Automated Syntactic Text Description Enhancement : the Thematic Structure of Discourse Utterances[1]

Jules Duchastel, Louis-Claude Paquin et Jacques Beauchemin
Université du Québec à Montréal

1. Introduction

Our work aims at the optimization of existing tools for computer-assisted description and analysis of textual data.[2] More specifically, we want to proceed to the thematic description of clauses and clause complexes of Quebec Budget speeches from 1934 to 1960. Our main objective is to enhance the work already done in this direction (Bourque and Duchastel, 1988)[3] by elaborating the analytic framework through a study of the thematic structure of these discourses. Having chosen a computerized lexicometric approach for content analysis, we believe that the relevance of lexical analysis can be greatly improved by means of syntactic and semantic description. Such description allows us to qualify lexical forms on the basis of morpho-syntactic and socio-semantic categories or even according to their position in the clause or clause complex. The data obtained from the material are what we call "qualified lexicons".

We will first set out the general context of our work by briefly explaining the research project on political discourse under the Duplessis Regime in Quebec (1936 - 1960) and giving a brief survey of the parsing strategy applied to the corpus. Secondly, we will present the theoretical background of thematic analysis and the operational model that we are using here. Finally, we will try to illustrate the relevance of such methodological work on research data.

2. General Context

Our general interest is in the analysis of political discourse in a socio-historical perspective. We have conducted research on various forms of political discourse originating in different Quebec

---

1 To be published in *Computer and the Humanities no. 26.*

2. See also "Automated Syntactic Text Description Enhancement : Determination Analysis", to be published in Research in Humanities Computing, Oxford University Press, 1991.

3. In the following pages, we make frequent reference to this book, in which we carried out a complete description of Budget speeches from 1934 to 1960. Our aim here is not to repeat those analyses, but to elaborate them with results obtained from the study of marked themes.

institutions over a period of nearly 25 years. We were concerned with the role of this discourse in the unfolding of political and economic transformation taking place between the Great Crisis of the thirties and the Quiet Revolution of the sixties. Were there any traces of these changes in discourse and how did these representations evolve? We were less interested in well structured ideologies than in mass discourse as it occures in public institutions. This explains the size (more than 5,000 pages) and the various forms of discourse (9 corpora, subdivided into a total of 18 collections). The political discourses in the strict sense of the term were Trone Speeches, Budget Speeches, the Constitutional, legislative and electoral discourses. The para-political discourse consisted of discourse from the Catholic Church, employers and unions.

The existence of such a large body of texts imposed the assistance of a computer. We were faced with the challenge of integrating a massive content analysis approach with a more sophisticated analytical framework. To comply with the objective of mass analysis, we adopted a lexical approach. As we already stated, however, the lexicons obtained through our extraction models were qualified on the basis of previous textual descriptions.

We first parsed every sentence of the corpora. The parser used was GDSF (Plante, 1980).[4] GDSF describes the syntactical surface structure of clauses. Déredec (Plante, 1980) is a language used to compile automats into LISP-interpretable codes. The automats of GDSF operate mainly on morphological categories which have already been semi-automatically tagged onto every word of the sequence.[5] We can define four main steps in the work of GDSF automats. The first step solves morpho-syntactic ambiguities. "Le" in French can either signify a determinative word for the noun, or a personal pronoun in the position of the object. The second phase establishes relations between constituants of the nominal and verbal groups. At this stage, the system establishes relations of determination in word phrases. The third level of the algorithm addresses the clause structure as such. Like most of the parsers of the late seventies, GDSF is normative in respect to the clause construction.

> "The most general model underlying the structural articulation detected by this grammar describes the interactions between two entities of the sentence : the noun and the verb." [6]

---

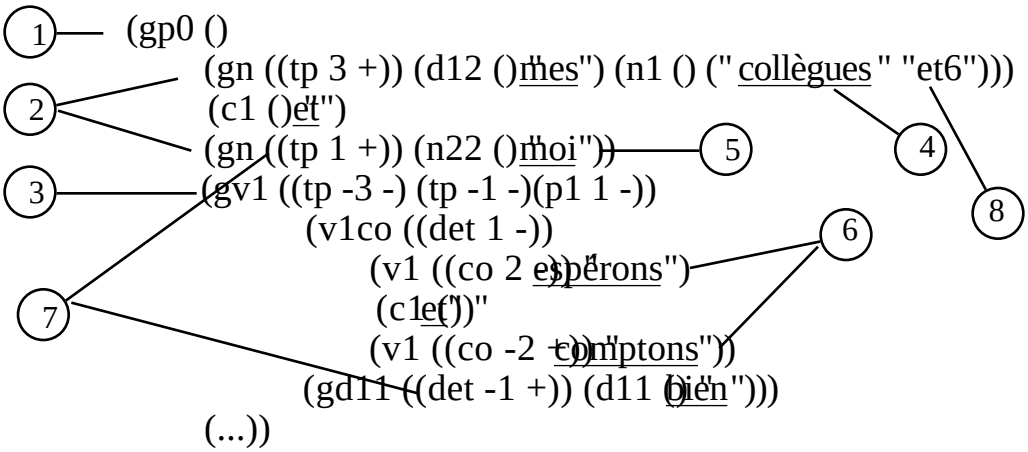
4. At the time the research was carried out, this heuristic parser was one of the most efficient for the French language.
5. For instance, in figure 1., we have D12, Nominal Determinant, N1, Noun, C211, preposition, D13, Adjective.
6. "Le modèle le plus général qui soustend l'articulation des structures dépistées par cette grammaire décrit des interactions entre deux entités de la phrase : les verbes et les nominaux.", (Plante, 1980, p. 1).

