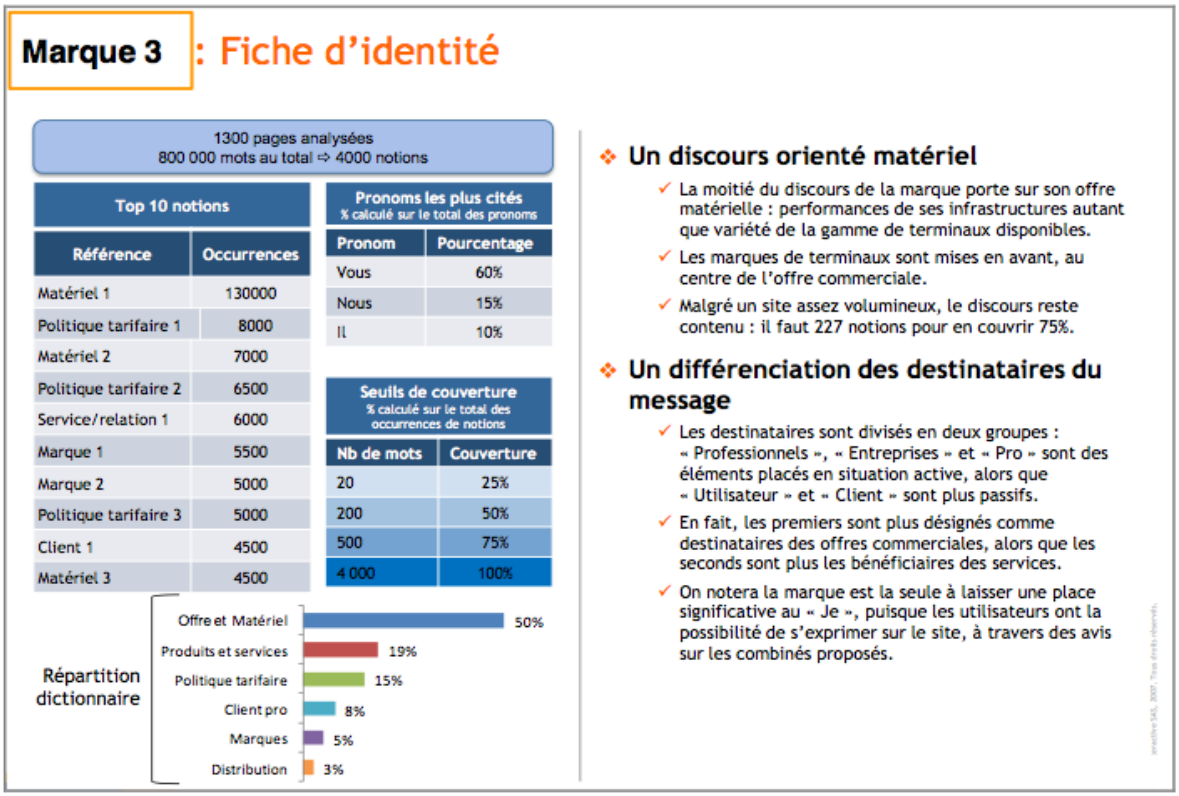

Figure 2: Identity card of a brand

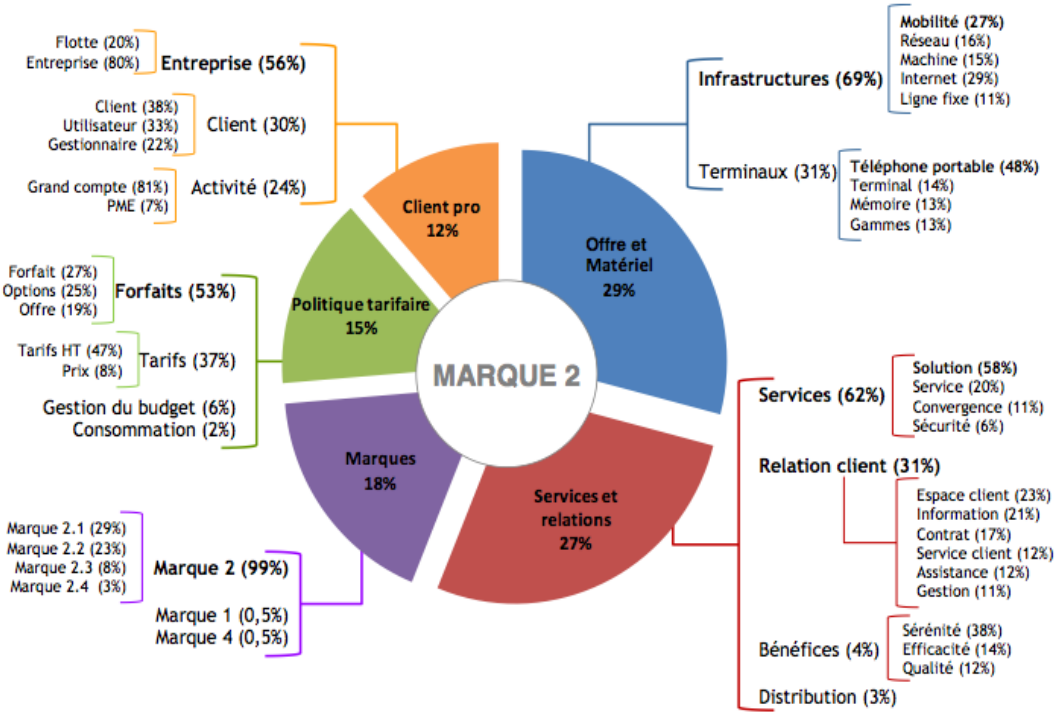

Figure 3: The distribution of equivalent classes defines in the thesaurus

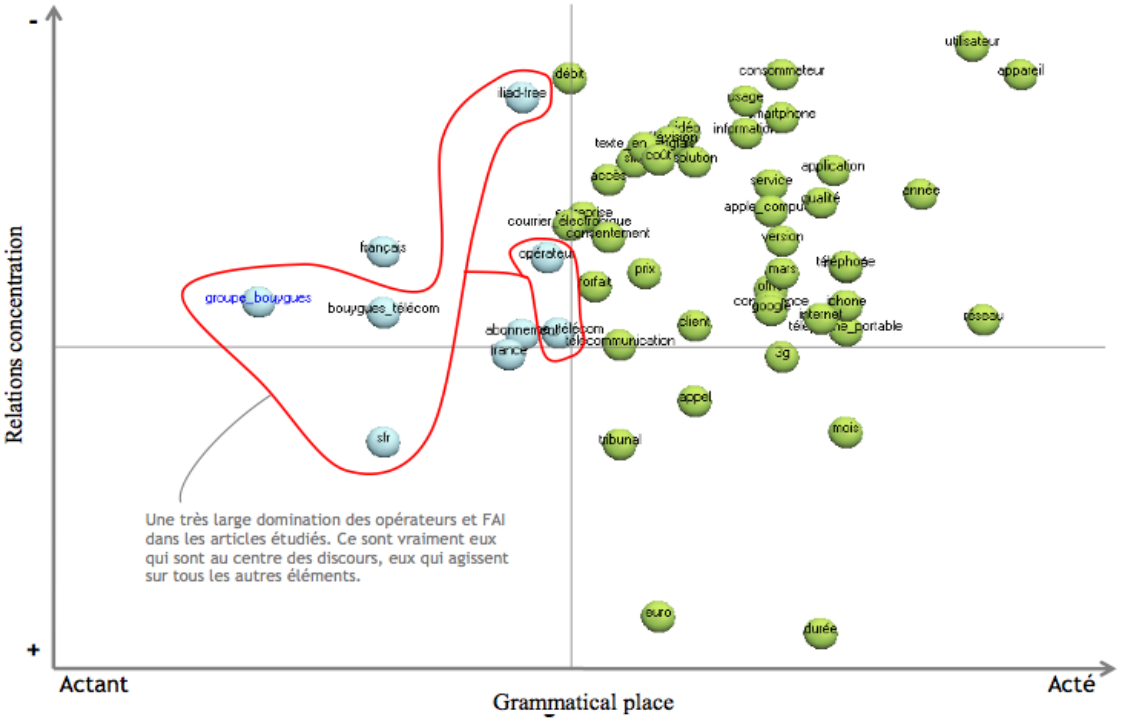

Figure 4: Chart of repartition for actor and acted references

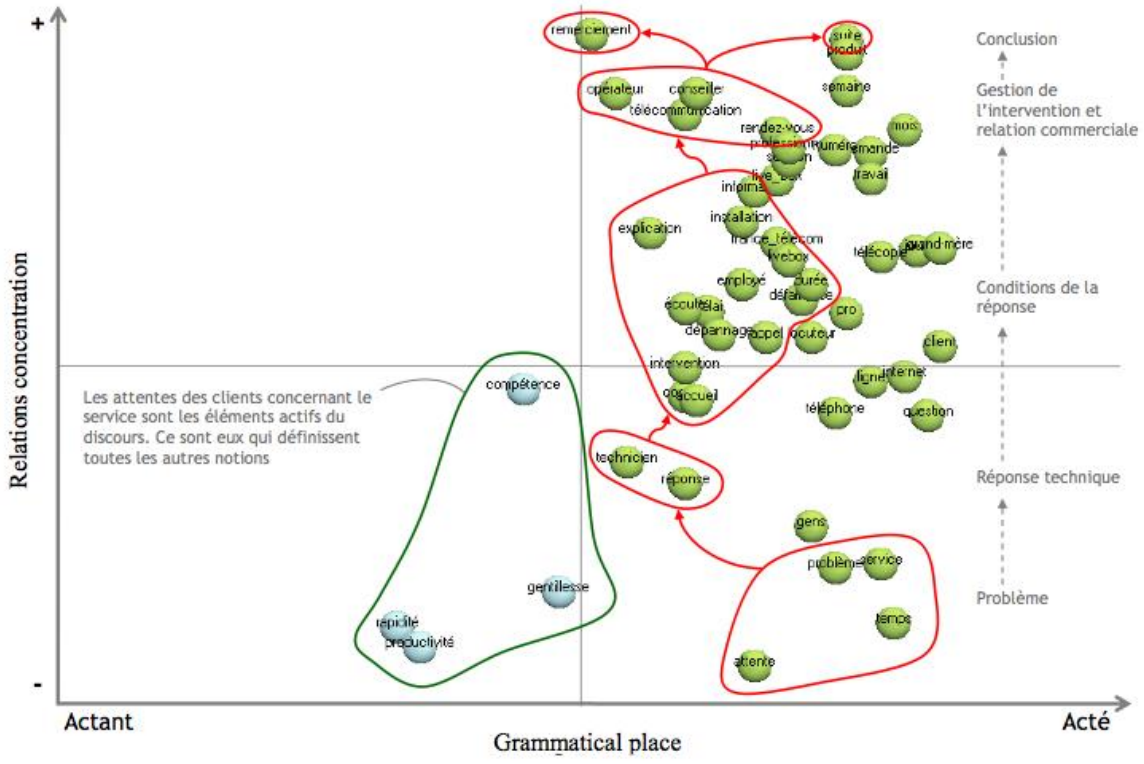

Figure 5: Map of the notion concerning the intervention of a technician on site

All companies are quoted in agent position, while the user is more often an object. We notice that the user seldom is a professional, which shows that generalist websites are not adjusted for professionals. "User" and "consumer" are submitted to companies, they have no real choice. They are used in varied contexts, in other words their relation are less concentrated. In general, all notions that matter to the material basis of the offer are on the right side of the chart (passive position), whereas words that refer to price policy and services tend to be in the middle of the chart. We notice that the words "tribunal" and "appeal" take place in an agent position, with weakly concentrated relations: This is explained by articles concerning Orange's issues with French justice. Relations are concentrated because the contexts are always similar. This analysis has a low interest, particularly because the corpus has not been defined precisely enough. Most of the interesting information comes from thematic analysis. This remark shows the pertinence of semantic analysis in that precise case. It also introduces questions regarding corpus definition.

3.2 Clients discourses analysis

The last step of the study concerned clients' discourses. The source we used was different: data had been collected in a quantitative questionnaire. This example can be seen as a new way to analyze open-ended questions. We have studied data in a double way. Firstly, we used a notions map, based on the Actant/Acted analysis, and then we used a linguistic analysis (linking lexical and syntactical analysis).

The map of notions enables us to analyze the answers in a double way, both thematic and discursive. Figure 5 is representing the map of the different notions taken into account, regarding a technician intervention.

