Valorisation d'une description syntaxique automatique: Analyse de la structure thématique des énoncés du discours

Par Jules Duchastel, Louis-Claude Paquin et Jacques Beauchemin
Centre d'Analyse de Textes par Ordinateur
UQAM

1. Objectif

Notre travail vise l'optimisation des outils de description et d'analyse de données textuelles par ordinateur[1]. Pour ce faire, nous disposons de divers progiciels qui ont déjà été appliqués sur un corpus de discours politiques. Il s'agit pour l'essentiel de modules de catégorisation morpho-syntaxique (CBSF) et d'analyse syntaxique de surface des phrases (GDSF)[2]. La description syntaxique ainsi obtenue a été effectuée sur l'entièreté du corpus (5,000 pages) et nous fournit la base pour des explorations et des analyses supplémentaires.

Nous nous arrêterons dans le présent travail à l'exploitation d'une description de la structure thématique des propositions et des phrases des textes retenus. L'objectif est ici de revenir sur une première exploitation de cette description appliquée à un sous-ensemble du corpus — le discours du budget 1934-1960 au Québec — afin de comparer deux usages différents d'une même description. Cette dernière a été effectuée et analysée dans une premier ouvrage sur le discours politique duplessiste (Bourque, G. et Duchastel, J., 1988). Nous proposons

---

[1] Ce travail s'inscrit dans le cadre des travaux du projet SACAO (Système d'Analyse de Contenu Assistée par Ordinateur, Centre d'ATO, UQAM). SACAO vise l'intégration systématique de procédures existantes ou nouvelles de lecture assistée de données textuelles. Il offre à des utilisateurs, dans un environnement logiciel relativement intégré, divers modules de description, d'exploration et d'analyse de données textuelles, tout en leur laissant le soin de paramétrer ces procédures en fonction de leurs propres hypothèses de lecture.

[2] CBSF (Catégorisation de base syntaxique du français), progiciel conçu par Lucie Dumas du Centre d'ATO, permet de reconnaître la catégorie syntaxique des formes lexicales de la langue française. GDSF (Grammaire de surface du français), progiciel conçu par Pierre Plante du Centre d'ATO, est un ensemble de procédures, programmées en Déredec, dont l'objectif est l'obtention des structures de surface du français écrit.

maintenant trois étapes d'une démarche méthodologique qui devrait permettre de valider ou de corriger le travail déjà effectué, mais surtout d'enrichir la description de la structure thématique des textes. Sur la base d'une réévaluation des descriptions produites par l'analyseur syntaxique (GDSF), nous redéfinissons les procédures d'identification et de repérage des thèmes à partir d'un approfondissement de la notion de thématisation, nous comparons cette nouvelle description avec l'ancienne et en évaluons la pertinence heuristique sur la base de l'analyse du matériel qu'elle autorise.

2. La thématisation et la hiérarchisation

2.1 Une approche fonctionnaliste

Il ne nous appartient pas de faire l'histoire de la tradition fonctionnaliste remontant à Hjelmslev et à l'école de Prague ni de départager les diverses tendances des grammaires fonctionnelles, telles qu'elles sont développées par différents auteurs ou pour des langues différentes. Disons seulement que nous emprunterons à cette tradition ses prémisses et un certain nombre de découpages qui enrichissent le potentiel de lecture des textes. L'approche fonctionnaliste intéresse le lecteur-expert[3] d'abord parce qu'elle favorise la compréhension sémantique du texte. Par contraste avec les approches dites syntaxiques qui privilégient les enchaînements syntagmatiques, les grammaires fonctionnelles s'intéressent avant tout aux formes linguistiques, dans la mesure où leurs diverses configurations paradigmatiques produisent du sens. M.A.K. Halliday (1985), dont nous nous sommes largement inspirés, résume bien les choix axiomatiques à la base de cette approche:

> "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 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". p.XXVIII.

[3] Jacques Beauchemin et Louis-Claude Paquin (1988) définissent le lecteur expert comme celui "dont la lecture et l'analyse de textes constituent la principale activité...".