A clause will Thus be composed of at least a verbal group (GV) and a nominal group (GN). Finally, the fourth step establishes the relation between clauses within one sentence.

The text, as it is described, will now include syntactical information. The structure (contextual dependency relations) and information (different sets of categories) will appear in the LISP format of parentheses. It is then possible to generate structural trees representing nominal and verbal groups as they are articulated in clauses and clause complexes. Our lexical approach will benefit from this information on morpho-syntactical categories and contextual dependency relations, such as the theme/rheme relation, the determination relation and the verb completive relations (direct or indirect). The following figure gives an example of a GDSF description.

Figure 1. GDSF Description

"Mes collègues et mois espérons et comptons bien ..."



1.- Clause category
2.- Nominal groups
3.- Verbal group
4.- Noun
5.- Pronoun
6.- Verbs
7.- Contextual dependency relations
8.- Sociological category

As figure 1 also shows, we proceeded to a second level of description of our material which we will call the socio-semantic description. Using dictionaries, heuristics and coding, we categorized all the nouns and adjectives according to a set of socio-semantic (or sociological) categories. The system consists of 144 categories grouped in six families. The first three families correspond to institutionalized spheres : Economics (e.g. : budget, market, agriculture, industry, science and technology,...), Politics (e.g. : State, Constitution, Law,...) and Social Institutions (e.g. : Public opinion, social domain, labour relations, health,...). The fourth consists of general social categories (e.g. : sex, age, class,...). The fifth family refers to personal and social values (e.g. : Tradition,

progress, liberty, work, order,...). The final one regroups semantic functional words (e.g. : fundamental, better, increase,..). This categorization arose because exploratory models, when applied to sociological categories rather than words, often reveal unsuspected regularities. On the the other hand, the system must also work on the words themselves, so that no information is lost.

In conclusion, we have at our disposal many variables which can be activated in the analytical process. In this paper, we will concentrate our efforts on the thematic structure of discourse utterances. We make the assumption that the syntactical description of the components, whatever the level of stucturation, is of some interest for the analysis of meaning. In particular, the theme/rheme relation seems to be very productive in the comprehension of a text.

3. Theoretical Background

Theme has to do with position in the clause. Since antiquity and throughout the Western philology, words have been expected to appear in some order in a sentence. This expectation is persistent. The Greek term "hyperbaton" meant a disruption in the words'order.[7] The description of this figure given by Fontanier is the following :

> "...word arrangement opposite to the order in which the ideas follow one another in the analysis of tought(...) the subject is stated after its modifiers or after the verb."[8]

In more modern theories, the notion of word order has been replaced by the notion of position. The position is a distinctive feature of the word. The inversion refers to a move from one position to another. In all cases, the idea remains that the first position in a clause seems to be privileged and that inversion of positions constitutes a meaningful figure.

For our purpose, we have chosen the functionalist approach. This tradition goes back to Hjemslev and the Prague School which showed much interest in the semantical comprehension of texts. M.A.K. Halliday (1985), from whom we borrowed our theoretical frame[9], summarizes the axiomatical choices underlying the functionalist approach :

> "In general, therefore, the approach leans towards the applied rather than the pure, the rhetorical rather than the logical, the actual rather than the ideal, the functional

---

7. "Hyperbaton est transcensio quaedam uerborum ordinem turbans." in *Sedulius Scotus in Donati Artem maiorem*, Migne, p. 384.
8. "...arrangement de mots inverse relativement à l'ordre où les idées se succèdent dans l'analyse de la pensée (...) le sujet se trouve énoncé après ses modificatifs ou après le verbe." (Fontanier, 1827, p. 284).
9. We must mention the existence of such approaches, for the French language, as for example Bureau, 1976, 1978.

> rather than the formal, the text rather than the sentence. The emphasis is on text analysis as a mode of action, a theory of language as a means of getting things done."[10]

This approach is not oriented towards the identification of a universal structure of language, but is rather interested in the comprehension of natural speech. It is based on meaning, but as a grammar it is an interpretation of linguistic forms. In the work cited, Halliday has chosen to work on the clause, which is a higher unit in the constituent structures of language. His interpretation of grammatical structures is based on the principle that linguistic items are multifunctional. He defines three principal kinds of meaning in a clause, each of them being expressed by certain functional configurations. The first is the theme function that refers to the clause as a message. "The Theme is what the message is concerned with" (Halliday, 1985, p. 36). The second function is the grammatical subject of the clause as an exchange. The subject is responsible for the success or failure of the exchange. The third function is the actor in the clause as the representation of a process. The actor does the deed in the process. All these functions may or maynot coincide, but the meaning of the clause depends on them. The scope of our contribution will be with respect to the thematic structure.

4. What is the Theme?

We must first define the notion of theme and identify the significant constituent structure in which it is localized. Spontaneously, one would define the theme as the object of a message, the reason for communication. Again, "the theme is what the message is concerned with". In Prague School terms, rheme corresponds to what is said about the theme, the argument. The only way to go beyond this definition is to determine the operational definition of theme. Before getting to this model, we must first identify the significant constitutent structure of the theme.

The question is raised by Halliday himself. Even if his book is essentially dedicated to the study of the theme in the clause and the clause complex, he suggests that the thematic organization appears in "different guises throughout the system of the language, with manifestations both above the clause and below it" (Halliday, 1985, p. 56). Below, the nominal group and the verbal group seem to incoporate the thematic structure. Above, paragraphs are organized along what he calls the topic sentence. The question then arises about the degree of generality of the theme and its localization in smaller or larger parts of discourse. The content analysis tradition does not bother with any formal

[10] p. XXVIII.

