Classification and Categorization in Computer Assisted Reading and Analysis of Texts

Jean Guy Meunier, Université du Québec à Montréal
Dominic Forest, Université du Québec à Montréal
Ismail Biskri, Université du Québec à Trois-Rivières

published

Lefebvre, C. Cohen, H. (2005) *Handbook on Categorization*, Elsivier 2006

1. Introduction

1. 1. CARAT: general presentation

In the early 1960, computer appeared as a revolutionary tool for computing mathematical symbols. Even the numerical ones!

> *"Scientists in AI saw computers as machines that manipulated symbols. The great thing was, they said, that every thing could be encoded into symbols, even numbers. "* (Newell, 1983, p. 196)

Surprisingly though, the real impact of the computer technology was not processing numerical symbols but a whole class of other types of symbols, such as the natural language ones.

Indeed the computer has been then widely applied to text processing so much that today most of the processing done by computer is applied to text (Internet, e-mails, documents, etc). Since then, research in cognitive sciences and information technologies (IT) has had an important impact on the reading and analysis of texts. And the particular field of the humanities has greatly gained from this.

Since now practically fifty years, technologies for computer assisted reading and analysis of text (CARAT) have penetrated the various humanities and social sciences disciplines (Hockey, 2001). One finds them in philosophy (McKinnon, 1968, 1973, 1979; Meunier, 1997; Lusignan, 1985, Floridi, 2002), in psychology and sociology (Barry, 1998; Alexa et Zuell, 1999a, 1999b; Glaser et Strauss, 1967; Jenny, 1997), in literature (Brunet, 1986; Kastberg Sjöblom and Brunet, 2000; Hockey, 2001; Fortier, 2002; Bradley and Rockwell, 1992; Sinclair, 2002; Rastier *et al.,* 1995; Bernard, 1999), in textual semiotics (Ryan, 1999; Rastier, 2001), in political sciences (Fielding and Lee, 1991; Lebart and Salem, 1994; Mayaffre, 2000), in history (Greenstein, 2002), etc.

From the encounter between these various disciplines and computer sciences and technologies has emerged an original research field called Computer Assisted Reading and Analysis of Text (CARAT) (Bernard, 1999; Hockey, 2001; Meunier, 1997; Popping, 2000). This research field is different from the artificial intelligence (AI) approach to discourse analysis (Hobbs, 1990) or automatic reading (Ram and Mooreman, 1999) where the objective is to have the computer simulate some type or other of "understanding" of a text in some specific application or process (inference, summary, knowledge extraction, question answering, e-mail routing, etc.). It is also different from information retrieval (Salton, 1989; Salton and McGill, 1983) or hypertext technologies (Rada, 1991) where the objective is to find documents from a particular query (or to navigate through documents). In CARAT, literary critics, philologists, content analysers, philosophers, theologians, psychologists, historians, and many other types of professional text readers (lawyers, journalists, etc.) require computer tools that assist them in their own and often personal reading and analysis expertise. And they cannot accept automatic text "interpretation" tools under any form whatsoever.

To assist the *reading* task, the computer technology offers more and more possibilities. Archives of electronic texts are now a common thing (*Oxford Text Archive*, *Brown Corpus*, *Perseus*, *Gallica*, *Frantext*, etc.). They are more and more critically edited, standardized (SMGL, HTML, XML, etc.) and can be explored with "intelligent" tools such as navigators, search engines, hypertexts, etc. (Condron *et al.,* 2001)

To assist the *analysis* process, the technology has also been very constructive. In fact, one could distinguish various generations of this technology. A first one (1950-1970) opened the era of the capture of the electronic text. A second generation (1970-1980) offered tools for the description and manipulation (extraction, concordance, statistics, lemmatization, etc.) of these electronic texts. A third one (1980-1995) started a standardized tagging (SGML, TEI initiative) and the linguistic processing (syntax, semantic, discursive and rhetorical, etc.) of the text. A forth one (since 1995) introduced sophisticated mathematical models (classifiers, neural nets, categorizers, etc.) on the text. There exist today a host of such tools: from the more hybrid to the very specialized, from the laboratory prototypes to the more industrial applications (*Concord*, *Word Cruncher*, *TACT*, *Cobuilt*, *Word Smith Tools*, *Lexa*, *Cosmas*, *Tropes*, *Intext*, *Alceste*, *Sphinx*, *Hyperpro*, *Atlas/TI*, *NU*DIST, etc*). CARAT has even received a buzz name in the commercial field applications: "Text Mining", for which there exist a multitude of commercial tools.

1. 2. Difficulties with the technology

Unfortunately, the traditional communities of the humanities and social sciences, except maybe for critical editions of texts, qualitative content analysis projects or historical archiving, do not recognize the full importance of this technology. One can invocate many reasons to explain these difficulties. A first one is the weakness of the ergonomics of these technologies that renders their learning and usage difficult for many users, particularly in the humanities and social sciences (Unsworth, 2000, p. 2). Except for a few systems, many technologies have often been developed by communities external to CARAT (linguistics, artificial intelligence, data processing, information retrieval, etc.) and for specific objectives. A second one relates to the limited sets of tools available for text analysis, where often a researchers is confined to text transcription, text encoding, lexical analysis (concordances, collocation), linguistic analysis (lemmatization, tokenzer, morphological taggers) or statistical analysis. Thirdly many commercial tools are often geared more to information retrieval than real assistance for reading and analysis of text. Finally, the technology is often a closed one. That is, it is proprietary, rigid and non-modular and, as often underlined, not really adapted to the dynamics of projects in CARAT.

We believe though that there are more profound theoretical reasons for these difficulties. There is a lack of understanding of the effective role played by this technology. The main stream community still entertains the wrong image that it is the computer that is doing some type of text interpretation. It has difficulty in seeing the computer as being just an assistant in these processes. There is also a misunderstanding of the linguistic aspect of what is a text, its reading and its analysis. Rastier (2001) readily reminds us how much a text as a whole is highly different from a sentence of a language, and must be approached with tools different from grammar and logic (Mann and Thomson, 1988). Finally, the classical technologies of CARAT, except for the ones of standards and protocols (Sinclair, 2002) have not really integrated the modelization tools of the software engineering languages as developed in the last 15 years (Braumbrough *et al.,* 1991; Lévesque, 1998) and which would allow modularity, reutilisation, exchanges, collaboration and even automatic code generation. So, if a renewal in the use of this technology is to happen then, one must have a more rigorous idea of what is reading and analyzing a "text" with the help of a computer. Therefore, a more theoretical and formal foundation of CARAT technology has to be explored.

1. 3. The nature of reading and analysing a text.

Indeed, the renewal has to better understand the real nature of what a text is. The term "text" is a very ambiguous one. In certain context, it may mean a paper object (vg: the Guttenberg Bible) in other ones, it may mean the abstract content (vg: the Bible as a text). Sometimes it is a linguistic object (vg: writing a text is different from writing a sentence) or a psychological object (vg: the understanding of a text).

A text is in fact is a complex semiotic object. Even if a text presents itself as a sequence of natural language symbols, it is more than a "linguistic" object in the sense of more recent linguistic theories. It is not a well form output of a formal grammar. Also, it is not identical to the discourse. Indeed the same text can be *read, pronounced* or *enunciated* in many different contexts (cf. the Bill of Right) and each time produce a different discourse effect.

A text contains different dimensions such as an argumentation, a narration, a description, a demonstration, a dialog, a theme, etc., all of which are not in themselves strictly grammatical phenomena, albeit they may present some strong regularities that belong to their genre (Rastier, 2001), their rhetorical structure (Mann and Thomson, 1988) or their logical structure (Hobbs, 1990). In fact, if a text was a language in the strict grammatical (algorithmic) sense, then, in learning to produce and recognize sentence and structures of a particular language, one would have also learned to produce and recognize textual structures. For instance if one would know how to produce and recognize Danish sentences, one would also know how to produce and recognize Danish legends!

