Analyse automatique de la structure thématique du discours pour la navigation documentaire

Frédérik Bilhaut

Analyse automatique de la structure thématique du discours pour la navigation documentaire

Frédérik Bilhaut
GREYC – Université de Caen
Campus II, Bât. Sciences 3, Bvd. du Maréchal Juin
14000 Caen, France
fbilhaut@info.unicaen.fr

Face aux problèmes d'ingénierie documentaire posés par la recherche d'information (RI), le document a longtemps été considéré comme une entité indivisible. Ainsi, la plupart des systèmes couramment utilisés sur le Web ou en gestion documentaire ne connaissent pas d'intermédiaire entre l'unité *mot* (ou *terme*) sur laquelle se basent les systèmes d'indexation automatique, et l'unité *document* qui sera retournée dans sa globalité à l'utilisateur. La conception du mot comme seule unité infradocumentaire est probablement liée à la prédominance des méthodes numériques et de surface, elles-mêmes légitimées par des contraintes calculatoires et de généricité. Bien que présentant un intérêt pratique indéniable, celles-ci ont rapidement montré leurs limites, et se sont inévitablement heurtées à des problèmes d'ordre sémantique bien connus. Cependant, alors que cette question motive une part importante de l'activité de recherche du domaine, elle reste généralement abordée en RI d'un point de vue lexical, sans que le statut du mot ne soit fondamentalement remis en cause.

La question est pourtant d'importance : dans un contexte de recherche ou de navigation documentaire, comment représenter au mieux le contenu informationnel d'un document ? Automatique ou non, la tâche d'indexation s'est traditionnellement bornée à produire une liste de *descripteurs* : mots-clefs, termes ou entités abstraites appartenant à une base conceptuelle. Alors que ces index décrivent les documents dans leur intégralité, les méthodes de type *text-tiling* [Hearst 1994] visent à découper le texte en segments thématiquement homogènes. Mais là encore, les segments obtenus sont représentés par des listes de descripteurs, et le modèle discursif se résume à une séquence linéaire de segments.

Face au besoin de caractériser plus finement le contenu des documents numériques, on assiste aujourd'hui à l'émergence du *Web sémantique*, accompagné de formalismes tels que le *Resource Description Framework* (RDF) ou les *topic-maps*. En spécifiant un modèle de représentation des méta-données, et la manière dont celles-ci peuvent être attribuées à des *ressources*, le Web sémantique vise la normalisation des formats d'échange entre les agents du Web. Mais il ne prétend pas répondre aux autres aspects problématiques de l'indexation : quelles « ressources » est-il pertinent d'indexer dans un document, sous quelle forme, et surtout comment produire *automatiquement* une indexation « sémantique » ? Les principes que nous proposons visent l'analyse automatique de la structure discursive des documents afin d'obtenir une indexation fine, applicable à la recherche documentaire habituelle mais aussi et surtout à d'autres fins plus spécifiques aux documents longs, telles que la navigation intradocumentaire ou le résumé automatique.

Nous nous basons pour cela sur la notion de *thème*. Le concept demeurant à la fois flou dans son acception courante, et extrêmement discuté par les sciences du langage, nous ne prétendons aucunement répondre à la problématique linguistique qu'il soulève, mais plutôt en donner une définition utile à la résolution des problèmes informatiques qui nous occupent, en nous limitant au texte informatif au sens de [Combettes et Tomassone 1988], genre textuel à la fois massivement représenté au sein des applications documentaires et appréhendable par un traitement automatique. Dans le champ de la linguistique, la notion de thème s'entend principalement à l'intérieur de la phrase ou de la proposition, et vise généralement à attribuer le statut de thème à une de ses composantes selon un modèle phrastique donné. Plus marginalement, la notion de thème textuel a également été abordée, par exemple par [Jones 1977] qui parle de « généralisation minimale d'un texte : un énoncé suffisamment général pour représenter l'intégralité du texte, mais assez spécifique pour représenter sa singularité ». Dans l'optique de la navigation documentaire, c'est bien une approche textuelle que nous devons adopter, notre objectif étant de décrire le contenu des documents dans leur intégralité. Ainsi la notion de thème devra-t-elle subir ici une double accommodation : elle devra d'une part s'appliquer au texte et non plus à la phrase, et d'autre part se matérialiser par un objet calculatoirement manipulable.

