Rubrique préparée par Fiammetta Namer

Université Nancy2 de Nancy, UMR « ATILF »
Fiammetta.Namer@univ-nancy2.fr

Frédérik BILHAUT (fbilhaut@info.unicaen.fr)

Titre : Analyse automatique de structures thématiques discursives - Application à la recherche d'information.

Mots-clés : traitement du langage naturel, analyse du discours, sémantique, recherche documentaire.

Title : *Automatic analysis of discursive thematic structures - Application to information retrieval.*

Keywords : *natural language processing, discourse analysis, semantics, information retrieval.*

Thèse de doctorat en Informatique, Université de Caen, UFR de Sciences, Laboratoire GREYC UMR 6072, sous la direction du Professeur Patrice Enjalbert. Soutenue le 14/06/2006.

Jury : Mme Marie-Paule Péry-Woodley (Pr, Université de Toulouse-Le Mirail & ERSS, présidente), M. Patrice Enjalbert (Pr., Université de Caen et GREYC, directeur), M. Benoît Habert (Pr., Université de Paris 10, rapporteur), Mme Adeline Nazarenko (Pr., Université de Paris 13, rapporteur), M. Michel Charolles (Pr., Université de Paris 3, examinateur), M. Philippe Laublet (MC, Université Paris 4, examinateur), M. Jacques Vergne (Pr., Université de Caen, examinateur invité).

Résumé : *Le problème de l'analyse thématique des textes, qui vise l'étude de leur structure selon des critères relatifs à la répartition de leur contenu informationnel, est d'une importance capitale dans le contexte de l'accès assisté à l'information. Quel est le sujet d'un texte ou d'un passage ? À quel propos apporte-t-il de l'information ? Comment cette information est-elle répartie dans le discours ? Telles sont les questions auxquelles on doit pouvoir répondre pour sélectionner les documents pertinents relativement à une requête, pour aider à la navigation dans des documents longs, ou encore pour produire un résumé d'un texte. Cette question n'est cependant pas toujours abordée explicitement en ces termes : nombre de travaux en recherche d'information (RI) ou en traitement automatique des langues (TAL) gardent certaines distances avec la notion de thème, et les apports sur ce*

même sujet de disciplines telles que la linguistique ou les sciences de l'information restent relativement peu considérés.

Notre travail a consisté à envisager l'apport de théories de cet ordre dans le cadre de l'analyse automatique de la structure thématique du discours. Tout en conservant une réelle visée applicative, il s'agit plus précisément de se concentrer sur la réalité linguistique et documentaire que recouvre la notion de thème, plutôt que sur celles des contraintes opératoires qui prévalent habituellement. Nous nous confrontons ainsi à d'autres problèmes tout aussi épineux, puisqu'il n'existe pas à ce jour de théorie consensuelle concernant la notion de thème au niveau documentaire ou même discursif. Il ne s'agit donc pas seulement d'opérationnaliser des modèles descriptifs, mais de chercher à recenser parmi les théories existantes des éléments susceptibles de servir au mieux les objectifs applicatifs de la recherche d'information. C'est le but que nous nous sommes fixé, en tentant de conserver en parallèle une vue sur les approches théoriques de la notion de thème et une visée applicative marquée.

La première partie de notre mémoire s'attache à présenter un parcours bibliographique autour de la notion de thème, aussi bien en sciences de l'information qu'en linguistique et en TAL. Nous croisons à cette occasion les notions de structure informationnelle, de topique discursif, de progression thématique, de pragmatique discursive ou encore de formalisation logique de l'à-propos. Nous concluons cette partie par un bilan visant à faire émerger un certain nombre de lignes de forces qui sous-tendent notre approche de la notion de thème en tant qu'objet discursif, sémantique, structuré, et attaché à la structure des connaissances liées à un domaine.

