Documentation et bibliothèques

L'analyse documentaire et les langages de spécialité : un filon à exploiter ?
Subject Analysis and Sublanguages: A Worthy Theme?
Análisis de documentos y lenguas de especialización: ¿un filón que debe explotarse?

Patrick Cossette

Volume 38, numéro 2, avril–juin 1992

Analyse et gestion de l'information textuelle

URI : https://id.erudit.org/iderudit/1028614ar
DOI : https://doi.org/10.7202/1028614ar

Aller au sommaire du numéro

Éditeur(s)

Association pour l'avancement des sciences et des techniques de la documentation (ASTED)

ISSN

0315-2340 (imprimé)
2291-8949 (numérique)

Découvrir la revue

Citer cet article

Cossette, P. (1992). L'analyse documentaire et les langages de spécialité : un filon à exploiter ? *Documentation et bibliothèques*, *38*(2), 96–102. https://doi.org/10.7202/1028614ar

Résumé de l'article

Depuis plus de vingt ans, les recherches sur les langages dits de spécialité ont ouvert de nouvelles pistes dans l'exploration du langage naturel. De contenu délibérément homogène pour mettre en valeur le message qu'ils véhiculent, ils sont identifiables tant par leur structure que par leur syntaxe et leur lexique. À partir d'une expérience portant sur des rapports d'analyse environnementale et réalisée avec le logiciel SATO, différentes classes de termes tirées de ces textes sont récupérées. En sondant ainsi les mécanismes de composition des documents grâce au principe des langages de spécialité, il est possible d'entrevoir des améliorations aux techniques d'analyse documentaire.

Analyse

L'analyse documentaire et les langages de spécialité : un filon à exploiter ?

Patrick Cossette
Assemblée nationale
Québec

Depuis plus de vingt ans, les recherches sur les langages dits de spécialité ont ouvert de nouvelles pistes dans l'exploration du langage naturel. De contenu délibérément homogène pour mettre en valeur le message qu'ils véhiculent, ils sont identifiables tant par leur structure que par leur syntaxe et leur lexique. À partir d'une expérience portant sur des rapports d'analyse environnementale et réalisée avec le logiciel SATO, différentes classes de termes tirées de ces textes sont récupérées. En sondant ainsi les mécanismes de composition des documents grâce au principe des langages de spécialité, il est possible d'entrevoir des améliorations aux techniques d'analyse documentaire.

Subject Analysis and Sublanguages: A Worthy Theme?

For more than twenty years, the research into sublanguages has opened new opportunities in the investigation of natural language. Deliberately homogenous in order to highlight the content, they are recognizable by their structures, syntaxes and vocabularies. Based on an experience using environmental analysis reports with the software known as SATO, different classes of terms pulled from the texts are retreived. In examining the mechanisms of document composition with sublanguages, it is possible to propose improvements to the techniques of subject analysis.

Análisis de documentos y lenguas de especialización: ¿un filón que debe explotarse?

Desde hace más de veinte años, los investigadores sobre los lenguajes especializados han abierto nuevas huellas en la exploración del lenguaje natural. Dado su contenido deliberadamente homogéneo, para poner de relieve el mensaje que transmiten, esos lenguajes especializados son identificables tanto por su estructura como por la sintaxis y el léxico. A partir de una experiencia realizada sobre los informes de análisis ambiental y obtenida con el programa SATO, se recuperaron diferentes clases de términos extraídos de esos textos. Al investigar de este modo los mecanismos de composición de documentos, gracias al principio de los lenguajes de especialización, es posible entrever las mejoras producidas en el análisis de documentos.