Bien qu'aucune théorie linguistique ne puisse s'insérer immédiatement dans ce cadre particulier, il nous paraît toutefois important de relever quelques principes qui semblent applicables au-delà de leur contexte premier. Il faut bien-sûr évoquer la conception du thème en tant que « sujet » d'un segment textuel, « ce sur quoi il porte ». Cette conception, proche de l'acception courante du terme, se retrouve dans les théories linguistiques en termes d'*à propos* ou d'*aboutness* [Halliday 1985, Lambrecht 1994]. Elle présente un intérêt évident en RI, puisque la tâche d'indexation vise précisément à produire une représentation du contenu informationnel d'un texte, et s'applique assez naturellement à des segments plus importants que la phrase, par exemple sous la forme d'un thème textuel comme évoqué plus haut. Nous pouvons également considérer la distinction du statut de *nouveau* ou *donné* attribué à un référent : l'information *donnée* appartient aux connaissances que le locuteur attribue à l'interlocuteur au moment de la réception du message, alors que l'information *nouvelle* est celle que le locuteur souhaite porter à la connaissance de celui-ci. Qu'il s'agisse d'une connaissance externe ou introduite par le discours lui-même, le *donné* ou *connu* constitue un « socle » permettant au lecteur de s'approprier l'information nouvelle. Nous considérerons ici qu'il constitue également le pivot du processus d'accès à l'information, en tant que point d'accès pour l'utilisateur vers l'information qu'il recherche. Toutefois, alors que les modèles linguistiques décrivent l'évolution, au fil du texte, de la *familiarité supposée* [Prince 1983] de l'interlocuteur avec les référents du discours, les processus cognitifs mis en jeu (inférence, activation) paraissent difficilement exploitables automatiquement. Notre objectif d'opérationnalisation nous conduit donc à ne considérer (au moins dans un premier temps) que les connaissances externes au discours lui-même. Cette approximation nous semble acceptable dans le cadre du traitement des documents informatifs ou techniques, qui présupposent souvent un certain nombre de connaissances de la part de l'interlocuteur, autour desquelles s'articulent les informations nouvelles. Citons enfin deux notions linguistiques auxquelles nous ferons référence plus loin. Tout d'abord la notion de *thème multiple* chez [Halliday 1985], qui permet d'envisager le cas où le processus de thématisation implique plusieurs composantes. Si elle est définie à l'intérieur de la phrase, nous verrons qu'elle peut également être intéressante au niveau textuel. Nous envisagerons également la fonction de *scene-setting* que [Chafe 1976] attribue au *topique*, ce dernier étant considéré comme « cadre spatial, temporel ou individuel dans lequel se tient la prédication principale », par opposition au *sujet* qui constitue le « point d'ancrage » de la prédication.

