CONTENT AND LEXICAL ANALYSIS:
A QUALITATIVE PRACTICAL APPLICATION

ISRC Working Paper 070498, April 1998

Henrique Freitas
Visiting Researcher at ISRC, University of Baltimore (USA)
Associate Professor, GESID - PPGA - Escola de Adm. da UFRGS and CNPq Researcher (Brazil)
Docteur "nouveau régime" – Université Pierre Mendès France (Grenoble, France)
hf@ea.ufrgs.br

Jean Moscarola
Director of Ecole de Commerce, Vente, Economie et Gestion, Université of Savoie, Annecy (France),
Doctor in Management (Université de Paris), Sphinx’s software Author
moscarola@univ-savoie.fr

Milton Jenkins
Professor and Director of ISRC, University of Baltimore
Ph.D. on Management Information Systems, University of Minnesota (USA)
mjenkins@ubmail.ubalt.edu

Abstract
This paper presents content analysis and lexical analysis: starting with general notions, it emphasizes their application and use. Its main contribution on is the organization and selection of the topics, and especially the illustrative cases. First, the technical stages of the content analysis are described, pointing out its value as a research instrument. Second, seven levels of lexical analysis are presented in a creative and evolutionary way, considering the use of computer software. Since the AIS 97 and ICIS 97 conferences, the IS world community has been debating the use of qualitative research by itself or preceding quantitative research studies. The potential contribution of these methods of data analysis will be made clear.

Keywords
Lexical Analysis - Content Analysis - Qualitative and Quantitative Data Analysis - Sphinx® Software

1 This is an working-paper prepared in cooperation between the ISRC (Information Systems Research Group, University of Baltimore, USA), the IREGE (Institut de Recherche en Gestion, Université de Savoie, Annecy, France) and the GESID (Grupo de Estudos em Sistemas de Informação e Apoio à Decisão), a research group from the School of Management (PPGA/UFRGS, Porto Alegre, Brazil), coordinated by Professor Henrique Freitas, whose focus is the discussion and accomplished meetings in Brazil, in France and in The USA, between Professor Jean Moscarola (France) and Professor Henrique Freitas, from 1994 to 1998; besides several contacts through Internet.

The software Sphinx® served as support to the accomplishment of the content and lexical analysis in this study. It is a software to help survey research and quantitative-qualitative data analysis, designed by Professor Moscarola and partners, and being improved with the support of the Brazilian team (Professor Freitas and his partner M. Cunha Jr.). The authors strongly acknowledge Mrs. Oveta Popjoy, Assistant Researcher at the ISRC, for the editing of this last text version. The research assistants CNPq of GESID/PPGA/UFRGS, Amarolinda Costa, Ângela Werner and Ionara Rech, as well as doctorate student Mirian Oliveira, and M.Sc. student M. Cunha Jr., also collaborated in an early edition of this text. This work is part of a research project and is supported by several Brazilian Government Research Agencies (CNPq, Fapergs, Propesp/UFRGS, PPGA/EA/UFRGS, and CAPES/COFECUB), by FULBRIGHT (USA), by Gereg (Université de Savoie, France), and by the ISRC/Univ. of Baltimore (Maryland, USA).
1. Introduction

Document analysis, which includes content analysis and lexical analysis, follow classic methods like the judicial and sociological research. It presents a major common rational characteristic, being more or less intuitive, personal, and subjective. Like others, the historical method has validity problems, such as the authenticity of the text, interpretation validity, and the truthfulness of statements of fact. Among others, all of these analyses, have the defect of a non-systematized work, depending only on the value and competence of the researcher.

With the proliferation of the means of communication (TV, radio, etc.), and especially of advertising on the Internet, some things became more obvious: the need to look for meanings, relationships, or even laws to measure influence factors; and the need to foresee certain events. A modern method of documents analysis was born from the necessity to accomplish studies in new directions: content analysis!

The general notions presented in this paper are mainly set in the texts of Grawitz (1976, 1993) and Gavard-Perret & Moscarola (1995), even though we can find some literature concerning this subject, which is still not so developed in the American as it is in the European academical world. We understand that, in the context of international cooperation, it is not our role to reinterpret the well-known, world-wide American literature, but mainly to present an illuminated reading of European and South American research in this field.

Except for Krippendorf (1980), the American literature is very recent. Weber (1990) authored a short paperback book entitled Basic Content Analysis. Silverman (1993) described methods for analyzing talk, text, and interaction, in a book entitled Interpreting Qualitative Data. Miles and Huberman (1994) authored a book Qualitative Data Analysis. The English Journal of Applied Management Studies published two papers on applying content or lexical analysis: Crouch and Basch, “The Structure Of Managerial Thinking” and Oswick, Keenoy, and Grant, “Managerial Discourses...”. MIS Quarterly also published a paper on content analysis during 1997, showing a clear new direction among our researchers and editors toward qualitative studies marking a new era of content and lexical analysis. The French literature is also receiving a breath of fresh air from Bardin (1996), and Lebart and Salem (1994). Any researcher conducting studies using these techniques would do well to take a look at the European literature.
In Section 2, general notions of content analysis are presented, and its application and use are emphasized in Section 3. In Section 4, the technical stages of content analysis are presented, also highlighting its value as a research instrument. The seven levels of lexical analysis are presented in a creative and evolutionary way in Section 5, considering the use of computer software. Having been inspired by qualitative data collected in research, some of the analysis techniques are illustrated in Section 6. The usefulness of, and interest in, content analysis in the context of qualitative research is presented in Section 7.

2. General Notions about Content Analysis

In this section, the definition of content analysis is described, as well as its characteristics, its sources, types, and trends.

2.1 - Definitions of Content Analysis

The resources that the social sciences offer for our reflection are essentially composed of oral and written communications. Oral communications include text of speeches, annotations and reports of interviews, and conversations. Written communications include official texts, newspapers articles, letters, everything in the literary field, even history and politics. During the course of research, we gather data to understand, to explain opinions, conduct, or actions. These data are almost always of verbal origin. The actions, when we apprehended them came from a context of words. Speeches and addresses are documented in words.

From our point of view, it is very important that the researcher in social sciences be able to analyze these data in a scientific way and not be satisfied to have a casual impression. The novelty of the modern procedure of content analysis consists in substituting the impression (a personal opinion, almost a guess), for more standardized procedures, quantification, in every case transforming raw documents into data for scientific analysis. For such task, the text will be decomposed; that is, it will be studied as a function of the words that it contains or the ideas it represents, the latter being chosen for reason of its relationship with the objective of our research (for example, “How many times did a political candidate use the term ‘social justice’ during the electoral campaign?”).

Berelson (apud Grawitz, 1993, p.534) defines content analysis as: “… a research technique for the objective, systematic and quantitative description of the obvious content of the communications, having for objective to interpret them”. In a general way, Grawitz (1976) complements this definition saying that we should consider dispensable the term “quantitative”. We would add to her comments that this “quantification” may be an easy way to compute and that we should seriously consider this possibility. We need to spend some time exploring the contents exactly to better understand what is “not so obvious”.

3
2.2 - Qualities and Sources of Use

Grawitz (1993, p.534) describes the characteristics which Berelson attributes to “objective, systematic, and quantitative” as following:

- **Objective:** The analysis should proceed according to the pre-established rules, obeying guidelines sufficiently clear and precise in a way that allows different analysts, working on the same content, to obtain the same results. This means, even if the analysts are in different places, at different times, and on different research teams, they agree on the aspects to analyze, the categories to use, and the operational definition of each category.

- **Systematic:** The whole content should be ordered and integrated in the chosen categories as a function of the objective pursued. Elements of information related to the objective should not be overlooked.

- **Quantitative:** It is to evidence the significant elements, to calculate its frequency, etc. This condition is dispensable; since when using qualitative data analysis certain analysts look more for themes than for exact measures of importance.

