The peculiarity of Canada—a country composed of two major national groups, spatially, linguistically and culturally differentiated—is strongly reflected in its social sciences. Canada contains two entirely distinct academic communities, one solidly bound to the Anglo-American context and the other more attached to French and Continental intellectual traditions. But in some cases, this dual national character can give rise to fertile interactions and exchanges. In the particular field of computerized analysis of qualitative data, we have to deal with both of these realities: on one hand, separate developments have taken place in English-Canadian and French-Canadian universities, following respective trends in sociological theory and methodology; on the other hand, certain “mixed” endeavors at the institutional level as well as reciprocal influences between individual scholars have sometimes furthered an interesting convergence. In this paper, we will present an overview of the use of qualitative software in Canada, stressing above all what we believe are its distinctive features vis-à-vis other countries, but knowing that it is impossible to portray in these few pages an exhaustive picture of the “Canadian situation.” The outline will, inevitably, be partial.

In Canada, words have to be defined before being used in public matters, and this also applies to scientific issues. It is not only a question of semantics and translation in a bilingual setting, but rather a question of interpretation: sometimes, even the word “sociology” does not seem to have the same meaning in either part of the country! In this context, when we touch on the subject of Computer-Assisted Qualitative Data Analysis (CAQDAS), a problem arises from the phrase “qualitative data analysis”: what exactly are we talking about? If we consider that the adjective “qualitative” in the expression “qualitative data” refers to a specific theoretical and methodological approach (i.e., “qualitative analysis”: grounded theory, ethnomethodology, the Chicago school, interactionism, etc.), then CAQDAS is a particular way of using computers to analyze verbal transcripts or written materials. However, if we consider the expression “qualitative data” to be equivalent to “non-numerical data,” in the sense that we emphasize the discursive nature of this data, then CAQDAS should be understood as the

* Ce texte a été publié dans Current Sociology, 1996; 44; 259, ages259-278. Il s’agit ici du fichier qui a été soumis pour publication. La mise en page peut différer de celle que l’on trouve dans la revue.
computerized approach to objects of language. That is why we will introduce the generic (and more neutral) term “textual analysis”—the study of meaning in textual sources as a specific type of scientific (e.g., sociological, anthropological) inquiry—to account for what represents in Canada a quite diverse reality.

The first part of the paper deals with facts about textual data computing in the Canadian context. The second part considers two software packages “made in Canada” which embody a “lexical,” or word-based, approach to textual analysis (SATO and TACT). We will also refer briefly to other software tools dedicated to the automated syntactic and semantic description of texts. In the third part, we discuss conceptual problems and challenges stemming from the use of computers by sociologists (and other social scientists) interested in the study of natural language sources.

Textual data computing in Canada

In Canada, researchers who work with language as data are slowly turning towards computers for help with the maintenance of their corpora. In certain cases, the computer becomes an aid at the stages of coding, retrieving, and even interpreting of the data. But, as everywhere, there is still much resistance and the process has yet to develop solid epistemological and methodological foundations if it is ever to become a well-established trend in the humanities. As we will see, the debate is just beginning to unfold, and the actual penetration of CAQDAS into the scientific community seems to be rather limited. Yet, it is important to point out that the tradition of computerized textual analysis, which is well grounded in the Canadian context, may or may not act as a facilitating condition for the design and distribution of qualitative analysis programs. We can venture to say that the notion that computers can manipulate symbols and do not necessarily imply “mathematization” already has a good grip on many social science departments across the country. Indeed, the use of computers to study textual sources begins very early in Canada, with the development of concordance retrievers by philosophers and other specialists in the field of hermeneutics. Already during the seventies, many scholars doing research in literature and communications—but also several sociologists and anthropologists—became interested in the possibilities of computerized textual analysis. Obviously, linguists were to become an indispensable part of the multiple projects that emerged in both English- and French-speaking scientific communities, ranging from grammatical analysis to narrative studies, from automated translation to artificial intelligence.

There are at least three institutional examples that illustrate quite well this tradition of computerized textual analysis in Canada: the Service d’Analyse de Textes par Ordinateur, the Centre for Computing in the Humanities,
and the Centre for the New Oxford English Dictionary and Text Research. In all cases, a strong interdisciplinary commitment was at the root of the project, and the idea that software designers and users should work together to link tool conception with theoretical reflection. Currently based in the Department of Linguistics at the Université du Québec à Montréal, the Service d’Analyse de Textes par Ordinateur's primary objective since 1983 has been to build computational environments for the analysis of natural languages, particularly French, which can be applied to discursive objects by researchers with different frameworks and approaches (syntactic analysis, content analysis, cognitive analysis, etc.) Among these environments, we can cite as key examples: FX (a computational language used for parsing as well as for programming knowledge-based systems) and SATO (Système d’Analyse de Texte par Ordinateur). The Centre for Computing in the Humanities was founded in 1985 within a cooperative program run by the University of Toronto and IBM Canada and is based in the Faculty of Arts and Science at that university. Among its many activities and services, the Centre offers courses, workshops and conferences on programs and techniques of research in humanities computing, publishes a newsletter and various working papers, houses a library of electronic texts and develops software packages such as MTAS (Micro Text-Analysis System), STRAP (Structural Analysis Program), and TACT (Text Analysis Computing Tools). More oriented toward the area of lexicography, the Centre for the New Oxford English Dictionary and Text Research was established in 1985 as an integral unit of the Computer Science Department at the University of Waterloo (Ontario) to address the requirements of computerizing the Oxford English Dictionary. Over several years, the Centre has developed a close collaboration with researchers in the humanities and social sciences who have served as users and evaluators of software and methods of text processing and information retrieval. In all three cases, special attention has been paid to the potential of computer-aided analysis of non-numeric data, an approach in which the machine is used to enhance the work of the investigator, not to replace it. In this matter, CAQDAS, which is a comparatively recent development in the wider field of textual data analysis, has traveled a very different path to face the same fundamental challenge: how to create software tools that increase rigor and efficiency while maintaining the methodological integrity of the data analysis process.