The chart in figure 5 gives several information types. Firstly, we notice that terms which are in an agent position designate the clients' expectations regarding technical interventions. It also appears that all these notions are in the lower side of the chart: They are used in more concentrated contexts. These expectations are the starting point of most answers: several sentences begin with these words, which are obvious and central for the respondents. Notions placed in object position allow us to understand a different type of discourse logic. In the beginning of the questionnaire people seem to have a problem: main notions ("waiting", "time"), are always associated. The perception of technical answers belongs to a more diversified context: there is a variety of issues, answers and perceptions. This is even more obvious concerning the commercial relation (appointment making, contact with a consultant). Finally, conclusions of the intervention (thanks or waiting for a continuation) take place in varied contexts.

This analysis, completed by syntactical and semantic analysis, allows an understanding of the way respondents are implied in their answer. Thus, we notice that adjectives ("competent", "fast", pleasant", "good", "efficient", "professional", "clear" and so on) and verbs ("fix", "solve", answer", "satisfy" and so on) that are used show strong expectations towards technical support, but these expectations are often deceived. This is highlighted by the use of opposition connectors (contrasting judgment: "but", "in spite of", "however") and of intensity and negation modalities (60 percent of modalities). Injunction verbs ("must", "improve", "can", "have to" and so on) are associated with adjectives and verbs that underline the fact that clients expect a change from a firm that do not satisfy their needs. Through the linguistic operators, we can see a personal involvement (showed by modalities, but also by adjectives and connectors) of the clients in their relation with the firm. These linguistic clues also show the clients' feeling that their dissatisfaction is not taken into account (injunctions to change). Thus, the technical relation with the brand seems to be the place of a personal, or even emotional, involvement with clients. That is why technical support is a sensitive part of the client relation.

This type of analysis can be a supplement of a more traditional opened-ended questions' coding, that determines generic themes, but does not highlight linguistic stakes. For that matter, semantic analysis can be used for coding. On the one hand, it allows gaining time; on the other hand, it enables building more precise and exhaustive coding patterns (taken the entire corpus into account, not only an extract of verbatim).

4. Conclusion and further research

The method used in the article has allowed us to analyze the discourses of twenty brands belonging to varied sectors and to compare possibilities of client relations, on a deeper level. Finally, it enabled us to better understand our clients' discourses.

Clients' discourses analysis has permitted us to compare brands discourses to the clients' feelings in their contact experiences with the firm. Open-ended questions have been analyzed on a deeper level, since brands discourses were more appropriate to this type of analysis. On websites, language is generally poor: for example, sentences are often non-verbal, which is a problem for a syntactic analysis or to determine the agents and objects. Answers to open-ended questions are different. They look most often like correct sentences, built according to a precise grammar. That is why it is possible to analyze them on a

deeper level. Therefore, this data has been studied according to a double approach: building a notions map (chart of agents/objects notions) and semantic analysis linking lexical and syntactical aspects. The double approach has given us an understanding of the heart of clients discourses, to analyze the way they are implied in the discourses about (or to) the brand. As well as finding the expectations of the clients we have understood the way these expectations are expressed, and above all the way people handle issues and find resolutions. A graphical approach enabled us to understand the general process of discourses, whereas the analysis of syntactical forms (specific verbs, adjectives, connectors and so on) permitted us to understand how clients are personally implied in their discourses.

4.1 Automated semantic contribution to opinion and discourses understanding

From a scientific and methodological point of view, automated semantic analysis enables us to gain more detailed and deeper understanding. The previous example concerning open-ended questions, with semantic analysis tools can build coding pattern taking into account all responses, and not only a sample of answers. Thus, the coding pattern is more precise because it is based on a more exhaustive view on information.

Automated semantic allows approaching discourses in a different way than traditional content analysis. This enables it to become enveloped in the message sender. Semantic analysis exceeds in a way content analysis and it takes the content of the discourses ("dictum") and its form (how it is said) into account. Semantic analysis gives a more complete view on discourses because it takes into account syntactical constructions, modalities and usage of adjectives, as well as all other words that enables discovery of the speaker's personality and the discourse enunciation context. This highlights the way the speaker is implied in his or her own discourse, in an emotional or argumentative way. It allows placing discourses and speakers in wider groups. Barthes (1984) states that: "every speech belongs inevitably to a dialect." (Barthes 1984, 439). This means that discourses are never the speech of just one individual: Each individual shares a part of its individuality with other people and with other people of its group(s). This is called inter-subjectivity (Larsson, 2008). Semantic-pragmatic analysis should enable us to reach this discourse inter-subjectivity, which means taking into account the way the speaker keeps to a context, with interactions, and history. This should enable extract specificities of groups, defined before analysis (according to "objective" criteria like gender or age) or defined by analysis (by recognition of regularities and disruptions of discourse).