Une deuxième partie est consacrée à la description de systèmes et de modèles de traitement automatique des langues s'appuyant sur ces principes. Nous décrivons tout d'abord des travaux, réalisés au sein du projet ***GeoSem****, portant sur l'analyse sémantique du discours géographique à des fins de RI.. Par la suite, nous décrivons deux axes de recherche qui ont résulté de ce projet, le premier concerne l'****analyse automatique des cadres de discours****. Nous proposons une première solution au délicat problème de l'analyse automatique de la portée des introducteurs, en présentant une méthode qui permet de reconnaître automatiquement les bornes des cadres temporels. Le second axe concerne la notion de* ***thème composite****, que nous avons développée dans le prolongement des travaux précédemment décrits, et qui constitue un modèle pour l'analyse thématique d'une certaine variété de structures discursives liées à la notion d'univers de discours. Nous présentons le modèle en lui-même avant de décrire la méthode d'analyse thématique automatique qui en découle. Nous envisageons les liens entre ce modèle et des concepts existants tels que l'encadrement du discours, la structure informationnelle ou encore la théorie de la structure rhétorique, avant d'introduire une notion d'****axe sémantique*** *que nous posons comme pivot entre l'organisation des connaissances d'un domaine et la structure thématique des textes qui s'y rapportent.*

*La troisième et dernière partie décrit la plate-forme **LinguaStream**, que nous avons développée parallèlement aux travaux précédemment évoqués afin de faciliter leur élaboration, et qui est devenue une plate-forme générique pour le traitement automatique des langues. À travers cette plate-forme, qui se veut avant tout un « laboratoire virtuel » pour le TAL, nous proposons un certain nombre de principes méthodologiques applicables aux problématiques de l'annotation des documents électroniques et surtout à la constitution de procédés d'analyses complexes fondés sur la formalisation de modèles d'ordre linguistique.*

URL où la thèse pourra être téléchargée :
- http ://www.info.unicaen.fr/~fbilhaut (thèse)
- http ://www.linguastream.org (plate-forme)

Marie-Laure GUENOT (mlg@lpl.univ-aix.fr)

Titre : Éléments de grammaire du français pour une théorie descriptive et formelle de la langue.

Mots-Clés : syntaxe, développement de grammaire, linguistique descriptive, linguistique formelle, Grammaires de Propriétés (GP).

***Title :** elements of french grammar for a descriptive and formal theory of natural language.*

***Keywords :** syntax, grammar development, descriptive linguistics, formal linguistics, Property Grammars (PG).*

Thèse de doctorat/PhD en Sciences du langage, Université de Provence (Aix-Marseille I), UFR LACS, Laboratoire Parole et Langage (LPL) sous la direction de Philippe Blache (Directeur de recherches CNRS). Soutenue le 07/12/2006.

Jury : M. Philippe Blache (DR., LPL, directeur), M. Sylvain Kahane (Pr., Université Paris 10 & MoDyCo, rapporteur), M. Jean-Yves Morin (PR, Université du Québec à Montréal, rapporteur), M. Henri-José Delofeu (Pr., Université de Provence et DeLIC, examinateur), M. Christian Rétoré (Pr, Université de Bordeaux I & LABRI, examinateur), M. Eric Wehrli (Pr., Université de Genève, examinateur)

***Résumé :** Dans cette thèse nous proposons un modèle de grammaire basé sur une théorie originale de la langue, et représenté formellement. Elle s'articule en trois parties. Dans la première partie nous dressons un bilan des positions théoriques sous-jacentes à un certain nombre de grammaires formelles coexistantes, afin de*

faire ressortir quatre éléments dont la conjonction fait de notre proposition une approche nouvelle : le non-générativisme, la non-modularité, la non-lexicalisation et la multi-dimensionnalité. Nous présentons ensuite le formalisme des Grammaires de Propriétés (GP), que nous avons utilisé pour représenter notre grammaire. À la suite de cela nous introduisons notre modèle de grammaire, basé sur les choix théoriques précédents et formalisé en GP, et nous proposons quelques formalisations d'autres modèles afin d'illustrer les possibilités de GP et l'originalité de notre modèle. Dans la seconde partie nous proposons un ensemble de descriptions syntaxiques du français fondées sur notre modèle et constituant un noyau de grammaire ; nous y présentons notamment les constructions nominales, verbales, adjectivales, propositionnelles, ainsi que les entassements paradigmatiques (coordinations et disfluences). Enfin, dans la troisième partie nous illustrons le fonctionnement de notre proposition avec l'analyse de quelques phénomènes syntaxiques, dont notamment le traitement des pronoms clitiques dans les constructions verbales, et celui des coordinations et des disfluences. Ce travail apporte à toute une partie de la linguistique descriptive une validation par son expression formelle, et à la linguistique formelle l'intégration de descriptions syntaxiques jusqu'ici non encore prises en considération. En outre, elle apporte une validation de GP en tant que formalisme linguistique en montrant ce qu'il permet par sa souplesse de représentation.