We consulted several specialists from Western, Central and Eastern Canada (the provinces of British Columbia, Ontario, Quebec, Nova Scotia) to gather their impressions of the extent of CAQDAS's distribution in their respective academic environments. In general terms, the opinion is that there is definitely a growing interest in the use of qualitative packages (specially NUD•IST, but also THE ETHNOGRAPH) by researchers and graduate students, but that this does not necessarily reflect a broad impact on qualitative research as a whole. In teaching,
several workshops dealing with specific programs are held occasionally in many universities but, even if some methodology courses may include a chapter on computerized analysis of textual data, most often curricula do not acknowledge its relevance. As for the profound influence that CAQDAS could have on methods and epistemology in the long run, it is still very difficult to assess its full extent. At this point, it is possible only to infer some of its potential effects from a variety of—often rather emotional—reactions to the use of qualitative and textual software one can find today in the Canadian social science community. Our correspondents told us that, between those colleagues who are reluctant to let the machine take over even the simplest clerical tasks and those easily carried away by the computer's unlimited abilities to multiply data descriptions, there is a wide range of attitudes including, notably, fear and fascination. Even if we find these perceptions exaggerated, they reflect a context of change, where the prospective gains—e.g. enhancement of methodological capacities, more “scientific legitimacy”—nonetheless pose a certain threat to the raison d'être of qualitative analysis, whose approach is primarily anti-positivistic. Being “close” to the material is something most qualitative researchers feel to be essential to their approach. The computer's often pre-defined formal operations (even if these allow the user to set or adjust their parameters) can be seen to intervene between the investigator and the empirical data. Moreover, the use of computers leads very rapidly to the accumulation of information, so that small case-based qualitative projects suddenly grow larger, with increasing numbers of records to handle. That is why Carl Cuneo, of McMaster University (Hamilton, Ontario), observes among sociologists who are reluctant to adopt CAQDAS or who adopt it critically not only the fear of quantifying qualitative data analysis, but also the fear of “losing touch” with the data (Cuneo, 1995). In fact, it would seem that most qualitative researchers who do use computers still prefer the simplest tool: a word processor. Carl Cuneo thinks that the reason for this is the fact that “so many sociologists and others doing qualitative research have had a negative reaction to computers, and they have found it much easier to do a straight keyword search in the word processor they knew and were working with rather than having to go out and buy a new, unknown program” (Cuneo, 1995). Conversely, Paul Sabourin, from the University of Montreal, points out that social scientists who “discover” qualitative analysis only when qualitative programs become available (“they will study unstructured data now that they can use computers”) often expect to get a ready-to-use methodology within a software package (Sabourin, 1995).

It is clear that the use of CAQDAS can have a strong impact on the scientific legitimacy and recognition by colleagues of research (and thus in such matters as the access to research funds). One very important issue raised in current debates on the use of CAQDAS is, as Janet Salaff, at the University of Toronto, puts it, that the question
of how much “scientism” is possible or necessary with these new methods will certainly remain open for a long time (Salaff, 1995). This means that the epistemological line marking the boundaries between quality and quantity, interpretation and description, etc., will probably have to be redrawn or at least reconsidered. But it also means that the role of qualitative researchers could be redefined within the institutional context. Thomas Regan, of Acadia University (Wolfville, Nova Scotia), believes that the growing use of computer programs should indeed “include fostering a greater respect for qualitative researchers, extending the teaching of qualitative research for those who give time to the mechanics of research” (Regan, 1995). The inevitable development of complex sets of techniques and procedures associated with the use of software tools will tend to shake up what many people consider a sort of craft. Actually, William Winder, of the University of British Columbia (Vancouver), anticipates that for the next few years, most researchers will use computerized techniques “behind the scenes,” and will not necessarily acknowledge the source of their insights because, in the humanities, “an idea that comes from a person is more valorized than one from a machine.” This means that any insight that is too directly linked to the computer could be disqualified by the following logic: “since the computer produces mechanically and furthermore produces in place of the researcher, then the researcher's insights are at best mechanically produced and of a purely technical nature, at worst a kind of rehashing of what the computer has already produced” (Winder, 1995). Thus, researchers who invest much of their time and creativity in the conception and application of computer tools will be seen as doing the work of a technician. That’s why, in William Winder’s opinion, “researchers will be coy about their computational sources,” at least in the near future.