This means that reading and analysing a text is not strictly an algorithmic process (Eco, 1965, 1992). We would add that it is probably not even a computational one (Meunier, 1997). Reading or analysing a text is more akin to complex procedures of heuristics and pattern recognition strategies than to pure grammatical and algorithmic procedures. This access to the textual content is in fact a complex interpretative process rooted in various dimensions such as linguistic and mundane semantic structure, inference, pragmatic memory, culture, social interaction, knowledge repositories, etc. Hence, it is in the nature of a text not to reveal itself in totality. Every text is polyvalent. It varies its meaning on reception (Jaus, 1978) and reader (Eco, 1965) and does not necessarily reveal itself in a first reading not even in a first analysis (cf. the Bible or the Koran). The hermeneutic tradition has been repeating this for at least the last century (Schleiermacher, (1987); Heidegger, 1962; Gadamer, 1996; Habermas, 1984, 1987). And contemporary theories of reading indulge in this thesis (Iser, 1985; Fish, 1980; Barthes 1970 Gervais 1993). If in reading and analysing text in the preceding perspectives we wish assist by computer this process, we must then try to "follow" as close as possible the strategies used by the human readers and analysers. And even more so if they are experts such as the ones encounter in the humanities and social sciences.

There exist many strategies for a good and deep reading of a text. Finding them is the aim of many psychological, philological and computational theories of reading. Among these many strategies, two important ones are often used: a classification process and categorisation process. Indeed, a reader or an analyst does not just follow the natural course of a text, line by line, chapter by chapter. He aims at finding some underlying, inexplicit pattern or structure of some sort that are called in various disciplines*, a theme, a concepts, an argument,,* etc. Text analysis is in one sense a second order reading that classifies or categories textual objects.

In the classification process, the reader, through various clustering, regrouping or reorganising techniques aims at discovering classes of relevant textual objects on the ground of some type or other of similarity criterion. For instance, in literature, one may want to regroup all the relevant passages of Shakespeare on a particular theme (JEALOUSY, for example). The various text passages may or may not express the theme explicitly. It may be just emerging at the beginning of the corpus, and only well exposed and expressed at the end of it. It is here that a classification procedure becomes a highly relevant tool for analysis purposes. It is, in fact, an implicit practice in many lexical, semantic, narrative or stylistic analysis.

In the *categorisation* process, the reader takes the analysis farther. He uses the classification procedures and results, but imposes on them some structural organisation. For instance, the expert reader of Shakespeare may have found classes of text objects that all express some similar content. On these classes the reader may which then to categorize them explicitly in terms of a category such as: JEALOUSY, LOVE, HATE, PASSION or DESTINY. He then may wish to offer some structuring of these themes, in order to build a relation between them (such structuring relation could be, for example, (DESTINY (PASSION (LOVE), (HATE) (JEALOUSY)))).

In realizing these classification and categorisation procedures an expert reader will call upon many heuristics that are part of his expertise, cultural background, erudition, objectives, etc. And each reader and expert has its own patrimony of techniques, procedures, best practices and they underlie his personal and subjective interpretation. In fact, they belong to his signature as an expert. And it is obvious that this highly complex interpretative process cannot be translated in easy and clear cut algorithms. And it cannot be easily applied and used by a computer so that the reading and analysis becomes automatic! But it does not follow that the computer cannot be useful in assisting these processes. Indeed, it can play a productive role, but only if it faithful to this interpretative process and well situated in the cognitive activity of the reading and analysis of a text. In this perspective, CARAT appears relevant. It has in fact become an important tool in various social sciences and humanities disciplines that rely on rich reading and analysis of text. In this context, CARAT offers a set of algorithmic tools assisting the construction of interpretative paths in reading and analysing texts. And scholars in these fields accept this computer technology if it really is an assistant in their own intellectual inquiry of a text and if does not try to substitute itself to them.

2. Definitions of classification and categorization for CARAT.

If CARAT is to be an assistant in the process of reading and analysis of text, one must better define the concepts of classification and categorization. Is classification categorization and vice versa? Unfortunately, such a definition is not easy to give for these two concepts which received a host of definitions in various scientific discourses. Let us distinguish some of the most important differences between these two concepts because, in the scientific literature, the concepts of *class* and *category* are often amalgamated. A category is a class and vice versa. But whatever the names used, one must distinguish the underlying concepts. A first concept sees a class or a category defined as a the result of some formal operation : the partition of a set of objects. This operation must offer the conditions under which a set of such objects are related among themselves in order to become a class. This is usually the definition used by mathematicians and logicians.

The second concept sees a class or a category as a symbolic *label* that is the name given to the class produced by whatever classifying process. This is the concept used in linguistics, information science and often in computer science. Here a category is often a labelled class.

The third concept defines a class or a category as a cognitive state – pertains to the agent that cognitively performs the classification and the categorization processes. It is the center of interest of many psychological, anthropological and computer sciences enquiries.

The forth concept understands a class or a category as a morphism that is as a one type or other (vg hierarchy) of structure that classes may entertain among themselves. It is the interest of structural (geometrical) and mathematical modelling of classes.

These conceptual distinctions lead to various types of scientific enquiries and albeit they are often related among themselves, they must not be confused. The problem of differentiating the varieties of *roses* is a different problem that the one of naming them. And this is not identical to the problem of situating *roses* in the flower species (taxonomy). This is also different from the psychological cognitive process by which non biologist or an experienced gardener can identify that flower. Even though we could always use the terms *classes* and *categories*, the important thing is to see in each case is if they are applied to the same or different operations. And hence, pertains to identical or different enquiries. And the solution to one problem can not be a solution to the other

3. Text classification and categorization

Given these various distinctions between class and category, let us now see what is the case for CARAT. Here, we can distinguish two different research programs that exist effectively and have received attention in the past decade.

3. 1. Text classification

A first one – text classification – pertains mainly to the first concept of classification that is classification as clusters, sets, etc. on objects that are here textual entities (sentences, documents, books, Internet sites, etc.).

> "*Text classification is the automated grouping of textual or partially textual entities.* " (Lewis and Gale, 1994)

In more technical terms, text classification is defined as an operation that is applied to textual entities on which equivalent classes are built. Classification in hence a process by which textual information is clustered together according to some criteria. A query on the Internet is a simple and basic example of this. A query recalls all the textual entities (sites) that contain the particular words of the query.

But the more interesting classification techniques are procedures were the criterion is slightly more complex that the one of the query classification. For instance, in a large text corpus (the Shakespeare corpus, for example) one may want to regroup all the relevant textual entities talking about different themes.

The research objectives in text classification are mainly to find techniques that build these classes either in a top down manner (by a list of predefined conditions) or in a bottom up manner (by discovering inductively the conditions under which the textual entities are similar). The classification systems deliver sets of textual entities. They never produce the labels that could either name or label the classes themselves.

3. 2. Text categorization

The second research program is the categorization one. It pertains more to the second definition of category, albeit, implicitly, it also relates to the third and forth definitions. As Sebastiani (2002, p. 1) clearly mentioned it, text categorisation is "*the activity of labeling natural language texts with thematic categories from a pre-defined set*. " In other words, text categorisation is "*the process of assigning entries from a predefined vocabulary.* " (Ruiz and Srinivasan, 1998)

Here, the aim is to add to the class built or found, labels or symbolic tags. Usually, as defined above, the set of tags is predefined and belongs to some system of classification. This is a top down approach. But in certain cases, the labels themselves are not predefined and must be discovered by some means or other. This is more typical of ontologies and taxonomies applications.

Hence, text categorisation is more than text classification. Categorization relies on a classification process, but adds to the classes some labels, that is, it attributes to them some type or other of predefined tags. These tags describe some aspect of the "semantic" content of a class of textual entities. It is in the choice of the labels that the cognitive and structural dimensions of categorisation are called upon.

For instance all similar segments of text classed together could be tagged by category terms such as LOVE, HATE, JEALOUSY, etc. They could also be tagged as PASSION, EMOTION or ACTION. And so doing one sees that there is an implicit cognitive and structural dimension at work. There is a difference in tagging a set of textual entities as HATE or as PASSION or as ACTION.

3. 3. Computer text classification and categorization

Both of these research strategies can be realized manually. In fact, it has been done for centuries in such manner. And it is still the most popular procedure today. Librarians still classify and categorize books manually. Most of the thematic analysis in scholarly literature or philosophy is done in such manner. Even qualitative content analysis of interviews, testimonies or historical documents follows the traditional procedures of reading and analysis.

But more and more, these classification and categorization techniques are being explored in a computational horizon. Indeed some research programs aim at finding and evaluating various algorithms that can realize in a computational fashion both the classification and the categorization of textual entities.