Les linguistes et autres « scaphandriers » de la langue se consacrent avec acharnement à observer, décomposer et interpréter les mécanismes linguistiques du langage naturel. S'ils y mettent autant d'efforts, c'est que le langage qu'ils traitent est peut-être naturel, mais surtout qu'il est fort complexe. L'une des avenues les plus constamment fouillées au cours des vingt dernières années tire son origine d'une supposition très simple : en travaillant sur des corpus de textes issus de milieux professionnels spécialisés (médecine, génie, etc.) ou ancrés dans une écriture stéréotypée (livres de recettes, manuels d'instruction, etc.), les chercheurs espèrent réduire les dommages sémantiques causés par l'énorme masse d'ambiguïtés véhiculées par le langage. Ces textes répondent à des messages dont la portée se veut immédiate ; les auteurs peuvent se permettre de dépouiller et d'uniformiser le vocabulaire, la syntaxe et la structure, sachant que les lecteurs auront les connaissances nécessaires pour effectuer les rétablissements qui s'imposent. Ils offrent ainsi une meilleure prise aux travaux d'exploration. Ces sous-systèmes linguistiques ont été baptisés « langages de spécialité », ou « sublanguages » dans la langue de Shakespeare.

Aussi spécialisées en linguistique ou en intelligence artificielle soient-elles, les recherches sur ces langages peuvent apporter une intéressante contribution à l'avancement des techniques en sciences de l'information. Elles s'adressent tout particulièrement aux spécialistes qui sont aux prises avec les barrières du traitement et du repérage en texte intégral.

Au cours des études de maîtrise à l'École de bibliothéconomie et des sciences de l'information[1], nous nous sommes attardé à ces rapports entre l'analyse documentaire et les langages de spécialité.

Les langages de spécialité

Sans le savoir, nous fréquentons tous des langages de spécialité dans nos activités domestiques ou professionnelles. Nous les retrouvons sous la forme d'un livre de recettes, d'un manuel d'entretien aéronautique, d'un rapport d'activités boursières dans un quotidien, ou encore d'un rapport médical. Qu'ont en commun tous ces textes d'origines aussi diverses ? Ils sont tous assimilables à un mode de communication exclusif à un groupe d'usagers évoluant dans un domaine

1. Cette recherche a été réalisée sous l'habile et généreuse supervision de Suzanne Bertrand-Gastaldy.

de référence restreint: « When natural language is used in a sufficiently restricted setting, we may be justified in calling the resultant forms a sublanguage »[2]. L'émetteur y insère son discours dans un cadre qui est volontairement le plus transparent possible pour les usagers-récepteurs. Dans un tel contexte, la langue n'est pas un multiplicateur de sens, mais au contraire elle s'éclipse au profit du contenu le plus univoque possible, quitte à s'écarter des règles de grammaire traditionnelles. Si les raccourcis linguistiques sont souvent la manifestation la plus spectaculaire des langages de spécialité, c'est qu'ils n'ont de raison d'être que pour faire dominer le fond sur la forme.

Cette caractéristique se traduit selon les textes par des régularités de différents ordres:

- ordre lexical: des séries de noms, de verbes, d'adjectifs et d'adverbes dominent le lexique;
- ordre syntaxique: certaines tournures grammaticales sont fréquemment réemployées (impératif, mode passif, longues énumérations de noms communs, surabondance ou pénurie de certaines catégories grammaticales, prédominance de certains temps de verbes, style télégraphique, etc.);
- ordre sémantique: la polysémie étant la grande ennemie que combattent tous les langages de spécialité, la portée sémantique des termes les plus représentatifs est limitée le plus possible à un seul sens;
- ordre structurel: jusqu'à maintenant peu abordée par les chercheurs, la disposition du tissu textuel d'un langage de spécialité doit à lui seul en permettre la distinction. L'étude des rapports qu'entretiennent les phrases entre elles peut servir de révélateur pour saisir la complexité d'un langage de spécialité (anaphores, macro-structures, etc.).

Figure 1
Exemples de langages de spécialité

Usage de l'impératif et pronominalisation:

« Éplucher les oignons. Les couper en petites rondelles. Les incorporer au mélange. Saupoudrer avec... »

Synonymie des noms et des verbes d'action:

« La Bourse a vu son cours chuter de trois points ce matin. Cette baisse est attribuable à ... Certaines compagnies ont néanmoins affiché un gain appréciable en progressant... »

C'est sur le plan de la catégorisation lexicale et des régularités grammaticales que les chercheurs ont le plus de succès à décrire un langage de spécialité:

> *A sublanguage is characterized by distinctive specializations of syntax and the co-occurrence of domain-specific word subclasses in particular syntactic combinations*[3].