A thorough survey of the major Canadian periodicals in our field (The Canadian Review of Sociology and Anthropology, The Canadian Journal of Sociology, The Canadian Journal of Criminology, Sociologie et sociétés, Recherches sociographiques, Criminologie, and Cahiers de recherche sociologique) allowed us to confirm that, so far, articles addressing any aspect of the computerized analysis of qualitative materials are extremely rare. Also, articles signed by Canadian authors in American and European methodological journals such as Qualitative Sociology, Quality and Quantity, Sociological Methods and Research, and Bulletin de méthode sociologique are exceptional. This certainly means that CAQDAS still has a long way to go. However, the controversy is already surfacing. A recent debate in Society/Société—the Canadian Sociology and Anthropology Association's newsletter—gives us a hint of the questions that are currently being raised by sociologists in Canada. In February 1995, Ellen Wall, at the University of Guelph (Ontario), wrote a scathing commentary about a tri-university workshop on NUD*IST that had been organized by Carl Cuneo and held on an interactive video link between
McMaster University, the University of Guelph, and the University of Waterloo. In her own words, she felt “discomfited” by the “thought that this software for qualitative data was being presented and accepted with so little critical comment” (Wall, 1995: 13). Concerned by the fact that “structure is all-important for the NUD•IST operators,” Ellen Wall criticized the idea that data management can solve problems, since “manipulating text units and connecting nodes in schematic models will never replace the intuitive knowledge that emerges during fieldwork experience” (Wall, 1995: 14). She closed the commentary by pointing out “the irony of qualitative sociologists’ taking up such linear, reductionist methods in this era of trans-disciplinary research” (Wall, 1995: 14). The point made by this qualitative sociologist was straightforward: knowledge obtained by search-and-retrieve procedures is categorically different from that gained through intuition and holistic understanding. Two responses to Ellen Wall’s short article appeared in the following issue of the newsletter. Denis Harrison, at the Université du Québec à Hull, questioned what he considered to be a “technological determinism” in her commentary (i.e., the notion that a program “creates” things) (Harrison, 1995: 16). As an experienced user of NUD•IST, he pointed out that the computer does not change anything in the way qualitative analysis is done: beginning with strong research questions, the analyst categorizes the data through successive readings and continues to make adjustments rather than moving linearly. The machine enhances the analyst’s work by mechanically performing non-interpretative operations, and this is by no means contradictory to qualitative inquiry, where the researcher has to organize and classify large amounts of information. There is of course the risk of letting the tool impose its own logic. Raewyn Bassett, Sue Cox and Ulrich Rauch, at the University of British Columbia, admit that the use of computers "has subtle but important implications for the researcher's sense of autonomy," and share Ellen Wall's concerns about the dangers of excessive rationalism and scientistic conceptions of the research process (Bassett, Cox and Rauch, 1995: 19). They underscore, however, the fact that data management activities are different from those associated with data analysis, and that the "systems of data management should therefore be seen as resources that enable the researcher to effectively engage in a whole range of critical, hermeneutic or other interpretative approaches to qualitative data analysis" (Bassett, Cox and Rauch, 1995: 23). While it is true that a researcher should not "overwork the data" and perform analyses "that have very little substantive knowledge to contribute," they challenge Ellen Well's impression that the use of computerized approaches lead inevitably to the deformation of the "fundamental principles of qualitative methodology" (Bassett, Cox and Rauch, 1995: 21).
Canadian scholars are thus beginning to acknowledge, adapt or react to the advent of new computerized techniques for the analysis of textual data. Some of them—sociologists, anthropologists, political scientists, critical theorists—had already explored one of the first meeting points of language and computing: lexical-based textual analysis. This approach to qualitative data involves a theorization of the discursive phenomenon in itself. We will see how this is shown in the design of two programs that have been in use in Canada since the mid-eighties.

**Lexical-based textual analysis**

Several traditions of textual analysis in the humanities, such as discourse analysis and literary criticism, have held the premise that form, as much as content, is a key dimension of all natural language objects. This is clearly expressed in the classical distinction between the enunciated object and the enunciating act, pertaining respectively to the study of the referential scope of communication and to the study of communication as a socially situated event. Much importance is then attached to the form of the text and, above all, to the distribution of its linguistic units. The speaker's choice of words and the words' combination in phrases and sentences are to be examined thoroughly in order to gain an understanding, not only of the message of the text, but also of the making of the text. Such is the philosophy that underlies the two programs that we will present here.

In this perspective, a corpus of written material can be considered as a set of lexical data, that is to say, as an array of words. But then, working with words as units implies dealing with their “identities” both in form and content. On one hand, different words in a corpus may convey a similar meaning (“vessel” and “ship” can designate the same object), refer to a common thematic area (“employment” and “money” have to do with the economic sphere), or belong to the same semantic field (“freedom” and “justice” evoke each other as the basic values of a liberal society). On the other hand, the same word may have more than one signification: a “society” is a “community,” a “company,” but also an “association,” etc. The distinction linguists make between type (the term itself as an entry in the dictionary) and token (its multiple instances of use in actual utterances) is useful to cope with this complexity. Every word in a textual data base has a form (a specific string of characters) as well as an address (a unique position) in the linear sequence that the whole corpus constitutes. With the aid of the computer, a researcher is able to scrutinize types and tokens simultaneously. For example, a list of all words (types) used by a particular speaker or present in a specific type of surrounding can be automatically generated (together with certain numerical information such as frequencies and distribution.) Once certain relevant words
have been picked out from the list, the researcher can explore all their occurrences (tokens) by observing them in context. What is crucial in this approach is that the “paradigmatic” structure of the text (the axis of word selection, as opposed to the axis of word combination, or “syntagmatic” structure) becomes visible. When looking at a specific segment of the corpus, the analyst can immediately inquire about the paradigmatic status of any of its words: Is the term “vessel” consistently preferred to the term “ship” by this particular group of speakers? Which are the most frequently recurring adjectives that qualify the word “money” when used to denote wealth? Is there an overall tendency to mention “freedom” and “justice” but not “freedom” and “solidarity” in the same sentences? and so on.