There are not many tools for content analysis, associating the qualitative with the quantitative. In the case described in this paper, the software used is Sphinx®, which offers several ways to navigate (or surf) the text to identify themes, to define new variables, or data provided by the researcher, which can be treated quantitatively. This software also performs crossing or two-way analysis with other contextual or more objective variables or data, facilitating more robust and precise inferences and conclusions. Also, it provides a range of tools for the codification of text, generating new data, through which we can identify dimensions of analysis, and even typologies starting from the base of the original text.

Content analysis can be a good technique to use in all information systems research, which is documented in written texts (official documents, books, newspapers, personal documents), in voice or image recordings (radio, television, etc.); or in other activities, that could be decomposed, such as a management meeting, or the use of any professional’s time. All material especially generated for psycho-social research (group discussions, depth interviews, and meeting reports, etc.) can also be subjected to content analysis. We have not yet found the limits to the application of content analysis and related software. The Internet provides a wealth of “free” data for researchers and curious data analysts to conduct a variety of investigations, which could generate unique and powerful information and could even lead to useful and rich conclusions.
2.3 - Types of Content Analysis

Since different sources of data are available, we can use a lot of models to explore, process and analyze data. Grawitz (1993, p.536-538) develops three different approaches: exploration versus verification; quantitative versus qualitative; and direct versus indirect.

Verification Versus Exploration

Grawitz (1993, p.536) insists that, in the social sciences, a clear distinction should be made between document analysis having as an objective hypothesis verification and that analysis whose objective is exploration or hypotheses definition. In hypothesis verification, the objective is precise and the results can be quantified. In hypothesis definition, the analysis is less rigorous and systematic, following rules and techniques which cannot be standardized, and it is based intuition and experience. Grawitz observes that when we want to classify certain relevant elements in a text or document in order to quantify them, we must know whether we are trying to elaborate or to follow our intuition to the detriment of any and all systematization. When there is a great mass of data, and we do not have hypotheses or clear research questions to serve as a guideline in the exploration of the text, the analysis should be conducted knowing that maybe we will be leaving essential (and unexpected) elements outside the the scope of the study.

Quantitative Analysis Versus Qualitative Analysis

According to George (apud Grawitz, 1976, p. 592), an important difference is that quantitative analysis try to accumulate the frequency of themes, words, or symbols, while qualitative analysis is based on the presence or absence of a given characteristic. Other qualitative-quantitative dichotomies: impressions versus systematization, hypothesis instead of verification, and flexibility as opposed to rigidity. What is important should be clear in each type of analyses. The “number of times” a data element occurs is what counts for the quantitative analysis, while a novelty, interest, theme, or subjective attribute is the object of qualitative analysis. We always have a dilemma: to adopt a few representative categories, or to regroup deliberately the data in a small number of categories, but sacrificing information.

Direct Analysis Versus Indirect Analysis

The direct quantitative analysis, more commonly used, consists of counting the responses just as they appear; the indirect quantitative analysis can sometimes, beyond what is had as clear and obvious result, to obtain for inference, even that the author left implied. This type of information (choice of the words, rhythm of speech, etc.) can be developed deeply, but demands time. We can so, starting from a quantitative analysis, to look for a subtler interpretation, or what is “latent” under the explicit language used in the text.
The indirect interpretation, which comes in addition to what is said, is not something only related to the qualitative world, it can be supported perfectly in a quantified context. Grawitz (1993, p.537) reinforces that this type of information can still be more precious when it is instrumental communication (destined to produce a certain effect on the receiver), in opposition to the communication purely representative (which informs about the state or situation of the originator or emitter). In the practice of content analysis, it is very important to know whether the message is representative (when we are satisfied in evidencing the apparent sense, without looking for “something more”) or instrumental (when we should vary the analysis strategy, always looking for its objective).

2.4 - Some trends... and the Internet

The technique of content analysis obtained flexibility with the technology progress, and it was adapted to various objectives and to different stages of research. Here we have some of the points discussed by Grawitz (1976), and we add the “Internet era”.

Several objectives: Using a text analysis of a psychotherapy interview, the researcher will seek to reach an instrumental interpretation, normally using indirect approach, of psychological and qualitative nature, of the patient’s unconscious elements. A content analysis of a newspaper will usually be direct, representative, quantifying the several items and developing the objectives of the newspaper and of the clientele which it serves.

Different stages: For hypothesis lacking a priori and precise objective of measure of one or another factor, the analyst will begin by familiarizing himself with his base documentation, reading it several times. This stage is unavoidably qualitative and intuitive; it will help to imagine categories that will be good to range several elements of the text and, after that, to quantify them.

The qualitative and quantitative worlds are not (or should not be) mutually exclusive; and yes, they are supposed to complement each other. The qualitative analysis supplies the necessary categories initially to a quantitative and rigorous content analysis. This is a controversial discussion going on in the IS world community (AIS 97 and ICIS 97).

Some Trends and Sources: The Internet is becoming a revolution in terms of data access. The problem remains ... which data? So, we still need a variety of user-friendly techniques and tools to “navigate or surf” better and more quickly around millions or billions of data and records, and maybe, only maybe, have the information we need. So many different issues and challenges, technological and organizational, will call for many more discussions. Everybody will be able to perceive when the real good information is on our screen. Everybody know the tools? Are we being educated to better perceive this new world logically different from the mainly quantitative one?
3. Content Analysis Application and Use

We can apply this kind of data analysis technique in several ways, which can involve different kinds of source, storage or media: oral communications, like TV, radio, discussions or debates, meetings, etc., or written communications, newspapers, books, papers, interviews, questionnaires, etc. Let's try to better understand this communication factor and also the field of application regarding this technique.

Concerning any communication, we will be in presence of an originator or emitter that throws a message possessing content and form, seeking to reach an objective, and addressed to one or several receivers. Who speaks? To say what? To whom? How? With what result? This is a simple outline proposed by Lasswell (apud Grawitz, 1976, p. 598) for the communication, which seems useful for the content analysis, the subject proposals covering again the group of problems.

**Who speaks?** ... or the study of the emitter (originator), where two situations are possible: the emitter reacts more or less to an incentive controlled by the observer, as in a guided or driven interview; or, more frequently, the analyst does not have any relationship with the emitter, and he looks at that based on the content of the emitted informations. Maybe we could want to identify the author or his characteristics, starting from the text-source.

**To say what?** It is to study the characteristics of the message content, to recognize the nuances and different directions of the content, to compare themes, speeches, slogans, or even the evolution of the author’s texts. Frequently, it is about to compare materials of different texts. The content analysis contributes strongly to the adapted elaboration of an interview, facilitating the code of the subjects of open-ended questions, or the reports of a meeting or group discussion, facilitating a systematization.

**To whom?** ... The receiver is determined, to who goes the message; could be interesting to identify even what is latent and not only clearly explicit (the study of a message of such leader is addressed to his own organization, but in fact he wants to pass a message to the competitors).

**How?** It is the study of the way which the text looks for producing an impression or message. **It is essentially something qualitative and subjective, that we take a look under a quantitative point of view.** The focus are the elements that contributes to produce this impression: choice of words, repetitions, composition of the sentence, etc. The categories will always be qualitative, but the analyst will quantify the associated data.

**With what result?** The point is to know the effect of the message on the receiver, or we want, with the right elements and in the best way, to influence a given public in a certain sense; or still, whether the strategies of the originator are clear, to foresee what he looks for and his implicit objectives. That is, sometimes it is explanatory or elucidative, some times is more predictive (or indicative).
The multiplication of the studies about communications, with the search of identification and understanding processes (besides the analysis of the explicit or “ordinary”), they took to consider in the analysis, besides the subjects above, the influence of the environment, even on the emitter as on the receiver, to measure the effect of a private communication. The content analysis can lift or same to solve great problems in the sociology area and social psychology (Grawitz, 1976, p. 603). As well as in different areas of management, as we can see in many of the applied references at the end of this paper.