Cette approche ne s'intéresse pas à l'existence d'un universel de la langue, mais aux formes de réalisation des pratiques langagières dans la parole ou le texte. Elle vise la description de formes linguistiques qui produisent du sens. Bien que se concrétisant dans des formes très variables, autant du point de vue de la construction, de l'opérationnalisation que de la validation des descriptions proposées par divers auteurs, elle permet la comparaison de résultats potentiellement complémentaires. A titre d'exemple, la richesse de ces descriptions se révèle dans la superposition de diverses configurations fonctionnelles dans le modèle proposé par Halliday pour la langue anglaise. En effet, pour lui, le même élément ou constituant d'une phrase peut occuper plusieurs fonctions dans le cadre de sa participation à différentes configurations systémiques qui la caractérisent. C'est ainsi qu'il introduit trois configurations principales: la configuration idéationnelle ou sémantique qui définit la fonction de l'acteur; la configuration interpersonnelle ou énonciative qui implique une fonction de sujet; la configuration textuelle ou de contenu qui renvoie à la fonction de thème. Dans tous les cas, le discours est reconstruit selon chacun des trois systèmes qui contribuent, directement et de manière complémentaire, à la production du sens. Nous nous intéresserons particulièrement dans le cadre de cet exposé à la structure thématique.

2.2. Thématisation

Deux questions doivent d'abord être examinées: comment définir le thème et à quel(s) niveau(x) de la séquence énonciative doit-on le localiser? L'idée plutôt consensuelle à la base de la notion de thème est qu'il représente "l'objet du message", pour reprendre la terminologie de Halliday[4] (1985). Quelle que soit la définition qu'on retiendra du message, on peut supposer qu'il nous dit quelque chose à propos d'un objet matériel ou conceptuel. Le thème, c'est l'objet dont on parle, le propos, ce qu'on en dit. La tradition de l'école de Prague a d'ailleurs défini le doublet conceptuel Thème/Rhème pour évoquer ces deux aspects du message. Bien que relativement élémentaire, cette définition rend pourtant compte de l'intuition de tout analyste du texte soucieux d'en connaître ou d'en indexer le contenu.

La seconde question concerne davantage l'amplitude du contexte (proposition, phrase, paragraphe ou texte) à l'intérieur duquel on tentera d'identifier

[4]"The theme is a function in the clause as a message. It is what the message is concerned with: the point of departure for what the speaker is going to say." (Halliday 1985), p. 36.

un ou plusieurs thèmes. Elle s'accompagne d'une question complémentaire concernant la pertinence de hiérarchiser divers niveaux thématiques selon leur localisation dans la structure énonciative. Si l'on écarte les méthodes qualitatives[5], pour ne retenir que celles qui s'appliquent à formaliser les descriptions dans des environnements informatiques, on peut distinguer deux approches. D'un côté, l'approche fonctionnelle considérera que le thème est, avant tout, lié à la proposition. C'est par extension logique que, s'il y a thème à ce premier niveau d'organisation syntaxique, on considère qu'il doive y avoir également thème de l'unité englobante: la phrase ("clause complex", chez Halliday), puis le paragraphe et, pourquoi pas, le texte dans son ensemble. Précisons, cependant, que chez Halliday, ce ne sont que les deux premiers niveaux qui ont fait l'objet d'une description, alors que les deux derniers sont avant tout évoqués à partir d'une projection toute théorique de la méthode. D'un autre côté (Marandin, 1988), nous avons affaire à une perspective tout à fait différente. Toute pertinence est niée au thème propositionnel et la description vise l'identification du thème de discours — thème configuré, thème inféré —. Mais comme le dit, lui-même, Marandin (1988), cela représente un "travail dont on a déjà mesuré à quel point il nécessite encore un long travail empirique et conceptuel mettant en jeu des champs différents: description linguistique, description textuelle, description des modes et des formes de la compréhension", p. 27.

Nous privilégions l'approche fonctionnelle, dans la mesure où elle est compatible avec la méthode de lexicométrie "qualifiée" que nous avons adoptée. Mais au-delà de sa commodité, l'approche fonctionnelle nous semble porter ses fruits au niveau de la description des contenus du discours et cela, sans que nous sous-estimions les limites actuelles que rencontre la lexicométrie dans la prise en compte de la présence massive des locutions nominales[6] dans le discours et du phénomène de l'anaphorisation[7].