One can argue that a lexical approach is not at all incompatible with typical qualitative activities that require “closeness” to the raw data. To chose the word rather than the segment as the basic recording unit does not imply adopting a reductionist view of text as mere information to be schematized: both qualitative and quantitative word-based procedures can be used to gain insight into different aspects of the enunciated object (semantic associations, situationally-determined vocabularies, etc.) and the enunciating act (the inscription of the speaker in his own utterance, the operations of logical modalisation, etc.) through non-linear readings. For example, in contrast with traditional content analysis techniques, lexical categorization should be conceived as an interactive search-and-coding process. The access to the data is thus performed through—and not substituted by—referential categories that emerge and evolve during the categorization process. But, precisely because of their strong lexical orientation, both Canadian programs for textual analysis differ from what is usually considered as CAQDAS software (see Weitzman and Miles, 1995).

TACT (developed by John Bradley and Lidio Presutti at the University of Toronto) is a textual analysis system designed to retrieve occurrences of a word, a word pattern, or a word combination to sort frequencies of words and phrases, to calculate several kinds of type-token statistics, to produce a ranking of collocates to a word by their strength of association, and to list all repeating fixed phrases. TACT is especially intended for the study of literary works, but it is also a useful tool for the analysis of other types of materials, such as political speeches, legal documents or transcribed conversations. With the assistance of TACT, a researcher can target, one at a time, a subset of relevant words within any textual database by browsing through an alphabetical list or by setting out a query according to a pattern of characters. Queries may contain refinements called “selectors” that specify proximity (two or more words found together within a user-specified span of words), similarity (in spelling), frequency of occurrence, and a condition related to whether or not words or patterns have one or more tag
attributes (lemma, part-of-speech or conceptual labels) previously marked up by the user. Once a set of words has
been selected by whatever means, it can be saved as a “group.” The selected word groups can then be displayed in
five ways: a keyword-in-context (KWIC) concordance, a variable-context concordance, the whole text, an
occurrence-distribution graph, and a table of collocates. Displays are linked so that, for example, the researcher
can go directly from a position in a distribution graph to the text it represents (McCarty, Centre for Computing in
the Humanities WWW home page).

This kind of approach relies on the assumption that no fundamental opposition exists between the analysis of
qualitative materials and the systematization of certain analytical procedures: a case study of a literary work
carried out by a critic still falls into the category of qualitative analysis—all meaningful units are considered in
their surrounding context—even if a certain degree of automation and statistical calculation is introduced. By
means of lists, graphs, and maps, TACT gives the user the possibility of measuring the extent to which the lexical
data fall into a pattern at some “level of significance.” This is accomplished with “position-based displays” that
reorganize the words of a text “topographically” so that “we can see where the words occur and do not occur,” as
well as the study of word associations and, eventually, “semantic networks” (Lancashire, 1993). As we shall also
see with SATO, the use of statistical techniques in textual analysis such as descending-frequency indexes and
collocational tables (lists of words that co-occur with a designated “node” word) does not amount to plainly
“quantifying” qualitative data.

SATO (developed by François Daoust at the Université du Québec à Montréal) was conceived as an
interactive tool allowing the user to freely combine and adjust different tasks into a flexible research design.
Besides the standard features of a textbase manager and a code-and-retrieve program, this software system has the
particular capacity of supporting the annotation of the text in its two dimensions, syntagmatic and paradigmatic,
through a system of type-level and token-level, numeric or symbolic “properties.” This means that every lexical
item, either in an index (as a type) or in the text itself (as a token) can be categorized “manually” or with
specialized, user-defined algorithms. The researcher can follow a specific trail (looking for certain terms,
frequency ranges or syntactic constructions), and attach, modify or remove categories from words. Moreover,
dictionaries can be built from a text and applied to another, so corpora can be updated, augmented and compared.
One interesting option is the use of general frequency dictionaries, which enable the user to assess the lexical
“originality” of a given text by calculating the distance between expected and observed frequencies. But, even if
SATO includes these and other “variable-oriented” capabilities, it must be distinguished from “lexicometric
programs.” Mostly used in France, but also quite popular in Spain and Italy, these programs are deeply rooted in the idea of “treating words like numbers.” They resort to multidimensional statistical methods like factor correspondence and hierarchical classification analysis, where the user has almost no control over what happens between the input of the raw data and the output of the results. SATO, on the other hand, should be seen as a “toolbox” enabling the user to gain access to the text in an “iterative mode,” that is to say, by moving back and forth, by going from types to tokens, and from categories to words and vice versa.

Like most text retrievers, SATO allows the user to launch search requests based on Boolean logic and elaborate wildcards. But because of its “lexical orientation,” it is specially adapted to include various syntagmatic and paradigmatic conditions in its searches. For example, a command could instruct the program to retrieve all segments with a particular morpheme, but only when it is followed by a verb (syntagmatic condition) and its frequency in each speaker’s discourse is greater than 10 (paradigmatic condition). The concordance hits can have a variable context: a given number of words before and after the node word (from one to several hundreds), the sentence (from punctuation mark to punctuation mark), the paragraph, the page, the whole document, etc. Then, the researcher may, for instance, create a subtext (or “domain”) with all the retrieved chunks and contrast its vocabulary with the rest of the corpus in order to identify a specific lexical pattern. The questions that matter in this kind of approach are, for instance: Why has this speaking subject recurrently chosen this specific type of phrase? Which signifier (the form) has been privileged to communicate a signified (the content)? How can we measure the “sensitivity” of these lexical preferences to particular situational or thematic constraints? These questions take on great importance in the perspective of discourse analysis (Armony and Duchastel, 1995).