4. Content Analysis Technical Stages and Value

As in almost every research, the first step is the idea itself, as well as its objective. The means of obtaining an answer is to ask a question, as in an interview. As the questionnaire allows us to guide an interview, the research analyst will select the categories to guide the analysis based on the data in the documents. Grawitz (1993, p.543-558) describes the main technical stages of content analysis, pointing out that we first need to choose categories, and then we need to start the main analysis, and evaluate content analysis as a research method (Sections 4.1, 4.2 and 4.3, respectively).

4.1 - Choosing the Categories

**Definition, Creation or Formulation:** The categories are the basis on which the content will be classified and eventually quantified. In the case of a survey, the categories can already be foreseen in code form. The content analysis should usually allow the appearing of variables and influence factors that we ignored in the beginning of the work. The categories, when we do not have a precise idea, should be defined based on the content. In an exploratory survey, the categories constitute themselves the object or context of analysis allowing us to select the data in the text; while in search of hypothesis verification, we usually have pre-defined categories. Grawitz (1976, p. 605) advises that we should “prolong the period of tests and establish several categorization systems or code before adopting an outline (a structure or system) of categories”. The categories should be derived from the original document, which is the object of analysis, or from some general field, or from activity in which the subject interferes. From the answers, in case it is an interview; from objectives, intentions, beliefs,, in the case of a text; all that still considering absent elements that could be relevant.
**Characteristics:** The choice of categories is the most important step in the content analysis; the categories are the connection between the research objectives and the results of the content analysis. The value of the analysis is contingent upon the legitimacy of the categories. The objectives of the research will guide the selection of what will be quantified. The categories should be exhaustive, mutually exclusive, objective, and pertinent. Exhaustive means the categories should include as much as possible of the text, if we put all categories together, we can make a judgement regarding the whole text, or we can have an idea concerning the whole text. Mutually exclusive means the same elements cannot belong to several categories. Objective means that the categories will have such clear characteristics that different analysts will obtain the same results with the same text. Pertinent means the categories relate the content of the text to the objectives of the research.

**Excesses to avoid:** One excess to avoid is the imposition an overly rigid outline for the analysis from the start that will prevent the comprehension of the complexity of the content. A second excess to avoid is preparation or definition of the outline in a superficial way, focusing on the explicit and ignoring the latent content. A third excess to avoid is choosing too detailed or too numerous categories, almost reproducing the text in order not to lose anything. The fourth excess to avoid is establishing categories too broad to let us sufficiently distinguish the elements. The goal is to approach an optimum number of categories or an optimum percent of the whole.

**Standardization:** It may be helpful to adopt a certain number of categories applicable to several cases; for example, age, gender, profession, religion, nationality, highest level of education. However, there is no algorithm for choosing categories that is appropriate for all cases. Researchers in some academic disciplines have adopted standards; such as, the marketing research under guidance from the American Marketing Association (AMA) and the European Association with the same purpose (ESOMAR). It is unlikely that two different analysts using the same content with the same starting point will obtain the same result. To be satisfactory, the outline of analysis should follow a general conceptual scheme accepted by peers. A standard illudes us because the specific content changes with each new research study.
4.2 - Problems with Quantification

Ambiguity, complexity, and variety of sources are problematic for content analysis.

**The sample problem:** These problems do not exist in the cases we studied. However, when the object of analysis is a text in its totality, or a series of speeches, when it is something as vast as several years of a newspaper, or as varied as advertising, we should define the research by choosing a certain type, a certain topic, or a certain section (only the editorials of a newspaper, for example). Or, then it cannot, instead of limiting the theme, to limit a sample that is representative of the whole. **Rules to select a sample:** (1) Define the universe to which the generalization is to be applied; (2) Make sure that each unit of this universe has a well-known probability of being contained in the sample; (3) Select a sample that is independent of all correlation among the units of the universe; and (4) Select a large enough sample to minimize sampling error.

**Representative Characteristics:** Perhaps we could define the sample as a function of the formulated hypotheses, but it would not be very scientific, and it would not include all of the factors involved. Berelson (apud Grawitz, 1976, p. 613) notes that the sampling problem implies three decisions with regard to analysis of communications: (1) **Choose the source.** The choice of the source depends on the objectives of the research. If the analyst wants to generalize conclusions from the results of the content analysis, he or she should be able to formulate the reasons for defining a limited scope of the universe. (2) **Choose the number and date of the messages.** How many messages constitute a representative sample? Over what period of time? The theme can determine the time, but not the period. The chosen unit of the universe under study should have a well-known probability of being contained in the sample, and of being independent of all correlation among the units of this universe. (3) **Observe the content.** The sample of the sample is difficult to define, given the variety of objectives; which drives us frequently to use “flexible” methods accepting in a same study lifted up sample types starting from different sources, different periods and even different contents. We should be careful at this point; but overall it is important to be honest in the research report.

**Units of Quantification:** If we plan to quantify occurrences of words in documents, we set up indexes to retain for classification and determine the length of the data elements, according to which the software will decompose the content. It is necessary to distinguish among the registration unit, the context unit, and the numeration unit.

The **registration unit** is the first unit of analysis. It is also the narrowest and of variable size. It can be a word, smaller unit, the number of occurrences, and in what context it was used. It can be a theme, significant fragments corresponding to the idea that recovers one of the categories. It is a less precise element than the word. It can have three lines or four pages, it will be then difficult to obtain results of great fidelity due to the great difference of an analyst’s perception from other for a same content. It can be an item, which means the total content of a book, a film, or a speech.
The context unit is subtle and more flexible and does not seek protection in a rigorous quantification. It is the widest content segment to which we refer when we want to understand the registration unit. For example, if the registration unit is the word, the unit of context can be the sentence.

The numeration unit is the document for which we will count the occurrences; the number of lines, for example; or the number of square centimeters in a newspaper article, etc. While the registration unit indicates the way of decomposing the content (significant element), the numeration unit implies the form of measuring these several “pieces”. Lacking a precise measure, we fall back on a numeration unit. To mark the difference between two themes, one appearing fifteen times in three lines, and the other appearing two times in one hundred lines, it is necessary to have a numeration unit (the number of lines, for example) different from the registration unit (the number of times the theme appears). The numeration unit refers to the space or the time: paragraphs, lines, centimeters, even minutes (radio, TV, etc) of registration.

4.3 - The Value of Content Analysis as a Research Instrument

Content analysis is a refined technique, therefore delicate, and it demands much dedication, patience, and time to satisfy the investigator’s curiosity, in addition to intuition, imagination to notice what is important, and creativity to choose the categories. At the same time, the analyst should have discipline and perseverance, rigidity when decomposing content or to count results of analyses. These are only some of the main aspects that Grawitz (1993, p. 553-558) points out, along with fidelity, logical validity, inference, and empirical validity.

**Fidelity:** By definition, content analysis should be objective, and the results should be independent of the measurement instrument, being convenient to minimize the differences in points of view among the analysts. But this is an old problem in the social sciences, the fidelity cannot be faced in the same way according to it is a quantitative analysis of a clear and obvious content or if it is an analysis more qualitative where we look for to identify latent intentions, where it is worth more the presence or absence of an element and not its frequency. In this case, it will be driven to minimize the subject of the fidelity. In the quantitative analysis, the analyst will treat at first the frequency of the elements, not mattering so much with the validity of isolated data.

**Logical validity:** Does the instrument measure what it intends to measure? An analysis is valid when the quantified description that it offers concerning the content (object of the study) is significant for original problem and it reproduces faithfully the reality of the facts that it represents. Of course, it is essential condition of the representativity of the sample and that the inherent technical conditions to each stage are observed satisfactorily. The more limited or defined the objective, the easier it will be to confirm the validity; and it is easier to demonstrate validity in a quantitative analysis than in a qualitative one.

