Papers on Social Representations
Volume 20, pages 37.1-37.47 (2011)
Peer Reviewed Online Journal
ISSN 1021-5573

[http://www.psych.lse.ac.uk/psr/]

Text Mining Methods for Social Representation Analysis in Large Corpora

JEAN-FRANÇOIS CHARTIER

Université Du Québec À Montréal (UQÀM).

Laboratoire d'Analyse Cognitive De l'Information (LANCI)

JEAN-GUY MEUNIER

Université Du Québec À Montréal (UQÀM).

Laboratoire d'Analyse Cognitive De l'Information (LANCI)

With mass text digitization (digital libraries, web, etc.), a huge amount of empirical data is now available for scientific inquiry. In social sciences and humanities, the use of statistical text mining methods to analyze these data has become unavoidable. Saadi Lahlou proposed in the mid-90s a coherent framework for the application of these methods to the study of social representation in large corpora. However, despite this initiative, text mining methods have remained marginal in this research program, partly due to a poor understanding of its methodological and theoretical assumptions. There are still many analyses which confound the software with the method. This paper presents an overview and a formalization of a statistical text mining method for the study of social representation, using Lahlou's works as illustrations. The goal is to look into the software black box while analyzing the steps and the formal operations involved. The linguistic and methodological assumptions are made explicit and alternative algorithmic operationalizations are highlighted.

DATA AND METHODS

Fifty years ago, Social Representation theory found its original formulation in Moscovici's study on the social representation of PSYCHOANALYSIS (Moscovici, 1961). The theory has greatly evolved since its beginnings (e.g. Moscovici, 1988; Doise & Palmonari, 1986; Jodelet, 1989; Abric, 1994; Wagner et al., 1999; Bauer & Gaskell, 1999; Marková, 2003; Flament & Rouquette, 2003; Voelklein & Howarth, 2005). Its evolution has always been supported by rich methodological reflections. The range of the methods of analysis is quite broad (Doise et al., 1992; Breakwell & Canter, 1993; Moliner et al., 2002; Abric, 2003a; Flick & Foster, 2008). Among these methods, descriptive, comparative and inductive approaches are largely dominant. These methods, which we could qualify as bottom-up approaches, are contrasting with the more deductive, or top-down approaches, founded on axiomatic or formal grammar (Meunier, 2002, p. 229-230).

Inductive methods are based on the empirical analysis of observables, regularities or patterns that have a meaning relative to the theory's hypotheses. In the field of social representation (SR) studies, these observables often come in the shape of language contents: interview or life story transcripts, answers to a questionnaire, free associations, news articles, or other similar forms. Different methods of analysis have been developed for the various types of observables: characterization techniques, similarity analysis, prototypical analysis, content analysis, factor analysis, etc.

In the last few years, new observables have become accessible to researchers in social sciences and humanities.

Massive Text Digitization And Change Scale Analysis: What Do You Do With A Million Books?

Mass digitization of text documents and library has made available a huge amount of empirical data of interest to the SR’s psychologist and sociologist, ranging from newspaper and dictionary articles, literary, religious, policy, legislative, historical documents to the entire social web’s contents (i.e. blogs, Facebook, Wikipedia, Twitter, etc.). Moreover, with the emergence of major digitization projects like Google Books and Open Content Alliance, Gregory Crane's (2006) now famous question becomes more crucial than ever: "what do you do with a million books?" This is an issue that social scientists must also address.

This digitization brings a change of scale in the amount of available empirical data and radically changes the kind of analysis that can be done. Among most spectacular examples, Michel et al. (2011) are currently conducting a quantitative study of world culture over five centuries and covering more than five million books, that is, about four percent of all books during that period!

This huge amount of data does not lend itself easily to traditional analysis. In the field of SR studies as elsewhere in social sciences, humanities and cognitive sciences, there is an increasing use of computational methods for statistical text analysis. The general assumption underlying the use of these methods is summarized by McNamara:

> "Large text corpora combined with computational techniques for analyzing these corpora allow scientists to extract meaning from text and, by consequence, to explore various aspects of the human mind and culture that manifest in text." (McNamara, 2011, p. 4)

These methods are developed by different communities, including statistical natural language processing (Manning & Schütze,1999; Jurafsky & Martin, 2000), data mining (Fayyad et al., 1996), pattern recognition (Theodoridis & Koutroubas, 2009), information retrieval (Manning et al., 2008), computational linguistics (Mitkov, 2003) as well as lexical statistics (Lebart & Salem, 1994). In computer sciences today, they are denoted by the general expression "text mining methods" (Weiss et al., 2005; Hotho et al., 2005; Feldman & Sanger, 2007).

Besides, these text mining methods (TMMs) are already used in many areas. There is of course automatic discourse analysis which has been a pioneer (Pêcheux, 1969), computer assisted reading and text analysis (Meunier et al., 2005; Popping, 2000), sociological analysis (Demazière et al., 2006), literary analysis (Reinert, 1993, Yu, 2008), socio-semantic network analysis (Roth & Cointet, 2010), organization analysis (Carley & Diesner, 2005), corpus linguistics (Stubbs, 2002; Rastier, 2011), scientometrics (Glenisson et al., 2005), cognitive sciences (McNamara,

2011) among many others. Researchers in social sciences and humanities using computational methods, especially in the field of text analysis, are more and more aware of the methodological, epistemological and cognitive issues underlying this technology (McCarty, 2005; Meunier, 2009). In the near future as well as in the long term, the importance and use of TMMs is likely to grow significantly.

WHY WERE LAHLOU'S PAPERS IMPORTANT FOR SOCIAL REPRESENTATION STUDY?

Early attempts at integrating these types of TMMs into the field of SR study date back to the mid-90s. Saadi Lahlou, in his doctoral thesis (Lahlou, 1995a) and in a series of related publications (Lahlou, 1992, 1994, 1995b, 1996a, 1996b, 1998, 2003, Beaudouin & Lahlou, 1993), was among the first scholars to demonstrate a coherent way to use TMMs within the SR theoretical framework.

Beyond the case study of the author, which was the SR of EATING from a dictionary, the contribution we highlight here is, as stated by Lahlou himself, the opening "[of] a new field of linguistic material to psychosocial investigation on a large scale" (Lahlou, 1996b, p.17), as well as "a family of empirical solutions to the question of meaning" (Lahlou, 1996a, p.63).

For his research, Lahlou used the software ALCESTE designed by Max Reinert (1983, 1986, 1987, 1990). ALCESTE is a lexicometry application that can be considered as a software implementation of what, in this paper, is referred to as 'TMMs'. To our knowledge, almost all SR studies relying on TMMs have been done using this same implementation software (e.g. Gaffié et al., 1998; Kronberger & Wagner, 2000; Viaud, 2002; Dany & Apostolidis, 2002; Licata & Klein, 2002; Kalampalikis, 2003; Kalampalikis & Moscovici, 2005; Alba, 2004; Viaud et al., 2007; Garnier et al., 2007; Colucci & Montali, 2008; Manetta et al., 2009; Geka & Dargentas, 2010; Caillaud et al., 2011; Gilles et al., 2011). Notwithstanding empirical case studies, methodological analysis since has been modest. They consisted mainly in piecemeal modifications in the ALCESTE software's settings, modifications, for the most part, already considered by Lahlou.

We believe that the "family of solutions" that are the TMMs didn't have in SR's field the resonance they have had elsewhere in social sciences, humanities and cognitive sciences. One reason may be that the framework and underlying assumptions of the TMMs haven't been properly explained. The relevance of TMMs for the study of SR is still difficult to assess because the kind of analysis they make possible remains partially understood. The methods have usually been described in a metaphorical language, too bonded to the ALCESTE software's settings, which may had the effect of hiding crucial assumptions or blurring the logic of some algorithms involved.

The Software Is Not The Method

Some circles within the humanities and social sciences are still wary about computational methods and TMMs particularly. A few years ago, Kelle (2000) reported how some researchers considered the use of computers as a source of methodological alienation. In the field study of SR, some have pointed out how these TMMs trigger in the analyst a sentiment of losing the control over the analytical processes. Buschini and Kalampalikis (2002) refer to this as "numeric behaviourism", that is, a methodological attitude consisting in seeing the computer as a "black box" in which one enters raw data on one side, and waits for ready-to-use results on the other side. However, alienation and numeric behaviourism aren't inherent consequences of the use of a computer. They are the consequences of certain methodological practices.

It seems that a source of some persistent misunderstanding lies in the confusion between the method and the software. Some methodological discussions in literature have been limited to the software parameters and setting. The misunderstanding is that the software is only the tip of the iceberg of the method. For instance, ALCESTE is a great software, with a solid reputation, but the discussion around its parameters is not exhaustive in terms of methodological analysis of TMMs for SR study. Moreover, in the case of commercial software such as ALCESTE, the calibration of most of the parameters remains inaccessible to the user because he does not have access to the source code of the program. The values of these parameters are fixed upstream by the software designer, upon whom the user hence depends.

ALCESTE is not a method per se, but a particular software implementation of the method. The method is, however, independent from the software. The method is a set of functional operations, and for any given operation, there are always several algorithms that can perform the computation and many softwares that can implement each of these algorithms. Figure 1 shows these different analytical levels of computational methods.

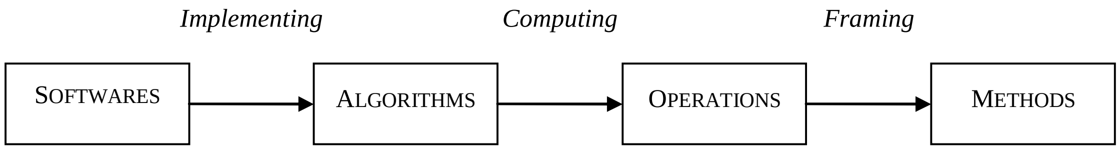

Figure 1. The different analytic levels of computational methods.

The diagram in Figure 1 illustrates that we can break down a method into the different operations it implies. These operations can be formally described as a function with the following form:

$$f : X \to Y$$

An operation takes a certain type of data as input *X* and generates as output another type of data *Y*. This operation is a 'black box' only if we do not take the time to look inside in order to deconstruct it. The transformation taking place between *X* and *Y* is a calculation using algorithms. Many algorithms can compute the same function. They are not neutral or equivalent, they each imply different hypotheses. Choosing an algorithm is a decision pertaining to the researcher as well as the decision to implement it in the software. There too, many softwares can implement the same algorithm, yet they are not equivalent, as they can have different parameters and calibration.

These distinctions are important. Not only is the software different from the method it implements, one must also not confound the software with the algorithm and the algorithm with the functional operation. They are different levels of description of the method. The functional level is the description of the logical constraints upon the information processing operation framing the method. The algorithmic level is the description of the computation of these operations. The software level is the implementation, that is, the description of the parameters' calibration of these algorithms[1].