This approach, that places context in the middle of the analysis, is not only linked to the semantic approach. It is a global approach for information and intelligence studies. This can be summarized in Floridi's (2007) "subjectivist interpretation of relevant information", which implies that information relevance can be understood only when it comes to an exchange.

Semantic analysis permits a qualitative approach of wider samples. Hitherto, qualitative research is limited to questions with relatively small samples. This limit is practical: The qualitative questioning of a large amount of people is costly, in term of fieldworks but also when it comes to analysis and interpretation time. Computer tools (particularly in automated semantic analysis) enable us to analyze answers for wider qualitative samples. As a result, we can plan to hit the experience saturation threshold, regarding the moment when all experiences on a subject can be considered.

4.2 Technical limits

From a technical point of view, the main limit of semantic analysis is the difficulty to adapt it to spontaneous discourses language. For example, an online forum is a discourse place. On a forum dedicated to a firm, the firm's name is obvious and the participants do not name it often. Instead they use the third person ("it" or equivalent). This raises some questions: how can we automatically spot the posts that speak about the firm? We cannot determinate that every "it" refers to the brand. How can we take the speaking moments into account? Discussions are defined by interruptions, people speaking when they may not. Finally, how can we manage interruptions in debates?

Speeches are not put together in a logical way, but in a chronological way. It is not always the logic of the debates that determines the apparition order of speeches; it is more often the writing time of the contribution. A contribution do not necessarily make reference to a previous post, it can answer to a question that has been asked a few messages before. In simple sentences, anaphora management is a problem. On a discussion forum, the anaphora referent is not even in the previous clause. In the "style" point of view, authors tend to "write as they speak", they use abbreviations, forget capital letters, punctuation signs and make orthographic, grammatical or syntactical mistakes. It is difficult for software that has been developed on formal language models to analyze informal or incomplete formulations, as we can find on forums like Twitter and Facebook.

This is why it appears essential to include abbreviations in terminological dictionaries. It is not possible to include all incorrect orthographic

forms in dictionaries. The introduction of automatic orthographic correctors, or at least of a tool that tries to compare unknown forms to lemmatized form, seems to be a solution.

To analyze such texts in natural language, it is necessary to begin by editing the text, which can be time consuming work.

In this time consuming aspect, editing the texts can be compared to another step of analysis: corpus constitution. When analyzing great quantities of texts semantically and with computer assistance, we adopt a corpus linguistic logic. In this field, sources determination is essential. The grouping of texts in a corpus is a first semantic approach. When we choose the elements of the corpus, we propose a first step of interpretation, linked to our context. The sources must be coherent, and have a representative dimension. For example, if we chose to analyze a brand image through what is said on forums, it is a first choice. This involves considering if the chosen forums are representative for what is said on forums in general, or even on the Internet, not to say that it is representative of what all consumers of the brands think. This interpretation depends on the scope that is adopted. That is why the corpus must be determined, often by an exchange between the analyst and the client. In a context of business, this need can be a limit to introduction of automated semantic since it involve stakeholders spending time on determining the corpus.

Other technical limits appear when we decide to analyze information coming from the Internet. An efficient way to analyze web pages is to investigate and save them in order to analyze them a second time. This can be made difficult by limits linked to websites structure and to the way they are created.

On a web page, how can we identify relevant information? Heuristics exist and in addition to useful information, a web page often contains a navigation menu, advertisements, hypertext links to others articles, legal information and so on. For example, the navigation menu contains several HTML links, advertisement in links and pictures, and legal information can be found on all pages. The message is rich in meaning and poor in hyperlinks.

Automation of websites investigation raises the question of information hierarchy. During the analysis, it is difficult to know the audience of a page and to organize all pages in a hierarchy. Thus, we can ask if information on a page that is often visited (for example a home page) has the same value as information placed on a page with few visitors.