**The inference:** This point deserves special attention because sometimes one expression has more than one interpretation, even positive and negative interpretations, depending on its context.
Empirical Validity: Is the study prediction just or precise? A difficult question to answer. Instead of convictions, prudence and humility are recommended when drafting the conclusions. However, Grawitz believes that the experience and the analyst’s training will countersign the value of its analyses.

5. Placing in evidence certain potentialities, applications and limits of the textual data analysis tools

Our aim is to evaluate the usefulness of content analysis and lexical analysis. For example, content analysis can be used to analyze in depth each person’s or group’s specific expression involved in a debate. The work on an organized “lemmatized” text allows us to use grammatical categories -- nouns, verbs, adjectives-- to organize the first impression of the content of the text.

The “lemmatizor” or “lemmatizator” is an English translation of the French “lemmatiseur” (used by Professor Jean Moscarola and his research team from Université de Savoie, France) representing a computerized tool that aids the analyst marking in the text nouns, verbs, adjectives, names, etc. in the text. We will use the term text “lemmatized” to describe the text that was already processed with the “lemmatizator”.

The banality (or triviality) of lexical indicators can be shown in the specificity that guides the selection of the most trivial sentences, or the most significant or original words, from any group or category. With these resources, we can also build ad hoc indicators to evaluate the degree of commitment regarding the actors or members of discussion groups.

In view of the progress in computer software these analysis techniques are easy to use. We can easily “surfing” or navigate inside a content, and we can quickly have some statistics regarding the lexicon (all the words within a text with their frequency of appearance). A great variety of investigation and even new readings are possible using the software meantime available to the researchers and analysts, it should not mask or hide the difficulty of the inherent final interpretation to this activity. Dedication and patience, allies to curiosity, are the most important characteristic of the researcher in this field.

Evolution of Textual Data Analysis: Content analysis tools make available to the researcher new work methods combining:

- the statistical lexical and the data analysis methods,
- the syntactic analysis,
- the lexical surfing and the reading assisted by computer.

The software Sphinx Lexica® (1997), one of the available software tools, integrates these approaches in the same environment (see Main Functions in the Appendix), and it returns the reader to the text instantly. It distinguishes seven levels of analysis based on Moscarola (1995, p. 7-8). Independent of the computer platform, and the technique of exploring the qualitative data, the concepts presented here can contribute to a better way of doing content analysis.
Seven Levels of Lexical Analysis:

**Level 1: The summary lexicon approach.** Reduce the body of the original text to the “high” or top of the lexicon (for example the top 25), which means that examining only the most frequently occurring words we can have us an idea of its content. By lexicon we mean the list of all words of the text used with the number of times that each occur.

![Figure 1. The Summary Lexicon Approach](image1)

**Level 2: The controlled lexicon approach.** Reduce the text to its lexicon (each word of the text and its frequency), and to control, through lexicon surfing or navigation, the validity and the foundations of the interpretations elaborated, starting from the lexicon. See Figure 2. By “verbatim” we mean that we can, if needed, export or transfer some of the contents to a report or file in order to use them to illustrate something. For example, export to a text file all comments, or opinions, or phrases that contain the word “love”. It depends on the analysis unit.

![Figure 2. The Controlled Lexicon Approach](image2)
Level 3: **The Selective Lexicon Approach.** Work on a reduced lexicon after having eliminated or deleted the tool words (definition of those words without a role or meaning something useful, all that we can ignore, like prepositions and articles, etc, the whole stored in a specific dictionary), concentrating then our attention on nouns, verbs, adjectives; everything by use of dictionaries and the “lemmatizator”. We can also create dictionaries with the words that we have a special interest (like positive or negative arguments, or words regarding a special field, like health, or other). We can also group words in a new word or concept (with a new name to the group or even with the name of the most frequent word).

Figure 3. The Selective Lexicon Approach
Level 4: Lexicon Statistics And Text Quantification. The Research of the lexicon characteristics, to establish the statistics of the words of the text according to an external non-textual variable, like... “department” (marketing, human resources, research and development, others), gender (male / female), age (under 20, 20 and over), or level of education (graduate, undergraduate), for example.

Figure 4. Looking for Lexicon Characteristics
**Level 5: Codifying the Lexicon and Generating Measures.** Describe the text by a nominal variable, created by a codification based on the presence in the answers of a group of selected among the words from the lexicon. We can do this kind of procedure many times, each time with a different purpose, each time generating new variables, closed this time (and not open-ended), which we can also cross with others like gender, age, etc., in a bi-variate analysis procedure.

**Level 6: Quantification of text.** Calculate the number of times certain words occur, starting from a “lemmatized” text and already with the tool words taken out of the lexicon (by taking out of the text the words that do not interest us using dictionaires, for such a task using special user-friendly procedures or “clicks”). It can allow us to know if certain opinion is “unique” or if it is a regular or common one.

---

**Figure 5. The Lexicon Codes and Measures, and the Text Quantification**
**Level 7: The Multivariate Analysis of the Textual Data.** Data analysis made on the variables extracted from the lexicon. We can apply the methods of multivariate analysis using the new variables (lexicon tables): factor analysis or automatic classification can be done. We can even consider the variables in relation to other contextual variables (like gender, education, age, etc), and thus integrate classical data analysis with the textual data analysis.

![Diagram](image)

**Figure 6. Data Analysis of Variables Extracted from the Lexicon.**

The access and the “surfing” regarding the text or document data or the qualitative research methods is then facilitated. The tools are now powerful. However, the analysis of textual data does not change the meaning of the data. Even if analysis reduces the “noise” in the data, the reduction of the long, and sometimes annoying readings, the faster production of conclusions and reports drives everybody to an *interpretive* reading. *Interpretation* is often dangerous since it is fast and it appears to be falsely objective. We should so be careful in our investigation and conclusions. That is a Moscarola’s observation (1995).
6. Practical Application of Qualitative Analysis in Management Field

In order to illustrate some of the content analysis applications described in the previous section, we will now analyze different cases, the first one concerning a diagnosis regarding Quality Control in an industrial environment in a South Brazilian State (Section 6.1); and others involving information to better define IS planning in an organization (Section 6.2), involving decision-makers’ profile (Section 6.3), involving a American presidential election debate (Section 6.4), and still some other examples (Section 6.5). Our goal here is to illustrate only the content analysis and lexical analysis principles, and not to further develop these cases.

6.1 - Evaluating Program Quality in an Industrial Environment: 120 South Brazilian Enterprises

We extracted some qualitative data from research accomplished in 1994 for SEBRAE/RS and FIERGS (IEL), two non-profit organizations representing almost all small, medium and large enterprises in the State of Rio Grande do Sul (South of Brazil). This study was coordinated by the Graduate Program of the Management School, called PPGA/UFRGS (Ruas, Pires, Freitas, Cunha & Antunes, 1994).

The research, whose goal was to assess the Quality Programs in the industry of the State of Rio Grande do Sul (RS)/Brazil, lasted four months involving 120 different companies (30 small, 30 medium and 60 large), with interviews concerning the industrial director, and also with the director of quality program (when the enterprise had one), or the director of human resources. Almost 300 different variables were stored after interviews, almost one-third text or open-ended questions. The subject submitted to this content analysis illustration is relative to only one of these questions, which is “the environment monitoring in that the organization interferes”. The industrial executive or manager was questioned on “which are the approaches used to monitor -- or look at -- the market?”, if the company accomplished such a task. Even though the vocabulary is “Brazilian”, the subject remains worldwide.

The first analysis that we can do links the Level 1, summary lexicon; with the Level 2, the controlled lexicon approach; it is illustrated in Figure 7. By Level 1, we mean go from the pure text to the lexicon that the software helps us to build. So, since we have a “lexicon” generated, we can know some details regarding these data, like the body of the text (formed by all the answers done or emitted by the 120 interviewees), the whole body of the text is constituted of 374 words, 184 different words constituting the “lexicon”. Concerning the Level 2, it can be analyzed through “lexicon surfing”.