As shown, computational methods like TMMs imply numerous decisions and the researcher can control the process entirely. It seems that the order in which some researchers perform the methodological analysis of TMMs in the SR's field, which is to start with the software instead of the functional and the algorithmic level of the method, yields some confusion, as software choice hence determines the method, while it should be the other way around. Metaphorically, one could say that the software becomes the tree hiding the forest.

In this context, it seems altogether reasonable to conclude that methodological analysis of TMMs for SR study has remained so far marginal. This is rather surprising, given that the use of these methods is said to become unavoidable in the very near future. In light of this, it seems that a proper assessment of the current situation must be made before going further.

In this paper, we present TMMs as used in the study of SR using Lahlou's works as illustrations. Though our discussion will not be bounded to Lahlou's own stance, nor to the stance of others, regarding software calibrations and settings. Our goal here is to look into the software black box while giving an abstract presentation of the TMMs, the formal operations involved and highlighting different algorithmic operationalizations.

But before presenting the formal framework, we explore the linguistic model upon which it relies. This model is rarely made explicit by psychologists and sociologists of SR and by computer scientists alike.

THE VECTOR SPACE MODEL

The linguistic model behind the use of TMMs is the *Vector Space Model* (VSM) (Gärdenfors, 2000; Widdow, 2005; Sahlgren, 2006; Jurafsky & Martin, 2000, p.643; Manning & Schütze, 1999, p.539; Salton et al., 1975). This model is based on two strong assumptions that reduce observable linguistic relations in large corpora to co-occurrence relations and equivalence relations.

[1] These distinctions are similar to very classical distinctions make in artificial intelligence, although in a different context (e.g. Marr, 1982).

Meaning Is Use

The first assumption is related to a Wittgensteinian insight: meaning is built through language in use. More precisely, it has been given a proper linguistic formulation by Firth and Harris through the distributional hypothesis:

> "You shall know a word by the company it keeps." (Firth, 1957, p.11)

> "Meaning is more easily stated as a property of word combinations (or of words in combination) than of words by themselves. […] Once we see that the meaning of a word in a particular occurrence depends on its environment, the categorization of meaning-ranges can be replaced by a categorization of the environing words." (Harris, 1991, p.325)

This first assumption suggests that meaning may be studied through the way people use words in combination with other words in their discourses. This is what, about meaning, is directly observable in discourse, especially language uses embodied in texts. Classically in linguistic, this can also be referred to as the syntagmatic relations between words. In the VSM, however, all syntagmatic relations are reduced to word co-occurrence relations observed through parts of discourse like phrases, sentences or other kinds of text segments that compose a corpus.

In this paper, we use the expression "part of discourse" to refer to these observables, but formally in the VSM, observables are text segments represented by co-occurrence pattern of m words. Therefore, a part of discourse may take the form of a vector:

$$\vec{p} = (w_1 ... w_m) \qquad (1)$$

In which w_t represents the weighted value of the word or term t in the part of discourse $\vec{p}$.

A set of n parts of discourse is then noted in this form:

$$E = \{\vec{p}_1 ... \vec{p}_n\} \qquad (2)$$

Formally, *E* represents a multidimensional vector space. The number of dimensions of this vector space corresponds to the number *m* different words in the set of parts of discourse.

This vector space can also rely on a graphical projection, where each vector $\vec{p} \in E$ can be plotted as a coordinate in a multidimensional space. For example, in Figure 2 are plotted thirty three-dimensional vectors. One can imagine that they represent thirty parts of discourse made from the same three words *x, y, z,* with different weighting values.

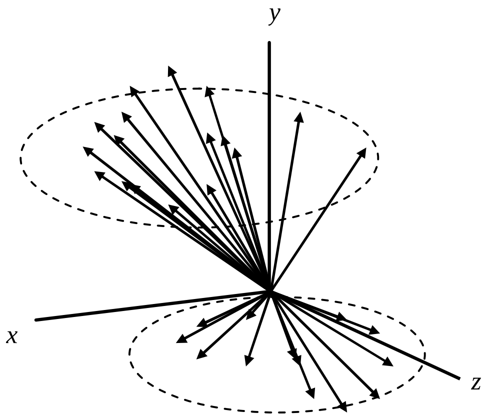

Figure 2. Example of a three-dimensional vector space of thirty word co-occurrence patterns (parts of discourse) grouped in two classes.

Meaning Is Differential Value

The second hypothesis is complementary to the previous one. It is related to a Saussurian insight: meaning is differential value in 'la langue'. This hypothesis is about equivalence relation between parts of discourse. Such equivalence means that two parts of discourse are, at least partially, substitutable one to each other with regard to a class of meaning. Classically in linguistic, this can also be referred to as paradigmatic relation. This kind of relation is not directly observable in discourses, but it can be induced from a large sample of language uses, for instance, text segments in corpora.

In the VSM, these relations are interpreted in terms of similarity or distance calculation. The similarity or distance between two parts of discourse is measured from their word co-occurrence patterns through a large corpus. Put differently, parts of discourse are interpreted as (more or less) semantically equivalent because they share similar word co-occurrence patterns; they're close in the vector space:

> “Two documents with similar index terms are then represented by points that are very close together in the space, and, in general, the distance between two documents points in the space is inversely correlated with the similarity between the corresponding vectors.” (Salton et al., 1975, p.613)

> “Proximity of vectors in the space […] corresponds to semantic similarity.” (Schütze, 1993, p.2)

Formally, semantic similarity between two chunks of discourse is measured with a metric function as:

$$d : E \times E \to \mathbb{R} \qquad (3)$$

This function assigns to every pair $(\vec{p}_i, \vec{p}_j) \in E \times E$ a real number, referring to either the distance or the proximity between two vectorized parts of discourse $d(\vec{p}_i, \vec{p}_j) \in$ $\mathbb{R}$. If we rely again on a graphical projection, it can then be said that the spatial proximity between the coordinates of word co-occurrence patterns becomes a metric of their semantic similarity. In the abstract example of Figure 2, one could conclude, because of their proximity in the vector space, that the thirty word co-occurrence patterns make two classes of meaning or two semantic clusters.

In sum, it’s essentially on the basis of these two linguistic assumptions — meaning as usages and differential value — and their mathematic formalizations — co-occurrence patterns and proximities in a vector space — that TMMs were initially conceived and applied.

THE METHOD: MAPPING SOCIAL REPRESENTATIONS FROM TEXTS

Lahlou has contributed to the development of a computational method for SR analysis based on TMMs. According to the author, this method aims at "identifying in the discourse, classes of statements like, which can be regarded as expressions of a common core of meaning" (Lahlou, 1995a, p. 140-141, our translation), what he also called elsewhere the "building blocks" of the SR (Lahlou, 2003, p.46). Furthermore, Lahlou specifies that "each class is considered as a basic nucleus of the representation, characterized by typical lexical traits." (Lahlou, 1996b, p.278).

Lahlou sees the method as an automatic inductive clustering process which draws a semantic map of the SR from and conveyed by a given corpus. The methodological approach can be summarized as follows: parts of discourse in the given corpus are language use instantiations of the SR; some of these parts of discourse (i.e. text segments as phrases, sentences, etc.) share semantic similarity because they have similar word co-occurrence patterns; given this, one can group various similar parts of discourse in equivalence classes; the semantic map so induced, taken as a whole, can be very useful for scientific inquiry of the SR.

As such, the method follows a very classical cascade pattern. It has three main phases, each including two or three steps: the first one is the data collection phase of the SR, which involves collecting a corpus of texts and extracting the set of relevant parts of discourse wherein the SR under study is instantiated or communicated; the second one is the modeling phase of the text data related to the SR, which includes the parts of discourse lexical content vectorization and the calculation of the semantic proximity/distance between parts of discourse; the final phase is the analysis of the vector space related to the SR, which involves automatic induction of the semantic classes, the extraction of salient class' lexical contents and, finally, the categorization process (see Figure 3).

In the following sections, we will present in detail the method. For each phase, we suggest a formal description of the underlying operations. ALCESTE is the most used software from researchers in the SR field of study, so we will emphasize the analysis of its algorithmic operationalization. But, we will also suggest other operationalizations through alternative algorithms. Furthermore, for some important algorithms, we shall discuss the calibration of their parameters.

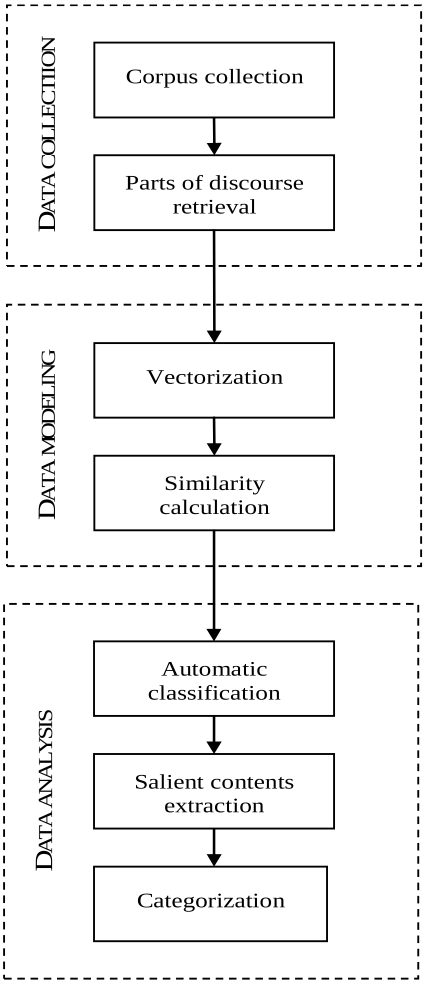

Figure 3. The three phases and its seven steps of a text mining method for social representation study.

1. DATA COLLECTION

The first phase is the collection of the textual data pertaining to the SR studied. As in any other methods of SR study, its importance is critical, as it entirely determines upstream the results of the treatment (Abric, 1994, p.59).

1.1 Corpus Collection

The first step in the collecting phase consists in gathering a body of documents in which the SR under study is to be found and retrieved. This corpus is composed of texts that can be taken from a variety of sources, e.g. newspapers, web sites, encyclopaedias, institutional archives, historical documents, literature, etc. In heuristic terms, the diagram in Figure 4 proposes a typology of the different corpora.