As a concluding remark concerning software development in Canada, we should point out that, aside from lexical programs, other “linguistic” computational tools have also been made available in the past decade to researchers in the humanities and the social sciences. We will mention only two of these tools, used in various anthropological and sociological research projects. The first tool, DEREDEC (later to become FX, a much more sophisticated version), is a programming language developed in the late seventies by Pierre Plante at the Université du Québec à Montréal. It lets the researcher build complex systems of textual description such as, for example, a grammatical model that permits automated detection of specific syntactic structures. This “machine” can then be applied to a corpus to identify syntagmatic patterns such as contextual dependency relations between words on the basis of their relative position in the clause (“determined” and “determining” words, i.e., which words define and specify which) (Duchastel, Paquin & Beauchemin, 1992, 1994). The second tool, DISCAN
(Computer package for discourse analysis), has been developed by Pierre Maranda at Laval University (Quebec). It does standard content analysis (using a custom-made thesaurus), which serves as input to a discourse analyzer, based on Markov chains, which translates the semantic mapping of the text into a net of more or less strong probabilistic links at the syntagmatic level (Maranda, 1992). The kind of approach put forward by these tools, sometimes considered too linguistic-oriented by social scientists interested in language as data, should not be dismissed without further consideration. We shall argue, in the next section, that working with text requires the analyst to look into the laws peculiar to human discourse.

The sociologist, the computer and the text: some epistemological considerations

We have seen that in Canada the use of CAQDAS has been quite moderate so far, though it begins to spread as more sophisticated computer software is becoming widely accessible in user-friendly versions. As we have seen, Canadian researchers have acquired, before turning to qualitative computer devices, some experience in the specific fields of lexical-based and natural language computing. We believe that most debates that have already taken place in those domains are currently being taken up again in the context of CAQDAS distribution. That is why we will now develop some conceptual and epistemological considerations about the computerized study of discourse concerning both natural language and qualitative-oriented research.¹

Two observations have guided our thought. The first concerns the growing importance, in sociology, of discourse either as a privileged form of access to social reality or as an object of knowledge in itself. The second is the growing utilization of—or expressed willingness to use—computerized methods for textual or discourse analysis. We will point out how the lack of reflection on the epistemological and methodological prerequisites for discourse analysis has nurtured the resistance towards the computerized approaches to textual analysis. Defining discourse analysis as the relation between hermeneutic and explanation procedures will help us resolve the dilemma confronted by the researcher who adopts computer-assisted methods.

We will first examine the problem of discourse analysis in the general context of a shift in the social sciences from the production paradigm to the language paradigm. This change in perspective results in different positions being taken on the nature of discourse and on the appropriateness of defining formal procedures for its analysis. We will then be able to address the dialectical relation between interpretation and explanation and to examine the necessity of some degree of formalisation as a prerequisite to any computer utilization. We will then examine various types of attitudes that prevail about the use of computer methods, propose what we think is
necessary to any computer design for textual analysis in the social sciences and, finally, indicate the limits inherent in computer-assisted analysis.

*Interpretation and explanation as part of the analytical process.*

Gyorgy Markus (1986) defines two major paradigms in social science theories. On one side, the production paradigm, mostly represented by Karl Marx, conceives society as a set of rules of production and reproduction of social relations within the institutions. On the other side, the language paradigm emphasizes the “intersubjectivity” and its mediation through language. This second paradigm is constitutive of several approaches in the social sciences (psychoanalysis, symbolic interactionism, ethnology, cultural studies, etc.) For Markus, even if both paradigms still coexist, language is taking on more and more importance in the understanding of social phenomena. This tendency explains the importance of discourse as a privileged object of study. But it also raises the question about the various methods for analyzing different kinds of data. Markus notes that it is not relevant to consider production phenomena as objective and, conversely, to look at language as an exclusively subjective matter. All theories are theories that attempt objectivity, that is to say, they try to explain as much as to interpret social phenomena as historical forms of human and social experience.

Not all the researchers will agree that language or its principal social manifestation—discourse—can be objectified. On the contrary, some social scientists will either ignore this objective aspect of discourse or sustain its irrelevance. In the first case, there is not much to say as long as researchers are not aware of the social nature and autonomy of discourse. In the second case, the position is well established in an epistemological and theoretical frame defined along hard relativist and constructionist lines (Denzin, 1988). From this viewpoint, no reduction of complexity is admitted and no descriptive or inferential devices can be set up. Analysis becomes a fully hermeneutic task without any kind of methodological constraints.

As long as discourse is considered an object amenable to an analytical process, it is possible to proceed with our discussion. Before we examine the relation between explanation and interpretation, we will identify three sets of oppositions that structure the conception of discourse as an object. The first opposition centers on content analysis versus discourse analysis. Besides the fact that most—though not all—content analysis favors quantitative strategies while discourse analysis is more qualitative-oriented, both methods employ different kinds of formal procedures for the description and explanation of the data. Furthermore, both approaches are
complementary in the sense that no discourse analysis is unconcerned by content and no content analysis is unaware of the formal dimensions of discourse.

The second opposition resides within the discourse analysis approach and between two traditions. The Anglo-Saxon tradition favors everyday discourse situations or verbal interactions within localized situations, while the French school of discourse analysis studies mostly standardized public discourses (political, media, scientific, etc.). Both perspectives resort to the same kind of paralinguistic methods (enunciation analysis, semantics, pragmatics, etc.), but the French tradition has tended to favor global interpretations about discursive and ideological formations (as shown, for instance, by the works of Michel Pêcheux and Michel Foucault) while the Anglo-Saxon tradition has generally offered research conclusions only at the level of phenomenological and ethnographic considerations. Translated into methodological terms, the goal is to situate the analytical process somewhere between a large-scale strategy that uses few descriptive devices on large corpora and an in-depth strategy of complex investigation of the many dimensions of one localized situation. In both cases, the equilibrium between the interpretative and explanatory dimensions of the analysis will be affected.