2 The authors acknowledge Professor Roberto L. Ruas (PPGA/UFRGS, Brazil) for the data concerning this case, and especially acknowledge Marcus Cunha Jr., M. Sc. Student at Management School, PPGA/UFRGS (Brazil) for his support in elaborating a first Brazilian draft of the data analysis concerning this item 6.1, which was published by FREITAS, CUNHA Jr and MOSCAROLA (1997). A Correspondence Analysis literature review can be found on CUNHA Jr. (1997).
The objective is to have a good argument of the specific context in which interfere each one of the words used in each emitted sentence. For example, the word “price” and the word “prices” combined or grouped appeared 19 times. To understand in which way or sense this word was so used compared to the other words, the body of the text was traveled or “surfaced”, that is, each one of the answers, by means of the software, which allowed to decide for the “grouping” of all these words in only one that expressed its idea (“#prices”, in the case, where the sign “#” meaning group of words). It could be possible at this second level to have all the opinions or answers that used the word “price” in their content, and then create a new file or transfer to a report these opinions: this kind of “surfing” would then show us that the most of them (16 from 19) answered right away “price”, and that only 3 answered “price and quality”, for example.

![Figure 7. Screen of Lexicon Surfing, Sphinx Lexica®](image-url)
When accomplishing this grouping of the same words (singular and plural, for example), or with same meaning (customer and consumer, for example), it starts already to operate by selective lexicon approach, following the logic of Level 3 (Figure 3). In this case, the aim is to focus on the investigation in a limited number of arguments, being excluded, through an appropriate tool, the words that don’t contribute to the analysis we want to do. We can simply mark and delete all words we don’t want to keep in the list of words (the lexicon), or we can ask the software to mark in the lexicon all words we have already stored in some especial dictionaires and then delete them from the lexicon. We can group all the “same” ideas or words with the same meaning or because they are written in a different way, etc. We can, at this Third Level, mark a certain word (or group) to obtain a “verbatim” (which is the group of interviewee’s sentences that contain the words selected for analysis). If we marked in the lexicon the word “salespersons” (rearrangement of “salespersons” and of “representatives”, in the case of our analysis) and we requested the “verbatim”, the result would be four cases:

| #salespersons (4) |
| --- | --- | --- | --- | --- | --- |
| interviewee 1: Through a questionnaire to the customers, and of the REPRESENTATIVES. |
| interviewee 8: Comments, the customers’ word, SALESPERSONS, etc. |
| interviewee 67: Meetings with REPRESENTATIVES of sales. Sit. market and relationship with consumers. don’t have formal questionnaire; 95: survey to the customers and vendors. |
| interviewee 73: Acceptability by the customer REPRESENTATIVES. |

Of course, each of the Levels already described and illustrated in Figures 1 through 6 could have much more functionality in each of the software package available on the market. Like the “lemmatization” in Level 3, which means that we should be able to use a tool asking to have on the screen only the verbs the respondents used to answer a certain question, which could be useful.

For the illustration purposes, let us consider, that all the exercises of rearrangement of words were accomplished, resulting in the following cast of words (each one representing a certain amount of ideas, or even answers), all of them potentially meaningful for the objective of the investigation being done. The symbol “#” represents that the words were rearranged, and the number between parentheses after each word represents the number of times that each one of them was mentioned by the sample.

<table>
<thead>
<tr>
<th>#Prices (19)</th>
<th>#Customers(15)</th>
<th>Market(12)</th>
<th>#Sale(10)</th>
<th>Participation(9)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality(9)</td>
<td>#Research(8)</td>
<td>#Competition(7)</td>
<td>#Financial(7)</td>
<td>Productivity(7)</td>
</tr>
<tr>
<td>Product(5)</td>
<td>#Benchmarking(3)</td>
<td>#Salespersons(4)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

20
Several kinds of process models could be applied to this keywords list, which could be important for the analysis. We will describe how to proceed, using some of these techniques of qualitative analysis in Level 4, looking for lexicon characteristics (Figure 4), and Level 5, codifying the lexicon and generating measures (Figure 5).

Levels 4 and 5 allow the search for a better understanding of the opinions, the phenomenon, or in this case the variable that we want to analyze. A better understanding that can be obtained by the single crossing (two-way analysis) with an external variable: in this case, we will consider the size of the involved organizations: small, medium, and large. This approach claims for a codification, or the creation of a new variable with multiple answers, denominated of “evaluation of the competition approach”, and whose answer categories were the words in the top of the list (lexicon), each one with the identification if YES or NOT it was present in the answer of each one of the interviewees.

Once this task was done, took place the two-way analysis or crossing between the variables “size” and “approach of evaluation of the competition”, being opted immediately by the representation in a factor chart or map (CA - Correspondence Analysis; Burt, 1950; Carrol, Green and Schaffer, 1986, 1987; Hoffman and Franke, 1986; Cunha Jr., Freitas and Slongo, 1995). Our intention is only to demonstrate the interest that can have in associating qualitative techniques with quantitative ones. In fact, different techniques of data analysis are being developed to allow the quantification of categorical variables in studies in different fields.

“The CA (bi and multivariate) can be pointed out due to its flexibility in operate with data of the categorical type, without violating mathematico-statistical principles of the techniques. This property allow the analyst or researcher to use the own resources of these techniques as well as use them as the base for its integration with others developed for quantitative data” (Cunha Jr., 1997). It is a special kind of MDS - multi-dimensional statistics, which can be interpreted as a MCA (main component analysis).

The relations between the variables can be identified based on the associations between each one of the “correspondent unities”, or between row-points and column-points. We are then able to represent the whole data in a chart; if the sum of the first two factors or axes is (almost) 100%, it means that it is not necessary to look for a 3rd or 4th factor or axis because we have already (almost) all the information represented in the factor chart. The authors already cited, developed some interesting concepts and approaches concerning factorial analysis of qualitative data, with graphical interpretation, representing interpoint distances comparisons, especially with a geometric interpretation of CA (Greenacre and Hastie, 1987), or perceptual mapping using the basic structure matrix decomposition (Fox, 1988). Anyway, it is perceptual and intuitive maps (Hair, 1994).
Concerning the “names” of each one of the axis, is never so easy because we are working with different kind of data. Normally, it will be “a posteriori” that we will have better conditions to define names, looking after the positive and the negative contributions for each factor. In the chart below, for example, we have a main axis explaining or containing more then 87% of the information in the two-way or contingent table, so it can be defined like “size of the organization”, going from “small” (in the left) to “large” (in the right). But in almost all cases, we will need to spend some time looking for details, trying to point out something based in the geographical position and in the proximity of the categories, inferring then something interesting, or even going back to the original data to only then find an adequate name to each axis. It will be very difficult unless we have “a priori” definitions like “price-quality”, or still like “tradition-innovation”.

![Figure 8. Correspondence Analysis of the Crossing: Size Versus the Market Monitoring Approaches.](image)

Really “quickly” analyzing the Figure 8, we can see that: the large companies are using more easily of indicators that can be quantified (productivity, participation, financial), while the small companies and the medium size ones are using sources more related to qualitative factors (customers, salespersons, competitive, etc). All the analyses based on the texts were accomplished without a previous reading of the originals, demonstrating the productivity gain we can have using the tool adapted to this kind of analysis. It can be done in a very very short period of time by the manager or analyst.

Let us finally see what we can do in the Levels 6 and 7 (Figures 5 and 6) of lexical analysis, mainly regarding data analysis based on the new data generated from text data, and still quantification.