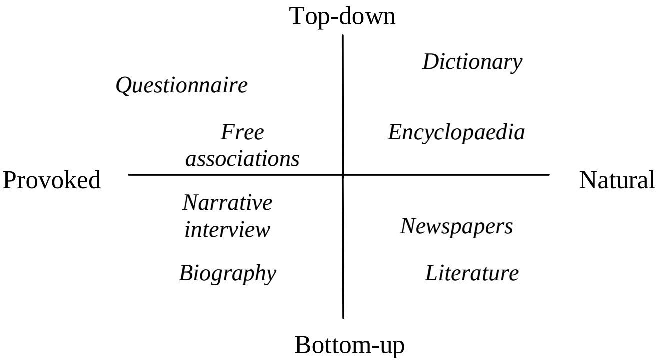

Figure 4. Typology of different corpora.

This corpora typology articulates itself around two axes. These axes are continuums rather than dichotomies. The distinctions they suggest are important, as different types of corpora sometimes necessitate different types of treatment. The first axis is about the conditions of the documents production. It relies on a distinction made by Bardin (2003, p.248) between provoked and natural corpora. Corpora are distinguished along this axis between those that require the researcher intervention in their production conditions and those that do not. The second axis is that of the contents organization. It relies on a distinction found in computer and information sciences (Weiss et al., 2005). This axis distinguishes corpora organized in a pre-defined meta-structure, upon which contents are indexed (i.e. questions from a questionnaire, dictionary entries, etc.) from corpora that do not have this type of structure.

In a first step, a corpus is a set of *N* documents which we note in the following way:

$$D = \{d_1 \ldots d_N\} \tag{4}$$

Each document $d = \bigcup p$ can be seen as the conjunction of several parts of discourse *p*, wherein each $p = \bigcup t$ is made up of the set of words or terms *t* occurring within it. The length of the part of discourse is a theoretical decision usually relying on grammatical unit or punctuation markers. The parts of discourses which compose a document may be the set of all phrases, statements, sentences, paragraphs or another kind of text segments it contains.

Corpus construction is a complex task. This step is similar to the way classical content analysis is used in the SR's field. While there is no foolproof procedure for gathering documents, there exists 'good practice guideline' that can be followed like relevance and homogeneity criteria. We will not go into details about these issues, as they already have been the subject of thorough discussions in other fields of social sciences in general, as well as in the field of study of SR in particular (e.g. Henry & Moscovici, 1968; Bauer & Aarts, 2000; Moliner et al., 2002, p. 43).

Let us only take note that criteria of relevance consist in identifying pre-requisites the documents need to meet in order to be admissible to a corpus. These pre-requisites are, among other things, the conditions of documents production. These conditions are criteria a researcher needs to make explicit when she makes the hypothesis that some documents – e.g. press articles – are a valid empirical source for the SR study.

An example of this is found in Lahlou's study of the SR of EATING. His corpus *D* was made up of a specific document: a French language dictionary (i.e. Le Grand Robert). It contained 100,000 entries. How is a dictionary pertinent? Lahlou explained that "[the dictionary] contains social knowledge on the world, sedimented in language. Encyclopaedias and dictionaries are the depositaries of human culture" (Lahlou, 2003, p.41). According to Lahlou, the relevance principle was the following: we "consider the dictionary as a collective subject, which we will interrogate […] as if the dictionary was a spokesperson of our culture." (Lahlou, 2003, p.42)[2].

[2] However, we can observe that this hypothesis – that a dictionary may be thought of as a spokesperson of a culture – is not one that TMMs can verify. Such verification needs other methods. For example, in his study of the SR of EATING, Lahlou compares his analysis on the dictionary with another analysis he made on a corpus of free associations gathered from 2,000 subjects.

The homogeneity criteria refer to the coherence and the systematicity of the corpus. Types of documents that do not have the same conditions of production, such as interview transcripts and news articles, or historical archives and web pages, will not be mixed in the same corpus (unless there are very good reasons to do so). Documents also have to be historically homogeneous. It might be necessary to divide the corpus in chronologically segmented pieces in order to do comparative analysis. Document homogeneity might also include the language or document encoding[3].

1.2 Parts Of Discourse Selection And Retrieval

Collecting a corpus is a step to be distinguished from the selection and the retrieval of the content to be analyzed. The latter consists in extracting from the corpus D only the content that has a thematic link with the object of study. In other words, the goal here is to select a sub-corpus, made up of the set of parts of discourse thematically linked to the studied SR.

For example, when Lahlou analyzed the SR of EATING in a dictionary, it was obviously not all its 100,000 entries that were linked thematically to the object, but just some parts of it. To mine the SR, he exploited the analogic organisation of the dictionary: for each entry, this dictionary provides a list of associated terms as synonyms, analogs homonym and derived terms. The parts of discourse selection and retrieval operation Lahlou designed consisted in the extraction from the dictionary of the associated terms of first order (i.e. the definition of the word “manger” and the set of definitions of its associated terms e.g. “absorber”, “avaler”, “consommer”) and the extraction of the associated terms of second order (i.e. the associated terms of the first order set e.g. “déglutir”, “engouffrer”, “apprendre”) (Lahlou, 2003, p.42-43). Thus, from the 100,000 entries in the dictionary, he kept only 544 of them; the entries which are explicitly associated with the word “manger”.

[3] In most softwares, it is necessary that the corpus be all in the same language since algorithms only process chains of characters. For example, two chains such as “minister” in English, and “ministre” in French, are two entirely different words for the software even if their meanings might be very close for a human. Some algorithms also calculate on the basis of morpho-syntactic patterns specific to a given language. For similar reasons, documents should also be written with correct spelling in order to ensure that there is no noise in the calculations. Another criterion, more technical yet unavoidable in a numerical context, is the document encoding in a format compatible with the software.

Depending on the kind of documents gathered, this step is not always necessary. This was the case of Kalampalikis and Moscovici (2005) who analyzed the representation of MARXISM in a corpus of 100 verbatim transcripts of interviews (≈800,000 tokens). Because of the conditions of documents productions, they kept them all. But what if they worked on newspapers or novels? What if instead Lahlou worked on the SR of EATING in newspapers? He couldn't have simply gathered all the content of the papers in which the word “eating” appears (a single article can contain contents of radically different thematics). He would have to choose the parts of discourse (the text segments) in the corpus thematically linked to the SR.

A last example: Sainte-Marie et al. (2011) analyzed the concept of EVOLUTION in Darwin’s *Origin of Species* with TMMs (what they called “conceptual text mining method”). The corpus *D* was the book, but the set of parts of discourse linked thematically to the concept was a small subset of *D*. The extraction of this subset from *D* is a non trivial task that may demand complex operations and algorithms.

Formally, one can see this operation of selecting and retrieving relevant chunks of discourse from documents as a function like this:

$$f : D \to D' \tag{5}$$

D is the corpus defined in (4) and $D' = \{p_i \ldots p_n\}$ is the subset of *n* parts of discourse retrieved from *D*. The function defined in (5) can use many algorithms to select and retrieve phrases, sentences, statements, paragraphs or another kind of text segment in the corpus that is thematically related to the SR.

A very simple algorithm that can be useful in text mining is what classical humanists have called ‘concordance’ and what corpus linguists have called ‘the KWIC index’ (i.e. ‘key words in context’) (Stubbs, 2002, p.61). In the case of SR analysis, the pivot of the concordance may be a word or a group of words that the researcher sees as the canonical lexical anchors of the SR studied. It is also possible to use more sophisticated machine learning algorithms to retrieve from a corpus parts of discourse linked to very specific thematic contents (Sebastiani, 2002). Computer scientists have recently developed numerous methods which will be useful for psychologists and sociologists who analyze SR in large corpora. For example, methods were developed to automatically find and extract expressions of opinions and emotions in texts (Liu, 2010). Chartier et al. (2012) have developed a machine learning method to find and retrieve axiological statements from a corpus. Sainte-Marie et al. (2011) have developed a method to automatically find and retrieve parts of discourse linked to a specific concept even if the concept has no canonical lexical anchor in the corpus.

When the corpus is modest in size, the selection and extraction of parts of discourse can be done manually through an analytic process close to the classical content analysis. When the corpus contains several thousand documents, computational methods become necessary to accomplish the task.

2. DATA MODELING

The second phase is the data modeling. This consists in formalizing the semantic space formed by the parts of discourse into a vector space. This comes in two steps. The first is the vectorization of the parts of discourse collected from the previous step. The second consists in calculating, with the help of a metric, the relations of proximity or distance between the vectorized parts of discourse.

This phase is not a modeling of the SR in itself, but rather a mathematical model of the empirical data. In order to understand the difference between those two, TMMs can be compared with other SR methods of analysis. In TMMs, the data modeling in a vector space is analogous to graph modeling as done in similarity analysis method (Flament & Rouquette, 2003). It’s also analogous to a factor analysis method as done in inter-individual variations analysis (Doise et al., 1992). Hence, it is important to note that a vector space is not a model of SR, nor a graph or factor axes. Still, they are all mathematical formalizations very useful to the study of SR.

2.1 Data Vectorization

The first step in data modeling phase is vectorization, that is, the construction of a vector space from the parts of discourse thematically linked to the studied SR and extracted in the previous step. It’s done in two sub-steps: the relevant word selection and their weighting according to their importance.

This vectorization operation is formalized in the following way:

$$f : D' \to E \qquad (6)$$

It takes as input the set D' of parts of discourse defined in (5) and transforms it into a set of vectorized parts of discourse as we defined it in (1) and (2).

The first sub-step of the vectorization aims at selecting, among all the words contained in the parts of discourse selected in D', those to be used in the modeling process.

2.1.1 Relevant words selection

Until now, words have been generally considered as terms, that is to say, chains of characters separated by two spaces. Several alternative operationalizations of what is a word are however possible. It is possible at this step to apply various linguistic transformations such as lemmatization or stemming. Lemmatization converts terms to their lemmas while stemming algorithm converts terms to their stem. Lemmatizing and stemming are two different transformations. Choosing different types of lexical units implies accepting different theoretical assumptions. Selecting lemmas as lexical units implies that no significant semantic difference exists between the lemmatized[4] form of a word and its gender or modal inflections. On the other hand, selecting stems[5] as the modeling lexical units assumes that there is no fundamental difference between words that have the same stem but different suffixes. In Lahlou's case study, for instance, the lexical unit was the stem.

[4] This is the infinitive for verbs, and the singular masculine for other words.

[5] For example, the stem ‘liber’ is the root of words like ‘liberation’, ‘liberal’, ‘libertarian’.

In order to make things simpler for the subsequent sections, we will keep on referring to ‘words’ or ‘terms’ in a general way, while emphasizing that this is a complex concept that can receive many operationalizations.

The total vocabulary from the set of parts of discourse D' is noted:

$$V(D') = \{t_1 ... t_M\}. \tag{7}$$

That is to say, $V(D')$ is the set of the M different words or terms occurring in the set D' of parts of discourse.

However, not all words or terms $t \in V(D')$ are relevant for the modeling phase. The goal of this first step is to select only the important words within the totality of the vocabulary. The construction of the vector space E therefore implies a sort of vocabulary filtering or reduction operation.