URL où la thèse pourra être téléchargée :
http ://www.lpl.univ-aix.fr/~guenot/Memoires/These-MLG.pdf

Erwan MOREAU (erwan.moreau@univ-nantes.fr)

Titre : Acquisition de grammaires lexicalisées pour les langues naturelles

Mots-Clés : apprentissage automatique, inférence grammaticale, modèle de Gold, identification à la limite, grammaires lexicalisées, grammaires catégorielles, langues naturelles.

Title : *Learning lexicalized grammars for natural languages*

Keywords : *automatic learning, grammatical inference, Gold's model, identification in the limit, lexicalized grammars, categorial grammars, natural languages.*

Thèse de doctorat en Informatique, Université de Nantes, Faculté des Sciences, département d'Informatique/UFR Sciences et Techniques, LINA-FRE CNRS 2729, sous la direction du Professeur Alexandre Dikovsky. Soutenue le 18/10/2006.

Jury : Mme Béatrice Daille (Pr, Université de Nantes, présidente), M. Alexandre Dikovsky (Pr., Université de Nantes, directeur), M. Makoto Kanazawa (Pr., National Institute od Informatics de Tokyo, rapporteur), M. Jean-Yves Marion (Pr., Université de Nancy, rapporteur), Mme Annie Foret (MC, Université de Rennes 1, examinatrice invitée), M. Christian Rétoré (Pr., Université de Bordeaux 1, examinateur), Mme Isabelle Tellier (MC HDR, Université de Lille 3, examinatrice).

Résumé : *L'inférence grammaticale désigne le problème qui consiste à découvrir les règles de formation des phrases d'un langage, c'est-à-dire une grammaire de celui-ci. Dans le modèle d'apprentissage de Gold, les exemples fournis sont constitués uniquement des phrases appartenant au langage. L'algorithme doit fournir une grammaire qui représente le langage énuméré. Les grammaires catégorielles sont l'un des nombreux formalismes existants pour représenter des langages. Kanazawa a montré que certaines sous-classes de ces grammaires sont apprenables, mais ses résultats ne sont pas applicables directement aux langues naturelles.*

Sur le plan théorique, nous proposons de généraliser les résultats de Kanazawa à différents types de grammaires. Les grammaires combinatoires générales sont un modèle flexible permettant de définir des systèmes grammaticaux à base de règles de réécriture. Nous démontrons dans ce cadre que certaines classes de langages sont apprenables. Dans un souci de généralité maximale, nos résultats sont exprimés sous forme de critères sur les règles des systèmes grammaticaux considérés. Ces résultats sont appliqués à plusieurs formalismes relativement adaptés à la représentation des langues naturelles.

Nous abordons également le problème de la mise en oeuvre de l'apprentissage sur des données réelles. En effet, les algorithmes existants capables d'apprendre des classes de langages intéressantes sont NP-complets. Afin de contourner cet obstacle, nous proposons un cadre d'apprentissage plus souple, l'apprentissage partiel : le contexte d'utilisation est modifié dans le but d'obtenir une complexité algorithmique plus réaliste. Nous testons cette approche sur des données de taille moyenne, et obtenons des résultats relativement encourageants.

URL où la thèse pourra être téléchargée :
http ://www.sciences.univ-nantes.fr/info/perso/permanents/moreau/publis/these_Erwan_Moreau_2006.pdf

---

Christophe BENZITOUN (Christophe.Benzitoun@up.univ-aix.fr**)**

Titre : Description morphosyntaxique du mot *quand* en français contemporain.