4.3 Renewal and methodological uncertainties

Automated semantic introduces challenges in some practices. The first of these challenges regards the gap between qualitative and quantitative fields. This approach can be compared to two types of methodologies. It can be considered as a qualitative methodology since its object is an unformed discourse. Automated semantic manages with quantitative processing matter and in order to process language through a computer, it must be transformed into computer data, which means mathematically managed. Automated analysis tools for language supply quantitative data; linguistic forms occurrences are represented as statistics, charts, tables and so on, which are quantitative representations. This is true when it comes to opinion mining, in which information is often envisaged as rating inferences, rankings and so on. Automated semantic analysis often has a qualitative part; it is possible and necessary, to come back to plain text. This qualitative comeback is a way to set highlighted linguistic forms back in their production context. This return to context is necessary to understand texts. Without it, the risk of misinterpretation is high. Quantification is not enough. Methodologies and tools of automated semantics are double-edge: qualitative material (discourse) is analyzed with a statistics and probabilistic logic, and allows results that are between the two areas.

The analyst using such tools has to master the two areas of the methodology. We exceed areas of market research (where we experimented) and of social sciences. This is likely to meet strong reticence (in each of these sectors).

The reluctance can be analyzed in two areas. The first one, which is the most obvious, regards the reliability and pertinence of the results. The trust and value of information coming from these types of tools can be questioned. The second reluctance is the fear of being compared with a computer, the fear that human intelligence could be belittled by the use of a computer tool. These two reluctances are linked. It seems necessary to understand that a tool cannot do anything without human intelligence, without human interpretation aptitude. A tool is only assistance for the analyst, who keeps his legitimacy as a decider and controller.

To understand the analyst role in a research process using automated semantics, a distinction exposed by Rastier (1994) can be used. As authors of this article we have adapted this distinction to our subject. Rastier (1994) analyzes understanding systems and distinguishes three steps: analysis, interpretation and understanding. He defines an understanding system as "every system that tries to pass from a syntactical tree to a semantic network and to make inferences inside this network." (Rastier 1994, 240). For him, there are three steps in the progression and at each step we can distinguish the role of the computer and of humans.

The first step, the syntactical tree, equates to *analysis*. It can be compared to morphological and syntactical analyses, which are the first parts of automated text analysis. This analysis is entirely done by software, which recognizes words and defines their relations.

The second step, semantic network, corresponds to *interpretation* and is performed by computer and human. The goal is to define a "signification", in the meaning adopted by Rastier (1994): "meaning became impoverished of context." (Rastier 1994, 240). Some software automates this step, like is the case with Tropes. This software uses two methods to determinate signification. Firstly, it extracts syntactical marks, modalities and so on, which organize the utterance and show interlocutors presence in discourses. It also classifies and organizes notions into a hierarchy based on its French language thesaurus. Thus, the software offers significance to each word, defining synonymy, hyperonymy or hyponymy links. For example, terms as "firm", enterprise" or "society" have a similar meaning: they belong to the equivalence class "Firm". This signification is abstract and polysemy risks are high because *interpretation* does not take context into account.

The last step, the understanding system, is *comprehension*. This step is completely mental, which means that it can only be human. It enables us to pass from "signification" to "meaning", to "create inferences inside the semantic network." The analyst uses all the elements extracted by the computer, the analyst makes comparisons and links them, in order to define the final meaning of the text. Thus, the building of a personalized thesaurus allows giving each word and each notion a specific meaning, relative to analysis context. The real value-added of the analysis appears at this level. Analysis is here fueled by the analyst's knowledge because analysts' own external data, external knowledge, memory and critical thoughts permit them to extract useful information from the text.

The usage of understanding systems, underline that computer and human intelligence are complementary. Software maintains assistance tools for analysts who remain the centre of analysis, since they are able to detect strategic information.