The assessment of the relevance and importance of a word may be based on grammatical or statistical arguments. Grammatical significance rests on the hypothesis that the semantics of parts of discourse is solely expressed through words with lexical meaning, i.e. nouns, adjectives, verbs or adverbs. Other kinds of words, such as articles, prepositions, and pronouns, are called ‘empty’ or ‘functional’ and are filtered.

The statistical significance of a word may be based on its information quantity. A word is considered highly informative if it is both representative and discriminative in a corpus. That is to say, idiosyncratic words like hapax (i.e. low frequency terms) and very frequent words such as the ones common to all parts of discourse, are considered uninformative and thus filtered.

The computation of function (6) can be accomplished using various algorithms which we refer to as filtering algorithms. A proper grammatical filter could simply consist in a list of such ‘empty’ words, sometimes called a ‘stop-list’. One can find many of these stop-lists on the web[6].

On the other hand, a statistical filtering algorithm may be based on the notion of ‘noise’ as referred by a calculation of entropy. Then, the information quantity of a term $t \in V(D')$ may be computed using a binary entropy function like:

[6] For a repertoire of stop-list for different languages: http://www.ranks.nl/resources/stopwords.html.

$$H(t) = -\Pr(t)\log_2 \Pr(t) - (1-\Pr(t))\log_2(1-\Pr(t)) \qquad (8)$$

Where $\Pr(t)$ is the occurrence probability of a term t in the set of parts of discourse D'. As it can be seen in Figure 5, term entropy is maximal when its probability of occurrence in a part of discourse is minimal or maximal. The general procedure in applying an entropy filter involves that the researcher fixes a minimal threshold α at $H(t)$ ranging between zero and one.

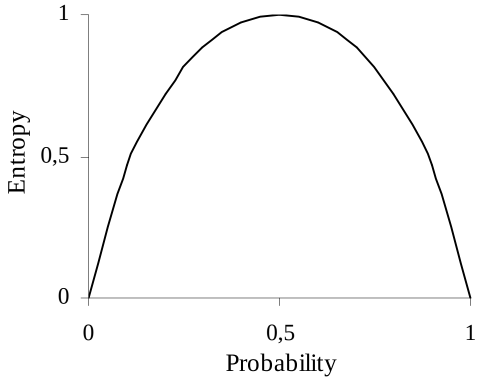

Figure 5. The relation between the entropy of a word and its probability of occurrence.

The construction of a reduced vocabulary $V(E)$ from the total vocabulary $V(D')$ can be summarized in the following way:

$$V(E) = \{t_1 ... t_m\} = \bigcup_{t \in V(D')} \{t : t \notin L \,\&\, H(t) > \alpha\} \qquad (9)$$

In which L is a stop-list, m is the number of words selected for the parts of discourse modeling into a vector space E and $|V(E)| < |V(D')|$.

An operation of word selection that uses a stop-list and a statistical filter can significantly reduce the size of the vocabulary. For example, in Lahlou's analysis of the SR of EATING, the total vocabulary $V(D')$ contained 16,896 different words. After the various filtering steps, the reduced vocabulary $V(E)$ has only 828 different words (Lahlou, 1995a, p.170).

The second sub-step of the vectorization is weighting words $t \in V(E)$ according to their importance for each vectorized part of discourse $\vec{p} \in E$.

2.1.2 Word weighting

Vectorized parts of discourse $\vec{p} \in E$ as defined in (1) are formed in attributing a weighting value w_t to each word t of the selected vocabulary $V(E)$. The weighting sub-step's goal is to rank words according to their semantic importance for each part of discourse.

The term weighting in a part of discourse is generally determined following two principles. The first one is representativity: a word in a part of discourse is heavily weighted when it is highly representative of its content. The second factor is discrimination: a word is heavily weighted in a part of discourse when it allows for discriminate its content from the rest of the corpus.

In Lahlou's study, which was using the software ALCESTE, the weighting was binary $w_t \in \{0,1\}$. This means that the value of a word was $w_t = 1$ if it was present in a part of discourse, or $w_t = 0$ if it wasn't. Another possible operationalization may consist, for example, in weighting a word according to its relative frequency in a part of discourse.

However, binary weighting and frequency weighting are solely based on a representativity principle. If we look elsewhere in the literature on the same topic we find that a classical weighting technique combining both representativity and discrimination principle is the $tf \cdot idf$ coefficient (i.e. term frequency × inverted document frequency). It can be calculated as follows:

$$w_{t,i} = tf_i \cdot idf_t = tf_i \times \log \frac{n}{df_t} \tag{10}$$

In which tf_i is the frequency of the term $t \in V(E)$ in the vectorized part of discourse $\vec{p}$, df_t is the number of parts of discourse the term *t* appears in, and *n* is the total number of parts of discourse in *E*.

The $tf \cdot idf$ is to be interpreted as follows : if a word has a high frequency in a given part of discourse and occurs in a limited number of parts of discourse within the rest of the corpus, this word has a high weight value for this particular part of discourse. Inversely, if this word is not frequent in a given part of discourse and is present in many parts of discourse of the corpus, this word hence has a low weight value. Several alternative ways to compute the weighting value can be found in the literature (Harman, 2005).

2.2 Similarity Calculation Between Parts Of Discourse

The second step in the data modeling phase is the calculation of similarity relations among every parts of discourse selected to form the vector space E of the SR. The goal of this step is to associate to each pair of vectorized parts of discourse a value representing their degree of similarity (or difference).

This step may be formalized in the following function:

$$f : E \to (E, d) \tag{11}$$

Function (11) takes a vector space *E* as input and associates to it a metric *d* as defined in (3).

A metric is an algebraic operationalization of the concept of semantic equivalence defined in the introduction. The computation of these similarity relations is at the root of the discovery of the SR's classes of meaning.

Similarity is computed according to the terms retained in the previous sub-step. Two vectorized parts of discourses that share the same vocabulary (i.e. similar word co-occurrence patterns) are interpreted as semantically close to each other. This is why the previous word filtering step is so crucial. Without this pre-selection of relevant words, the similarity calculation could lead to the conclusion that two parts of discourse are semantically similar because they share words such as ‘thus’, ‘the’, ‘have’, etc., which would obviously be fallacious.

In Lahlou's work, the metric used through the ALCESTE's algorithmic operationalization, was the ‘chi-square’. It is calculated as follows:

$$chi2(\vec{p}_i,\vec{p}_j)=\sum_{t=1}^{m}\frac{1}{w_{t,i}+w_{t,j}}\left(\frac{w_{t,i}}{|\vec{p}_i|}-\frac{w_{t,j}}{|\vec{p}_j|}\right)^2 \tag{12}$$

Wherein $w_{t,i}$ is the weight of the term *t* in the part of discourse $\vec{p}_i$, $m=|V(E)|$ is the number of term selected previously for the data modeling process and $|\vec{p}_i| = \sqrt{\sum_{t=1}^{m} w_{t,i}{}^2}$ is the norm or the length of the vectorized part of discourse $\vec{p}_i$. The $chi2$ metric is a measure of the distance between two parts of discourse. The distance is equal to zero when two parts of discourse are proportionally equivalent, it is small when they are similar, and high when the difference is important.

Another metric widely used in other algorithmic operationalizations, different that the one of ALCESTE, is the *cosine*:

$$\cos(\vec{p}_i,\vec{p}_j)=\frac{\vec{p}_i\cdot\vec{p}_j}{|\vec{p}_i|\cdot|\vec{p}_j|} \tag{13}$$

As before, $|\vec{p}_i| = \sqrt{\sum_{t=1}^{m} w_{t,i}{}^2}$ is the norm or the length of the vectorized part of discourse $\vec{p}_i$, $w_{t,i}$ is the weight of the term *t* in the part of discourse $\vec{p}_i$ and $\vec{p}_i\cdot\vec{p}_j=\sum_{t=1}^{m}\left(w_{t,i}\cdot w_{t,j}\right)$ the dot product between the two vectors.

The *cosine*, as the $chi2$, is scale invariant, a very interesting property, since parts of discourse are sometimes of different lengths. For instance, unlike the Euclidian metric (another widely used metric), which measures similarity between two vectors in terms of absolute value, the $chi2$ and the *cosine* measure the similarity in terms of proportions. Moreover, the *cosine* has the advantage over the $chi2$ that it can be interpreted more easily, since its values are normalized. The *cosine* is a measure of proximity (i.e. correlation) between two vectors. The *cosine* value is zero when two vectors are orthogonal and one if they are equivalent.

Different metrics can be used for computing semantic proximity. Dozens of different ones can be found in the literature (Ellis et al., 1994; Rajman & Lebart, 1998). The choice of a metric is an important methodological decision in terms of algorithmic operationalization and software implementation. Furthermore, this is also a complex and determinant theoretical decision for geometrically inspired cognitive models as the VSM is (see Gärdenfors, 2000).

The calculation of the similarity relations between vectorized parts of discourse thematically linked to the studied SR is the last step of the modeling phase. The third phase is the SR's data analysis.

3. DATA ANALYSIS

The third phase consists in the construction of the SR's semantic map, which involves generating a class structure from the parts of discourse, extracting salient lexical content from every class and categorizing its content. We propose to see the analysing phase as a process going from an extensional description, through automatic classification, of the semantic of the parts of discourse, to an intentional comprehension, through categorization, of this semantic. The extraction of the classes' salient lexical contents is an intermediary step between the two.

3.1 Automatic Text Classification

The previous phase results in the production of a vector space that represents the semantic proximity relations between all parts of discourse wherein the studied SR is linguistically instantiated or communicated. These relations of semantic proximity vary a lot. Certain parts of discourse are very close to one another in the vector space, whereas others are quite far. The hypothesis behind the automatic text classification step is that the distribution of parts of discourse lexical features is not random but structured. We assume that text clustering algorithms may help to discover subsets of chunks of discourse related to the SR semantic dimensions.

The goal of this step is to find the main semantic classes that best describe the similarity relations between parts of discourse. This step can be formalized as follows:

$$f : (E, d) \to P \qquad (14)$$

The function (14) takes as input a vector space and a distance function (E, d), and generates as output a partition $P = \{G_1 ... G_k\}$ of the vectorized parts of discourse $\vec{p} \in E$. A partition is a set of *k* classes in which each contains parts of discourse that are semantically equivalent, or at least, semantically similar to one another.