Nous allons maintenant décrire un système s'appuyant sur une approche du thème que nous défendons comme applicable au texte et manipulable calculatoirement. Nous nous appuierons dans un premier temps sur une expérience menée au sein du projet GeoSem[1], visant la réalisation d'un système de RI spécifiquement adapté à l'information géographique [Bilhaut 2003]. Celui-ci permet de mener une recherche basée sur trois aspects caractéristiques de l'information géographique : temps, espace et faits socio-géographiques dits « phénomènes ». Afin de répondre efficacement à des requêtes mêlant ces trois composantes (comme « le taux de scolarisation dans l'Ouest depuis 1975 »), le système procède à l'analyse sémantique des expressions spatio-temporelles ainsi qu'à une analyse du texte permettant d'obtenir une indexation « composite » de certains passages. En se basant notamment sur la théorie de l'encadrement du discours [Charolles 1997], le système délimite des segments caractérisés par un phénomène accompagné d'une localisation spatiale et/ou temporelle qui lui sont sémantiquement liés. Ainsi, le système indexera ces segments par des triplets tels que (le retard scolaire ; en Normandie ; de 1950 à 1960). Conçu en collaboration avec des géographes, ce mode de représentation du contenu tient son efficacité de son adéquation avec le mode de structuration propre à l'information géographique. Le système que nous proposons ici vise à généraliser la notion de *thème composite*, qui nous semble applicable à bien d'autres domaines.

Pour cela, nous formulons l'hypothèse que, pour un domaine de spécialité donné, il existe une interdépendance entre la structuration des connaissances et l'organisation du discours qui s'y rapporte. Plus précisément, on considérera que parmi les concepts d'un domaine, certains d'entre eux jouent un rôle particulier dans la répartition des connaissances, répartition que l'on pourra reconnaître dans la structure du discours. Ces *concepts structurants* vont ainsi *répartir* l'information, *situer* les autres concepts dans le champ du domaine. C'est manifestement le cas des axes du temps et de l'espace en géographie, où la plupart des faits sont à envisager dans un cadre spatio-temporel précis (bien que parfois implicite). Selon le même principe, le domaine scolaire sera habituellement organisé selon des niveaux (primaire, secondaire, supérieur) ou par statut (privé ou public), alors que le domaine politique pourra s'articuler autour des tendances politiques (de gauche à droite) ou des différents types d'élections (des cantonales aux présidentielles). En observant des corpus de différents domaines (scolaire, politique, horticole ou encore logistique), on peut constater que ces structures « ontologiques » jouent un rôle non négligeable dans l'organisation du discours, et que le modèle d'indexation composite évoqué plus haut y est transposable. On

1 Projet soutenu par le programme CNRS « Société de l'information », regroupant GREYC, ERSS, ESO et EPFL.

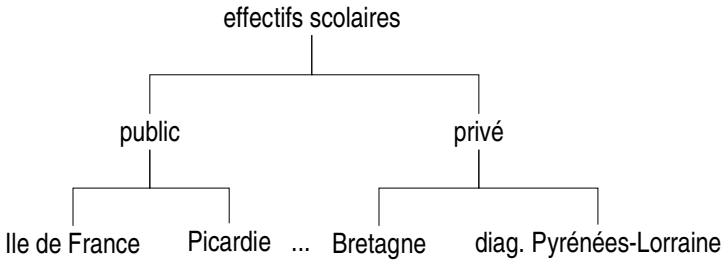
obtiendra par exemple des thèmes composites tels que (retard scolaire ; en 6ème ; dans le privé ; dans l'ouest) ou encore (J. Chirac ; élections présidentielles ; en 2002). On remarquera que, comme pour le domaine géographique, ces exemples se composent d'un élément central (retard scolaire, J. Chirac) accompagné d'un certain nombre d'éléments qui le situent dans son domaine (en Normandie, en 6ème, ...). Plus généralement, nous représenterons ces thèmes par des n-uplets (n | s_1 ; ... ; s_n) contenant un élément quelconque *n* du domaine (que nous nommerons *noyau*) suivi d'un certain nombre d'éléments structurants s_i (que nous nommerons *satellites*), par exemple (les instituteurs | dans le public ; en 1975). Remarquons que ce modèle n'impose aucune contrainte sur la nature de ces éléments, qui pourront se matérialiser indifféremment par des mots, des lemmes, des termes, des concepts, ou toute autre structure sémantique.