This research program has become quite a stable paradigm in itself and it is quite lively in the fields of information and knowledge management technologies (often called "text mining technologies") for which exist a rich and relevant state of the art literature (Sebastiani, 2002; Lewis and Gale, 1994; Kodratoff 1999).

Our own research program is to see if these algorithms can be successfully applied in the context of the reading and analysis or texts. In other words, we are trying to find out if they are useful in assisting reading and analysis practices in the humanities and social sciences. We intend here to expose in a heuristic manner how these approaches of automatic classification and categorisation can be heuristically applied to CARAT. It is not obvious how and for what purpose these approaches can be successfully applied to CARAT. In this field of application, these techniques cannot be an end in their self. No text reader and analyser would find it very useful to submit his corpus to a classification and a categorisation process just for the sake of it. But we believe that are useful if they are linked to the dynamic process of text interpretation.

In the field of the humanities and social sciences, these techniques have been known, mainly in the European academic circles as statistical discourse analysis. But they have not really taken roots yet in the American tradition of computer text analysis.

4. Methodology for text classifying and categorizing

Text classification and categorization rely on a similar methodology, at least in their first steps but they distinguished themselves in the later ones. Here, we present this technique in seven main steps.

4. 1. Steps 1, 2 and 3: From a text to a matrix

4. 1. 1. Step 1: Identification of units of information and domains of information

The first step consists in defining what will be for the classifying and categorizing algorithms the effective textual entities or textual input. We distinguish two different types of textual entities. Each of them will be required in further up the processes. The first type of textual entities is what we call the units of information (UNIFS). The second type is the domains of information (DOMIFS).

A) Units of information (UNIFS)

A first type of basic unit of information in a text is linguistically grounded (Kucera and Francis, 1967) and usually defined in terms of "words" or some of its linguistic variants. For instance we can find:

- *Basic word* (any sequence of letters separate by a blank)

  "John loves Mary" John, loves, Mary
- *Lemmas or stems* (words reduced to their basic forms)

Loved → Love (or Lov)

Intermediately, Intermediate, Intermediates → Intermediate

- *Morphems* (words that have been distinguished according to their syntactical categories)

  Cry (Name) vs. *Cry* (verb)

One very useful type of units of information is composed of complex words (phrases) grammatically grounded. These phrases can be words (such as *Jet set, Table set, Chess set, Table napkin, Chest pain*) or verbs (such as *Kick the bucket, step on,* etc.). Others will be purely based on sequences of words or collocations (word grams) (Choueka, 1988). For instance in the sentence "John loves Mary today", one could define as units of information "John", "loves", "loves Mary" or "Mary today".

Another type of units of information can be sensitive to syntactic or semantics informations. For instance, the type word "house" could be distinguish as "house$_{\text{(name)}}$"or "house$_{\text{(verb)}}$".

A second basic type of units of information has recently been used. They are called *n-grams* and are composed of *n* symbolic characters (Damashek, 1989; Cavnar and Trenkle, 1994). For instance, in the sentence "John loves Mary" a three gram will take sequences of 3 letters ("Joh", "ohn", "hn ", etc.). These sequences of *n-grams* can also be constrained under some probabilities (such as Bayesian approach (Church, Gale, Hanks et Hindle,1989). That is, not all sequences of letters are retained; only the one with a certain probabilistic characteristic.

As one can see, all these units of information require to be automatically identified by some algorithm. These are what tokenizers, parser (syntactical, morphological or semantical) do. Naturally, each algorithm will encounter many complex problems. Be it words or *n-grams*, the choice of the right units of information depends on the aim of the operation.

Each choice of units of information has it own advantages and draw backs. If the units of information are all basic words with all their flexional variety, the list of units of information is more respectful of the language but the list easily build up into tenths of thousand words. Often, the choice of one or other form of units of information can depend on the size of the corpus to be process. The internet is not the same as a domain limited corpus.

B) Domains of information (DOMIFS)

The second type of information to be identified in this first step is the domains of information (DOMIFS) (the fragments of texts on which a classifier will be applied). These fragments could be phrases, sentences, paragraphs, pages, chapters, documents, or fixed sequences of words.

To identify concretely in a corpus these domains of information some type or other of algorithm is needed. In continuous texts this will be realised by a segmenting program that defines operationally on the corpus what is to be retained. In information retrieval these domains of information will often be different documents or web pages. In CARAT, the domains of information are usually paragraph.

What is the best fragmentation possible and for what purpose (a page, a document, a title, a sentence, a line, etc.)? The answer lies in the aim of the classifying procedure. And here only repeated experiments will determine which length or type of fragments fits best for a particular type of analysis. In our own experiments, we have privileged medium size paragraphs (50 to 1000 words). This seems the average length into which an author has time to introduce a theme and develop it.

Another problem is the algorithm or the set of rules necessary for identifying the chosen fragments. The information retrieval (IR) tradition has the easy way in taking the document as a basic fragment. Even more easily is the "abstract". Unfortunately, it is not so for ordinary plain text. Indeed, in the fields of humanities, classical texts are becoming standardized through XML tags. This helps delimiting the text but render the processing for classification and categorization more cumbersome.

Also in these domains, it not as clear as one thinks as to what is a line, a paragraph, a page, more so a sentence or a phrase. Each choice brings its share of problems and computer processing costs.

4. 1. 2. Step 2: Cleaning and filtering

The next step consists in cleaning the corpus and filtering some of its information units (units of information or domains of information). The aim of this step is to produce a set of most relevant information units.

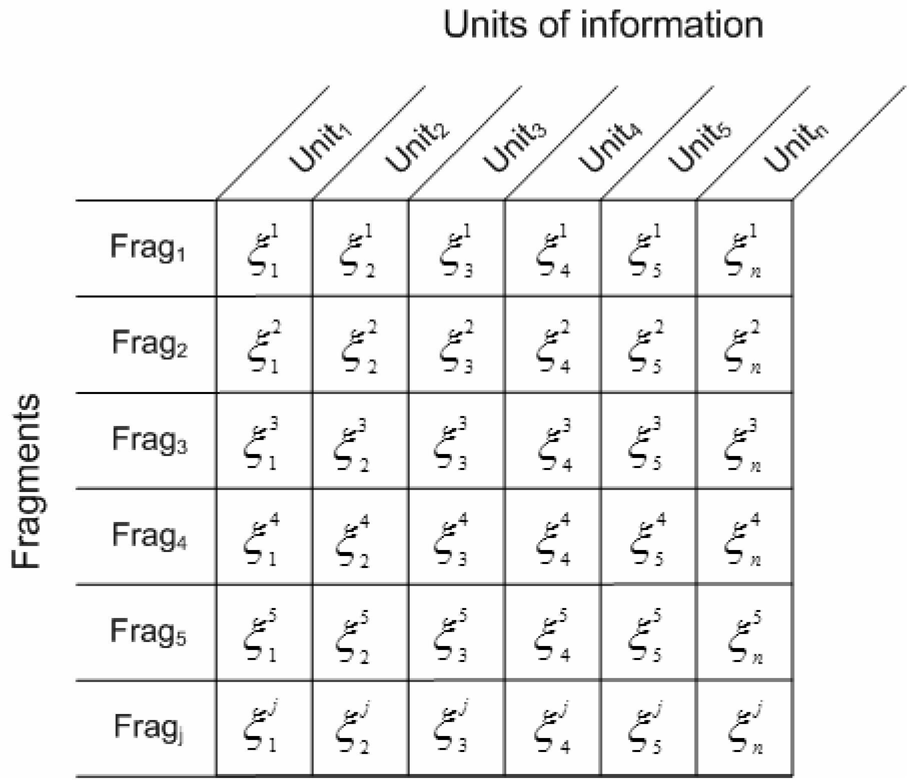
For instance, in a classical edition of a literary book, one may not wish to keep as basic units of information the title page, the word "chapter", etc., that is all the meta informations that a good edition will integrate.

Because of the mathematical structures of certain classifying techniques, it may not be relevant to keep very low frequency words. It is also usual to clean the text from its functional words (stop words) (articles, preposition, punctuation marks, etc.).

Even though this step is not theoretically very complex, it is necessary, at least for the classification and categorization processes, because they reduce considerably the numerical dimensions to be process. But it will often be a big surprise for the novice in CARAT to find out how long and complex this cleaning and filtering are. It will call upon many decisions, not all of them evident. Moreover, in some cases, du to the idiosyncrasies of some of these decisions, the operations may have to be done manually.