Lorsqu'il y a fréquence élevée de termes sémantiquement voisins et de mêmes cooccurrences[4] répétées à l'intérieur d'une disposition syntaxique récurrente, les spécialistes admettent alors l'existence d'un langage de spécialité. Et ce, d'autant plus que les règles grammaticales employées détournent la grammaire traditionnelle vers des usages inhabituels. En termes imagés, les langages de spécialité peuvent être considérés comme des « micro-climats » textuels dont la singularité consiste à être à la fois uniques et hautement prévisibles. L'intérêt de ce terrain d'étude pour la recherche est alors manifeste. En limitant leurs recherches à des textes bien circonscrits, les chercheurs s'exposent moins aux pièges de la langue et à ses vicissitudes.

La recherche sur les langages de spécialité

C'est à la fin des années 1960 que germèrent les premières intuitions sur les curieuses métamorphoses de la langue dans certains textes. Au premier chef, elles sont dues au linguiste américain Zellig Harris[5] et à ses théories distributionnelles et transformationnelles. Celui-ci prétendit que la dimension sémantique de certains textes est invariablement inscrite dans les modèles représentatifs de co-occurrences de ces mêmes textes. Grand théoricien, Harris est toujours aujourd'hui une référence obligée, mais son attrayante hypothèse restait tout de même à vérifier.

Au cours des années 1970, la communauté des chercheurs intéressés s'est rapidement agrandie. Cette période est dominée par l'équipe du

Figure 2
Échantillon des recherches du *Linguistic String Project*

Patient	Verbe	Partie(s) du corps	Symptôme(s)	Examen	Diagnostic
Il	ressent	estomac	douleurs		
	révèlent	abdomen		radio-graphies	ulcère d'estomac

2. Richard Kittredge, « Sublanguages », *American Journal of Computational Linguistics*, vol. 8, no. 2 (April-June 1982), 79.
3. Naomi Sager, « Sublanguage: Linguistic Phenomenon, Computational Tool », in Ralph Grishman and Richard Kittredge, *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, Hillsdale, Lawrence Erlbaum Associates, 1986, p. 1.
4. Une cooccurrence est le fait pour deux mots de se retrouver ensemble dans un même contexte déterminé par l'analyste (la phrase, le paragraphe, par exemple).
5. Zellig Harris, *Structures mathématiques du langage*, Paris, Dunod, 1971, 248 p.

   Zellig Harris, « Discourse and Sublanguage », in Richard Kittredge and John Lehrberger, *Sublanguage: Studies Language in Restricted Semantic Domains*, Berlin, de Gruyter, 1982, p. 231-236.

Linguistic String Project de l'Université de New York sous la direction de Naomi Sager. Le calibre des travaux menés par ce groupe est tel qu'ils sont demeurés des modèles du genre[6]. Les rapports médicaux et pharmacologiques sur lesquels ont travaillé ces chercheurs ont littéralement été passés à la moulinette. L'objectif ultime de la démarche de Sager était la constitution de ce qui a été nommé le modèle informatif d'un langage de spécialité. Les classes de mots et les relations que celles-ci entretiennent entre elles y sont carrément substituées aux mots du texte. À cette fin, les chercheurs ont d'abord extrait de phrases similaires les classes sémantiques de mots relatives à certaines catégories (parties du corps humain, symptômes, types d'examens, etc.) selon que ces termes étaient des verbes, des noms ou des adjectifs. Ensuite, ils ont distingué sur le plan syntaxique les classes de sujets des classes d'objets. Il en est ressorti des modèles de phrases reprenant de façon systématique les mêmes classes de mots. Enfin, ils ont fractionné les structures syntaxiques complexes en propositions simples. En quelques tableaux, l'équipe du *Linguistic String Project* a pu ainsi saisir le modèle syntaxique et sémantique des rapports médicaux.