The other thing that automated semantic transforms is the way speakers are considered by analysts, particularly in market research. By putting discourses in the middle of interests, it highlights the exchange between the person who questions and the one who answers. In Internet discourses, there is an exchange, at least implicit, between a speaker and a receiver. This point of view allows placing people in the group(s) where they belong. In traditional analysis, particularities of targets are highlighted: these targets are defined by objective criteria like age, gender or product consumption. In our new point of view, we consider publics that belong to diverse social groups, have a history, and live in a specific context. We put the knowledge of the discourse sender in the middle of our questions, which implies other questions, particularly in relation with the collection of consumer discourses on the Internet. We often ignore people who are speaking on the web and who they could represent. It could be interesting to question the identity of those Internet users, and the criteria that should be chosen to define this identity. Should these criteria be the same as in "real" life, or should they be different ones? Being an Internet user speaking on websites, is it not the beginning of an identity? This question about validity of an analysis concerning people we know nothing about can be seen as a limit of that method.

Setting up an automated semantic analysis solution is costly. That must not be ignored. This is an investment of research and development. Buying a tool, taking time to discover software and train employees is an investment and setting up an automated semantic analysis solution is at least a middle-term investment. This can be complicated in a sector such as market research since visibility often does not go above a few months.

These remarks added to previously quoted technical limits, also underline that the tool choice may not have been as relevant as previously thought. Today, powerful solutions exist, which manage efficient technical limits. For future analysis, it would be efficient to develop a partnership with a firm that develops software. In that case, market research institute could concentrate on its core work, on its value-added; analysis; and entrust software firms with technical issues.

Regarding these limits, automated semantic for opinion analysis must stay a complementary methodology, which can help existing methodologies. It assists these methodologies in two ways. First, it allows a faster and easier processing for specific steps (open-ended questions, qualitative numerations and so on). It also permits a new point of view on problems processed, in addition to traditional content analysis methodologies.

References

BARTHES R. 1984. *Le bruissement de la langue. Essais critiques IV*, Seuil, Points Essais, 439 p.

BEAUDOIN J. 2005. *L'opinion, c'est combien? Pour une économie de l'opinion*, Village Mondial, 237 p.

BOURDIEU P. 1984. « L'opinion publique n'existe pas », in *Questions de sociologie*, Les Editions de Minuit, Reprise, pp. 222 - 235

CARDIE C. 1997. "Empirical Methods in Information Extraction", *AI Magazine*, vol 18,

CONDAMINES, A. 2007. « L'interprétation sémantique de corpus : le cas de la structuration de terminologies », in *Revue française de linguistique appliquée*, XII-1, Juin, pp. 39 - 52

DEMAZIERE, D. (Ed). 2006. *Analyses textuelles en sociologie – Logiciels, méthodes, usages*, PUR, Méthodes, 219 p.

FLORIDI, L. 2007. *A Subjectivist Interpretation of Relevant Information* », in PICHLER, A. and HRACHOVEC, H., *Wittgenstein and the Philosophy of Information*, Proceedings of the 30. Ludwig Wittgenstein Symposium, vol. 1

FUCHS, C. (Ed). 1993. *Linguistique et traitement automatique des langues*, Hachette-Classiques, HU Linguistique, 303 p.

GEYKEN, A. 2008. « Quelques problèmes observés dans l'élaboration de dictionnaires à partir de corpus », in *Langages*, 171, Septembre

GHIGLIONE, R. (Ed). 1998. *L'analyse automatique des contenus*, Dunod, Psycho Sup, 168 p.

JENNY, J. 2004. « Quali / Quanti – Distinction artificielle, fallacieuse et stérile ! », *1er congrès de l'AFS*, Groupe RTF 20, Session n°4, 25 février, consultable à l'adresse http://testconso.typepad.com/files/jenny-quanti-quali.pdf (le 8 novembre 2010)

LARSSON, B. 2008. « Le sens commun ou la sémantique comme science de l'intersubjectivité humaine », in *Langages*, 170, Juin, pp. 28 - 40

MARC, X, TCHERNIA, J. (Ed). 2007. *Etudier l'opinion*, PUG, 260 p.

MARTIN, R. 2001. *Sémantique et automate*, PUF, Ecritures électroniques, 190 p.

PANG, B. and LEE, L. 2008. Opinion mining and sentiment analysis, *Foundations and Trends in Information Retrieval*, 2 (1-2),

RASTIER F., CAVAZZA M., ABEILLE A. 1994. *Sémantique pour l'analyse. De la linguistique à l'informatique*, Masson, 240 p.

TAMBA I. 2005. *La sémantique*, PUF, Que sais-je?, 128 p.