The debate between qualitative and quantitative approaches constitutes the third opposition. It is interesting to note that it crosses the two other oppositions and makes it possible to explain the ins and outs of the question of formalisation. Formalisation should not be posed as a question of quantity as against quality. As long as we do not accept a hard constructionist position, both qualitative and quantitative approaches can entail, at least partially, formal representations and procedures. The opposition between qualitative and quantitative approaches has long rested on Wilhelm Dilthey's distinction between the natural sciences, which foster the explanation of material phenomena, and the spiritual sciences oriented towards global comprehension of human behavior. Paul Ricoeur (1986) criticizes this opposition and, while admitting that these sciences differ in their object, he does not think that they are different in their common goal of objective knowledge. If such is the case, all sciences aim at a greater comprehension of phenomena through procedures of explanation and interpretation.

This leads us to the problem of reconciling explanation and interpretation as knowledge procedures compatible with science. Let us remember that even for Dilthey, the global comprehension of a phenomenon was not equivalent to its interpretation. To comprehend an object, one has to systematize the applicable procedures of interpretation. Ricoeur (1986), citing Dilthey, writes: “We call comprehension the process by which we know something ‘psychological’ in nature with the help of perceptible signs which are its manifestations …”
In this context, interpretation is considered the art of understanding signs and cannot be confused with the global comprehension of the phenomenon. It is on this basis that Ricoeur invites us to consider both explanation and interpretation as two moments of the reading of a text, leading to its full comprehension.4

Thus both interpretation and explanation can be defined as research procedures contributing to the construction of knowledge. Explanation derives from scientific reasoning based on the description of social forms and their mutual relations. While interpretation consists of producing meaning about phenomena, it is not completely independent from explanation. Consequently, we will distinguish two levels of interpretation: local and global. We will call local interpretations the procedures to stabilize rules of interpretation in the research process. For example, explicit rules for coding in context can be considered local interpretation procedures. An interpretation of the signification of a unit is necessary to assign it a code. The production rules in expert systems can also be considered local interpretation rules. They constitute inferential processes based on interpretation. These interpretation rules are essential to scientific reasoning and they should at least be clarified if not formalized.

Global interpretation, however, consists of giving meaning to research results. In this sense, interpretation escapes, to a certain extent, from the problematic of objective knowledge. A gap always exists between the results obtained through the research process (descriptions, local interpretations, inference processes, and so on) and the global interpretation formulated in terms of the researcher's theoretical framework. Global interpretation deals with the fundamental question of bridging theory and the empirical world. Our concern here is to master the local interpretation procedures so as to stabilize rules of interpretation in the analytical process.

At this point, it is possible to conclude that the use of computers is compatible with discourse analysis insofar as we agree on two matters. First, at a very general level, the shift from production to language must not be interpreted as an unidirectional move from objective to subjective knowledge. Language can and must be studied through objective procedures. Secondly, the requirement of objectification, translated into explanation and interpretation procedures, is consistent with the use of computers. The many discourse analysis approaches must then be examined through their descriptive and inferential devices so that these procedures can be eventually formalized into algorithms of some kind.
Using computer assistance in textual analysis

Using the computer requires the researcher to comply with a certain amount of formalisation and automation. This requirement will vary as much as it does in other discourse analysis methodologies. Computer use does not modify the methodological questioning of discourse analysts, but it leads them to specify the elements of their analytical process. It also helps to show the intrinsic limits of formalisation in a given research design. The computer is useful as long as formal units can be identified and described, whatever the level of complexity such description might hold. However, in our view, formalisation and automation can not reflect all levels of complexity and permit the full understanding of discourse. Indeed, the full automatic understanding of a text would imply that the machine was able to decipher any message through the analysis of the graphic, linguistic, cognitive, or pragmatic structures of discourse and of their mutual relations. The best argument against the pretensions of artificial intelligence is the absence of this kind of synthetic approach in any existing discourse analysis tradition. Again, computers do not produce knowledge that is not previously thought of. It can only help to clarify the formal operations of research and enhance task-oriented capacities.

Before presenting what we believe is the most appropriate approach in using computers for textual analysis, we will outline four types of responses to the use of the computer. The first type is characterized by hostility or, at least, indifference. Discourse is conceived as a complex phenomenon that cannot be split into different components. This response has two versions that refer to two different discourse analysis traditions. Some qualitative or comprehensive approaches convey hostility to any use of a computer aside from word processing, because of a hermeneutic or empathic posture that disqualifies all sorts of formal description and pre-defined procedures. Other researchers will show a lack of interest in the use of computers because of the procedures they favor. Because it applies a great number of devices at the many levels of complexity of the text (syntactic, semantic, pragmatic, etc.), the use of a computer would be more demanding than rewarding.

A second type of response is somewhat paradoxical. It is characterized by a naive enthusiasm for any computer system proposing some kind of textual analysis. This ingenuousness is often based on a lack of discourse theorization discourse and on a poor conception of the methodological underpinnings of the analytical process. Researchers will then opt for the first accessible computer program hoping that it will produce research results. To the extent that there are no model-free computer procedures, this attitude might lead to an uncritical interpretation of these results.
The utopian attitude shares with the previous response a certain degree of ingenuousness to the extent that it is more or less based on the faith that artificial intelligence will succeed in simulating the human process of understanding and generating language. This attitude also implies that the complexity of language can be reduced and that every structural level of discourse can be analyzed with the help of formal descriptions and computer algorithms. The utopian type is rarely found in the social sciences. Aside from the linguistic and cognitive sciences, only a few micro-sociological research projects contribute to this perspective. This could be explained by the fact that sociologists or other social scientists look at discourse from the other end, that is to say in its full complexity.