4. 1. 3. Step 3: The matrix

In the third step, the textual units and fragments are transformed into a matrix using the Vector Space Model (Salton and McGills, 1983; Manning and Schütze, 1999) (Figure 1). Here, each segment with its information units is seen as a vector where the informational content is translated as a property of this vector. The values given to each entry of the matrix represent the attribution of this property to the vector. The type of values given depends on the classifying model chosen (e.g. presence, absence, fuzziness, weighting, etc.). All functional and subjectively non-relevant words can also be eliminated from the text. This can be done either manually (according to a goal set) or automatically (using rules or a dictionary).



Units of information

Fragments

	$Unit_1$	$Unit_2$	$Unit_3$	$Unit_4$	$Unit_5$	$Unit_n$
$Frag_1$	ξ_1^1	ξ_2^1	ξ_3^1	ξ_4^1	ξ_5^1	ξ_n^1
$Frag_2$	ξ_1^2	ξ_2^2	ξ_3^2	ξ_4^2	ξ_5^2	ξ_n^2
$Frag_3$	ξ_1^3	ξ_2^3	ξ_3^3	ξ_4^3	ξ_5^3	ξ_n^3
$Frag_4$	ξ_1^4	ξ_2^4	ξ_3^4	ξ_4^4	ξ_5^4	ξ_n^4
$Frag_5$	ξ_1^5	ξ_2^5	ξ_3^5	ξ_4^5	ξ_5^5	ξ_n^5
$Frag_j$	ξ_1^j	ξ_2^j	ξ_3^j	ξ_4^j	ξ_5^j	ξ_n^j

Figure 1. The matrix *fragments of text – units of information.*

The matrix can be a direct count of the presence of the units of information in each segment. But as the literature has shown, they are many variants possible.

Thus the whole text is transformed into a vectorial space. This allows the use of all the mathematical tools that can be applied on this space. This is the force of the classification and categorization approach. The text is not parsed linguistically but mathematically. From here on, classification and categorisation techniques are different in their methodology albeit slightly related.

4. 2. Steps 4 and 5

4. 2. 1. The classification process

In the classification procedure, the forth step consists in applying some mathematical or logical (rule based) "regrouping techniques" to the set of vectors (the matrix). These techniques are numerous. In the literature of mathematical text classifiers, many types of classifiers have been proposed and explored. All of which have their parameters; hence, fecundity and limits. One successful implementation of these types of models has been, in the information retrieval community, the SMART system (Salton, 1989). Among the most classical ones, are found the statistically oriented techniques (clusterizers, correlators, factorial analysis (Reinert, 1994), the principal component analysis (Benzecri, 1973), the K-Means (Hart, 1998), the Neural Networks classifiers (Grossberg *et al.,* 1991; Kohonen, 1982, 1997), the genetic algorithms (Holland, 1975), the Markovian fields classifiers (Bouchaffra and Meunier, 1995). Recent important variations of these models are to be found in Latent Semantic approaches (Deerwester *et al*. 1990), support vector machine (Cristianini and J. Shawe-Taylor, 2000). Also, some more probabilistic techniques have been tested, such as Bayesian classification or machine learning techniques (Sebastiani, 2002; Kodratoff, 1999). Finally, many recent research projects are exploring the potential of hybrid techniques. A good example of such techniques is found in Nauck (1999). Nauck developed as neuro-fuzzy classifier which combines a traditional neural net with fuzzy logic concepts and techniques.

These techniques group together textual segments (domains of information) under some criterion or other of similarity. At the end of the process of classification, the system offer to the reader classes of similar segments based on similarity criterions. All of these mathematical techniques have been applied to textual information processing and discourse analysis[1]. Apart from some limits, they have given quite positive results and they compare very positively with the more linguistically oriented techniques. Their great advantage is the time processing economy. They have also shown to be essential in processing large textual corpora.

[1] Each two years, an international conference specially adapted to social sciences and the humanities using statistical and mathematical techniques applied to computer text processing is held (*Journées internationales d'Analyse statistique des Données Textuelles (JADT)*).

Normally the classification process is the last step. But in CARAT, one must go farther in the explorations of the classes found. So in a fifth step, the classes produced are analysed and some of their properties are extracted depending of the objective pursued. The main objective of this fifth step is to find in the classes of segment some way or other to filter them out, to summarize them and present them in some transparent way. In this process, one could focus, for example, on the semantic information present in a particular class of segment (the main topic, the semantic field, etc.).

This is realized through logical and mathematical operations (vg. statistics, set theoretical, decision or semantic rules, etc.) applied either to domains of information of the classes or to their lexicon. The means by which the reading and analysis of the content of the classes is discovered is not a well defined procedure. It is an ongoing and important research problem. It is here that the interpretative analysis is really at work. It is from here that the true content analysis of the text really emerges.

4. 2. 2. The categorization process

In the categorization procedure, the forth step consists in defining from a set of predefined categories the labels that will be used for "tagging" the fragments of text (each vector of the matrix). This set of tags is taken as a working hypothesis for the expert reader and it can originate from various sources (thesaurus, classification systems, content organisation, dictionaries, taxonomies, etc.). The system is then trained by simulation: the computer system follows an expert reader that manually tags a sample set of segments. From this tagging, the categorizing algorithm then "learns" what "counts" as exemplars of a particular tag. There exist a variety of algorithms to realize this learning of categorization. Some are purely mathematical (perceptron, Rocchio's algorithm, etc.); some are a mix of rules govern decision trees (cf. the Ripper algorithm); some are typically machine learning strategies (Michalski, 1983) grounded on various external tools such as Word Net.

In the fifth step, once the learning is consolidated, the results are then projected on the whole text. By itself, the system then tags the rest of the segments of the text into each one of the categories. This is realized through the matrix built in the second step. The categorization techniques are then applied to the matrix. That is to say, because the system "knows" to which category a sample vector belongs, it finds all segments that are similar to the ones it has learned and automatically projects the label on them.

4. 3. Step 6: Navigation

Once the corpus has been correctly classified or categorized, the system has to offer mapping techniques (Barry, 1998) that present in a convivial manner the organisation of the classes of textual entities, their lexicon or list of categories (Spence and Press, 2000; Fayyad, Grinstein and Wierse, 2001).

Albeit theoretically basic, these techniques offer a great assistance in the interpretative activity of reading and analysis.

4. 4. Step 7: Evaluation

Finally in the sixth step (for both classification and categorisation), evaluation measures (precision, recall, accuracy, etc.) are applied to the results. These evaluation techniques, implemented in algorithms, have been (and are still applied to) many information processing domains. They are used in information retrieval, knowledge engineering, knowledge extraction, natural language processing, ontology building, data and text mining, etc. For instance, in information retrieval techniques, relevance feedback, precision and recall have become the norm. For some classifiers, comparative strategies have also been applied such as the Van Risjsbergen's measures that compare the elements of classes.

We believe that in CARAT this type of methodology is difficult to apply, for, albeit all appearances, reading and analysis of text is not an objective process but is highly subjective. The problem is related to the complexity of the interpretative activities of reading and analysis. As we have already mentionned above, because of the great varieties of text interpretations, it is not easy to find standards that could serve as parameters for these evaluations.

For instance, what set of hyperlink could be objective as to be taken as a benchmark? Personal experiences show that one navigates through web sites often in a haphazard fashion. This is not often a result of the poverty of the technology but the effect of the discovery process a reader carries with him. For this, we think that a more suited evaluation process can be given and be more faithful to the idiosyncratic strategy a particular user follow, in his reading and analysis of text.

Often, the benchmarks used are the classical interpretations given to some important textual corpora. In the following section, we shall illustrate, from our own researches, some heuristic application in CARAT.

5. Applications in CARAT

We will now illustrate how the classification and categorization techniques presented above can be heuristically applied in the field of CARAT. We have chosen a subset of cases taken out from our own past experiences. They show, we hope, how these computer techniques assist the process of reading and analysis of texts.