Lahlou gives the following definition to a text classification operation:

> “The statements are classified by analogy and contrast, on the basis of their lexical content. This gives classes that contain statements. Similar statements are classified together in one class, and as different as possible of the statements of the other classes.” (Lahlou, 1996a, p.77, our translation)

This joins several other canonical definitions:

> “Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters).” (Jain et al., 1999, p.264)

> “Clustering algorithms group a set of documents into subsets or clusters. The algorithms’ goal is to create clusters that are coherent internally, but clearly different from each other.” (Manning et al., 2008, p.349)

> “[…] text classification is defined as an operation that is applied to textual entities on which equivalent classes are built. Classification is hence a process by which textual information is clustered together according to some criteria.” (Meunier et al., 2005, p.962)

Common to these definitions is the general idea that parts of discourse (i.e. text segments) sharing lexical features are grouped together in equivalence classes, while those that are different are separated. The classification follows three conditions:

(i) $G_i \neq \varnothing$, no class is empty;

(ii) $\bigcup_{G_i \in P} G_i = E$, the set of classes and the vector space have the same extension;

(iii) $\bigcap_{G_i \in P} G_i = \varnothing$, all classes are disjoint.

Finding these equivalence classes is a nontrivial task. There are an exponential number of possible partitions of the same set of parts of discourse[7]. Clustering algorithms are heuristics that are designed to assist the analyst in the discovery of these classes. A clustering algorithm is used to induce (construct) a partition from the parts of discourse that will express or approximate the semantic classes of the SR. In other words, this is an optimization problem. These algorithms seek for the partition that either maximizes the inter-class inertia or minimizes the intra-class inertia, or both simultaneously. There are dozens of algorithms suitable for computing the classification operation (Xu & Wunsch II, 2005; Berkhin, 2006; Jain et al., 1999). We show two of them.

In the ALCESTE*'s* algorithmic operationalization Lahlou used in his research, the algorithm used to compute the function (14) is a kind of descendant hierarchical classification (DHC) technique. DHC algorithm consists in recursively splitting a cluster into two sub-clusters. The process is illustrated by a dendrogram in Figure 6:

[7] For example, there are 42 different possible partitions of a set of 10 elements and 190,569,292 different possible partitions of a set of 100 elements!

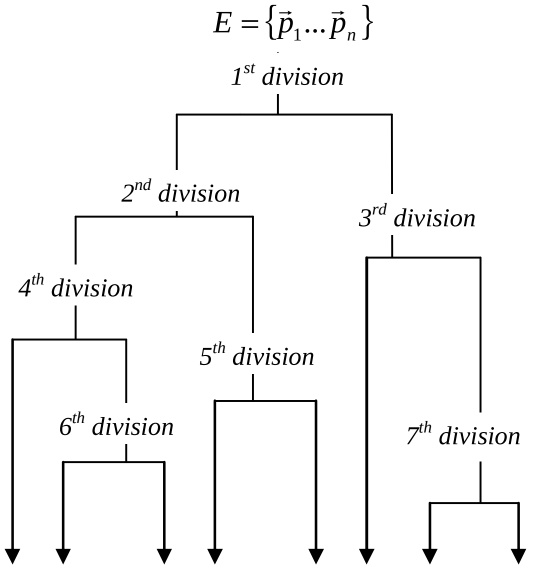

Figure 6. Dendrogram representing a descending hierarchical clustering of a set of parts of discourse.

This kind of algorithm implies three calibration decisions, that is, there are three parameters in this kind of algorithm that need to be defined by the researcher. The first decision is about the choice of a selection criterion with regards to the class to be submitted to a division process. The first iteration submits the whole set of parts of discourse to a bipartition, but from the second iteration, the algorithm needs one or more criteria in order to select the class that is going to be divided into two sub-classes.

The second decision is about the choice of an optimization criterion that the algorithm seeks to satisfy during the division process of a class into two sub-classes. The optimization objective of the DHC algorithm is finding the best splitting that maximize inter-class inertia. This objective is a function that can be formally written out in the following way:

$$\underset{(G_i^k, G_j^k) \in \Omega(G_k)}{\arg\max} \; d(\vec{\mu}(G_i^k), \vec{\mu}(G_j^k)) \qquad (15)$$

$\Omega(G_k)$ is the set of possible divisions[8] of a class G_k into two sub-classes (G_i^k, G_j^k), $\vec{\mu}(G_i^k)$ is the centroid of a sub-class G_i^k and $d(\vec{\mu}(G_i^k), \vec{\mu}(G_j^k))$ is a metric as defined in (3), (12) and (13).

The third calibration consists in fixing the maximal number of iterations of the algorithm, that is, the maximum number of divisions it makes. Each divisive process *i* creates *i+1* sub-clusters. For example, Figure 6 shows the process as being repeated through seven iterations, thereby, generating a partition of eight classes.

The implementation of this algorithm in ALCESTE is founded on the following calibration setting. Firstly, the selection criterion is the cluster size. At each iteration, this is the class containing the greatest number of parts of discourse which is divided into two sub-classes. Secondly, the metric used in function (15) is the $chi2$ as defined in (12). Thirdly, the maximum number of iterations is by default fixed at 15, meaning that it is possible to generate up to a maximum 16 classes (Reinert, 2002).

Another suitable algorithm for automatic classification is the flat clustering algorithm k-means (KM). This algorithm is very different than the one in ALCESTE, but it is perhaps the most widely used across disciplines. This algorithm is an iterative process that clusters together parts of discourse which are nearest to the same point of attraction in the vector space. The points of attraction are centroid vectors recalculated at each iteration until stabilization. Figure 7 illustrates a very simple example of the process.

[8] The number of possible divisions of a class into two sub-classes is exponential. For one class containing *n* parts of discourse, there are $2^{n-1} - 1$ possible divisions (Edward & Cavalli-Sforza, 1965). In practice, as soon as $n > 20$ it becomes extremely difficult to exhaustively search for the best possible division. Heuristics are then used that, without guaranteeing that the best bipartition will be arrived at, allow a good approximation. In the *Alceste* software, this heuristic is a factor analysis algorithm.

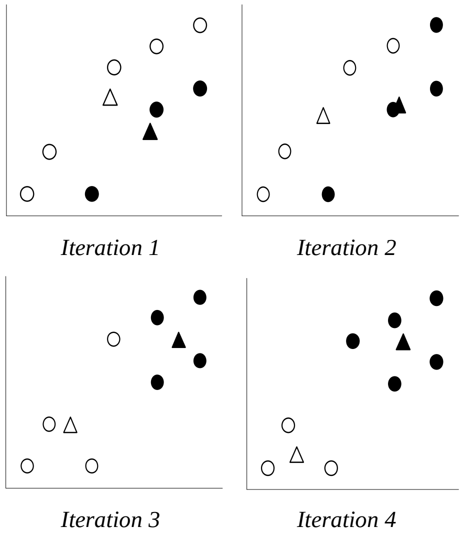

Figure 7. An example of a k-means classification of eight parts of discourse into two classes. The two centroids are represented by triangles, the eight parts of discourse by circles.

The KM algorithm implies two calibrations. The first one is the choice of the optimization criterion that the algorithm seeks to satisfy during the clustering process. The optimization objective of the KM algorithm is finding the best clustering that minimize intra-class inertia. This objective can be formalized by the following function:

$$\arg\min_{P} \sum_{i=1}^{k} \sum_{\vec{p} \in G_i} d(\vec{p}, \vec{\mu}(G_i)) \tag{16}$$

$P = \{G_1 ... G_k\}$ is a partition of *k* classes, $\vec{\mu}(G_i)$ is a centroid and $d(\vec{p}, \vec{\mu}(G_i)$ is a metric as defined in (3), (12) and (13).

Initializing the coordinates of the *k* attraction points in the vector space is the second calibration requested by the KM algorithm. At the first iteration, random values are usually assigned to them. Subsequently, their values correspond to the closest parts of discourse centroids. The number *k* is fixed between 2 and $n-1$, where *n* is the number of parts of discourse in *E*.

The automatic classification step’s goal is identifying classes of meaning in discourses through which the SR is expressed. This is a determinant step of the method. Lahlou suggests interpreting these classes as being “cognitive elements” forming the studied SR. Many clustering algorithms can compute the clustering function (14). Furthermore, each algorithm is a different heuristic and involves important theoretical decisions that could influence the results (Estivill-Castro, 2002).

The automatic classification step groups together the similar parts of discourse and separates the different ones. The technique allows for an extensional description of the semantic classes of the discourses in which the SR is embodied. The second and third steps of the analysing phase are geared towards the intentional comprehension of the content in these semantic classes.

3.2 Extracting Salient Lexical Content From Every Class

The second step of SR analysis consists in extracting salient lexical content from each class produced in the previous step. The salient lexical content is the set of words strongly associated with the parts of discourse grouped together in a particular class. The goal of this step is to find, for each class, the set of words that best characterize lexically its semantic. It consists in offering to the analyst a general linguistic representation of the characteristic content of the classes. One way, among many others, of doing this is to extract, from every class, their salient lexical features.

This operation can be formalized as follows:

$$f : P \to Q \tag{17}$$

This function takes a partition $P = \{G_1 ... G_k\}$ as input and extracts from it the set of salient contents $Q = \{T_1 ... T_k\}$, where $T_i = \bigcup t$ is the set of words selected to characterize the class G_i.

Several expressions have been used to refer to these salient contents. Lahlou, building up on Reinert’s work (Reinert, 1993), speaks of ‘lexical worlds’. In lexicometrics, researchers rather speak of ‘lexical specificities’ (Lebart & Salem, 1994). In the field of information retrieval, salient words are used as ‘cluster labelling’ (Manning et al., 2008, p.396). In structural semantic, these words may pertain to what is called ’isotopies’ (Rastier, 2011).

In our context, it is rather the role given to these salient lexical contents within the method that really matter. This extraction process is an intermediary step between an extensional description and an intentional comprehension of the SR semantic map. These salient words are statistical clues. They are selected with the help of an association coefficient calculated between a word and a class[9]. The researcher can use these clues at the subsequent step in order to categorize the classes' semantic content.

In practice, little more than ten salient words per class are extracted in order to characterize its semantic content. For example, in his work on the SR of EATING sedimented in a dictionary, Lahlou proposes that, with a DHC algorithm, six classes can be made with the 544 parts of discourse selected for the analysis. Of these six classes, here is a fragment of the salient (French) words extracted:

$T_1 =$ *{désir, faim, appértit, soif, satisfaire, envie, convoit, assouvi, rassasi, avidité…}*

$T_2 =$ *{touch, attrape, prendre, main, nez, attaqu, embrass, baise, joue, mordre…}*

$T_3 =$ *{viande, pain, aliment, fruit, pat, légum, animal, cuire, tranch, bouill…}*

$T_4 =$ *{repas, table, restaur, plat, dîne, cuisin, déjeuner, invit, serv, buffet…}*

$T_5 =$ *{connaître, bon, sentir, aim, agréable, emploi, goût, possed, vivre, est…}*

$T_6 =$ *{rempl, épuise, encombr, ronge, sature, consum, détruire, approvisionn, sujet, absorb…}*

Figure 8. Extraction, done by Lahlou with the *khi2* coefficient, of the salient words from each of the six classes of the SR of EATING. The studied discourse is sedimented in a French language dictionary. Words are reduced to their stems. Only the ten most important words are presented.