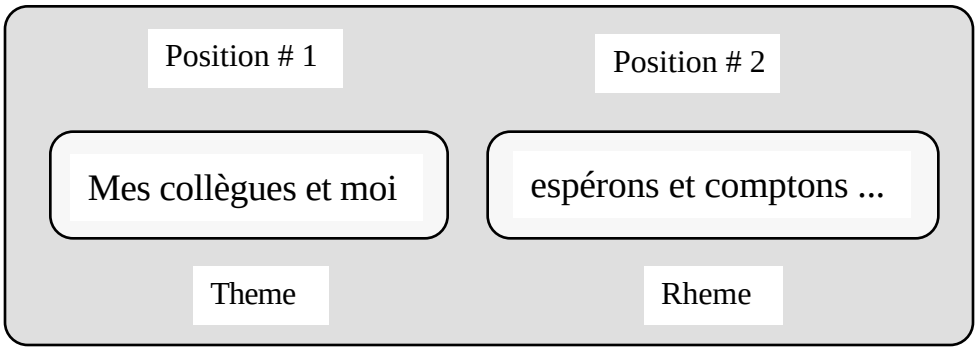
definition of the Theme's context. The theme is identified on the basis of the ideas expressed, independently of any linguistic structure. This approach is based on the researcher's intuition and therefore no reproduction of the experience is possible. In the field of computational linguistics, some theorists contest the very existence of the clause theme and propose formal strategies to identify what shoud be called the discourse theme (Marandin, 1988). As this author puts it, there is still a lot of empirical and conceptual work to be done in different fields of research : linguistic descriptions, textual descriptions and descriptions of different modes and forms of comprehension, before it is possible to design any algorithm that would identifiy the theme of the discourse (Marandin, 1988).

In the interim, without denying the relevance of identifying many levels of themes, we have adopted the clause approach. This approach is congruent with our lexicometric orientation. The assumption, in lexicometric studies, is that words, as units, are significant material. The recurrence of some words can be interpreted in a meaningful way. The study of the thematic structure of the clause will make it possible to qualify some of these words which occupy the theme position in the clause. Our proposition is that the lexicon of thematic words will reveal in some way the thematic structure of the text.

5. Operational Model

Our first task is to identify the theme in the clause. Halliday tells us that "the Theme can be identified as that element which comes in first position in the clause" (Halliday, 1985, p. 39). He adds that the position in itself does not define the theme, but is the "means whereby the function of theme is realized, in the grammar of English (Halliday, 1985, p. 39). We can say that the function of theme is realized in the same manner in French. Figure 2 illustrates the theme/rheme position in a clause from our corpus.

Figure 2. The Theme/Rheme Positions

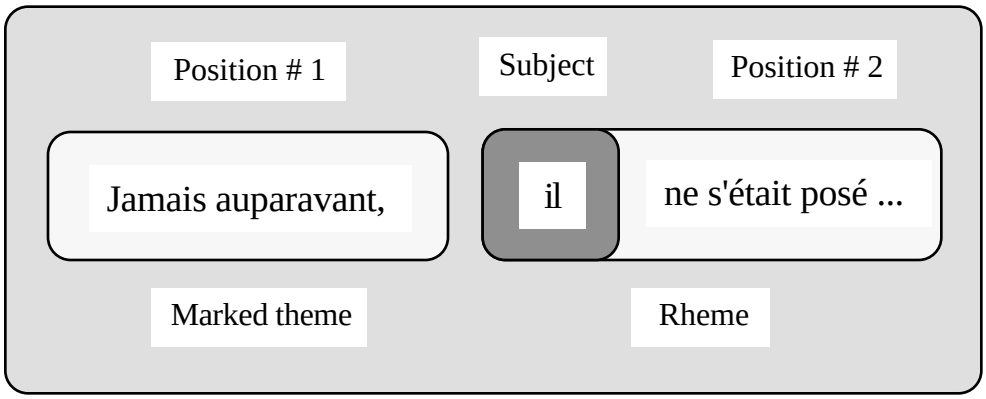


From an operational point of view, the problem lies in the delimitation of the first position in a clause. Halliday deals extensively with this aspect for the English language. The theme could be formed of either one element - one nominal group, adverbial group or prepositional phrase - or of two or more elements forming a single complex element - a phrase complex -. These are what Halliday calls a simple theme. In some cases, the thematic structure becomes much more complex, as in the instance of nominalization.

We use GDSF to determine the elements in first position in the clause. The theme does not necessarily correspond to a simple word. In the case where the first position is occupied by functional words not functioning as grammatical subjects or complements, then the subject or complement immediately following the first position is considered as part of the theme (then called multiple theme). We could say that the theme is the semantic elements in first position in the clause.

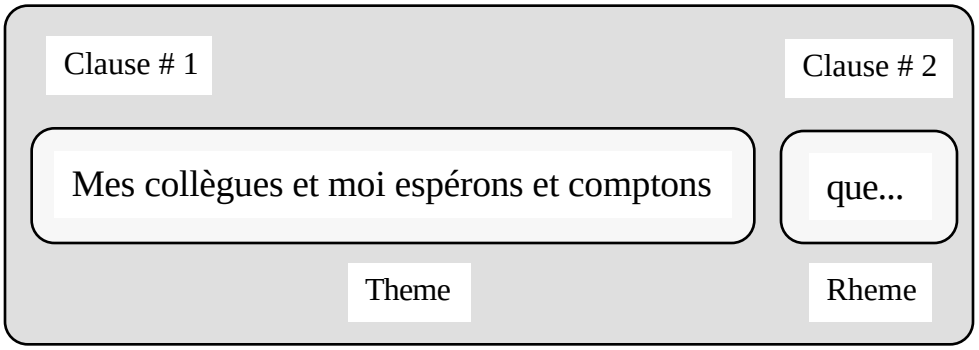
At this point, a second question arises. Is there any hierarchy among the multiplicity of themes as they appear in each clause? We identified two principles that enable us to produce such an ordering of themes. The first is what Halliday calls the marking of the theme. The second refers to the organization of and interaction between clauses in the clause complex. Halliday tells us that in English, in a declarative clause, the theme is usually conflated with the subject. He then talks of the unmarked theme. In any instance where this conflation does not occur, we are in the presence of a marked theme. The most frequent case of marked theme is the adverbial group that comes before the subject, but the most significant one is the complement in first position. This is because the complement, being a nominal, would have subject potential, but nevertheless has been made thematic without being the subject. In French, the expected sequence of a clause is Subject, Verb, Object (Dubois et al., 1970). Consequently, every time the first position is not occupied by the grammatical subject, we will consider the theme to be marked. Figure 3 illustrates a marked theme.