The fourth type can be called pragmatic. It consists in using computers because of their automated or assisted procedural capacities. This attitude fits well with the present state of the art in textual analysis, where there is no unified theory or methodology. Computers are then predominantly used for their capacity to generate repetitive and in-depth descriptions and analysis of data. While computing consists primarily in the automation of procedures, it can also be regarded as an aiding device in the analytical process. As long as we can define some fundamental procedures in the research process, the computer can be programmed to facilitate these procedures in the light of the researcher's theoretical perspective.

Adopting this realistic approach, we will now propose a model for the utilization of computers in textual analysis in the social sciences. We favor a hybrid approach that links automated and assisted procedures. If some mechanical tasks are always performed by computer programs, we must be assured that the underlying models are well understood so that the researcher will never lose control over the analytical process. That is why we prefer a pragmatic approach using tools that enable us to enhance our reading ability by providing quick access to the contents of great bodies of texts and by increasing precision and regularity (Duchastel et al., 1995).

Our starting point for this model is the fact that any analytical process will resort to a set of fundamental operations independent of any theoretical or methodological perspective. These operations are common to any discourse or content analysis. As far as these operations can be translated into computer procedures, we think that they will enhance the researcher's analytical capacity. The procedures fall into four logical types: data management, data description, data exploration and data analysis.

Data management procedures range from textual scanning and word processing to database or full text management systems with greater or lesser indexation and navigation functions. The first methodological operation consists in defining a set of formal or content units that will become the objects of the consecutive
procedures. Textual data management systems allow for the identification, segmentation and handling of different levels of textual units (word, phrase, sentence, part of text, thematic segments, etc.)

Depending on your approach, it is possible to identify a great number of data description devices. It is probably impossible to completely distinguish the operation of description from that of analysis. Some descriptions are less analytical in the sense that they propose an empirical classification of data and do not refer to any kind of theory. Some others are part of a theoretical model representing the investigated phenomena. In this case, the description of data is inevitably linked to the analytical process. We nevertheless distinguish data description from data analysis. Both groups of procedures have some elements in common, but can be distinguished both in theory and in practice. Description refers to the characterization of units while analysis focuses on the reasoning involved in any cognitive process.

Coding, which corresponds to the operation of categorization, is the most general process and precedes any application of rules or identification of relationships. It has become usual to differentiate between two types of coding, referential and factual (Seidel & Kelle, 1995). Referential coding is a heuristic device that gives the researcher better access to the material through a number of given categories. Factual coding, defined as the conceptual representation of phenomena, leads to more complex analytical descriptions. Coding words or thematic segments with socio-semantic categories is the most common example of the referential type. Morpho-syntactic categorization is a good illustration of the factual type, as far as this type of categorization is based on a linguistic theory and can lead to a more complex description of the data (the syntactic structure of sentences). In many fields of discourse analysis, representation models can be translated into more factual categories and structural descriptions — as is the case with certain approaches to pragmatics or to argumentation.

But whatever the perspective, all researchers need to categorize their data in a referential or factual manner. Most qualitative researchers will favor the heuristic approach, which gives them a thorough access to the data. Some will seek assistance for theory construction or validation. Many computer programs offer coding functions that authorize varying degrees of complexity in the representation of data. Researchers should be aware that these functions may only be partially adapted to their own needs, and should take into account the level of units that can be coded (words, parts of text, etc.), the facility to code in context, the presence of automated procedures for coding, the iterative capacity of the system, the type of representation favored (tagging, hierarchical structure, etc.)
The third type of procedure is data exploration. Exploration refers not only to information retrieval, but also to the systematic query of analytical descriptions of the data. Exploration regroups all procedures aiming at retrieving words or textual segments, gathering coded units or identifying relations or structures on the basis of previous analytical descriptions. Hence the computer can facilitate the identification of either regularities or isolated but significant data. It enables comparisons between elements of the text, by taking advantage of its capacity to run through the whole data, both extensively and in depth. These procedures are usually automated, and their efficiency depends on both the relevance of the preceding descriptions and the quality of the procedural rules.

As we have stated before, the analytical process can be seen as a combination of explanation and interpretation operations. Some researchers may prefer to remain close to the raw data and make minimum use of formalized analytical procedures. They will use computer programs for their capacity to manage, describe and explore the data. In other instances, researchers may chose to make their reasoning explicit. Whatever kind of analytical path is chosen (hypothesis examination, production rules in expert systems, parsers, statistical analysis), certain rules will have to be formalized.

The textual data analysis is a logically independent operation, but we will find it at every stage of data processing, with the exception of the data management step. We will illustrate this by an example of different analytical procedures present at the three other levels of the process. As we have argued, some descriptions are more analytical in their nature depending on the existence of formalized rules. These rules are usually programmed into automated devices. We may take the example of syntactic parsers. On the basis of a previous automatic or semi-automatic morpho-syntactic categorization, such parsers construct, with the help of automatons, a representation of the syntactic structure of sentences. If linguists tend to consider the descriptions produced by this kind of device to be the result of the analytical process itself, social scientists tend to view them as contributing to the description of their data. Applied at the stage of data description, they could enrich the characterization of the text units and enable the researcher to increase the complexity of the rules of exploration.