On remarquera que plusieurs modèles linguistiques décrivent des structures comparables au niveau phrastique. On pourra notamment mettre en parallèle la structure noyau/satellites avec l'opposition entre *sujet* et *thème multiple* chez [Halliday 1985], entre *topique* et *sujet* chez [Chafe, 1976], ou encore entre *topique* et *thème* chez [Dik, 1980]. Au niveau textuel, nous pensons pouvoir nous appuyer sur plusieurs mécanismes extra-phrastiques permettant de bâtir des structures similaires au fil du discours, au sein duquel une structure thématique en noyau/satellite se dessine phrase après phrase. En premier lieu, le contexte extra-linguistique peut intervenir, comme par exemple la date de rédaction ou un cadre spatial implicitement associé au document. Grâce à l'adoption de formats normalisés de spécification de méta-données (cf. Dublin Core, RDF, XMP), la prise en compte de ces aspects par des processus automatisés est envisageable. Un autre aspect concerne la structure logique des documents, et tout particulièrement les titres qui fournissent des informations thématiques importantes. Celles-ci sont souvent exploitables facilement grâce à l'adoption de formats de documents électroniques favorisant la représentation des aspects logiques et sémantiques (cf. TEI, DocBook). Enfin, certaines structures discursives, par leur fonction de répartition de l'information, peuvent spécifier des satellites thématiques valant pour des segments entiers. Nous nous intéressons particulièrement au modèle de l'encadrement du discours, qui décrit comment un segment textuel dit *cadre* (potentiellement composé de plusieurs phrases) peut être soumis à un critère d'interprétation donné par une expression détachée en initiale de phrase dite *introducteur*. Nous envisageons enfin la complémentarité de ces modèles textuels avec les apports des méthodes numériques.

Grâce à l'analyse de ces différentes structures dans les textes, nous délimitons des segments en les caractérisant par des n-uplets thématiques, l'ensemble formant une *structure thématique* arborescente. La figure ci-dessous reproduit un exemple de segment textuel accompagné de la structure thématique que l'on peut lui associer. On entrevoit immédiatement les applications d'un tel résultat : en plus d'affiner le processus classique de RI en retournant des passages précisément délimités, il permet d'envisager des systèmes de navigation intradocumentaire ou même de synthèse automatique.

L'explosion des effectifs scolaires

Dans l'enseignement public, elle s'accélère *en Île-de-France*, *en Picardie*, *dans le Centre*, ainsi qu'*en Provence* ; elle reste modérée *dans l'Ouest et le Nord*. [...] L'enseignement privé enregistre des baisses d'effectifs *en Bretagne*, où il est fortement implanté, ainsi que dans les académies de *la diagonale Pyrénées-Lorraine*, où son audience est par contre traditionnellement réduite [...]



Il est cependant difficile d'envisager une analyse automatique réalisant cette tâche sans aucune ressource. En effet, même si divers indices de surface (du type connecteurs logiques) peuvent souvent être exploités, on observe également que certains modes d'organisation discursive s'appuient uniquement sur les connaissances propres au domaine, sans faire appel à des marques explicites. Il en va ainsi dans l'extrait ci-dessus, où l'introducteur « Dans l'enseignement public » fait écho à « L'enseignement privé » qui apparaît simplement en tant que sujet grammatical. Sans avoir connaissance de l'opposition privé/public, il semble difficile de découvrir la structure que nous cherchons (même si la présence du premier introducteur est un indice important). Pour cette raison, le système que nous développons s'appuie sur modèle simple de représentation des connaissances, où les éléments structurants du domaine peuvent être regroupés en classes que nous appelons *axes sémantiques*. En s'appuyant sur

ces connaissances, la détection automatique de la structure thématique devient envisageable, même si l'analyse des marques explicites reste nécessaire. À l'évidence, le choix du recours à des connaissances extérieures est lourd de conséquences, puisqu'on on s'interdit le traitement de texte tout-venant. Nous nous limitons donc aux applications pour lesquelles le coût de la constitution des ressources est acceptable relativement aux fonctionnalités offertes par le système. Mais la constitution des ressources reste problématique : il est coûteux et complexe de modéliser des connaissances *ex-nihilo*, alors que le résultat obtenu est souvent peu générique car spécifiquement lié à une application ou une approche particulière d'un domaine donné.