Figure 3. Marked Theme



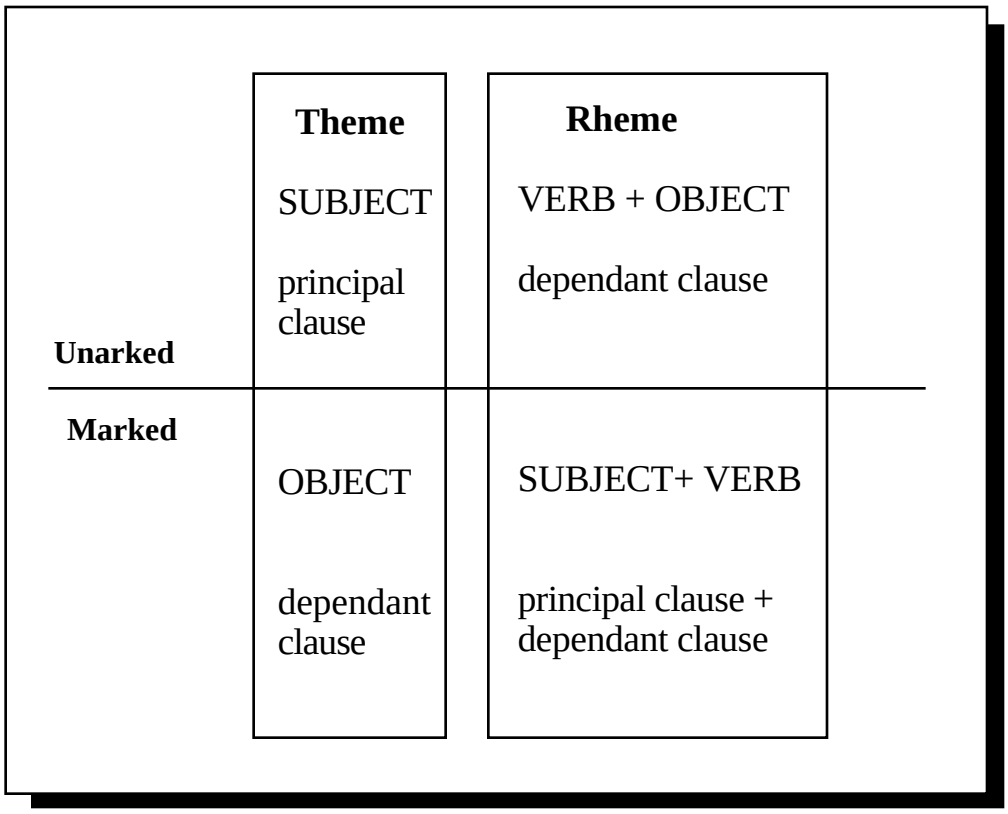
The organization of and interaction between clauses in the clause complex constitute a second criterion allowing us to give greater weight to some of the identified themes. At the clause complex level, it is possible to define a thematic element. The first clause constitutes the theme of the clause complex. A clause complex is usually formed of a head clause and one or more dependent clauses. The head clause usually occupies the first position in the clause complex and will be said to occupy the thematic position. Figure 4 illustrates a clause complex theme.

Figure 4. Clause Complex Theme



In all cases where the first clause of a clause complex is a dependent one, the clause complex theme will be considered as marked. Figure 5 shows the different positions of grammatical elements of the clause and clause complex which determine to the marked-unmarked nature of the theme.