The Levels 6 and 7, of lexicon statistics, or text quantification, and mainly of multivariate analysis, allow a lot in the search of the better understanding of opinions (variable that we want to analyze) through the calculation of “lexicon intensities” of certain fields, starting from a debugged text (from which we already took out the words without clear meaning for the analysis in subject). It can take place multivariate analysis thought the crossing of variables originated or based on the lexicon with other contextual variables.
In Figure 9, we have a Main Components Analysis (MCA, Moscarola, 1991, 1993; Lagarde, 1995), where the existent associations are verified between the lexicon intensities of the selected words and the branch of companies activity. The lexicon intensity is an index calculated by a data analysis software, starting from the number of words in the answer to a subject and the number of repetitions of the selected word in this same answer, being the intensity lexicon the relationship word repetition and size of answer.

Let us consider the branch of activity of the involved organizations for example: metallurgy, chemistry, etc. An analysis of the Figure 9 allows us to verify which of the main indicators of evaluation of market are being used in the involved branches. The circles allow to visualize the existent groupings.

![Figure 9. Main Components Analysis Involving Branch of Activity of the Company and the Lexicon Intensity of the Selected Words.](image)

Some potential application of the theoretical concepts concerning content and lexical analysis were illustrated with the “Quality Program” example. We can still explore the concepts with some appropriate tool, fitting to the researchers the conduction of the process focused in their research objectives. We present some other cases, but only to show the reader some of the several and different potential applications of this quanti-qualitative method and tools.
6.2 - Content Analysis, crossing qualitative and quantitative data to improve the organizational IS planning

In a large organization, we could ask some questions to people from each department or managerial level, and then try to point out what are the main critical factors of success, or the pitfalls, or the most important barriers, for example. Let's keep the “pitfalls”. They would answer freely, and then we entry the answers, and do a “content analysis” as indicated in the Figure 10 below (left picture): we read the answer and create a new code from there, generating a new variable (closed this time, no more open-ended like the original data), the one-way analysis sorted is also illustrated below (right picture). This new objective data, we can cross with the “management level” or other variable, as illustrated below, the contingency table, and the factor chart; considering that we can “clean out” some of the elements of the table before transfer the results to the report.

### The text variable and the new code from there...

<table>
<thead>
<tr>
<th>Category</th>
<th>Code</th>
<th>No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lack of resources at the highest levels</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Users should accept</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Management level</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Production units</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top management</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IT staff</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### The new variable with the main ideas (one way)

<table>
<thead>
<tr>
<th>Potential Pitfalls</th>
<th>No. ans.</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lack of resources</td>
<td>48</td>
<td>16.72%</td>
</tr>
<tr>
<td>Lack of input by all users</td>
<td>39</td>
<td>13.59%</td>
</tr>
<tr>
<td>Departments’ needs are different</td>
<td>32</td>
<td>11.15%</td>
</tr>
<tr>
<td>Inflexible</td>
<td>29</td>
<td>10.10%</td>
</tr>
<tr>
<td>Politics</td>
<td>24</td>
<td>8.36%</td>
</tr>
<tr>
<td>User/Management acceptance/assistance</td>
<td>19</td>
<td>6.62%</td>
</tr>
<tr>
<td>Not up-to-date/time constraints</td>
<td>19</td>
<td>6.62%</td>
</tr>
<tr>
<td>Lack of leadership/bad mgmt</td>
<td>14</td>
<td>4.88%</td>
</tr>
<tr>
<td>Incompatibilities among departments</td>
<td>14</td>
<td>4.88%</td>
</tr>
<tr>
<td>Centralization/imposition</td>
<td>13</td>
<td>4.53%</td>
</tr>
<tr>
<td>No cooperation/coordination</td>
<td>13</td>
<td>4.53%</td>
</tr>
<tr>
<td>Lack of communication</td>
<td>12</td>
<td>4.18%</td>
</tr>
<tr>
<td>Ignoring existing infra/project</td>
<td>11</td>
<td>3.83%</td>
</tr>
</tbody>
</table>

### Crossing the new variable with the “management level”

<table>
<thead>
<tr>
<th>Potential Pitfalls</th>
<th>Top management</th>
<th>IT staff</th>
<th>Production Units</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lack of resources</td>
<td>33.33% (6)</td>
<td>13.33% (2)</td>
<td>24.39% (40)</td>
<td>24.37% (48)</td>
</tr>
<tr>
<td>Inflexible or politics</td>
<td>16.67% (3)</td>
<td>20.00% (3)</td>
<td>28.66% (47)</td>
<td>26.90% (53)</td>
</tr>
<tr>
<td>Not up-to-date/time constraints</td>
<td>16.67% (3)</td>
<td>0.00% (0)</td>
<td>9.76% (16)</td>
<td>9.64% (19)</td>
</tr>
<tr>
<td>Incompatibilities among departments</td>
<td>0.00% (0)</td>
<td>20.00% (3)</td>
<td>6.71% (11)</td>
<td>7.11% (14)</td>
</tr>
<tr>
<td>Lack of leadership</td>
<td>11.11% (2)</td>
<td>6.67% (1)</td>
<td>6.71% (11)</td>
<td>7.11% (14)</td>
</tr>
<tr>
<td>No cooperation/coordination</td>
<td>5.56% (1)</td>
<td>20.00% (3)</td>
<td>5.49% (9)</td>
<td>6.60% (13)</td>
</tr>
<tr>
<td>Centralization/imposition</td>
<td>5.56% (1)</td>
<td>6.67% (1)</td>
<td>6.71% (11)</td>
<td>6.60% (13)</td>
</tr>
<tr>
<td>Lack of communication</td>
<td>0.00% (0)</td>
<td>13.33% (2)</td>
<td>6.10% (10)</td>
<td>6.09% (12)</td>
</tr>
<tr>
<td>Ignoring existing infra/project</td>
<td>11.11% (2)</td>
<td>0.00% (0)</td>
<td>5.49% (9)</td>
<td>5.58% (11)</td>
</tr>
<tr>
<td>TOTAL</td>
<td>100% (18)</td>
<td>100% (15)</td>
<td>100% (164)</td>
<td>100% (197)</td>
</tr>
</tbody>
</table>

Illustrating the two-way analysis

24
Figure 10. One-Way, Two-Way, and Correspondence Analysis from Content Analysis, One Example
6.3 - The decision-maker profile: Brazil, France, and USA, using multivariate CA with categorical data

In a cross-cultural exploratory survey study, involving respondents and researchers from Brazil, France, and USA, we collected 72 different data elements from each one of the 134 respondents (50 each from France and Brazil, and 34 from the USA), 24 being open-ended questions. We used multivariate correspondence analysis with categorical data to learn something concerning the perception the managers had about the decision-making process (Macadar, Costa, Freitas, Becker, and Moscarola, 1997). It is a study based on Hofstede (1991), national culture, and on Simon (1987, 1997), decision-making process; since the new global reality is presenting many questions (Thurow, 1996).