5. 1. Thematic analysis

Our first application for CARAT illustrates how the classification approaches can be used in a thematic analysis of philosophical texts. Indeed, one of the main basic philosophical interpretation tasks is conceptual analysis as developed under a certain theme. For instance, a philosopher may want to explore the concept of "TRUTH" in Descartes'writings, and discover the various semantic fields in which it operates. Computer wise, such a thematic analysis task is more complex than information retrieval. The reader does not know before the analysis what will be the span of the lexicon in which this theme is expressed in the corpus. So the classification approach can be a very heuristic tool to support the discovery and exploration of a particular theme.

In this experiment, we have applied a classification approaches to a French philosophical text (Descartes'*Discours de la méthode*). The original untouched text (except for basic printing noises) contained 21 436 words. The text was submitted to the various steps of the preceding classification methodology. At the fifth step, the process produced classes of segments according to certain similarity criterion. In this experiment, we have chosen the ART1 neural network classifier. From the classes obtained, the lexicon of the segments was extracted. The figure below shows the lexicon for each thematic class.

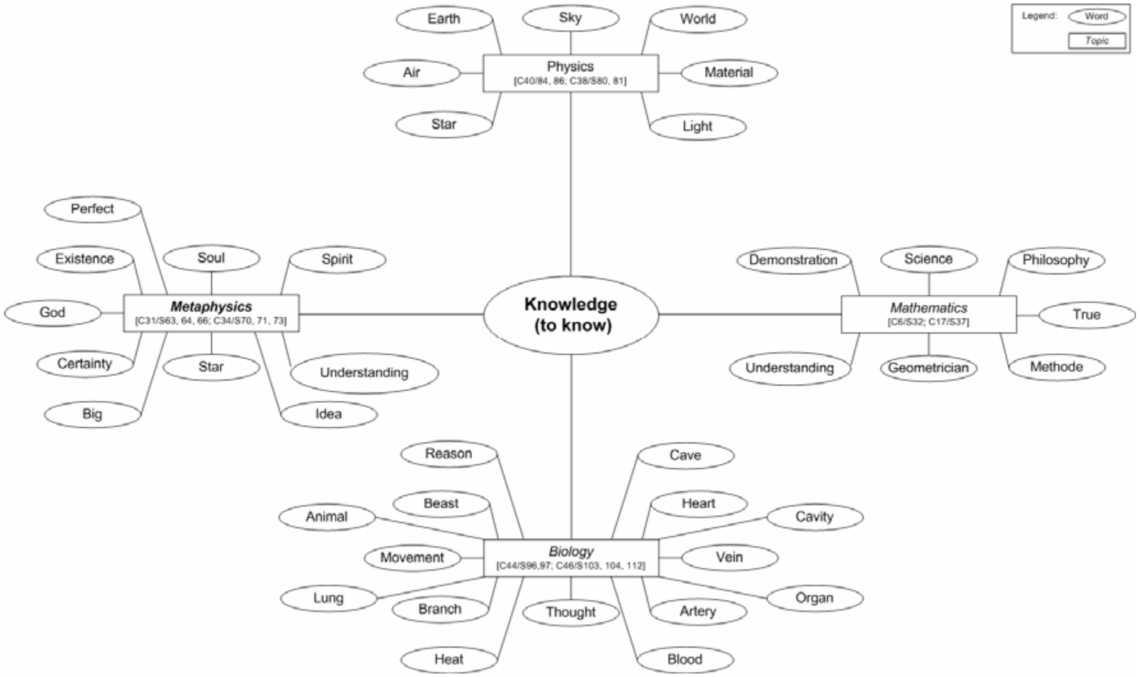


Figure 2. Thematic

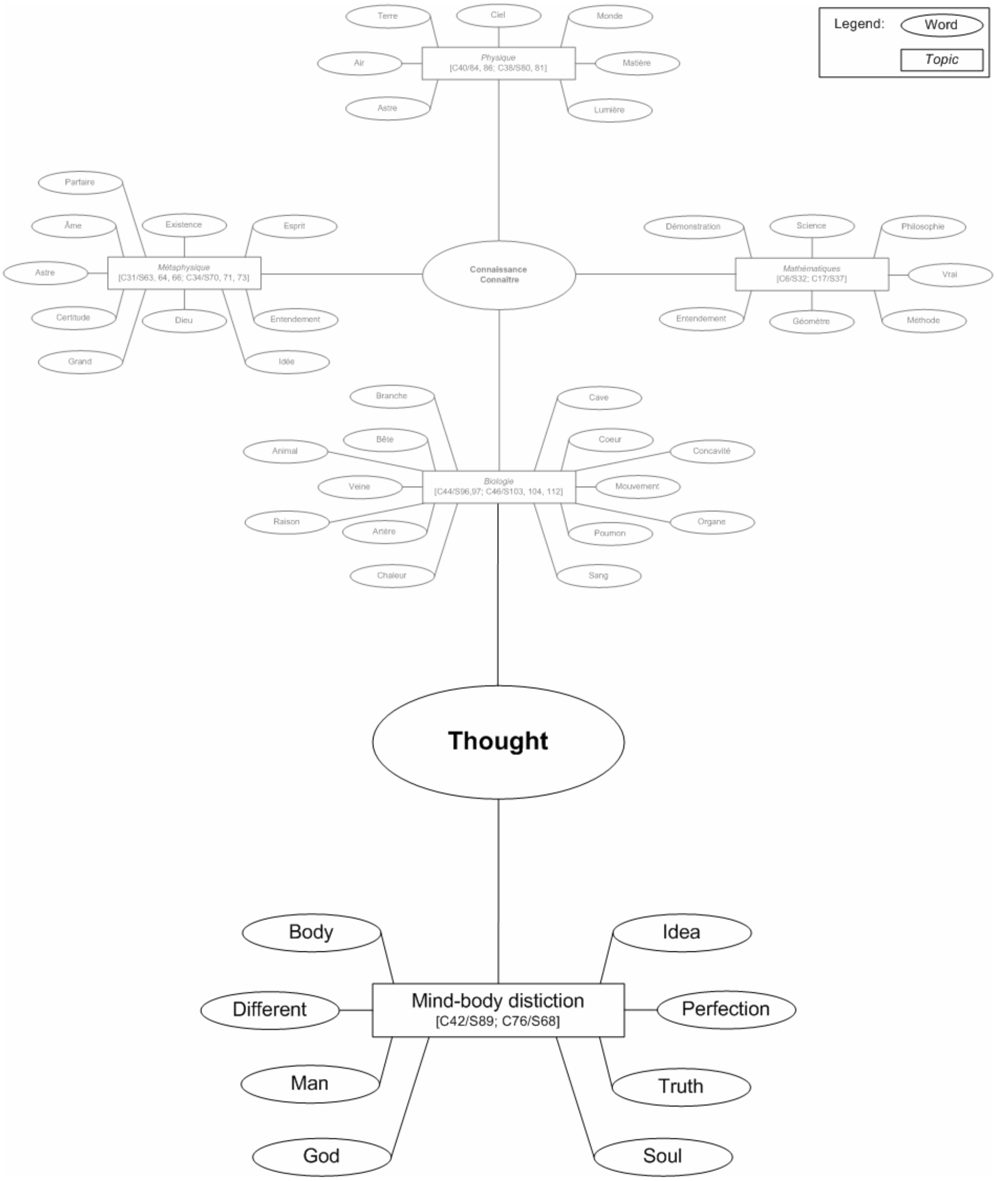
exploration

The analysis starts by choosing a word that is present in many classes. This becomes the leading “thematic word”. In the following illustration, we have chosen the word (or concept) “connaissance” (“knowledge”) which appears in many classes.

The figure 2 show that, in fact, the word “connaissance” is used in various semantic fields. For instance, in one of the class, the vocabulary surrounding the term “connaissance” relate to the metaphysical question in Descartes’writings. In another class, in operates more in a mathematical environment. It is here that the expert reader introduces his interpretative abilities. The computer has only classified the segments of text under certain similarity conditions that are seen as expressing dimensions of a semantic context.

With this preliminary classification function, the reader can then “direct” his analysis according to the chosen theme. And from then on, he can explore other different themes found in Descartes’philosophy. For instance in the following example (Figure 3), he may switch to another concept inside a particular class (e. g. “pensée” (“thought”)) and start a new thematic path. This strategy is then applied to the entire text. Other results (Forest, 2002; Forest and Meunier, 2000) have shown that this strategy could be applied to many other concepts.

Figure 3.



Other path in thematic exploration.