Figure 5. Summary of the Marked - Unmarked Nature of the Theme



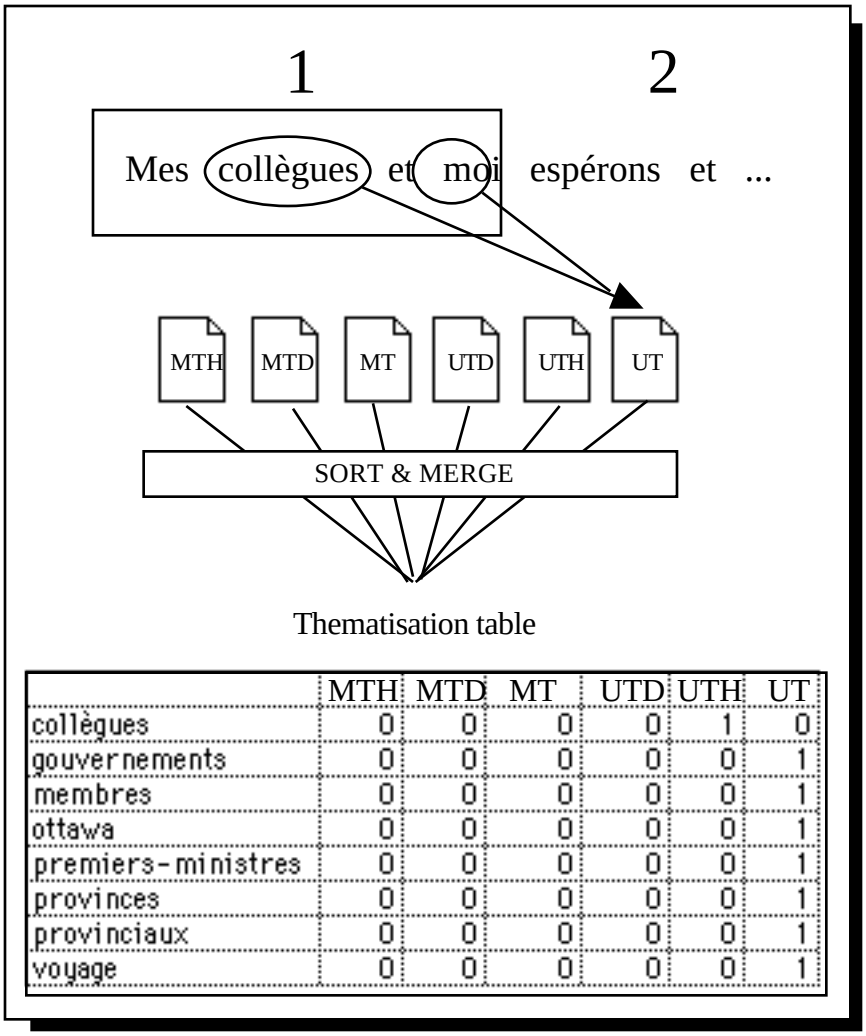
If we combine these two criteria for theme marking, that is, inversion of word or clause order, it is possible to define six decreasing theme values :

1) A marked theme of a dependent clause in first position (MTD)
2) A marked theme in a head clause in first position (MTH)
3) All other marked themes (MT)
4) An unmarked theme of a dependent clause in first position (UTD)
5) an unmarked theme of a head clause in first position (UTH)
6) All other unmarked themes (UT)

In conclusion, we will give a brief overview of the procedure designed for Thematic analysis. Using previous GDSF descriptions of the corpus, we first proceeded to a close evaluation of its output. The clause complex descriptions were recursively scanned and clause constitutents were checked. A few ill-formed descriptions were discarded. We also discarded all relative and participial clauses because, being enclosed in nominal groups, they could not be interpreted in the same manner as other clauses. The remaining output was then processed by an algorithm identifying the position of all lexical elements in the clause and clause complex according to our typology. They were then extracted and put into the appropriate files. The resulting six Tables were sorted and a lexicon was

computed for each. Those lists were finally merged and transferred to a database to be analyzed. The algorithm is in Le-Lisp[11] and was processed on a VAX 750.

Figure 6. Construction of a Thematization Table



	MTH	MTD	MT	UTD	UTH	UT
collègues	0	0	0	0	1	0
gouvernements	0	0	0	0	0	1
membres	0	0	0	0	0	1
ottawa	0	0	0	0	0	1
premiers-ministres	0	0	0	0	0	1
provinces	0	0	0	0	0	1
provinciaux	0	0	0	0	0	1
voyage	0	0	0	0	0	1

6. Producing a New Thematic Description

Our first objective was to enhance a previous syntactic description of the thematic structure of discourse utterances (Bourque et Duchastel, 1988). The main improvement consisted in introducing different categories of thematic position. This would permit us to introduce a better interpretation of the thematic content of the corpus. A secondary effect of this new description was

11. Le-Lisp, version 15.2, J. Chailloux, INRIA, France.

the reduction of the total number of lexical elements defined as themes. The decrease was evaluated at 23.55% for the whole corpus with small variations according to the different families of sociological categories in thematic position. The principal explanation for this phenomenon is the exclusion of themes in the relative and participial clauses. This factor does not affect the significance of the global distribution insofar as we posit that relative and participial clauses are randomly distributed. The other explanatory factor is that the first algorithm retained all the words related to the subject phrase, while the second algorithm is able to distinguish the theme from the subject. Some substitutions took place when we applied the second algorithm. Unmarked themes were identified, in contrast with the first algorithm where the grammatical subject was considered as the theme of the clause. Some of these substitutions explain the decrease in the number of themes identified, but we believe that we now have a more precise description

Of greater interest is the result that we were able to validate our typology of thematic categories. Table 1 illustrates the distribution of the different categories of themes on the pretest corpus.[12]

Table 1. Distribution of Thematic categories

	Marked Theme	**Unmarked Theme**	**Total**
Dependent first position	0.21%	4.01%	4.22%
Head first position	14.52%	58.44%	72.96%
Others	1.98%	20.84%	22.82%
Total	**16.71%**	**83.29%**	**100%**

At first glance, table 1 confirms two propositions stated above. First, the conflation of theme and grammatical subject is the usual case. Secondly, the head clause normally occurs in first position in the clause complex. 83.29% of all themes are unmarked, meaning that they occupy both the thematic and subject position. The first position in the clause complex is occupied by the head clause in 72.96% of all cases.

But a closer look at this table indicates that there are great differences in importance between some of the thematic categories. Those differences could be further investigated to evaluate the appropriateness of our ordering of thematic values. This is not, however, our purpose here. From a lexicometric standpoint, we have decided instead to merge some of those values in order to obtain a more significant typology. The first category would regroup every instance where a marking activity seems to take place. This is the case for all categories of marked theme in the clause,

12. The analysis that follows was made on a sub-corpus composed of all the Budget Speeches from 1934 to 1960.

whatever the position of this clause in the clause complex; that is, the marked themes in the dependent or head clause in first or any other position. Considering that a dependent clause in first position is also marking the theme, we include this value in the marked theme category as well. The second and third categories remain as they are. In summary, we have : 1) the marked theme (20.72%), 2) the unmarked theme in the head clause (58.44%) and 3) the unmarked theme in dependent clauses (20.84%). In the following data analysis, we will concentrate on the marked theme.