---

[5] Ici, nous faisons une distinction entre ce que, d'un côté, nous appelons, par commodité, les méthodes qualitatives, qui ne visent pas la formalisation, et, d'un autre côté, les méthodes qui se définissent par leur systématicité, pour ne pas dire leur formalisme. Dans le cas des analyses qualitatives, les unités de contexte sont indéterminées apriori, variables et, en fait, délimitées par la présence du thème. En effet, les méthodes "qualitatives" identifient le thème comme objet du discours à partir de l'intuition et du jugement du lecteur-expert et, selon les approches, considèrent que les thèmes peuvent s'emboîter les uns les autres, comme les unités qui les contiennent, s'emboîtent elles-mêmes pour former des organisations de plus en plus complexes.

[6] Tout en disposant de modules informatiques pour le repérage et le marquage des locutions nominales, nous ne les utilisons pas ici parce que nous aurons

L'hypothèse lexicométrique pose que la plus ou moins grande récurrence des unités de signification dans le discours produit du sens. L'analyse thématique propositionnelle permet d'étiqueter tout mot exerçant une fonction thématique dans la proposition et, comme nous le verrons plus loin, de complexifier l'étiquetage selon une typologie de fonctions thématiques. C'est ce à quoi nous faisons allusion quand nous parlons de lexique "qualifié". L'analyse des données ainsi produites nous permet d'accéder à une structure thématique complexe qui nous dira, en effet, de quoi parle le discours.

2.3. La hiérarchisation thématique

La définition du thème implique que nous opérationalisions les modalités de son repérage. On reconnaîtra le thème à la première position qu'il occupe dans une proposition. Ce critère de la première position s'appuie sur le postulat que dans l'usage de l'anglais ou du français, le locuteur énoncera au point de départ de toute proposition l'objet de son message et qu'il le développera par la suite. Il peut s'agir d'un groupe nominal (GN), d'un groupe adverbial (GA) ou d'un groupe prépositionnel (GPr). On parlera alors de thème singulier — simple ou complexe[8] —.

Le critère premier d'identification du thème est donc de nature positionnelle, c'est-à-dire la première position dans la proposition. Ce critère ne permet cependant pas, à lui seul, de clôturer l'identification du thème. Rappelons que Halliday définit trois méta-fonctions sémantiques: idéationnelle (acteur), interpersonnelle (sujet) et textuelle (thème). La méta-fonction textuelle se réalise par la position des éléments dans la proposition. Elle n'est cependant pas indépendante des autres. C'est pourquoi Halliday affirme qu'il doit toujours y avoir un élément idéationnel dans le thème. Ainsi, dans les cas où certains mots-outils occuperaient la première position de la proposition, tout en ne fonctionnant ni comme un sujet ni comme un complément direct ou circonstanciel, alors le sujet ou le complément qui suit

---

surtout l'occasion de travailler sur des catégories sémantiques projetées sur les textes.

[7] Nous posons, pour l'instant et à des fins méthodiques, que dans des ensembles larges de données textuelles, l'anaphorisation serait un phénomène plutôt constant.

[8] La discussion des multiples distinctions apportées par Halliday ne saurait être développée ici. Nous renvoyons plutôt le lecteur aux choix que nous avons effectués pour notre propre recherche, tels qu'ils sont définis dans la section 4.

immédiatement fera partie du thème. Il s'agit alors d'un thème multiple. Mais, dans tous les cas le thème de la proposition comprendra toujours le thème "topic" — c'est-à-dire le thème lui-même dans le cas du thème singulier ou son extension lorsqu'il s'agit d'un thème multiple —. Pour résumer, nous pourrions alors dire que le thème est le contenu qui occupe la première position dans la proposition.

Une fois résolu le problème de son identification, qu'en est-il de celui de sa hiérarchisation? Celle-ci concerne la variation potentielle de l'importance relative des thèmes occupant des places différentes dans les divers niveaux du discours. Nous répondrons qu'il existe deux axes sur lesquels peuvent se mesurer des différences. Le premier renvoie au marquage linguistique du thème et le second à la position occupée dans la hiérarchie des propositions au sein d'une même phrase.