There are many suitable association coefficients for computing the function (17). We show two simplified arithmetic ways to calculate such coefficient.

Let *a* be the number of parts of discourse in the class G_i that contain the word *t*; *b* the number of parts of discourse that contain the word *t* but aren't member of the class G_i; *c* the number of parts of discourse that do not contain the word *t* but are member of the class G_i; *d* the

[9] A word will be more associated to a class if it is very common to it, and very rare in other classes of the partition.

number of parts of discourse that do not contain the word *t* and aren't member the class G_i; and $n = a + b + c + d$ the total number of parts of discourse in the partition *P*.

The work of Lahlou with the software ALCESTE relies on the $khi2$ coefficient (not to be confused with the $chi2$ metric):

$$khi2(t,G_i) = \frac{n \times (ab - ac)^2}{(a+b) \times (a+c) \times (a+d) \times (c+d)} \tag{18}$$

Again, if we look elsewhere in the literature on the same topic we find that another coefficient widely used is the "information gain" (Manning et al. 2008, p.252):

$$\begin{aligned} I(t,G_i) = {} & \left(\frac{a}{n} \log \frac{n \times a}{(a+b)(a+c)} \right) + \left(\frac{b}{n} \log \frac{n \times b}{(b+a)(b+d)} \right) \\ & + \left(\frac{c}{n} \log \frac{n \times c}{(c+a)(c+d)} \right) + \left(\frac{d}{n} \log \frac{n \times d}{(d+b)(d+c)} \right) \end{aligned} \tag{19}$$

These coefficients (18) and (19) attribute to a term $t \in V(E)$ and a class $G_i \in P$ a score representing the strength of the association between them. The higher the score, the more specific (statistically) the word *t* is to the class G_i. The more specific, the more it is justified (statistically) to consider the word *t* as a salient lexical clue for the categorization process and the semantic map comprehension.

The $khi2$ is a statistical independence test between a word *t* and a class G_i. The higher the score, the higher the confidence about the word's relevance one can have. The information gain coefficient measures the quantity of information obtained on class G_i through a word *t*. Its interpretation is sometimes easier than it is for the $khi2$ because the range of its values is normalized between zero and one: when $I(t,G_i) = 0$, the word *t* gives no information about G_i, inversely, when $I(t,G_i) = 1$, the word *t* gives complete information about G_i.

Different coefficients have different rationales. Salient words extracted with a given coefficient are not necessarily the same ones retrieved with another (Manning et al. 2008, p.257-

258). It's an important operation as the last step of the method – the categorization – depends on it.

3.3 Categorization Of Classes

Until now, the partition of the SR map has only been described extensionally by the complete enumeration of the class members and by the extraction of salient words. To complete the semantic mapping, the researcher must also intentionally define the classes. In order to accomplish that, he may attribute synthetic semantic categories to each class.

This last step refers in short to what Lahlou called the "art of comprehension". The approach aims to determine, by an abductive inference process, whether the salient terms identified statistically corresponds to a semantic consistency or not. In Lahlou's own words:

> "The process of classes' comprehension lies in the decision to consider the typical features of a class [i.e. salient words] represent all a unique 'idea' through which this class is identified." (Lahlou, 1995b, p.224, our translation)

Formally, the categorization process is a kind of function as this one:

$$f : Q \to C \tag{20}$$

$Q = \{T_1 ... T_k\}$ represents sets of salient lexical contents related to each classes as defined in (17) and $C = \{c_1 ... c_k\}$ is the set of categories inferred by the interpreter. Lahlou refers to these categories as 'paradigms', in the saussurian meaning of the word. These categories are external to corpus data. This is the analyst who infers and adds them during her interpretation.

According to Lahlou, we must emphasize that this step is an *interpretative* process, carried out by the researcher, and not by the computer[10]. Nevertheless, the process of

[10] This is a very complex issue, and in this paper, we can just highlight it. It is particularly on this topic that the 'manual' and the automatic text analysis approach often clash with each other. A hermeneutical attitude will defend the specificity and the incommensurability of text interpretation, while a computationalist attitude will often defend its algorithmic operationalization possibility. In the state of the art in computer sciences, it is not clear if there exists

comprehension may be described and Lahlou proposes to see it as an abductive inference process of three sub-operations (Lahlou, 1995b, p.225, 2003, p.57). The first operation is the recognition that the salient words T_i of a class G_i are signs of a same underlying semantic category c_i. The second sub-operation is the inference made by the analysts (and informed by his knowledge of the domain) of a plausible hypothesis about this semantic category c_i. This is about formulating a plausible intentional definition of the class G_i. Finally, the third sub-operation is an evidence accumulation process in order to corroborate the previous hypothesis. It's a verification that c_i really is a semantic category which plausibly underlies most of the salient terms $t \in T_i$ and most of the parts of discourse $\vec{p} \in G_i$, and only underlies them. The whole process is an iterative one: if a first categorization hypothesis turned out to be incorrect, the analyst must hence adjust it and start over.

The process of comprehension is repeated for each class of the partition. The operation consists in finding, for each class, the best tag synthesizing its semantic content. The 'art of comprehension' is obviously a much more complex hermeneutical process than what we are suggesting here. We are only discussing at an abstract level. We suggest that it may be illustrated as a sort of null hypothesis test, which seeks to minimize false-positive and false-negative errors.

Venn diagrams (a), (b), (c) and (d) from Figure 9 illustrate four types of categorization. Let us suppose that the two sets pictured in those four diagrams represent the extension of a class G_i and of a category c_i. Diagram (a) represents a categorization that is too general, thus producing false-positive error. This situation means that the analyst has, for example, defined class G_i by the category $c_i = \text{ANNIMALS}$, while $c_i = \text{MAMMAL}$ would have been more accurate. Inversely, diagram (b) illustrates a categorization that is too restrictive, thus producing false-negative error. In diagram (c), the analyst made a categorization that is too polysemic, thus causing both false-positive and false-negative errors. Finally, diagram (d) represents a perfect categorization, without error, in which the extension of the class G_i is the same as the extensions of the category c_i.

or not an algorithm that a computer could use to compute this function (20). It could be that the function (20) is not Turing-computable, i.e. not computable on a Turing machine (Meunier, 2002).

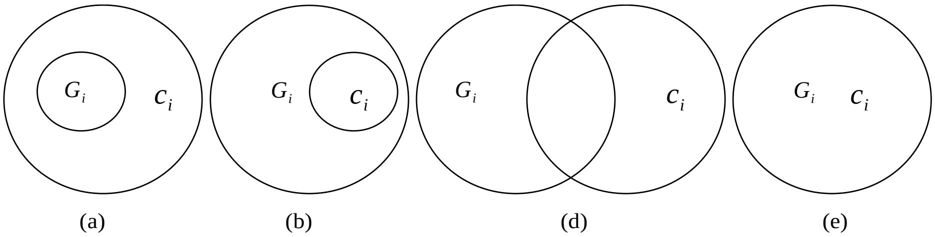

Figure 9. Venn diagrams illustrating the overlap between the extension of a category and of a class.

These are illustrations, and we can presume that a categorization free of errors is virtually impossible, and, in fact, is not even necessary for the analyst's comprehension process. Nevertheless, formally speaking, the categorization can be understood as a process seeking an optimization scheme. The rationale of this scheme is minimizing the functions (21) and (22):

$$\arg_C \min \frac{c_i - G_i}{c_i} \tag{21}$$

$$\arg_C \min \frac{G_i - c_i}{G_i} \tag{22}$$

Function (21) measures the false-positive error, while and the function (22) measures the false- negative error[11].

In his study of the SR of EATING, according to this process, Lahlou categorized the six classes of its semantic map as it's shown in the following Figure:

[11] In computer sciences, function (21) and (22) are called precision and recall tests.

f (désir, faim, appértit, soif, satisfaire, envie, convoit, assouvi, rassasi, avidité…) = LIBIDO

f (touch, attrape, prendre, main, nez, attaqu, embrass, baise, joue, mordre…) = PRENDRE

f (viande, pain, aliment, fruit, pat, légum, animal, cuire, tranch, bouill…) = NOURRITURE

f (repas, table, restaur, plat, dîne, cuisin, déjeuner, invit, serv, buffet…) = REPAS

f (connaître, bon, sentir, aim, agréable, emploi, goût, possed, vivre, est…) = VIVRE

f (rempl, épuise, encombr, ronge, sature, consum, détruire, approvisionn, sujet, absorb…) = REMPLIR

Figure 10. Lahlou's categorization of the six classes of the semantic map of the SR for EATING. Words in small capitals are the categories.

Then, Lahlou suggests to interpret the SR of EATING by the set of following categories:

SR of EATING = {LIBIDO, VIVRE, PRENDRE, REMPLIR, NOURRITURE, REPAS}[12]

The categorization is the last step of the method and only then can the analyst determine whether the SR’s semantic mapping is achieved or not. For various reasons, the mapping is not always adequate. Such an outcome may be of methodological origin: for instance, relevant parts of discourse selection criteria, relevant words selection criteria, similarity metric, classification algorithm and so on may, each or all, prove to be inadequate for a given task.

Furthermore, empirical data may themselves resist the analysis. The analysis may fail because the data itself lack structure, i.e. the SR’s parts of discourse aren’t organized through semantic equivalence classes. The method shown is an interpretive process computer assisted with various algorithms. But in the end, the analysis of a SR’s semantic map must always be validated by the scholar.

[12] We will not here comment the analysis and the interpretation Lahlou makes of the SR of EATING. We invite the reader to consult Lahlou's works for further details and comments on the interpretation of these results.

CONCLUSION

The purpose in this paper was mainly methodological: to formalize the TMMs used for SR analysis in large corpora in order to highlight the different operations and assumptions involved as well as the algorithms and the software settings, and their consequences on relevant issues.

We have presented a method in three phases. The first one is the empirical data collection. It implies gathering a corpus of documents D in which is retrieved a sub-corpus D' of parts of discourse thematically linked to the SR being studied. The second phase is the data modeling. It involves the construction of the SR's vector space E (involving itself the selection and the weighing of relevant words among all parts of discourse) and the calculation of similarity relations between all parts of discourse (E,d). The third and final phase of the method is the data analysis. It involves, firstly, an extensional semantic map description P through the SR’s parts of discourse automatic classification. Secondly, salient lexical contents Q are extracted from classes obtained from the partition. The analysis ends by an intentional comprehension of the semantic map through a categorization process of the classes’ contents C.