Pour cette raison, nous développons un système permettant de constituer semi-automatiquement les ressources nécessaires, par apprentissage supervisé d'axes sémantiques à partir du corpus. La quasi-instantanéité de ce processus permet de considérer les ressources obtenues comme un simple *point de vue* sur le domaine, spécifique à l'utilisateur, et évolutif. Tout en étant proche des systèmes d'extraction terminologique, ce système ne vise en aucun cas l'exhaustivité, et s'applique uniquement à extraire les termes *structurants* d'un domaine représenté par un corpus suffisamment large et homogène. On s'appuie pour cela sur l'hypothèse que même si en discours ces termes ne sont pas *systématiquement* employés dans des situations caractéristiques, ils le sont *fréquemment*. Plus précisément, nous détectons les termes présentant régulièrement certaines *fonctions discursives*, afin d'évaluer leur caractère structurant : présence dans les titres, en tant qu'introducteur de cadre, ou encore en tant qu'extension prépositionnelle de certains syntagmes nominaux. Le système procède dans un second temps au regroupement des termes structurants (par contexte commun fréquent), et produit ainsi des axes approximatifs qui sont présentés à l'utilisateur. Après validation et/ou modification, le modèle obtenu est directement utilisable par le module d'analyse des documents décrit plus haut.

Le système que nous proposons s'articule donc sur trois points. En premier lieu, nous définissons une notion de *thème composite* permettant d'indexer finement le contenu informationnel d'un document par sa *structure thématique*. D'autre part, nous proposons un modèle de représentation des connaissances permettant de guider l'analyse discursive, sous la forme *d'axes structurants* du domaine. Enfin, nous proposons un système d'*apprentissage supervisé* qui, combiné à une interface utilisateur, permet la constitution et l'édition rapides de ces ressources. À partir de l'indexation obtenue, les applications visées sont la recherche d'information mais aussi la navigation intradocumentaire et la synthèse automatique.

Références

F. Bilhaut, T. Charnois, P. Enjalbert, Y. Mathet, 2003, « Passage extraction in geographical documents », *New Trends in Intelligent Information Processing and Web Mining*, Zakopane, Poland.

W. L. Chafe, 1976, *Giveness, constrastiveness, definiteness, subjects, topics and point of view*, « Subject and Topic », Academic Press.

M. Charolles, 1977, « L'encadrement du discours – Univers, champs, domaines et espaces », *Cahiers de recherche linguistique*, 6, pp. 1-60.

B. Combettes, R. Tomassone, 1988, *Le texte informatif, aspects linguistiques*, De Boek-Wesmael, col. Prisme.

B. Grosz, A. Joshi, S. Weinstein, 1995, « Centering : a framework for modelling the local coherence of discourse », *Computational Linguistics* 21(2).

M. A. K. Halliday, 1985, *An introduction to functional grammar »*, Edward Arnold.

M. Hearst, 1994, « Multi-paragraph segmentation of expository texts », 32th Annual Meeting of the Association for Computational Linguistics.

K. Lambrecht, 1994, *Information structure and sentence form. Topic, focus and the mental representation of discourse referents*, Cambridge University Press.

L. K. Jones, 1977, *Theme in English expository discourse,* Edward Sapir Monograph Series in Language, Culture, and Cognition, 2, Jupiter Press.

M.-P. Péry-Woodley, 2000, *Une pragmatique à fleur de texte : approche en corpus de l'organisation textuelle*, Mémoire d'H.D.R, Carnets de grammaire, Rapports internes de l'ERSS.

E. F. Prince, *Toward a taxonomy of given-new information*, « Radical Pragmatics », Academic Press.