7. Analysis of the Thematic Structure

The enumeration of themes and their different values led us to define two ratios, a thematic ratio and a marked theme ratio.[13] The thematic ratio is the relation between the total occurrences of one word and the number of times it is thematized. The marked theme ratio is the relation between the total occurrences of one word when thematized and the number of times it is thematically marked. Table 2 gives a overview of the values of these ratios for every family of sociological categories.

Table 2. Thematic and Marked Theme Ratios for each family of sociological categories

Sociological families	Total Frequencies	Theme Frequencies	Thematic Ratio (%)	Marked Theme Ratio (%)
Economics	13627	3539	25.97	12.10
Politics	9798	2894	29.54	16.75
Social Institutions	3876	1002	25.85	20.96
Social Categories	3784	998	26.37	15.13
Values	4990	799	16.01	26.78
Functional Semantic Words	3486	528	15.15	21.59
Total	**39561**	**9760**	**24.67**	**16.41**

If we look at the thematic ratio, we can regroup the six families into three classes. First, there is a modal class at about 26% that includes economics, social institutions and social categories. A second class is slightly higher than first, that ism, the political categories at 29.54%. Finally, two families have much lower results : Values, at 16.01% and functional semantic words, at 15.15%. One expected result is the strong thematic ratio for political categories, which only confirms the political nature of the discourse and its desire to put the interventions of political actors and

13. In this paper we have not used any significance tests, our approach at this time being merely exploratory with respect to the value of a thematic description. A statistical analysis of more elaborated results would be the subject of another article.

institutions at the forefront. On the other hand, it is interesting to note that the values and functional words are less thematized, but have, at the same time, the highest marked theme ratio. Although these words are not what the discourse is about, when thematized, they are more often marked, which means that they are insisted upon. We might advance the hypothesis that this reflects the rhetorical nature of political discourse.

If we examine the contrasted indices for the economic categories, we can see that they have a modal ratio of thematic (25.97%), but the marked theme ratio is the lowest (12.10%). This can be explained by the fact that economic categories are a natural object for a Budget speech and therefore reasonably thematized, but are not the main purpose of the discourse, explaining the weak marked theme ratio. This reinforces the conclusions we reached (Bourque et Duchastel, 1988) concerning the major orientations of the Budget Speeches toward general political ends. The marking of the theme suggests that values, social and political institutions are emphasized rather than economics.

We shall now turn to some partial findings that were made on the basis of a more extensive thematic description of our data. We will concentrate on certain topics identified in earlier analyses and illustrate how the marking of the theme adds its own significance to interpretation of data. We will first show that the study of marked themes enables us to characterize political discourse as a specific type of discourse. We will then illustrate how variations of the marked theme ratio throughout the period confirm the pattern already discerned with respect to the peak of the Duplessis Regime (Bourque et Duchastel, 1988). Finally, we will see the way in which certain peculiar features of this apogee discourse are reinforced by taking into account the marked themes.

7.1. Budget Speech as a Specific Type of Discourse

The following considerations are preliminary and require confirmation through a more extensive study of different types of discourse from our corpus (e.g. electoral, legislative, religious,...). We will nonetheless try to illustrate, on the basis of certain findings about the marking of the theme, that Budget speeches are characterized by specific features which are typical. Political discourse is expected to set goals and state how they may be achieved. Some specific features indicate that this is the case here. The Budget speakers define the time and space of their actions. They assume responsibility for them. They present themselves as responsive to public opinion. Finally, they dramatize situations. Table 3 presents the thematic and marking indices for a selected group of words refering to the features just described.

Table 3. Thematic and Marked Theme Ratios for Words Featuring Attainment of Goals in Budget Speeches

Categories and "Words"	Total Frequencies	Thematic Ratio (%)	Marked Theme Ratio (%)
Time	6988	25.03	44.08
Space	4836	25.60	21.73
General Economy	2541	23.69	11.13
Financial Sector	1381	24.69	18.48
Public Opinion	633	20.54	36.15
"Grave"	3	7.32	33.33
"Difficile"	44	13.84	36.36
"Théorie"	16	25.81	37.50
"Possible"	7	5.38	57.14
"Pratique"	9	12.50	44.44
"Compétent"	6	13.33	33.33
"Courage"	9	19.15	33.33
"Réaliste"	24	16.33	37.50
"Énergique"	22	16.67	27.27

Time and space are fundamental categories of all discourses. Time and space localization are basic elements of the deictic referential system of all languages. We suggest that political discourse uses this system in a particular way. In the case of Budget speeches, the total frequency of each time and space is very important. In both cases the thematic ratio indicates a modal behaviour. In addition, the marked theme ratio for time has a very high value (44.08%). This can be explained if we look at the lexicon under the time category. The most frequent and marked thematic words are "année", "cours", "années", "ans", "mars", etc. This type of discourse must define the precise time localization of its actions, as in this example : "Au cours de la présente année,...". The same can be said about space. By far the most important word of the space lexicon is "Québec", followed by "Canada", "pays", "province", etc. The fact that these deictics are marked themes shows that the Budget speeches refer directly to the regulation of time in a specific spatial environment.

Budget speeches deal with economics. In table 2 we saw that in the family of economic categories there are a great number and, in one out of four occurrences, they occupy the theme position. On the other hand, the marked theme ratio is the lowest of all the families of categories. This trend is confirmed if we look at other non specific economic categories such as "general economy" or "budget". In both instances, the total number of occurrences is important (2541 and 890 tokens respectively), the thematic ratio is near the modal value and the marked theme ratio is very low (11.13% and 10.90). Some more specific economic categories are nevertheless of greater

importance from the standpoint of thematization. This is the case with "financial sector", which regroups words such as "prêts", "emprunts", "dettes" and "taux d'intérêts", etc. This category is important in number (1381), has a normal thematic ratio (24.69%), but also has the highest marked theme ratio in this family (18.48%). From this it appears that budget speeches stress categories that will permit measurement of government achievements. The words categorized under "financial sector" are thus a type of measurable variable of government efficiency.