5. 2. Categorical exploration of philosophical texts

Another application of the classification categorisation techniques in the field the humanities consists in finding in a corpus the various segments of a text that are representatives a certain conceptual categories. For instance, one could be interested in finding the various segments of text that are relevant for the thematic category "TRUTH".

In this particular application, the expert reader has the system learn the passages of a text that he believes correspond to a conceptual category that he intuitively thinks underlies a particular passage. This is done on samples segments of the text. Afterwards, the learning is projected on the whole corpus. This procedure is typical of the learning phase of classical categorisation techniques.

This strategy is taken because in many highly technically and abstract texts, one does not and cannot have before hand how an author will explore a particular theme. And the analyser would like to explore if some of his intuitions about the theme are really at work in the text. For instance what theme are the most important ones in the corpus? How is a particular theme realized in the corpus? Are there other sub-theme related to a chosen one? etc. Because of the complexity of the information processing involved the computer becomes a heuristic and assisting tool in exploring certain intuitive categorical theme on the text.

We have explored this strategy on Bertrand Russell's text *The Problems of philosophy* (De Pasquale and Meunier, 2003). The application consisted in having the system discover, through leaning, passages corresponding to specific categories (by opposition to terms) such as "KNOWLEDGE", "ETHICS", "SPIRIT", even though the passages did not contain these words.

After the learning process (realized on samples), the system, through a simple perceptron classifier, manages to find many sentences expressing the chosen categorical themes. These various segments are then presented to the expert for analysis and evaluation. For example, the following segment was found to be relevant for the category "KNOWLEDGE": "*In this respect our theory of belief must differ from our theory of acquaintance, since in the case of acquaintance it was not necessary to take account of any opposite. (2) It seems fairly evident that if there were no beliefs there could be…*" But the system also rejected the following segment: "*Some relations demand three terms, some four, and so on. Take, for instance, the relation 'between'. So long as only two terms come in, the relation 'between'is impossible: three terms are the smallest number that renders it possible. York is between London…*"

The matching of the sentence and the categories relies not only on the presence of certain words but on the simultaneous co-presence of certain related terms such as "acquaintance", "knowledge", "truth", "reason", etc.

This type of very simple dynamic categorical learning may not be suitable for information retrieval and indexation but is surely a rich assistance in the heuristic exploration of a thematic category in a humanity text.

5. 3. Content analysis

In social sciences, an important computer applications related to CARAT is qualitative content analysis of discourse. In this case, the computer mainly assists the analyser in ascribing a set of predefined categories to each relevant token expression of the text. This is for instance what is realized through the content analysis programs such as the Nu*dist or Atlas. But this procedure is time and energy consuming mainly when the corpus is large. In CARAT, the procedure is reversed. The computer is used as a heuristic tool for discovering the possible categories themselves and the links between them.

In our research, this procedure has been explored on an anthropological corpus (composed of verbatim of interviews of an Indian Canadian community): The Innus of Quebec.

Two types of analysis were developed. A first one was a classification analysis. It consisted in extracting the classes of segments specific to a particular chosen theme. This procedure is similar the precedent one and it is hence possible to discover unseen semantic relations between the expression of the segments.

The second type is an analysis by “attractor" categories in which one discovers the expressions around which many other tend to polarize, that is, to form an attractor theme.

Let us now briefly illustrate this methodology. The fist consists in choosing some particular classes of segments that seem to express some particular theme. This theme will be first the *attractor*. All the terms contained in these classes are then extracted and organized by a decreasing order of frequency for each class. Hence some terms will be more important or relevant than others in the classes.

In a second step, all the “winning expressions” are then translated into a set of categories pertaining to a thesaurus. For instance, the expressions “pot”, “marijuana”, “cannabis”, etc. are transformed into the common thesaurus category “DRUG”. This transformation allows the analysis to work on the categorical labels instead of the words themselves.

In a third step, just as in the thematic analysis presented above, relations are then discovered with a neural net (ART) and identified though the analysis of the lexicon of each thematic class that relates to the chosen main theme or attractor. In the following example, we show the results in a graph that reveals the semantic net for the attractor “DRUG”. In this graph, each circle represent a particular theme such as:

-HJ : Yound Humans Man or / 1HF Young Human women :
R : Religion

PV : vital Processes positive+ or négative+

G : Géography

ES : Émotions and sentiments positive + or negative -

A : Actions positive + or negative -

DC :Cognitive Domain

E : Évént positive + or negative

RS Sociales: Relations positive + or negative

T : Transformation

And the arrows in the graph indicates a discovered textual proximity relation between the thematic concepts in the corpus.

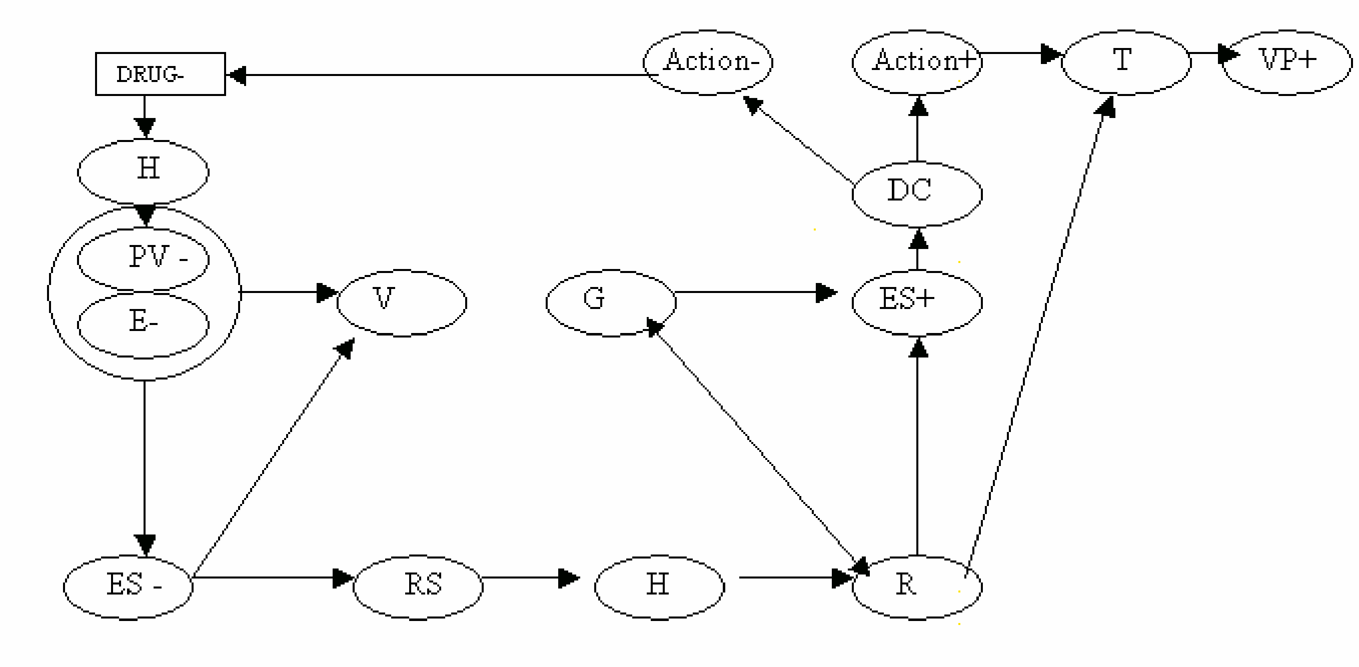


Figure 4 : Exploration of sub themes for the attractors : « DRUG »

The whole graph (figure 4) represents how the main theme of DRUG relates to other sub themes. For instance, starting from the top. The DRUG theme relates to the Young men (HJ) which in turn relates to it as a Negative vital process (PV-)and a Negative events (E-) etc.

Translated in more simple terms, an anthropologist sees in such this net the following thematic structure: It is his own interpretation: In this Innuit culture, relative to the question of Drugs, the elderly (the grand mother) (Mamit Innuat) uses the catholic religious rituals (prayers and pilgrimages) to influence the behaviour of the young members of the community in order that they abandon drug consumption and so that the they do not require the services of the methods and the vocabulary of the social services implemented by the Québec Goverment in the autochthones community[2].

As in the thematic analysis procedure, the computer does not produce this interpretation. It only produces some structuring of the textual data that renders more obvious important regularities in the corpus and to which the interpreter has to pay attention. It is on these regularities that the interpretation is couched.