Le marquage du thème se manifeste dans un écart par rapport au schéma syntaxique de la proposition qui n'est pas régi par des contraintes grammaticales. La séquence française de base est la suivante: SUJET + VERBE + OBJET. Selon J. Dubois et al. (1970, p. 33), la position est un trait distinctif du syntagme. Ils reprennent en cela la tradition classique. Dans la préface à son Traité sur L'ordre des mots dans la phrase latine, J. Marouzeau (1950, p. II) énonce d'emblée que "si, en latin, l'ordre des mots est libre, il n'est pas indifférent". En français, alors que l'ordre des mots est contraint, nous supposons que des effets de sens sont produits par l'usage de figures ou de procédés d'inversion. D'ailleurs l'inversion est répertoriée par P. Fontanier: il s'agit d'un : "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." (1827 p. 284). J. Dubois et al. considèrent que la séquence OBJET + VERBE + SUJET est une inversion parfaite (1973, p. 84).

Chez Halliday le marquage du thème se révèlera par la non-coïncidence de la fonction thème et de la fonction sujet dans une même unité. Comme la norme langagière, autant en anglais qu'en français, est d'énoncer d'abord le sujet, nous pouvons adopter ce critère. Ainsi, lorsque le groupe de mots occupant la première position de la proposition ne rempliera pas la fonction de sujet grammatical, il sera considéré marqué. Il pourra s'agir d'un GN complément antéposé, d'un GA ou GPr. Une autre forme de marquage se rencontrera dans les les constructions prédiquées, comme dans la forme "c x que p" (Courtine); le x sera alors tenu pour thème

marqué. Dans tous les cas, le marquage du thème est vu comme ajoutant un poids particulier à la valeur thématique d'une unité.

La seconde manière de considérer la hiérarchisation des thèmes propositionnels est de considérer la place qu'ils occupent dans la structure propositionnelle que représente une phrase. De la même manière que la première position dans la proposition permet de déterminer ce qui est thématisé, ne serait-il pas pertinent de considérer que la proposition principale d'une phrase — généralement en première position — implique que son thème est plus important que ceux des autres propositions? Cette question renvoie à une discussion étendue sur la représentation en grammaire fonctionnelle de la complexité des phrases (voir Halliday, 1985 et Bureau, date). Nous nous contenterons ici des distinctions classiques entre principale, coordonnées et subordonnées et ferons l'hypothèse que la valeur des thèmes propositionnels rattachés à chacun de ces types peut être ordonnée sur une échelle.

Aux fins de notre recherche, nous ordonnerons trois positions pouvant être occupées par un thème dans la structure propositionnelle: 1. le thème d'une subordonnée antéposée; 2. le thème de la proposition principale; 3. le thème de toute autre proposition. Le thème propositionnel de la principale en première position est, en général, considéré comme le thème de la phrase. Il peut cependant arriver qu'une proposition subordonnée soit antéposée à la principale. En raison de cette inversion de l'ordre attendu, son thème sera alors réputé de plus grande importance. Les thèmes de toutes autres propositions coordonnées et subordonnées devraient être considérés, dans l'ordre, de moindre importance. En raison de la complexité de la tâche de description, nous renonçons à les distinguer du point de vue du degré d'importance et les considérerons tous deux comme comportant une valeur thématique plus faible que les précédents.

Enfin, dans la mesure où certaines subordonnées (relatives et participiales) rattachées à un groupe nominal n'appartiennent pas directement à la structure propositionnelle de la phrase, puisque subordonnées à un groupe nominal qui est une unité inférieure à la proposition, nous ne retiendrons pas le thème de ces propositions. Nous obtenons ainsi

Tableau 1

Catégories de thèmes en fonction du marquage et de la localisation

	thème Marqué	Non-Marqué
Subord.antéposée	TMSA	TNMSA
Principale	TMP	TNMP
Autre	TMA	TNMA