The TMMs for studying SR may therefore be synthesized by a meta-function of this form:

$$F : D \rightarrow D' \rightarrow E \rightarrow (E,d) \rightarrow P \rightarrow Q \rightarrow C \qquad (23)$$

The whole method is an iterative process. Although our presentation may suggest a linear process, the method actually involves many loops, as it may demand, in many occasions, to step back, adjust parameters, and to proceed by trial-and-error. The researcher may opt for different operationalizations through algorithm selection, different implementation through parameters calibration, robustness evaluation and/or by comparing and/or combining the classification results obtained for each alternative.

Opening The Black Box

As we have said in the introduction (see Figure 1), the method must not be confused with the software that implements it. As such, its functional description differs from its algorithmic description, which in turn also differs from its concrete description. The functional level describes the formal operations realized at each step while the algorithmic level describes the way these operations are computed. At last, the concrete level describes the software used and its parameters, in order to implement these algorithms. There are always many ways for computing the same function, and many softwares that can implement it. All levels of description are crucial in understanding a computational method. Unfortunately, there are still many methodological analyses of TMMs which confine themselves to the concrete level of description.

We have proposed, for each of the seven functions which compose the meta-function (23), two possible algorithmic operationalizations – typically one from ALCESTE and an alternative. This means that the method presented in this article could be operationalized at least in 128 different ways (i.e. 2^7). Furthermore, one could find in the literature about 10 different algorithms satisfying the logical constraints for each function that made up the meta-function (23). As such, there are probably millions of different ways to operationalize the method (i.e. 10^7). These various operationalizations are likely not equivalent, and rely on various hypotheses which can have consequences.

When a scholar uses a ‘closed’, ‘turn key’ or proprietary software such as ALCESTE, he implicitly and sometimes unknowingly endorses the several and often hidden operationalization decisions made upstream by the software developer. Conversely, when the method is abstracted on it own, it becomes much more transparent, intelligible, and its genericity and flexibility can be better assessed. Assumptions and methodological decisions are made explicit; it can be more easily submitted to scientific evaluation.

Mining Digital Public Spaces

During our presentation, we have illustrated the various steps and operations involved in the method with the help of previous works from Lahlou (1994, 1995a, 1995b, 1996a, 1996b, 1998, 2003), and in particular his case study on the SR of EATING as sedimented in a common French language dictionary's discourse. Lahlou's pioneering works were among the most important in the development and the application of TMMs to the study of SR in large corpora. Given the abundance of digital textual data accessible today, there is no doubt that these methods are bound to gain in importance in social sciences and humanities. These data are press articles, encyclopaedias, literature, blogs as well as contents from the social web coming from digital public spaces. They are, as Lahlou said about dictionaries, areas where sedimentation of culture, social practices and cognitive activities happen. They are empirical (digital) traces left by a population. As such, these traces may be subject to scientific inquiry.

Shatter Traditional Boundaries

TMMs are computational methods. These methods shatter traditional distinctions. TMMs enable reproducible and falsifiable experimentation, but unlike laboratory experimentation, TMMs also have ecological validity, as it allows for natural data analysis, i.e. not provoked by the researcher’s intervention. As such, they share something with anthropological ethnography, but at a very different observational scale. At last, they are quantitative methods, but applied to meaning analysis, a domain long reserved to qualitative methods.

Towards A Three-Leveled Social Representation Analysis

We conclude by broadening the discussion towards a more general methodological challenge. The research program of SR is now relatively mature and this is noticeable especially in the methodological standards the scientific community has gradually imposed on itself. One of these standards has been proposed by Abric, namely the three levels of SR analysis.

According to Abric (1994, p.60, 2003b, p.376), SR analysis methods can be divided into three categories, depending on the level of analysis they enable. These levels are: SR content and category recognition; SR structure identification; and SR core identification. We conclude that the TMMs as used by Lahlou and others only achieve the first level of SR analysis.

Indeed, the method as presented here is relatively silent regarding the cognitive organization of the semantic categories in SR, and about its social organization. We believe that future research should be made in this direction, in order to develop TMMs enabling second- and third- level SR analysis.

REFERENCES

Abric, J.-C. (1993). Central system, peripheral system: Their function and roles in the dynamics of social representations. *Papers on Social Representations, 2,* 75–78.

Abric, J.-C. (1994). Méthodologie de recueil des représentations sociales. In J.-C. Abric (Ed.), *Pratiques sociales et représentations* (pp. 59-82). Paris: PUF.

Abric, J.-C. (1994). *Pratiques sociales et représentations*. Paris: PUF.

Abric, J.-C. (2003b). L'analyse structurale des représentations sociales. In S. Moscovici & F. Buschini (Eds.), *Les méthodes des sciences humaines* (pp. 375-392). Paris: PUF.

Abric, J.-C. (Ed.) (2003a). *Méthodes d'étude des représentations sociales*. Ramonville Saint-Agne, Eres.

Alba, M. (2004). El método ALCESTE y su aplicación al estúdio de las representaciones sociales del espacio urbano: el caso de la Ciudad de México. *Papers on Social Representations, 13*(1), 1-20.

Bardin L. (2003), L'analyse de contenu et de la forme des communications. In S. Moscovici & F. Buschini (Eds.), *Les méthodes des sciences humaines* (pp. 243-270). Paris: PUF.

Bauer, M., & Gaskell, G. (1999). Towards a paradigm for research on social representations. *Journal for the Theory of Social Behaviour, 29*(2), 163-186.

Bauer, M. W., & Aarts, B. (2000). Corpus construction: a principle for qualitative data collection. In M. Bauer & G. Gaskell (Eds.), *Qualitative researching with text, image and sound: a practical handbook* (pp. 19-37). London: Sage.

Beaudouin, V, & Lahlou, S. (1993). *L'analyse lexicale, outil d'exploration des représentations. Réflexions illustrées par une quinzaine d'analyses de corpus d'origines très diverses*. Paris : Crédoc, Cahiers de recherche, n°48, septembre 1993.

Berkhin, P. (2006). Survey of clustering data mining techniques. In J. Dans Kogan, C. Nicholas, C. & M. Teboulle (Eds.), *Grouping Multidimensional Data* (pp. 25-71). Springer: Berlin Heidelberg.

Breakwell, G. M., & Canter, D. V. (Eds.) (1993). *Empirical Approaches to Social Representations*. Oxford: Clarendon.

Buschini, F., & N. Kalampalikis (2002). La synonymie, l'analogie et la taxinomie. In C. Garnier, & W. Doise (Eds.), *Les représentations sociales: Balisage du domaine d'études* (pp. 187-206). Montréal: Éditions nouvelles.

Caillaud, S., Kalampalikis, N., & Flick, U (2011). The social representation of Bali Climate Conference in the French and German media. *Journal of Community and Applied Social Psychology*. doi: 10.1002/casp.1117.

Colucci, F. P., & Montali, L. (2008). Comparative application of two methodological approaches to the analysis of discourses. *Internationnal Journal of Multiple Research Approaches*, *2*(1), 57-70.

Crane, G. (2006). What do you do with a million books?. *D-Lib Magazine*, *12*, 3.

Dany, L., & Apostolidis, T. (2002). L'étude des représentations sociales de la drogue et du cannabis/ un enjeu pour la prévention. *Santé publique, 14*(4), 335-344.

Demazière, D., Brossaud, C., Trabal, P., & Van Meter, K. (Eds.) (2006). *Analyses textuelles en sociologie - Logiciels, méthodes, usages*. Rennes: PUR.

Diesner, J., & Carley, K.M. (2005). Revealing social structure from texts: meta-matrix text analysis as a novel method for network text analysis. In V. K. Narayanan & D. J. Armstrong (Eds.), *Causal mapping for information systems and technology research: Approaches, advances, and illustrations* (pp. 81-108). Harrisburg, PA: Idea Group Publishing.

Doise, W., & A. Palmonari (1986). *L'étude des représentations sociales*. Paris: Delachaux et Niestlé.

Doise, W., Clémence, A., & F. Lorenzi-Cioldi (1992). *Représentations sociales et analyses de données*. Grenoble. Paris: PUG.

Edward, A. F. S., & Cavalli-Sforza, L. L. (1965). A method for cluster analysis. *Biometrics, 21*(2), 362-375.

Ellis, D., Furner-Hines, J., & Willett, P. (1994). Measuring the degree of similarity between objects in text-retrieval systems. *Perspectives in Information Management, 3*(2), 128-149.

Estivill-Castro, V. (2002). Why so many clustering algorithms: a position paper. *ACM SIGKDD Explorations Newsletter, 4*(1), 65-75.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17*(3), 37-54.

Feldman, R., & Sanger J. (2007). *The text mining handbook*. New York: Cambridge University Press.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis, Special Volume, Philological Society* (pp. 1-32). Oxford: Blackwell,.

Flament, C., & M. L. Rouquette (2003). *L'anatomie des idées ordinaires: comment étudier les représentations sociales*. Paris: Armand Collins.

Flick, U., & Foster, J. (2008). Social Representations. In C. Willig & W. Stainton-Rogers (Eds.), *Handbook of qualitative research in psychology*. London: Sage Publications.

Gaffié, B., Marchand, P., & Cassagne, J.M. (1998). Positionnement droite / gauche et portraits de groupes politiques. Revue canadienne des sciences du comportement. *Canadian Journal of Behavioural Science, 30*, 36-48.

Gärdenfors, P. (2000). *Conceptual spaces: the geometry of thought*. Cambridge (Mass.): MIT Press.

Garnier, C., Marinacci, L., & Quesnel, M. (2007). Les représentations sociales de l'alimentation, de la santé et de la maladie des jeunes enfants. *Service social, 53*(1), 109-122.

Geka, M., & Dargentas, M. (2010). L'apport du logiciel ALCESTE à l'analyse des représentations sociales: l'exemple de deux études diachroniques. *Les Cahiers Internationaux de Psychologie Sociale, 85*(1), 111-135.

Gilles, I., Bangerter, A., Clémence, A., Green, E. G. T., Krings, F., Mouton, A., Rigaud, D., Staerklé, C., & Wagner-Egger, P. (2011). Collective symbolic coping with disease threat and othering: A case study of avian influenza. *British Journal of Social Psychology*. doi: 10.1111/j.2044-8309.2011.02048.x

Glenisson, P., Glanzel, W., Janssens, F., & De Moor, B. (2005). Combining full text and bibliometric information in mapping scientific disciplines. *Information Processing & Management, 41*(6), 1548–1572.

Harman, D. (2005). The history of IDF and its influences on IR and other fields. In J. I. Tait (Ed.), *Charting a new course: Natural language processing and information retrieval. Essays in honour of Karen Spärck Jones* (pp. 69-79). Netherlands: Springer.