6. The computer design: SATIM

As one can see, classification and categorisation techniques can be successfully used in CARAT. By its nature, such as methodology requires a very flexible computer platform because the various parameters of the methodology and the evaluation procedure cannot be easily defined and applied systematically. In order to realize this research program, all the preceding strategies must be realized through on some very flexible computer platform of some sort or other. There exist but few of of these platform specially dedicated for the humanities and social sciences [3] . We present here a brief review of our own in house program called SATIM which is a computer platform for high level design of computer assisted information processing. Its main function it to help design, produce and experiment complex processing chains on various types of information (text, multimedia, etc.). More specifically, it offers three levels of construction specialized either for a conceiver, a researcher and an end user. We can conceptualize the SATIM Platform as presenting three levels of architecture called the *workshop*, the *laboratory* and the *application*.

[2] The computer procedure has been compared in this case to a classical manual analysis of the corpus which had been done in another project. (Gagnon, 2004). The results were similar. They both gave the same structural organisation of the data. The main differences between both approaches are found, among others, in the amount of time for producing the same graph, either manually or with the computer

[3] Among the classical systems are ALCESTE (2004) . (http://www.image.cict.fr/index_alceste.htm). An emerging system is the D2K system applied to text analysis (T2k (2004) see: http://alg.ncsa.uiuc.edu/do/tools/d2k ; Statistically oriented systems for text processing are SPPSS (for text mining) TEXTPACK , SPAD T, SPHINX, , etc

6. 1. The workshop

This first level of the platform is an incubator and a repository for the various modules (functions) that are to be used in building analysis chains. This workshop deposits, integrates and manages various modules that a user may build or event acquire from various sources (if they are compatible). It serves hence as an integrated repository of a various computer tools for computer text analysis (lemmatizers, parsers, classifiers, etc). These modules are autonomous and independent. They can even have been built in various programming languages. This is a condition for maintenance and updating the workshop.

This SATIM workshop is in fact a special type of toolbox. It is not so much a set of modules (functions) but a set of tools for working on the modules. These tools allow them to communicate their input and output in view of a specific research objective or processing.

SATIM relies on three computers programming approach: the object-oriented design, the multi-agent and, most of all, the combinatorial functional design presented above. In its actual form, SATIM is in fact a huge relational database managing a variety of existing modules (lemmatizer, matrix builders, classifiers, statistical packages, etc.). This workshop is for the moment only accessible to expert programmers and its design is part of an ongoing software engineering project.

6. 2. The laboratory

The second level of the SATIM platform offers CARAT researcher tools to explore, in a transparent manner, various analysis chains (built from the precedent toolbox and repository). At this level, the user doesn't need to be a computer programmer, but he must be an expert in the field of a CARAT applications. Through an ergonomic interface, the user chooses the modules he wishes to see functioning in one or other of the analysis chains. He is assured that these chains are syntactically well formed that is, they form a complex algorithm. Their semantic (i. e. their meaning) depends specifically on the tasked aimed at. That is, even though the modules can be combined in a well form manner, they are not necessary semantically relevant for one particular task.

Two analysis chains have been built: Numexco and Grammexco. These two chains are classification chains on texts. One works with words as basic units of information, the other with n-grams. Other chains are being built as research goes on (Indexico for indexing, Ontologico for ontology maintenance, Thematico for thematic analysis, Categorico for automatic categorization of texts).

6. 3. Applications

The third level of the SATIM architecture is for the end user. If after many tests and experiments a particular successful chain is finally accepted, it can be wrapped up as an autonomous application, transparent to the end user and where only certain parameters are modifiable. It may have its own interface. And if a particular chain does not fit a specific goal objective and modules have to be changed, one must go back in the laboratory and experiment new chains.

7. Conclusion

We have presented in this paper how computer techniques of classification and categorisation can be used and applied in the field of computer assisted reading and analysis of texts. These techniques can diminish the burden the many researchers in the field of social sciences and humanities and even in various textual information technologies when they must go deeper in the content of these text so as to extract themes and organize possible navigation in them. It seems to us that it is only when more of these types of techniques will have been transformed into ergonomic systems that the social sciences and humanities will really adopt them. For the moment, there is still much more research to be done in order to better understand and model the process of reading and analysis of text.

8. References (liste des references mentionnées dans le texte)

Alexa, M. and C. Zuell. (1999a). *Commonalities, difference and limitations of text analysis software: The results of a review*. ZUMA arbeitsbericht, ZUMA: Mannheim.

Alexa, M. and C. Zuell. (1999b). *A review of software for text analysis*. ZUMA: Mannheim.

Barry, C. A. (1998). “Choosing qualitative data analysis software: Atlas/ti and Nudist compared”. *Sociological Research Online*, vol. 3, no 3. www. socresonline. org. uk/socresonline/3/3/4. html.

Barthes, R. (1970). *S/Z*. Paris: Seuil.

Benzecri, J. -P. (1973). *La Taxinomie. Vol. I. l’analyse des correspondances*. Paris: Dunod.

Bernard, M. 1999. *Introduction aux études littéraires assistées par ordinateur*. Paris : Presses Universitaires de France.

Bouchaffra, D. and Meunier, J. -G. (1995). *A Thematic Knowledge Extraction Modeling through a Markovian Random Field Approach*. 6th International DEXA 95 Conference and Workshop on Database and Expert Systems Applications, Sept. 19-22, London, UK.

Bradley J. and G. Rockwell. (1992). *Towards new Research Tools in Computer-Assisted Text Analysis*. Presented at The Canadian Learned Societies Conference, June, 1992

Braumbaugh, J. *et al*. (1991). *Object Oriented Modeling and Design*. Prentice Hall.

Brunet, E. (1986). *Méthodes quantitatives et informatiques dans l’étude des textes*. Paris: Champion.

Cavnar, W. B. and Trenkle, J. M. (1994). *N-Gram-Based Text Categorization*. Paper presented at the Symposium on Document Analysis and Information Retrieval. Las Vegas.

Choueka, Y. (1988) : "Looking for needles in a haystack", Actes RIAO, onference on User-Oriented Context Based Text and Image Handling, ambridge, p. 609-

Church, K., Gale, W., Hanks, P., & Hindle, D. (1989). "Word Associations and Typical Predicate-Argument Relations". *International Workhop on Parsing technologies,* Carnegie Mellon University, Aug. 28-31,

Condron *et al.* (2001). Digital Ressources for the Humanities. Mongantown: West Virginia University Press.

Cristianini, X. and Shawe-Taylor, J. (2000). *An introduction to support vector machines*. Cambridge: Cambridge University Press.

Damashek. M. (1989). “Gauging similarity with n-grams: language independent categorization of text”. *Science*, no. 267, pp. 843-848.

De Pasquale, J. -F. and Meunier, J. -G. 2003. *Categorisation techniques in Computer Assisted Reading and Analysis of Text (CARAT) in the humanities*. *Proceeding of the ACH/ALLC Conference (Computer and the Humanities)*, Netherlands: Kluwer Academic Publishers, vol. 37, no 1, p. 111 à 118.

Deerwester, S. *et al*. (1990). *Indexing by latent semantic analysis*. *Journal of the American Society for Information science*, pp. 391-407.

Eco, U. (1965). *L’œuvre ouverte*. Paris: Seuil.

Eco, U. (1992). *Les limites de l’interprétation*. Paris: Grasset.

Fayyad, U., Grinstein, G. G. and Wierse, A. (eds). (2001). *Information visualization in data mining and knowledge discovery*. San Francisco: Morgan Kaufmann Publishers.

Fielding N. G. and R. M. Lee, (eds). (1991). *Using Computers in Qualitative Research*. Thousand Oaks, CA: Sage.

Fish, S. (1980). *Is There a Text in This Class? The Authority of Interpretative Communities*. Cambridge: Harvard University Press.

Floridi, L. (2002). *The Blackwell Guide to the Philosophy of Computing and Information* Blackwell

Forest, D. 2002. *Lecture et analyse de textes philosophiques assistées par ordinateur : application d’une approche classificatoire mathématique à l’analyse thématique du* Discours de la méthode *et des* Méditations métaphysiques *de Descartes*. Mémoire de maîtrise, Montréal, Université du Québec à Montréal.