3. Le matériel

3.1. Nature du corpus

Notre travail actuel s'inscrit dans la continuité d'une recherche de longue haleine sur le discours politique déployé durant la période duplessiste au Québec (1936-1960)[9]. Nous nous sommes en effet intéressés au rôle du discours politique dans les transformations de la société québécoise de la crise des années 30 à l'aube de la Révolution tranquille. Notre questionnement portait sur les transformations observables dans le discours qui puissent témoigner non seulement de la modernisation des structures économiques ou politiques, mais également des représentations sociales. Nous avons donc cherché à réunir non pas les traces plus ou moins doctrinaires des idéologies de l'époque, mais plutôt à reconstituer le discours politique de masse tel qu'il émane et circule à partir de diverses institutions publiques. Cela explique la taille du matériel recueilli (5,000 pages) et sa diversité. Nous avons retenu plusieurs manifestations du discours politique en tant que tel (discours sur le budget, discours du trône, discours constitutionnel, discours legislatif, discours électoral), mais également les discours à dimension sociale et politique émanant de diverses institutions de la sphère publique (certains mandements des évêques, le discours d'action catholique, les discours syndicaux et patronaux).

Afin d'éclairer la démarche qui suit, il faut insister sur trois caractéristiques de notre corpus. Le discours politique représente un genre discursif particulier. Ce que nous pourrons dire de la structure thématique de ces discours ou de la pertinence de nos catégories de thème ne vaut que pour ce genre et cette époque donnée. Deuxièmement, notre corpus global est constitué de sous-ensembles qui renvoient

[9]Mettre les principales références de la recherche.

à des contenus idéologiques différents et, possiblement, à des formes discursives également variables. Il sera donc pertinent d'amorcer notre réflexion sur les variations entre ces formes. Troisièmement, nous travaillons sur de grands ensembles textuels. Cela présuppose la mise sur pied de stratégies informatiques dont l'objectif est double: donner accès à la connaissance de larges contenus à diffusion massive, tout en permettant une lecture de plus en plus raffinée des données.

3.2. Analyse de lexiques qualifiés sur la base de grands ensembles textuels.

Notre démarche se distingue de celle de la lexicométrie classique essentiellement par le travail de description que nous avons pratiqué sur nos données. Les systèmes informatiques retenus nous ont ainsi permis de produire ces descriptions et de constituer, à partir de différents modèles d'exploration des données textuelles, des lexiques "qualifiés" par la syntaxe ou encore à la sémantique. L'analyse des données (matrices de mots ou de catégories distribués selon divers plans de segmentation) relevait alors des outils statistiques consacrés.

Nous avons d'abord produit une description syntaxique des textes du corpus. Suite à la catégorisation morpho-syntaxique produite par CBSF, nous avons appliqué un analyseur syntaxique. Depuis la fin des années 70, plusieurs analyseurs syntaxiques (Parsers) ont été construits avec plus ou moins de succès pour la description de textes en langue naturelle. Deux approches principales caractérisent ces analyseurs. La première s'intéresse davantage à la simulation de modèles linguistiques formels qu'à la description de textes en tant que telle. La seconde met en oeuvre des stratégies heuristiques et vise une couverture beaucoup plus large. GDSF est un analyseur de ce type. Il identifie, pour toute proposition, des "relations de dépendance contextuelle": thème/propos, détermination, indications sur les compléments verbaux.

La seconde description renvoie à la catégorisation des nominaux et adjectivaux à partir d'une grille de 144 catégories en grande partie sociologiques, regroupées en 6 familles dont chacune est subdivisée à son tour en domaines. Les trois premières familles réfèrent à des sphères institutionnalisées: l'économie, le politique et des institutions particulières. La quatrième se rapporte aux principales catégories de l'univers social. La cinquième définit les principales valeurs véhiculées

dans le discours. La sixième réunit les notions qui exercent un rôle de foncteur sémantique. L'idée derrière cette catégorisation est de permettre le traitement regroupé d'éléments sémantiques jugés temporairement "équivalents". Notre système permet, par contre, de régresser en tout temps à un état antérieur du corpus et de traiter les mots non-catégorisés.