This case illustrates the “corresponding units”, and at the same time the row-points and the column-points, generated after each one of the categories of all variables involved in the multivariate correspondence analysis being done here: Figure 11 shows how a Burt (1950) table is built. The factor chart (Figure 12) is based on the Burt table. This is only to illustrate what we can do with this method and tools.

| L1 | L2 | C1 | C2 | C3 | V1 | V2 | V3 | V4 | V5 | V6 | V7 | V8 | V9 | V10 | V11 | I1 | I2 | I3 |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 81 | -  | 35 | 33 | 13 | 42 | 33 | 34 | 16 | 26 | 33 | 20 | 22 | 15 | 7  | 5  | 29 | 17 | 35 |
| 41 | -  | 12 | 17 | 16 | 22 | 16 | 11 | 11 | 9  | 7  | 12 | 9  | 3  | 2  | 11 | 14 | 16 |    |
| 35 | 12 | 47 | -  | -  | 24 | 15 | 14 | 7  | 18 | 23 | 17 | 14 | 10 | 6  | 0  | 16 | 6  | 25 |
| 33 | 12 | -  | 45 | -  | 23 | 26 | 28 | 10 | 13 | 12 | 6  | 10 | 11 | 0  | 7  | 12 | 16 | 17 |
| 13 | 17 | -  | 30 | 11 | 14 | 8  | 10 | 6  | 7  | 4  | 10 | 3  | 4  | 0  | 12 | 9  |    |    |
| 42 | 16 | 24 | 23 | 11 | 58 | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | 22 | 13 | 23 |
| 33 | 22 | 15 | 26 | 14 | -  | 55 | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | 24 | 13 | 18 |
| 34 | 16 | 14 | 28 | 8  | -  | 50 | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | 19 | 12 | 19 |
| 16 | 11 | 7  | 10 | 10 | -  | -  | 27 | -  | -  | -  | -  | -  | -  | -  | 8  | 7  | 12 |    |    |
| 26 | 11 | 18 | 13 | 6  | -  | -  | -  | 37 | -  | -  | -  | -  | -  | -  | -  | -  | -  | 10 | 11 | 16 |
| 33 | 9  | 23 | 12 | 7  | -  | -  | -  | -  | 42 | -  | -  | -  | -  | -  | -  | 11 | 12 | 19 |    |    |
| 20 | 7  | 17 | 6  | 4  | -  | -  | -  | -  | -  | -  | 27 | -  | -  | -  | -  | -  | -  | 8  | 6  | 13 |
| 22 | 12 | 14 | 10 | 10 | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | 11 | 13 | 10 |
| 15 | 9  | 10 | 11 | 3  | -  | -  | -  | -  | -  | -  | -  | -  | 24 | -  | -  | -  | -  | 10 | 3  | 11 |
| 7  | 3  | 6  | 0  | 4  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | 10 | -  | 3  | 3  | 4  |    |
| 5  | 2  | 0  | 7  | 0  | -  | -  | -  | -  | -  | -  | -  | -  | -  | 7  | 1  | 3  | 3  |   |    |
| 29 | 11 | 16 | 12 | 22 | 24 | 19 | 8  | 10 | 11 | 8  | 11 | 10 | 3  | 1  | 40 | -  | -  |    |    |
| 17 | 14 | 6  | 16 | 9  | 13 | 13 | 12 | 7  | 11 | 12 | 6  | 13 | 3  | 3  | 3  | -  | -  | 31 |    |    |
| 35 | 16 | 25 | 17 | 9  | 23 | 18 | 19 | 12 | 16 | 19 | 13 | 10 | 11 | 4  | 3  | -  | -  | 51 |    |    |

This is a frequency table (Burt’s table) for the 19 value labels. Non-responses have been excluded.

L1 : Language_Yes  L2 : Language_No  
C1 : BRAZIL  C2 : FRANCE  C3 : USA  
V1 : #decide_solve  V2 : #evaluate simulate  V3 : #choose  V4 : #think  V5 : #Act_to do  V6 : #manage or control  
V7 : #conceive_dev  V8 : #inform_debate_discuss  V9 : #heard_feel*  V10 : #plan  V11 : anticiper  
I1 : Individually  I2 : Both(Ind-Collect)  I3 : Collectively

Figure 11. The Burt Table Looking for Decision-Maker Profiles
Each axis on the factor chart in Figure 12 represents a factor. In this case, the table is not so strongly represented. The first two axes represent only 6% of the table information. The factor chart is based on the Burt table, and with these factors, we have shown some differences among the countries which deserve careful investigation. It is the beginning of an in-depth analysis that could be performed, and we can start by reading geographically the behavior or perception of each country regarding the three other variables (languages, verbs_most or actions, and individually-collectively). The multiple correspondence analysis is described by Hoffman, De Leeuw, and Arjunji (1994). Related techniques are described by Lebart, Morineau, and Warwick (1984). An example is developed by Strutton and Pelton (1994).

![Factor Chart](image)

**Figure 12.** The Factor Chart after a Multiple Correspondence Analysis: Finding Decision-Maker Profiles
6.4 - Data Mining with Speech Qualitative data: The Election Debate “Dole versus Clinton”

This is a very interesting field that will continue to be developed for far beyond what it is today. Using the Internet, we can very quickly download the speeches from a political, social, or whatever discussion, and make many different lexical analysis. Gavard-Perret and Moscarola (1995) analyze the Bush-Clinton USA presidential candidates debate. Later, Moscarola also explored some data regarding the Dole and Clinton elections.

We have at least two possible applications of content analysis and lexical analysis in this cases: the content itself or the main ideas, and some statistics based on the texts, or the speech-acts. We can know very easily how many words each one spoke, what was the average length in words of each speech or contribution in the discussion, whether they repeated each one of their main arguments, whether they said something “unique”, whether they were “original” in their speech, etc. We can even look only for the verbs, meaning the actions ... same thing regarding adjectives or nouns. We will then be able to identify the objects or figures of each speaker, their main keywords, their main messages to the people, their main ideas concerning some important subjects like education, social security, health, etc. and then compare them this way!

Are they really discussing something? Are they only playing a game where they have known very well what they would like to say long before the debate, which means were they ignoring their opponent during the discussion? We can find answers to these questions by making lexical analysis of their speeches! For example, if we divide the Dole-Clinton speech data into five equal parts, and cross it with the main arguments used by both candidates, we have the illustration in Figure 13.

![Image of Figure 13](image-url)
• Lehrer spoke 957 words, with a mean of 11 words in each intervention (phrase length),
• Clinton spoke 7,962 words, with a mean of 19.5 words in each intervention,
• Dole spoke 8,488 words, with a mean of 14.3 words in each intervention.

We can also point out that Clinton said “YOU” 52 times, and Dole by his side said “YOU” 157 times. So, what does it mean? All other words (I, WE, PEOPLE, AMERICA, AMERICAN) were spoken more or less the same number of times by each one of the candidates. This constitutes some interesting to further investigate: can we find a different content “behind” the “YOU” used by each one of them?

We can identify the main repeated segments or expressions very easily, they were for example:
• Clinton: worked hard, crime bills, middle east, better of than we were four years ago.
• Dole: economic package, people watching, United Nations, Mr President.

We had already divided the speech data into five equal parts, creating a new variable (PART). After that, we created another new variable, merging the options Dole and Clinton (variable SPEAKER) with each one of the five parts (variable PART), arriving at the illustration in Figure 14. This Figure show us that they never really discussed! In fact, if we link each one of the parts regarding Clinton, and doing the same with the parts concerning Dole, we can verify that each one gave the people his own message. There are no lines crossing their main ideas during the whole speech. Who could imagine that or figure it out quickly? How many television viewers even noticed that it was not a debate?

Figure 14. The Main Ideas in Each of the Five Parts of the Speech-Debate, and the Discussant: a Non Debate?

The chart in Figure 14 was also produced by the technique called correspondence analysis (CA). CA is explained by Benzécri (1973), and by Cheung (1994). CA theory and application were described and illustrated by Greenacre (1984).
6.5 - Other examples: Internal communication, Data mining, Hospital, Advertisement

Moscarola (1991, 1993, 1994, 1995) conducted several different studies, like the internal communication in a bank compared with the customers’ expectations (totally different!) by Bachelet and Moscarola (1995). You can ask access to these studies directly through the World Wide Web at “http://www.lesphinx-developement.fr”.

Data mining proceedings of AIS 97 and of ICIS 96 from Internet: The proceedings of several world conferences are now available via the Internet or on CD-ROM. For instance, the program, and even the abstracts and papers from the AIS 97 conference in Indianápolis, USA, and from ICIS 96 in Cleveland, USA. After import the text or ASCII data from Internet servers, we were able to define a database structure (with fields like “meeting”, “session”, “authors”, “title”, “abstract”, etc), and so have available this data in order to compare the main subjects of each conference. Also, we can have a better idea on “how worldwide they are”, as well as identify some subjects of our interest, as in a literature search. We distributed a short comparative report at the ICIS 97 in Atlanta, USA concerning this data analysis (available upon request from hfreitas@portoweb.com.br).