Mots-Clés : linguistique sur corpus, approche pronominale, quand, mots qu-,

syntaxe descriptive, liens inter-constructionnels.

Title : *A corpus based morphosyntactic description of quand in contemporary French.*

Keywords : *corpus linguistics, pronominal approach, quand, qu- words, descriptive syntax, clause linkage.*

Thèse de doctorat en Linguistique française, Université de Provence, UFR Lettres Arts et Sciences du langage (LACS), Équipe Description Linguistique Informatisée sur Corpus (DELIC), sous la direction du Professeur Henri-José Delofeu. Soutenue le 01/12/2006.

Jury : M. Frédéric Sabio (MC, Université de Provence, président), M. Henri-José Delofeu (Pr., Université de Provence, directeur), M. Bernard Combettes (Pr., Université de Nancy 2, rapporteur), M. Michel Pierrard (Pr., Vreie Universiteit Brussel, rapporteur), Mme Marie-José Béguelin (Pr, Université de Neûchatel, examinatrice).

Résumé : *Malgré la richesse de ses usages et les problèmes théoriques dont il est l'objet, le mot «* quand » *n'a jamais été au centre d'une étude spécifique. Or, d'autres mots assurant un lien tels que* où *(Hadermann, 1993),* si *(Delaveau, 1990),* que *(Deulofeu, 1999) ou* parce que *(Debaisieux, 1994) ont été les sujets de récentes études. Quelques auteurs parmi lesquels Sandfeld (1936), Chétrit (1976), Olsson (1971) et Borillo (1988) lui ont consacré des ouvrages ou des articles, mais le traitement de* quand *s'insérait dans des problématiques plus larges, pour les deux premiers (l'étude des « propositions temporelles »), ou plus restreintes, pour les deux suivants (l'étude des temps dans la « temporelle » et dans la « principale »). Cet item possède pourtant des fonctionnements intéressants ne relevant pas forcément de la « subordination circonstancielle ».*
En effet, on observe des exemples qui ne semblent pas poser de problème particulier et qui correspondent bien à la définition de la « subordination circonstancielle ».

1) *«* Bonsoir. Tu remettras la clef au concierge **quand** tu seras prête. Je n'attendrai pas ton bon plaisir. *» [Maupassant,* Bel Ami*]*

D'autres, en revanche, ne rentrent pas dans ce cadre. Dans 2), la construction introduite par quand *semble être en position « d'objet direct », ce qui ne va pas de soi pour une « subordonnée circonstancielle de temps ».*

2) J'aime **quand** on a notre propre étiquette, notre propre emblème et qu'on vend les produits touristiques venant du Québec. *[Hansard]*

Dans 3), la construction en quand *(*Quand-C*) s'insère dans une pseudo-clivée, dispositif syntaxique dans lequel la marque à l'initiale (*ce qui*) repère une position (ici sujet) normalement instanciée par l'élément se trouvant après* c'est *(Roubaud, 2000).*

3) Ce qui a été important et qui a fait choc, c'est **quand** on a réuni les partenaires sociaux, et que j'ai pu dire, au nom de tous les sidérurgistes, qu'aucun sidérurgiste ne demandait de nouvelles mesures d'âge. *[*Ouv-Emploi*]*

Or, dans cet exemple, cette construction n'a pas d'équivalent dans la position canonique de sujet, comme le montre l'exemple 3a) ci-dessous,

3a) ? **Quand** on a réuni les partenaires sociaux a été important.

contrairement à d'autres pseudo-clivées :

3b) Ce qui a été important c'est le partenariat avec l'ANPE. ⇔ Le partenariat avec l'ANPE a été important.

Dans 4), la Quand-C *ne peut modifier le verbe* penser, *comme dans* Je pense à toi quand je vois ta soeur mais modifie plutôt toi.

4) Je pense à toi **quand** tu étais petit *[ex. Jeanjean, 1984 : 133]*

Dans 5), les critères, autres que la marque morphologique, sont difficiles à trouver pour démontrer une relation de dépendance.