Government achievement relies on public opinion. Budget speeches will try to justify good or bad economic management through public opinion. If we look at the family of social intstitution categories, we note that "public opinion" comes in second place (633 tokens). Its thematic rate is not very high (20.54%), but its marked theme ratio is the highest in this family and one of the most important among all categories (36.15%). We are therefore justified in stating that reference to public opinion is not incidental, but is rather an active constituent of a validation process.

Table 3 enumerates the semantic functional words which reached the highest marked theme ratio. In many ways, they connote the dramatical aspects of this discourse. The problematical character of an economic situation is stated with words like "grave" (marked theme, 33.33%) or "difficile" (36.36%). Some other words suggest means to attain goals : "théorie" (37.50%), "pratique" (44.44%), "possible" (57.14%). Words like "compétence" (33.33%), "courage" (33.33%), "réaliste" (37.5%) and "énergique" (27.27%) refer to essential qualities required for problem solving. These features are the result of many layers of data analysis carried out on the Budget speeches and serve to illustrate the appropriateness of a refined description of the theme structure. When some words or categories have both high thematic and marking rates, they can be said to be the main reason for communication. From this perspective, the preceding features make it possible to describe the Budget speeches as a specific type of discourse. We would have to elaborate a comparison with other discourses in order to validate this specificity. Although we are unable to do so here, we are nonetheless aware of the possibilities offered by such textual descriptions.

7.2. Variations of the Marked Theme Ratio throughout the Period

The Budget corpus has been divided into seven sub-texts corresponding to the seven legislatures that occurred from 1934 to 1960. This division enables us to study the thematic variations from one period to another. We have characterized these periods according to the economic or political situation : the economic crisis during the first two legislatures in the thirties, the war period from 1939 to 1945, the peak period of Duplessis Regime from 1945 to 1953 and the decline of the Regime after 1954.

For each period, we have defined a global marked theme ratio for all categories. We are thus able to provide a general idea of the use of marked themes on a temporal comparative basis. A low ratio can be interpreted as typical of a low profile discourse which does not make great use of emphatic means of communication. A high marked theme ratio would mean, on the contrary, that discourse is taking a more active form to stress certain topics. Figure 8 illustrates the variation of the marked theme ratio throughout the period, from 1934 to 1960.

Figure 7. Variation of the Marked Theme Ratio Throughout the Period

Stated in more sociological terms, our hypothesis is that a high marked theme ratio signifies that discourse is repositioning itself and dealing with new social stakes. A high ratio would manifest the willingness to change the main topic of discourse or to emphasize particular aspects of old topics. Figure 8. shows that the highest rates of marking are in the three central zones. This corresponds to the war period and the Duplessis peak. It is also a period of great transformation. Apart from the economic modernization of Québec, taking place from the beginning of the forties, we find great concern over new forms of State intervention and Constitutional debates. Before this period, the discourse is quite stable. After 1953, the Duplessis Regime declines and the discourse is much gloomier.

These observations merely confirm our previous analysis (Bourque et Duchastel, 1988). In effect, the emphasis of discourse is changing during the period from 1940 to 1953. We can observe that the economic categories are declining in favour of the State, institutional and social categories. A

major turn in the topics of discourse is evident. The fact that theme marking also increases during the same period gives greater substance to this transformation.

7.3. Discourse Features from the War Era and the Peak of the Duplessis Regime

In this part we present some typical features of one specific period of the Budget speeches. The period is characterized by the strongest global marked theme ratio (15.43%). Duplessis is back in power after an interval of liberal government during the war. His regime is recommencing from a new foundation and discourse is expected to reflect the re-organization of Quebec's social and political scene and the advent of a new ideological arrangement, typical of this Regime. Again, the following observations serve to confirm previous results as stated in Bourque and Duchastel (1988). It is interesting to note hat the study of the marked theme ratio reconfirms the appropriateness of those conclusions. Table 4 compares the marked theme ratio for the 1945-48 span to the average ratio for the whole period (1934-1960). The figures show which categories seem to be more or less emphasized in the discourse. We have chosen, from our set of categories, some of the most significant ones according to the re-orientation process mentioned earlier.

With the exception of two singular categories, the marked theme ratio of retained categories indicates that all of them are receiving much more attention from the speakers. We have already established that the global rise of this ratio indicates that the discourse is calling attention to certain topics. We will try to illustrate in some detail how this attention is focussed.

Table 4 Marked Theme Ratio of some significant categories for the 1945-48 Period

Categories	Marked Theme Ratio	Average Marked Theme Ratio
Economical Family		
	16.34	12.10
General Economy		
Natural Ressources	27.27	11.13
Industries	18.72	7.89
Science and Technology	15.38	9.97
Agriculture	16.67	15.79
	6.06	12.85
Constitution		
Conflict	17.61	15.91
	37.50	24.39
Professions		
Population	30.77	12.39
Age Groups	27.78	14.29
Social Classes	25.00	18.87
Communities	16.67	13.13
	18.42	14.46
Political Parties		
Social Policies	20.73	20.49
Education	50.00	17.07
Culture	28.13	16.80
	46.67	24.43

Our first observation concerns the behaviour of the economic family of categories. Its global marked theme ratio, normally one of the lowest (12.10%), rises to 16.34% in this period. While this indicates that something is happening in the economic field, an examination of certain economic categories will give us more clues. "General economy" has a very high rate, confirming the economic turn of this discourse. At the same time, we note an significant decrease in "Agriculture" which was a main feature of Duplessis' class alliance prior to 1940. The fact that "Agriculture" is no longer "what the message is concerned with" only confirms our previous findings concerning the decreased significance given to this matter by the Duplessis Regime after the war. On the other hand, we notice that a set of categories closely linked to the industrial process has a much higher marked theme ratio than average. "Natural Resources", "Industries" and "Science and Technology" have, respectively, a ratio of 18.72%, 15.38% and 16.67%. This manifests the industrial concerns of the Government in a world which is modernizing at great speed.