3.3. le discours sur le budget

Le travail que nous effectuons actuellement vise l'optimisation des descriptions produites sur les divers sous-corpus afin d'en accroître le potentiel d'analyse. Le premier livre publié sur nos travaux porte avant tout sur l'analyse du discours sur le budget. Une des composantes de cette analyse repose sur le traitement thématique. Cela nous motive donc à revenir sur cette première analyse afin d'en mesurer l'écart avec la nouvelle analyse que nous proposons. Le discours sur le budget est une pièce essentielle dans le dispositif discursif du gouvernement. Discours d'orientation économique du gouvernement, prononcé normalement sur une base annuelle, il excède cette dimension économique et représente un des énoncés officiels de politique générale. Ce discours a été retenu dans son ensemble sur toute la période de 1934 à 1960. Nous appliquerons donc notre typologie thématique sur ce discours afin de vérifier son potentiel de description et sa pertinence empirique.

4. L'algorithme

Comme nous l'avons dit plus haut, la description de la structure thématique a été produite à partir de l'application de l'analyseur syntaxique GDSF. Nous avons d'abord validé deux aspects de la description syntaxique obtenue: l'identification et la structure des propositions de la phrase et l'assignation de la relation de dépendance thème/propos.

Dans le premier cas, l'analyse des propositions introduites par QUE ou QU' catégorisées complétives a été raffinée. Si elles étaient précédées d'un GN, elles étaient recatégorisées relatives. Cette distinction était nécessaire en vertu du principe d'exclusion des propositions relatives ou participiales rattachées au groupe nominal (GN) thématisé.

La seconde modification se rapporte aux définitions plus complexes que nous avons proposées du thème et de la hiérarchisation thématique. Dans GDSF,

est considéré comme thème, le premier groupe nominal (GN) au complet qui précède le groupe verbal (GV). Habituellement, il s'agit du sujet grammatical. Notre algorithme dépiste et distingue toutes les classes de thèmes que nous avons définies dans le premier tableau.

Le traitement se fait séquence par séquence; une séquence de texte est délimitée par 2 ponctuations fortes. La séquence comporte au moins une proposition notée (GP). Le GP est l'unité de base retenue pour l'analyse thématique; la séquence est l'unité supérieure qui, en gros, correspond à la phrase. Un GP est composé de groupes de mots ou syntagmes. Enfin, un GP est bien formé s'il comporte un groupe nominal (GN) et un groupe verbal (GP) dans le cadre de la théorie grammaticale sous-jacente à GDSF[10]. Rappelons que le thème de la proposition consiste en son premier groupe; par extension, le thème de la séquence est le celui de la première proposition.

Les séquences sont lues récursivement, GP par GP. En premier lieu, la nature des constituants des GP est inventoriée et ceux qui sont mal formés ne sont pas analysés plus avant. Les propositions relatives, rattachées à des GN, ne sont pas considérées. Les GP bien formés sont ensuite numérotés, ce qui permettra de distinguer le premier des autres : il s'agit du thème de la séquence. Puis, les GP subordonnés sont dépistés à partir de la présence d'une conjonction de subordination.

Enfin, le thème, soit la première position du GP, est analysé afin de déterminer s'il est marqué ou non. Le GN sujet est un thème non marqué, le GN complément, tel que repéré par GDSF, est un thème marqué. Tous les mots du GN sont considérés thèmes, sauf les relatives ou les participiales. Les groupes adverbiaux sont considérés thèmes marqués. Enfin, les subordonnants sont considérés thème, de même que le groupe qui suit immédiatement (thème multiple). La coordination n'entraîne pas de changement de position.

Un traitement particulier est réservé aux GP prédiqués en "c'est X que P". Les GP formés par le QUE, en vertu du réaménagement de GDSF, sont alors

---

[10] "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)

recatégorisés relatives et excluent la prise en compte de leur thème. Le GN qui forme le x est, par contre, traité comme un thème marqué.

5. Premières conséquences d'une nouvelle analyse thématique

En proposant une nouvelle description de la structure thématique du discours du budget, nous introduisons deux différences majeures par rapport à la première analyse proposée dans Bourque et Duchastel (1988). La première concerne l'introduction de plusieurs catégories de thème selon leur marquage et le niveau de leur localisation. La seconde consiste en une réduction du nombre de thèmes retenu par la deuxième description. Nous nous interrogerons, tour à tour, sur les conséquences de ces changements.