Harris, Z. S. (1991). *A theory of language and information: A mathematical approach*. Oxford: Clarendon Press.

Henry, P., & Moscovici, S. (1968). Problèmes de l'analyse de contenu. *Langage, 3*(11), 36-60.

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A brief survey of text mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology, 201*(1), 19-62.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys, 31*(3), 264-323.

Jodelet, D. (Ed.) (1989). *Les représentations sociales*. Paris: PUF.

Jurafsky, D., & Martin, J.H. (2000). *Speech and language processing: An introduction to natural language processing, speech recognition, and computational linguistics*. Prentice-Hall.

Kalampalikis, N. (2003). L'apport de la méthode ALCESTE dans l'analyse des représentations sociales. In J.C. Abric (Ed), *Méthodes d'étude des représentations sociales* (pp. 147-163). Ramonville Saint-Agne, Eres.

Kalampalikis, N., & S. Moscovici (2005). Une approche pragmatique de l'analyse ALCESTE. *Les Cahiers Internationaux de Psychologie Sociale, 66*, 15-24.

Kelle, U. (2000). Computer-assisted analysis: coding and indexing. In M. W. Bauer & G. Gaskell (Eds.), *Qualitative researching with text, image and sound. A practical handbook* (pp. 282-298). London: Sage,.

Kronberger, N., & Wagner, W. (2000). Key words in context: statistical analysis of text features. In M. W. Bauer, & G.Gaskell (Eds.), *Qualitative researching with text, image and sound. A practical handbook* (pp. 299-317). London: Sage.

Lahlou, S. (1992). *Si/alors : "bien manger" ? - Application d'une nouvelle méthode d'analyse des représentations sociales à un corpus constitué des associations libres de 2000 individus*. Paris : CRÉDOC, Cahiers de recherche, n°34. Avril 1992.

Lahlou, S. (1994). L'analyse lexicale. *Variances, 3*, 13-24.

Lahlou, S. (1995a). *Penser manger. Les représentations sociales de l'alimentation*. Paris, EHESS, Thèse de doctorat. [Online] http://hal.archives-ouvertes.fr/docs/00/16/72/57/PDF/THEDE100a.pdf.

Lahlou, S. (1995b). Vers une théorie de l'interprétation en analyse des données textuelles. In S. Bolasco, L. Lebart, & A. Salem (Eds.), *JADT 1995. 3rd International Conference on Statistical Analysis of Textual Data*. CISU, Roma, vol. I, pp. 221-228.

Lahlou, S. (1996a). La modélisation de représentations sociales à partir de l'analyse d'un corpus de définitions. In E. Martin (Ed.), *Informatique textuelle* (pp. 55-98). Institut National de la Langue Française, Paris: Didier Érudition. Collection Études de Sémantique Lexicale.

Lahlou, L. (1996b). A method to extract social representations from linguistic corpora. Japanese *Journal of Experimental Social Psychology, 36*, 278–291.

Lahlou, S. (1998). *Penser manger*. *Alimentation et Représentations sociales.* Paris: PUF.

Lahlou, S. (2003). L'exploration des représentations sociales à partir des dictionnaires. In J.-C. Abric (Eds.), *Méthodes d'étude des représentations sociales* (pp. 37-58). Ramonville Saint-Agne: Eres.

Lebart S., & Salem A. (1994). *Statistique textuelle*. Paris: Dunod.

Licata, L., & Klein, O. (2002). Does European citizenship breed xenophobia? European identification as a predictor of intolerance towards immigrants. *Journal of Community & Applied Social Psychology, 12*(5), 323-337.

Liu, B. (2010). Sentiment analysis and subjectivity. In N. Indurkhya, & F. J. Damerau (Eds.), *Handbook of natural language processing* (2nd Edition) (pp. 1-38).

Manetta, C., Urdapilleta, I., & Sales-Wuillemin, E. (2009). Étude des représentations en contexte: une méthodologie combinant l'analyse ALCESTE et la méthode des opérateurs de liaison. *Les cahiers internationaux de psychologie sociale, 84*(4), 81-105.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge: MIT Press.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge: Cambridge University Press.

Marková, I. (2003). *Dialogicality and social representations*. Cambridge: Cambridge University Press.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco, CA: W.H. Freeman.

McCarty, W. (2005). *Humanities Computing*. London: Palgrave.

McNamara, D. S. (2011). Computational methods to extract meaning from text and advance theories of human cognition. *Topics in Cognitive Science, 3*(1), 3-17.

Meunier, J.-G. (2002). Les théories constructivistes de la représentation sociale et la computationnalité. In C. Garnier , & W. Doise (Eds.), *Les représentations sociales: balisage du domaine* (pp. 227-240). Paris: Editions Nouvelles.

Meunier, J.-G. (2009). CARAT – computer-assisted reading and analysis of texts: The appropriation of a technology. *Digital Studies / Le champ numérique, 1 / 3*, octobre 2009, [Online] http://www.digitalstudies.org/ojs/index.php/digital_studies/article/view/161.

Meunier, J.-G., Forest, D., & Biskri, I. (2005). Classification and categorization in computer assisted reading and analysis of texts. In H. Cohen, & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (pp. 955-978). Amsterdam: Elsevier.

Michel, J. B., Shen, Y. K, Aiden, A. P., Veres, A., Gray, M. K., Team, B., Pickett, J. P., Hoiberg, D., Clancy, D., & Norvig, P. (2011). *Quantitative analysis of culture using millions of digitized books, Science, 331*, 6014, 176-182.

Mitkov, R. (Ed.) (2003). *The Oxford handbook of computational linguistics*. Oxford: Oxford University Press.

Moliner, P., Rateau, P., & Cohen-Scali, V. (2002). *Les représentations sociales: Pratique des études de terrain*. Rennes: PUR.

Moscovici S. (1988). Notes towards a description of social representations. *European Journal of Social Psychology, 18*, 211-250.

Moscovici, S. (1961). *La psychanalyse, son image et son public*. Paris: PUF.

Pêcheux, M. (1969). *Analyse automatique du discours*. Paris: Dunod.

Popping, R. (2000). *Computer-assisted text analysis*. SAGE.

Rajman, M., & Lebart, L. (1998). Similarités pour données textuelles. In S. Mellet (Ed.), *Actes des 4e Journées internationales d'Analyse statistique des Données Textuelles* (pp. 545-555). Nice: Université de Nice - Sophia Antipolis.

Rastier, F. (2011). *La mesure et le grain. Sémantique de corpus*. Paris: Honoré Champion.

Reinert, M. (1983). Une méthode de classification descendante hiérarchique: application à l'analyse lexicale par contexte. *Les cahiers de l'analyse des données, 8*(2), 187-198.

Reinert, M. (1986). Un logiciel d'analyse lexicale. *Les cahiers de l'analyse des données, 11*(4), 471-481.

Reinert, M. (1987). Classification descendante hiérarchique et analyse lexicale par contexte: application au corpus des poésies d'A. Rihbaud. *Bulletin de méthodologie sociologique, 13,* Janvier, 53-90.

Reinert, M. (1990). ALCESTE une méthodologie d'analyse des données textuelles et une application: Aurelia de Gerald de Nerval. *Bulletin de méthodologie sociologique, 26,* Mars, 24-54.

Reinert, M. (1993). Les mondes lexicaux et leur logique. *Langage et Société, 66,* 5-39.

Reinert, M. (2002). *ALCESTE, Manuel de référence*. Université de Saint-Quentin-en-Yvelines, CNRS.

Roth, C., & Cointet J. (2010). Social and semantic coevolution in knowledge networks. *Social Networks, 32*(1), 16-29.

Sahlgren, M. (2006). *The Word-Space Model*. Doctoral dissertation, Department of Linguistics, Stockholm University.

Sainte-Marie, M., Meunier, J.-G., Payette, N., & Chartier, J.-F. (2011). The concept of evolution in the origin of species: A computer-assisted analysis. *Special issue of LLC on DH2010, 26*(3), 329-334.

Salton, G. Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613–620.

Schütze, H. (1993). Word space. In S. Hanson , J. Cowan, & C. Lee Giles (Eds.), *Advances in Neural Information Processing Systems 5 (pp. 895-902).* San Mateo, CA: Morgan Kauffman,.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys. 34*(1), 1-47.

Stubbs, M. (2002). *Words and phrases. Corpus studies of lexical semantics*. Oxford: Blackwell.

Theodoridis, S., & Koutroubas, K. (2009). *Pattern recognition* (4th Edition). London: Academic Press.

Viaud, J. (2002). Multidimensional analysis of textual data using ALCESTE and the social representation of unemployment. *Revue européenne de psychologie appliquée/ European review of applied psychology, 52*(3/4), 201-212.

Viaud, J., Uribe Patiño, F. J., & Acosta Ávila, M.T. (2007). Représentations et lieux communs de la mondialisation. *Bulletin de psychologie, 1* (487), 21-33.

Voelklein, C., & Howarth, C. (2005). A review of controversies about social representations theory: a British debate. *Culture & Psychology, 11*, 431–454.

Wagner, W., Duveen, G., Farr, R., Jovchelovitch, S., Lorenzi-Cioldi, F., Marková, I., & Rose, D. (1999). Theory and method of social representations. *Asian Journal of Social Psychology, 2*, 95–125.

Weiss, S. M., Indurkhya, N., Zhang, T., & Damereau, F. J. (2005). *Text mining. Predictive methods for analyzing unstructured information*. New York: Springer-Verlag.

Widdows, D. (2005). *Geometry and meaning*. Stanford (USA): CSLI Publications.

Xu, R., & Wunsch II, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks, 16*(3), 645-678.

Yu, B. (2008). An evaluation of text classification methods for literary study. *Literary and Linguistic Computing, 23*(3), 327-343.

JEAN-FRANÇOIS CHARTIER is a Ph.D. candidate in Cognitive and Computer Sciences at the Université du Québec à Montréal (UQÀM) and holds a master degree in Sociology. He is currently a researcher at the LANCI laboratory (Laboratoire d'analyse cognitive de l'information). He is also a Doctoral Fellow of the Social Sciences and Humanities Research Council (SSHRC) and of the Fonds Québécois de Recherche sur la Société et la Culture (FQRSC). His research interests are the cognitive models in social sciences and the computational methods in computer sciences. E-mail: chartier.jf@gmail.com.

Since 1970, JEAN-GUY MEUNIER's research has been in Computer Assisted Reading and Analysis of Text (CARAT) in the Humanities. He is full professor at the Université du Québec à Montréal (UQÀM), director of the LANCI laboratory (Laboratoire d'analyse cognitive de l'information) and member of the International Academy of Philosophy of Science. His actual research is in Computer Assisted Conceptual Text Analysis. Email: meunier.jean-guy@uqam.ca.