At a more political level, Duplessis' discourse is mainly characterized by its constitutional struggle against Federal attempts to increase and centralize State intervention. The marked theme ratios show the traces of that struggle in both "constitution" (17.61%) and "conflict" (37.50%).

The marked theme ratio increase of the most important social categories also shows the redefinition process which appears to be taking place during this period. All constituents of society seem to be called upon. "Population" (27.78%), "Age Groups" (25.00%), "Communities" (18.42%) and even "Social classes" (16.67%) become the central topic of discourse as if it were necessary to rally all social forces around new objectives. "Professions" is the social category with the highest increase of all (from an average of 12.39% to a 30.77% ratio). This must be explained on its own. "Profession" refers here mostly to the petty bourgeoisie which was a central cog in the Duplessis hegemonic system. The economic changes and the content of constitutional debates could greatly affect this social class and, consequently, class mobilization is taking place in the form of marked thematization of professionals.

Finally, a look at some social institutions shows that three main institution ratios are greatly increasing. "Culture" and "Education" are traditionally under Quebec's jurisdiction. The constitutional debates during those years challenge this jurisdiction. The Budget speeches reflect these preoccupations by giving greater significance to those topics. Social Policies is a very sensitive point for the whole discussion which is taking place. The increase in the marked theme ratio (from an average of 17.07% to a 50.00% ratio) shows that it becomes one of the most important topics of discourse.

8. Conclusion

Our work flows in two main directions. First, as sociologists, we are interested in the description of political ideologies in their discursive dimension. Secondly, however, we think that it is relevant to design appropriate research tools to arrive at a more complete objective description of data, which is not to say that the interpretation will automatically become objective. Having previously used computer-assisted methodologies on political corpora, we have tried here to show how the optimization of existing tools for computer-assisted description and analysis of textual data can enhance the results produced in their less developed version. We choose to elaborate an analytic framework for the study of the thematic structure of discourse.

The question is to evaluate the relevance of this framework for the interpretation of discursive data. The results presented here cannot be considered as definitive. We have only tried to show how a more rigorous description of themes in clauses and the addition of a marked theme ratio could confirm previous findings with respect to the discourse of the Duplessis Regime. At that level, the data presented above confirms the principal trends of our previous analysis. It provides further

evidence of the significance attributed to certain topics by discourse. This is evident if we consider the marked theme ratio increase taking place during the period from 1940 to 1952. We have shown, in our earlier work, that the war era and the Duplessis peak constitute the densest period in terms of new ideological content. It is also true that the examination of the marked theme ratio increase of some of the more topical categories of this period provided further support for our earlier conclusions. Marked theme is a way of stressing the object of the message. Thus, not only do some categories have great significance, they are also treated in such a way that they gain significance. Finally, we note that our remarks concerning the typical form of political discourse will have to be validated through further research.

Overall, we are convinced that lexicometry has much to gain from systematic description procedures which enrich the possible interpretation of the meaning of words.

B i b l i o g r a p h y

Beauchemin, J., Paquin, L.C., "Apport des parseurs à l'analyse des données textuelles par ordinateur", in *La description des langues naturelles en vue d'applications informatiques*, Université Laval, Centre international de recherche en aménagement linguistique, Québec, CIRB, 1989, p21-31.

Bourque, G., Duchastel, J., *"Restons traditionnels et progressifs". Pour une nouvelle analyse du discours politique, le cas du régime Duplessis au Québec*, Montréal, Boréal, 1988, 399 pages.

Bureau C., *Linguistique fonctionnelle et stylistique objective*, Paris, PUF, 1976, 264 pages.

Bureau C., *Syntaxe fonctionnelle du français*, Québec, PUL, 1978, 246 pages.

Courtine, J.-J., "Analyse du discours politique", in *Langages*, No. 62, 1981, 128 pages.

Daoust, F., Duchastel, J., Dupuy, L., "Système d'analyse de contenu assistée par ordinateur (SACAO)_, in *La description des langues naturelles en vue d'applications informatiques*, Université Laval, Centre international de recherche en aménagement linguistique, Québec, CIRB, 1989, p. 197-210.

Dubois, J. et al. *Réthorique générale*, Paris, Larousse,1970.

Fontanier, P., *Les figures du discours*, Paris, Flammarion, édition 1977.

Halliday, M.A.K., *An Introduction to Functional Grammar*, London, Edward Arnold, l985, 1986, 1987, 387 pages.

Löfftedt, B., *Sedulius Scotus in Donati Artem maiorem*, Migne, 1977, 463 pages.

Marouzeau, J., *L'ordre des mots dans la phrase latine*, Paris, Les Belles Lettres, 1950.

Marandin, J.-M., "A propos de la notion de thème du discours. Eléments d'analyse dans le récit", in Le thème en perspective, *Langue française*, Paris, Larousse, No. 78, 1988, pp. 67-87.

Plante, P., *Une grammaire Déredec des structures de surface du français appliquée à l'analyse de contenu de textes*, Montréal, Service de l'informatique, Université du Québec à Montréal, 1980.