Tableau 2
Fréquence relative des thèmes selon leur catégorie

Thème	Marqué	Non-Marqué	
Subord./antéposée	0.21%	4.01%	
Principale	14.52%	58.44%	
Autres	1.98%	20.84%	
Total	**16.71%**	**83.29%**	**100%**

Constatons d'abord que la thèse voulant que le thème corresponde, dans la plupart des cas, au sujet grammatical est confirmée par les résultats globaux. En effet, 83,29% de toutes les occurrences thématisées ne sont pas marquées. Mais, une première analyse de ces données nous a poussés à reconsidérer la pertinence de garder intactes les six catégories différentes de thèmes, du moins pour les besoins d'une première analyse des données. En effet, les thèmes marqués dans les subordonnées antéposées ne représentent qu'approximativement 2 occurrences sur 1000 thèmes, alors que les thèmes marqués dans les propositions catégorisées "autres" comptent pour 20 occurrences sur 1000. Nous avons donc

retenu le principe de leur regroupement avec les thèmes marqués de la principale. Par ailleurs, il nous a semblé que les thèmes non-marqués apparaissant dans les subordonnées antéposées pouvaient être aussi considérés comme appartenant à la catégorie des thèmes marqués, selon le principe qu'un effet de marquage est produit, sinon au niveau de la proposition, du moins au plan de l'ordre attendu des propositions. Nous obtenons ainsi une catégorie regroupée (TMSA, TMP, TMA, TNMSA) comptant 20.72% de toutes les occurrences des thèmes. La catégorie des thèmes non-marqués de la principale resterait intacte, soit 58.44% de l'ensemble. Enfin, les thèmes non-marqués des propositions "autres" constitueraient une troisièmes catégorie, comportant 20.84% des occurrences thématisées. Dans les exemples d'analyse que nous développons par la suite, nous nous en sommes tenus à la catégorie regroupée des thèmes marqués.

Indépendamment des distinctions opérées entre diverses catégories thématiques, la nouvelle description a produit une réduction globale du nombre de thèmes repérés dans le corpus. En conséquence, les indices de thématisation compilés pour chacune des 5 familles de catégories sont en baisse. On trouvera au tableau 3, la fréquence en nombre absolu des grandes familles de catégories sociologiques, les fréquences de ces catégories lorsqu'elles étaient en position de thème (1ère et 2ème description) et les deux indices de thématisation calculés à partir de ces chiffres. Le tableau indique également le taux de marquage pour chacune de ces grandes familles de catégories sur la base de la nouvelle description.

Tableau 3: Les indices de thém.....(mettre en note de ce tableau que les Us7 et 8 sont exclus

La lecture du tableau 3 nous permet de mesurer une baisse de 23.55% de l'indice global de thématisation intervenue entre la première et la deuxième description. La baisse globale est largement explicable par l'exclusion des thèmes des propositions relatives ou participiales rattachées à des groupes nominaux. Cette baisse n'est cependant pas uniforme d'une famille de catégories à l'autre. Ainsi, par ordre décroissant, les pourcentages de baisse sont de 29.04% (famille des valeurs), 27.21% (famille de l'univers social), 25.26% (famille politique), 20.83% (famille économique), 15.94% (famille des institutions). Il faut ici introduire une seconde explication de ces baisses. En effet, le nouvel algorithme retient le premier groupe de mots du GP, qui, contrairement à ce que dépiste le premier algorithme,

n'est pas nécessairement le sujet. C'est ainsi que nous pouvons maintenant distinguer les thèmes marqués des thèmes non-marqués. En somme, entre les deux descriptions, intervient un jeu de substitutions. A certains thèmes non-marqués dépistés par le premier algorithme, sont substitués des thèmes marqués dépistés par le second. C'est donc dire que pour chacune des familles de catégories les baisses sont explicables, en partie, en raison de l'enfouissement de certaines occurrences dans les relatives ou participiales et, en partie, par les substitutions résultant de l'application du nouvel algorithme. Par exemple, le thème d'une proposition d'abord identifié comme une catégorie économique non-marquée pourrait être remplacée par une catégorie institutionnelle marquée.