Forest, D. et Meunier, J. -G. 2000. *La classification mathématique des textes : un outil d’assistance à la lecture et à l’analyse de textes philosophiques*. *In* Rajman, M. & Chappelier, J. -C. (eds.). *Actes des 5es Journées internationales d’Analyse statistique des Données Textuelles*, 9-11 mars 2000, EPFL, Lausanne, Suisse. Volume 1, pages 325 à 329.

Fortier, P. A. (2002). Prototype effect vs. rarity effect in literary style. In Louwerse, Max and Willie van Peer (eds.), *Thematics: Interdisciplinary Studies*, x, 448 pp. (pp. 397–405).

Gadamer, H. G. (1996). *Vérité et méthode : les grandes lignes d’une herméneutique philosophique*. Paris: Seuil.

Gagnon, D., "NUMEXO et l'analyse par attracteurs et par classes des entrées de l'ECHO (Encyclopédie Culturelle hypermedia de l'Océanie)", *Lexicometrica*. Numéro spécial : L'analyse de données textuelles : De l'enquête aux corpus littéraires, juillet 2004

Gervais B., (1993) *A l'écoute de la lecture*. VLB éditeur, Montréal.

Glaser, B. G. and Strauss, A. L. (1967). *The discovery of grounded theory. Strategies for qualitative research*. Chicago: Adline.

Greenstein, D. I. A. (2002) *Historian’s Guide to Computing*. Oxford Guides to Computing for the Humanities.

Grossberg, S. *et al*. (1991). “ART2-A: An adaptive resonance algorithm for rapid category learning and recognition”. *Neural Networks*, no 4, p. 493-504.

Habermas, J. (1984, 1987) *Theory of communicative action* vol 1, &2 Boston, Beacon Press

Hart, P. E. (1998). *The condensed nearest neighbour rule*. IEEE Transaction on Information theory, no. 14.

Heidegger M, (1962) *Being and Time* , San Franscisco :Harper and Row

Hobbs, J. R. (1990). *Literature and Cognition*. CSLI Lecture Notes, Number 21, Stanford, CA: Cambridge University Press.

Hockey, S. (2001). *Electronic Texts in the Humanities: Principles and Practice.* Oxford: Oxford University Press.

Holland, J. (1975). *Adaptation in Natural, and Artifial Systems*. Ann Arbor (Michigan): University of Michigan Press.

Iser, W. (1985). *L'Acte de lecture. Théorie de l'effet esthétique*. Bruxelles: Mardaga.

Jauss H. R. *Pour une esthétique* de la *réception,* trad., Gallimard, 1978.

Jenny, J. (1997). « Méthodes et pratiques formalisées d'analyse de contenu et de discours dans la recherche sociologique française contemporaine. État des lieux et essai de classification », *Bulletin de Méthodologie Sociologique,* vol. 54.

Kastberg Sjöblom, M. and Brunet, E. (2000). « La thématique. Essai de repérage automatique dans l'œuvre d'un écrivain ». In Rajman, M. et J. -C. Chappelier (dir. publ.). *Actes des 5^{ièmes} Journées internationales d'Analyse statistique des Données Textuelles (JADT)*. Lausanne, 9-11 mars 2000, vol. 2, p. 457-466. Lausanne : EPFL.

Kodratoff Y., (1999)"Knowledge Discovery in Texts: A Definition, and Applications, " Proc. ISMIS'99, Warsaw, June

Kohonen T. (1997) *Self-Organizing Maps*. New-York, Berlin: Springer Verlag.

Kohonen, T. (1982). *Clustering, taxonomy and topological Maps of Patterns*. IEEE Sixth International Conf. On Pattern Recognition, pp. 114-122.

Kucera, H. and Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence: Brown University Press.

Lebart, L. et Salem, A. 1994. *Statistique textuelle*. Paris : Dunod.

Lévesque, G. (1998). *Analyse de système orienté-objet et génie logiciel, concepts, méthodes et application*. Montréal : Chenelière/McGraw-Hill.

Lewis, D. D. and Gale, W. A. (1994). *A sequential algorithm for training text classifiers*. In W. Bruce Croft and C. J. van Rijsbergen (eds.). *Proceedings of the seventeenth annual international ACM-SIGIR conference on research and development in information retrieval*. Dublin: Springer-Verlag, pp. 3-12.

Lusignan, S. (1985). « Quelques réflexions sur le statut épistémologique du texte électronique ». *Computers and the humanities,* vol. 19, p. 209-212.

Mann, W. C. and S. A. Thompson. (1988). "Rhetorical structure theory: Toward a functional theory of text organization". *Text*, Vol. 8, No 3, p. 243-281,

Manning, C. and Schutze. H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge (Mass.): MIT Press.

Mayaffre, D. (2000). *Le poids des mots. Le discours de gauche et de droite dans l'entre-deux-guerres.* Maurice Thorez, Léon Blum, Pierre-Etienne Flandin et André Tardieu (1928-1939). Paris : Honoré Champion.

McKinnon, A. (1968). « La philosophie et les ordinateurs ». Dialogue, vol. 7, no 2, p. 219-237.

McKinnon, A. (1973). "The conquest of fate in Kierkegaard ». CIRPHO, vol. 1, no 1, p. 47-58.

McKinnon, A. (1979). "Some conceptual ties in Descartes'Meditations". Dialogue, vol. 18, p. 166-174.

Meunier, J. -G. (1997). « La Lecture et l'Analyse de Textes Assistées par Ordinateur (LATAO) comme système de traitement d'information ». Sciences Cognitives, no 22, p. 211-223.

Michalski, R. S. (1983). *A Theory and Methodology of Inductive Learning*. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach*. Palo Alto: Tioga.

Nauck, U. 1999. *Design and Implementation of a Neuro-Fuzzy Data Analysis Tool in Java*. Doctoral thesis, Braunschweig, Technische Universität Braunschweig.

Newell, A. (1983). Intellectual issues in the history of artificial intelligence. In F. Machlup & U. Manfreds (Eds.), The study of information: Interdisciplinary messages. New York: Wiley.

Popping, R. (2000). *Computer-assisted text analysis*. London: Sage.

Rada, R. (1991). *From Text to Expert text*. New-York: McGraw-Hill.

Ram, A. and Moorman, K (eds). (1999). *Understanding Language Understanding. Computational Models of Reading*. Cambridge (Mass.): MIT Press.

Rastier, F. 2001. *Arts et sciences du texte*. Paris : Presses Universitaires de France.

Rastier, F. *et al*. (eds). 1995. *L'analyse thématique des données textuelles : l'exemple des sentiments*. Paris : Didier Érudition.

Reinert, M. (1994). *Quelques aspects du choix des unités d'analyse et de leur contrôle dans la méthode Alceste*. In L. L. S. Bolasco and A. Salem (eds.). *Analisi Statistica dei Dati Testuali*, vol. 1, Rome: CISU, pp. 19-27.

Ruiz, E. and P. Srinivasan. (1998). "Automatic text categorization using Neural Networks". In Efthimiadis, E. (eds). *Advances in classification research, vol. 8: proceedings of the 8th ASIS SIG/CR classification research workshop*. New Jersey: Information Today, pp. 59-72.

Ryan M. L. (1999). "Introduction" and "Cyberspace, Virtuality and the Text". Both in *Cyberspace Textuality: Computer Technology and Literary Theory*. Ed. Marie-Laure Ryan. Indiana University Press. 1-29 and 78-107.

Salton, G. 1989. *Automatic Text Processing.* Reading (Mass.): Addison-Wesley.

Salton, G. et M. McGill. 1983. *Introduction to Modern Information Retrieval*. New-York: McGraw-Hill.

Sebastiani, F. (2002). *Machine learning in automated text categorization*. *ACM Computing Surveys*, vol. 34, no 1, p. 1-47.

Schleiermacher, F, (1987) Hermeneutique, Paris Lille (orig: 1808-1809)

Sinclair, S. (2002). « Humanities Computing Resources: A Unified Gateway and Platform ». COCH/COSH 2002. University of Toronto, May 26-28, 2002.

Spence, R. and Press, A. (2000). *Information visualization*. Boston: Addison-Wesley.

Unsworth, J. (2000). « *Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this?* ». *Symposium on Humanities Computing: formal methods, experimental practice.* London: King's College. May 13, 2000.