Some analytical processes can be developed at the exploration level. Expert systems illustrate the analytical aspect of certain exploration devices. These computer systems simulate expert reasoning. Certain facts being observed (the database), others being known (the knowledge base), an inferential engine based on production rules will generate the conclusions. This technology can be used at the exploration level of the research process. Interpretation rules could be systematically applied to the data to generate corresponding results. Logic
programming that helps in inductive or deductive analytical cycles (Sibert and Shelly, 1995) and hypothesis testing in qualitative data analysis (Hesse-Biber and Dupuis, 1995) are also good examples of this kind of analytical task performed at the exploration level.

Obviously analysis is the main activity at what we have called the data analysis phase of the research process. The results obtained during the previous phases are submitted to analytical procedures. For example, different statistical calculations can be carried out on the new data produced during the exploration phase.

We have shown that fundamental methodological operations can be translated more or less extensively into computer procedures. Depending on their own knowledge objectives, the researchers will use the computer at one or several phases of the analytical process and they will require more or less formalisation and automation. Since certain logical operations are common to all approaches, we have suggested that it is possible to select programs and functions most appropriate for the ends of each. But these choices are not without repercussions.

Let us return to our four types of response to the use of computers. Resistance to the use of computers is not inevitable. As long as an assistance-oriented approach is preferred to full automation, computers can become useful at many levels of research. Each type of research has to draw on the devices that seem most useful to its own purposes. Naive enthusiasm must be dismissed because researchers ought to reflect as much on the theoretical nature of discourse as on the methodological and procedural aspects of the analytical process. They must seek adequate information and understanding of the model underlying any kind of computer program. Results should not be sought for themselves, but be the by-product of a controlled process at every level of analysis. Utopianism must be questioned. Formal and automated procedures should not necessarily be considered in conflict with qualitative research. However, even if automatic systems pre-determine in part the results they produce, they should never remain hidden in black boxes. The realistic approach will then consist in using the most appropriate functions enhancing the rigor and capacity of the analytical process without compromising the epistemological posture adopted by the researcher.

Conclusion

Returning to the Canadian context, we have to conclude that qualitative data computing is still relatively uncommon if we limit the definition of qualitative research to grounded theory, interactionism and ethnomethodology. If, on the contrary, we consider qualitative analysis as the non-numerical study of symbols or, more specifically, the study of discourse and text, we then have to recognize that Canadian researchers have been
pioneers. There is no objective basis on which to evaluate the extent of CAQDAS's use in our universities and research centres or to compare it with other national contexts. However, our informal survey has suggested that only a modest number of Canadian researchers can be considered experts. Not many courses on CAQDAS are taught and very few research results in the field are published. On the other hand, the demand seems to be steadily increasing. This conclusion is based on two observations. The demand for counseling is incessantly growing in Canadian research centres dedicated to computer-assisted textual analysis and research proposals submitted to granting organizations increasingly refer to the utilization of CAQDAS.

Two reasons could be given to explain the relative under-utilization of CAQDAS and the absence of specific CAQDAS development in Canada. On one hand, we must be aware of the fact that the qualitative community strongly resists any methodological innovation that could imply a “positivistic” shift. As made evident in the debate in the Canadian Sociology and Anthropology Association newsletter, some qualitative analysts question any attempt to formalize, quantify or remove the researcher from the data. The other reason we can think of is the existence of a tradition in computerized textual analysis that has different theoretical roots from the qualitative approach. As the field of discourse analysis has been taken up by other disciplines and approaches, qualitative researchers might not have found what they were looking for in Canadian computer programs. However, we have tried to show that some of the fundamental questions involved in the use of computers have already been formulated in the textual analysis tradition and that discourse analysis, even if it varies from one approach to another, comprises basic operations that are common to all qualitative approaches as long as they are interested in the study of language.

Bibliography


CUNEO, Carl [E-mail message; sender: cuneo@mcmail.cis.mcmaster.ca] (1995, June 27).


Regan, R. [E-mail message; sender: tom.regan@acadiau.ca] (1995, July 6).


Sabourin, Paul [E-mail message; sender: sabourin@ere.umontreal.ca] (1995, July 6).

Salaff, Janet [E-mail message; sender: salaff@epas.utoronto.ca] (1995, October 5).


Winder, W. [E-mail message; sender: winder@unixg.ubc.ca] (1995, July 3).

. We are well aware of the fact that discourse and text should be distinguished. On one hand, the concept of discourse has a greater breadth than the notion of text. On the other, most analytical approaches and all computer-assisted analyses will give greater importance to the textual aspect of discourse.

1. As the point was made in the first section, discourse analysis is intrinsically associated with the qualitative perspective that seeks to gain access to reality through objects of language. In the following section, we will use the term discourse analysis to discuss general theoretical and methodological questions. We will refer to textual analysis when we focus on the concrete analytical process.
2. At the present time, part of the French school has shifted its interest towards the same kind of unstandardized objects that characterize the Anglo-Saxon school.

3. The translation from French is ours.

4. Ricoeur distinguishes discourse and text, the latter being the “decontextualized” version of the former. For the author, as long as we are considering the text, it is possible to explain it through the study of its internal relations and structures (he gives examples of structural analysis). When we “recontextualize” the text, we then have to interpret it in its relations with the speaker's or the listener's real world. We argue that explanation and interpretation procedures can be applied to the contextual aspects of the text, and up to a certain point, help recreate the complexity of discourse.

5. Renata Tesch (1990, 1991) defines a list of practical operations common to all types of qualitative analysis, though every approach favors different sets of operations. We regroup these operations into four logical categories.