Comment interpréter ces transformations? Dans le cas des thèmes des relatives ou des participiales, on peut affirmer que nous perdons les occurrences qui ont peu ou pas de valeur thématique. Dans le cas des substitutions, ces dernières se font, par définition, au profit des thèmes marqués. Dans tous les cas, les thèmes dépistés gagnent en importance dans la hiérarchie thématique. L'examen, de nouveau, du tableau 3 indique que l'ordre d'importance des indices de thématisation d'une famille de catégories à l'autre n'est pas modifié: Politique (respectivement: 39.52%; 29.54%); Univers social (36.23%; 26.37%); Economie (32.80%; 25.97%); Institutions (30.75%; 25.85%); Valeurs (22.57%; 16.01%). Mais, alors que la famille des catégories politiques tend vers la moyenne, les indices de thématisation des catégories économiques et institutionnelles ont moins tendance à baisser et ceux des catégories reliées à l'univers social et aux valeurs ont une tendance plus forte à la baisse. Tout ce que nous pouvons dire pour le moment, c'est que les catégories économiques et institutionnelles qui ont été dépistées par le premier algorithme était déjà dans une meilleure position dans la hiérarchie thématique, alors que les catégories de l'univers social et des valeurs étaient plus dispersées dans la trame du discours. La nouvelle description devrait nous assurer une meilleure base pour l'analyse.

Etudions maintenant les rapports qui s'établissent entre la thématisation et le marquage. A un niveau très général, on peut dire que les familles de catégories se distribuent dans trois classes du point de vue de leur indice de thématisation: une classe où l'on observe un comportement modal autour de 26% pour les familles économie, institutions et univers social; une classe légèrement au dessus de la première où l'on retrouve les catégories politiques; enfin, une classe très en dessous de la classe modale où se situent les valeurs. On peut faire ressortir

que le discours thématise avec plus de force le politique, ce qui est attendu dans le cas d'un discours politique. La faible thématisation des valeurs est peut-être également le fait de ce genre discursif. Ce qu'il importe de noter cependant, c'est qu'alors que trois des familles de catégories oscillent entre un taux de 15 et de 21% de marquage du thème, cette même famille des valeurs a presque 27% de ses thèmes qui sont marqués. De même, la famille où les thèmes sont le moins marqués est celle des catégories économiques (12.1%).

La famille des catégories économiques, bien qu'elle soit globalement très thématisée (25.97%), présente la plus faible proportion de thèmes marquée (12.1%). L'univers économique constituant l'arrière fond d'un discours sur le budget, on peut faire l'hypothèse que le travail discursif ne lui réserve pas de traitement particulier. Cet univers est omniprésent, attendu et prévisible. Il ne serait pas l'objet central de la "stratégie discursive". Bien qu'il soit la raison première pour laquelle se tient ce discours, il semble, cependant, qu'il devienne l'occasion de discourir sur toute autre chose. Ainsi, les familles extra-économiques auront tendance à être beaucoup plus marquées. Le comportement global de la famille des catégories de valeurs en constitue un exemple convaincant. Lorsque le discours thématise des catégories recouvrant des valeurs, il en fait l'objet d'une forte focalisation en les marquant. L'appel aux valeurs renvoie ainsi à un dispositif discursif nettement repérable en vertu duquel des considérations d'ordre philosophique, existentiel ou politique émergent à la surface du discours, se détache du bruit de fond de l'économique pour produire leurs effets.

Bibliographie

Beauchemin, J. et Paquin, L.C., "Apport de l'ordinateur à l'analyse des données textuelles", à paraître dans les actes du colloque La description des langues naturelles en vue d'applications informatiques, Université Laval, Centre internatioal de recherche en aménagement linguistique, décembre 1988.

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

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.

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

Duchastel, J., Dupuy, L., Paquin, L.-C., Beauchemin, J., Daoust, F., "SACAO, Computanional Applications for Textual Data Analysis in Social Sciences", in Advances in Computing and the Humanities. Content, Concepts, Meaning, Israël, Ben-Gurian University, 2 vol., à paraître.

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.

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

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.