5) Pécuchet venait d'en remettre la note à Bouvard **quand** tout à coup le tonnerre retentit et la pluie tomba *[Flaubert,* Bouvard et Pécuchet*]*

Enfin, 6) forme un énoncé autonome à lui seul.

6) **Quand** je pense que quelqu'un qui livrerait cet homme-ci gagnerait soixante mille francs et ferait sa fortune ! *[Hugo,* Quatre-vingt-treize*]*

Dans le cas des Quand-C*, les liens avec le contexte peuvent donc être variés, dans un spectre allant de la dépendance la plus étroite (ex. 2) à l'autonomie totale (ex. 6). Nous nous proposons d'étudier l'ensemble des environnements dans lesquels intervient quand, réservant une place particulière pour les cas généralement négligés dont l'étude suppose de nouveaux outils descriptifs autres que la « subordination canonique » et des investigations poussées dans des corpus tout venants et très volumineux tels que le web. Une partie importante est consacrée aux discussions portant sur la méthodologie et à la présentation du cadre d'analyse, ces deux dimensions ayant un intérêt tout à fait particulier dans l'étude du mot* quand *ainsi que dans ses rapports avec les autres « relatifs ». Cela se voit notamment dans la détermination de son statut catégoriel, quand il n'est pas interrogatif, fluctuant entre « pronom (ou adverbe) relatif » et « conjonction de subordination » (cette seconde analyse ayant les suffrages de la plupart des grammaires).*
Nous nous sommes passionné pour ce mot en apparence anodin, car, comme la plupart des sujets restreints, il permet d'éclairer de manière assez peu prévisible nombre de questions fondamentales : le marquage morphologique des relations syntaxiques, la pertinence de l'argumentation en syntaxe et de la terminologie, etc. De plus, il nous semblait que la connotation dont il est la victime, dès le latin face à cum *et encore de nos jours avec* lorsque*, demandait à être éclairée. Cet*

éclaircissement, notamment, nous a permis de mettre en évidence que les langues sont des systèmes complexes non homogènes dans lesquelles il est possible d'utiliser diverses stratégies pour exprimer (à peu près) la même chose. Les grammaires décrivent généralement l'un de ces systèmes ou les mélangent alors que leur fréquence respective est déterminée par les genres de textes. La question désormais est celle de savoir comment faire en sorte que les grammaires (formelles ou pas) puissent rendre compte de la variabilité des constructions syntaxiques.

Références bibliographiques

Borillo A.,, « Quelques remarques sur quand connecteur temporel », *Langue Française*, 77, 1988, pp. 71-91.

Chétrit J., *Syntaxe de la phrase complexe à subordonnée temporelle,* Klincksieck, Paris, 1976.

Debaisieux J.-M. , Le fonctionnement de parce que en français parlé contemporain : Description linguistique et implications didactiques, Thèse de Doctorat, Université de Nancy II, 1994.

Delaveau A., La conjonction si dans ses emplois interrogatifs et conditionnels, Thèse de Doctorat, Université de Paris VII, 1990.

Deulofeu J., « Questions de méthode dans la description morphosyntaxique de l'élément que en français contemporain », *Recherches sur le français parlé* n°15, 1999, pp. 163-198.

Hadermann P., *Étude morphosyntaxique du mot où, Champs linguistiques, Duculot,* Paris, 1993, Louvain-la-Neuve.

Jeanjean C., « Toi quand tu souris» : analyse sémantique et syntaxique d'une structure du français peu étudiée », *Recherches sur le Français Parlé,* 6, 1884, pp. 131-165.

Olsson L., Étude sur l'emploi des temps dans les propositions introduites par quand et lorsque et dans les propositions qui les complètent en français contemporain, Uppsala, Acta Universitatis Upsaliensis, Studia Romanica Upsaliensia, 1971.

Roubaud M.-N., *Les constructions pseudo-clivées en français contemporain, coll. Les français parlés* – Textes et Études, Paris, Honoré Champion, 2000.

Sandfeld K., *Syntaxe du français contemporain, Tome II : Les propositions subordonnées,* Librairie E. Droz, Copenhague,Paris,1936.

URL où la thèse pourra être téléchargée : www.up.univ-aix.fr/delic/perso/benzitoun/