Stumpf and Freitas (1996) conducted a two-year study inside a university hospital, using a Focus Group technique to collect data from six different groups of professionals: medical care and medical teaching; nursing care and nursing teaching; management; and research. The data were analyzed and used to define a new patient record and then to make important modifications in the hospital information systems. The results of this study were presented in Argentina, Australia, Brazil, and Canada. Savings were realized in the Hospital where the new patient record was adopted.

Domenjoz, Gavard-Perret, and Moscarola (1995), in a paper “Price and communication: how do they interrelate? an analysis of car advertisements published in English”, analyzed the content of 52 car advertisements using an observation protocol. They combined descriptive and subjective observations of the media and the visuals coded into a grid with exhaustive recording of all the text. Both data analysis and lexical analysis were applied in order to determine, first, how price is expressed (content, position, related vocabulary, context), and secondly, how price range determines specific types of messages or themes in relation to the media and the markets. The major conclusions drawn are that price is a multiform variable playing a definite role in the communication mix, and that price range has a major impact on the style and themes of advertisements.
7. Conclusion: The Usefulness of Content Analysis

Due to its complexity, it should be investigated in which cases content analysis or lexical analysis should be applied. According to Berelson (in Grawitz, 1976, p. 623), it should be applied to all cases that require great precision and objectivity. The comparisons and the evolutions form the main content analysis field. When sufficiently defined and detailed, content analysis allows us to pass of the simple description and to reach the objective of every scientific research: the discovery of explanations and causal relationships.

The value of content analysis depends on the quality of the conceptual elaboration done a priori by the researcher, of the exactitude with which it will translate itself in variables, of the analysis outline or categories, and also of the agreement among the reality to analyze and these categories. So that such analysis deserves to be accomplished, the subjects guided by the categories should show an interesting hypothesis and they should correspond to the selected material.

Anyway, it is time for “multiple methods” (Sawyer et al, 1997), it is more and more clear that we need be in touch with real situations using real data and looking for “Rurr” (really useful rigorous research), following Brown et al (1997) in order to build better relationships. It is also true that we look for relevance in our studies, as Manddviwalla and Gray discussed in a very recent paper published in the IRMJ (Information Resources Management Journal, v. 11, nr. 1, Winter 1998). In this paper, other authors share important points of view like: Beyond rigor and relevance (Robey, D. and Markus, L.), The application of IT research to organizations (Kavan, C.B.), and The challenge of relating IS research to practice (Senn, J.).

Overall, it is time to go ahead with many more qualitative studies, and to educate our managers, starting with our children, that the world is not only quantitative, but also qualitative. At least, what good quantitative study could not have been preceded by a qualitative one? Creswell (1998), and also Kirk and Miller (1986) offer us some concepts and discussions concerning qualitative research and mainly about reliability and validity in this kind of study. By the way, the the MIS Quarterly (v. 21, no. 3, September 1997) published three papers based on qualitative research in the same issue. Even though it was authored by Richard O. Mason, James L. McKenney, and Duncan G. Copeland, the fact that we had three of them in the same issue is a positive sign from the editors.

Mason (1997) offers us some good insights to start discussing “What is qualitative research?” and mainly “What should qualitative research be?” Like she said “Asking difficult questions”. It should be: “systematically and rigorously conducted, strategically conducted, yet flexible and contextual, should involve critical self-scrutiny by the researcher, or active reflexivity, should produce social explanations to intellectual puzzles, should produce social explanations which are generalizable in some way, or which have a wider resonance, should not be seen as a unified body of philosophy and practice, whose methods can simply be combined unproblematically, should not be seen as necessarily in opposition to, and uncomplementary to, quantitative research, should be conducted as an ethical practice, and with regard to its political context”.

31
Mason (1997) discuss five main questions regarding qualitative research which everybody should ask:

- What is the nature of the phenomena, or entities, or social ‘reality’, which I wish to investigate?
- What might represent knowledge or evidence of the entities or social ‘reality’ which I wish to investigate?
- What topic, or broad substantive area, is the research concerned with?
- What is the intellectual puzzle? What do I wish to explain? What are my research questions?
- What is the purpose of my research? What am I doing it for?

The use of techniques of qualitative data analysis and research is a latent theme in information science and scientific management communities (Willcocks et al, 1997; Kendall et al, 1997). We hope that this document constitutes a contribution to rescue some of its concepts and techniques, and to illustrate some of its potential applications.
References


IRMJ Information Resources Management Journal, (11, 1) Winter 1998 (Beyond rigor and relevance.) Robey, D. and Markus, L.; the application of IT research to organizations, Kavan, C.B.; The challenge of relating IS research to practice, Senn, J.).


MIS Quarterly v. 21, number 3, September 1997, (Three papers on qualitative research, from Richard O. Mason, James L. McKenney and Duncan G. Copeland).


Appendix - TEXTUAL DATA ANALYSIS
Tools or Software of General Use for Any Field Applications

1. It allows access, reading and analysis of texts of every nature!
A system in microcomputer (Sphinx® or other) it is destined to those that want to analyze texts, any that it is the origin of these... interviews, technical or literary documents, author’s documents, textual data bases (bibliographical, patents, etc).

2. To assist or support several objectives
- To revise, to examine, or to visualize a text or data base to have a quick but quite good idea of its content,
- To analyze the structure of the text,
- To seek certain elements of the text to produce strata,
- To make a systematic analysis of the content,
- To characterize or to identify certain elements of enunciation style,
- To elaborate thematic dictionaries,
- To build the index or summary of a document,
- To apply the methods of analysis of textual data to a certain text,
- To control certain characteristics of the text based on the construction of statistical indicators.

3. To be useful regarding several application fields: the analysis of surveys data and whole kind of interviews; the linguistic research, the lexicology and the terminology; the studies and literary creation, literary teaching and the linguistics; the document research; the competitive or intelligence business; the advertising creation; the journalism; the analysis of interactions of a group of studies or a group of top managers; the learning starting from the leaders’ speeches or political; the understanding of a collective process of decision (groupware meeting), the marketing, the HR, the R&D, the production, the end-users behaviour, and all other managerial activities or studies.

4. A tool with wide spectrum of application
- Import functions, edition (and cut out), allowing to structure the body of an existent and available text,
- Functions of lexicon analysis, automatical production and management of the lexicon of the text,
- Functions of lexicon surfing, with navigation in the text, starting from elements of the lexicon or of context variables, extraction of elements of the text in function of its content lexicon, search for (and identification) of context, verbatim production (extracts of the text following certain approach or criteria),
- Spell check functions of text and notes, verifying the text, marking in the text, grouping words or expressions/repeated segments, production of derived texts based on the text being analyzed,
- Functions of dictionaries management, with production, to enrich and to use dictionaries in graphic ways or composed expressions,
- Functions of index production, it looks for and defines the words composing the index or summary of a document, with automatical attribution of the numbers of corresponding pages,
- Functions of syntactic analysis, it looks for of the grammatical class of the words, lemmatization (consider the derived forms of the words, as plurals or conjugations of a verb, for its original form or root, singular or infinitive),
- Statistical functions, counting and unfolding of occurrences, construction of lexicon tables, calculation of lexicon indicative, calcul of specificities, characterization of contexts,
- Functions of textual data analysis, production of variables originated from textual data, factor analysis, classification, integrated treatment of textual data and of other kind of variables,
- Functions of database management, automatical modification of the analysis levels (text, paragraph, sentence, proposition or all other unit of sense) and restriction of the context variables in function of the selection (or cut out).

All these functions are available in a compatible PC software, for Windows®, in a friendly, flexible way, so that the user or analyst can have the autonomy and the peacefulness to exercise its curiosity and to elaborate analyses and inferences in depth.