Par exemple, à partir des deux phrases suivantes, il est possible de produire un modèle informatif partiel :

1. Il ressent des douleurs à l'estomac.

2. Les radiographies de l'abdomen révèlent un ulcère d'estomac. (figure 2)

En refaisant cet exercice sur une grande échelle, le *Linguistic String Project* a recueilli une somme d'information considérable. Conséquence pratique, leur analyse et leur catégorisation ont permis la conception de systèmes-maison automatisés de condensation automatique et de vérification de la séquence des actes médicaux. La voie était nettement ouverte pour les sciences de l'information.

D'autres branches de la linguistique computationnelle ont aussi capitalisé sur les langages de spécialité.

Les systèmes de génération automatique de texte linguistiquement bien formés ne sont possibles que s'ils se situent dans un sous-langage technique bien délimité. Il a été démontré que les langues en général sont trop vastes pour être traitées correctement (au niveau syntaxique et sémantique) par les systèmes actuels[7].

Chantal Contant a recouru aux nouvelles de la bourse pour concevoir un système de génération automatique de textes en français. Dans le même ordre d'idée, des chercheurs de l'Université de Montréal ont misé sur la réciprocité linguistique des bulletins de météorologie en anglais et en français pour concevoir le système de traduction automatique TAUM-METEO. Par la suite, ils se sont tournés vers les manuels d'entretien aéronautique. La preuve est maintenant faite que les textes en français et en anglais d'un même langage de spécialité ont beaucoup plus en commun que ceux provenant du langage courant.

En travaillant avec un moyen de communication balisé et archiprévisible dans son écriture, ces chercheurs ont mis à profit les mérites des langages de spécialité. Avec les travaux de Richard Kittredge, de l'Université de Montréal, et de John Lehrberger[8], toutes ces recherches contribuent à faire de Montréal un pôle important dans le domaine.

La multiplicité des travaux qui ont vu le jour depuis une dizaine d'années a fait passer le langage de spécialité du stade de concept prometteur à celui de terrain d'exploration privilégié du langage naturel. Les progrès de l'informatique repoussent toujours plus loin les limites de l'analyse de textes par ordinateur. Par le fait même, la liste des types de documents analysés et promus au rang de langages de spécialité ne cesse de s'allonger, jusqu'à en arriver peut-être un jour à une typologie. Les travaux se multipliant, ceux-ci débordent des cadres disciplinaires et engagent de plus en plus tous les professionnels intéressés aux sciences du langage. Et surtout, la portée des observations dépasse de plus en plus l'analyse lexicale et syntaxique. Elle aborde maintenant la grammaire des textes et l'analyse du discours. De nos jours, les chercheurs visent de plus en plus à interroger les propriétés exclusives des textes, comme cela s'est fait, depuis des décennies, avec la phrase. En effet, à l'instar des phrases qui le composent, le texte obéit à des règles lexicales, syntaxiques et sémantiques qui lui sont propres. La fonction de la grammaire des textes consiste à en faire l'étude.

À mesure que la qualité et la complexité des moyens mis en oeuvre s'améliorent, il est permis de croire que les applications en sciences de l'information se feront plus concrètes et se répandront à l'extérieur des laboratoires universitaires. Aux premières loges, les industries de la langue s'abreuvent de plus en plus à ces améliorations pour affronter les nouveaux défis qui les attendent. La fièvre que suscite l'intégration européenne a d'ailleurs incité nombre de chercheurs européens à s'intéresser aux travaux de leurs confrères américains. Cette conjoncture assure une certaine pérennité aux efforts initiaux de Zellig Harris et Naomi Sager.

---

6. Daniel Gordon and Naomi Sager, « A Method of Measuring Information in Language Applied to Medical Texts », *Information Processing & Management,* vol. 21, no. 4 (1985), 269-289.

   Ralph Grishman et al., « Grammatically-based Automatic Word Class Formation », *Information Processing & Management,* vol. 11 (1975), 39-57.

   Lynette Hirschman and Naomi Sager, « Automatic Formatting of a Medical Sublanguage », in Richard Kittredge and John Lehrberger, *Sublanguage : Studies Language in Restricted Semantic Domains,* Berlin, de Gruyter, 1982, p. 27-80.

   Naomi Sager, « Computer Analysis of Sublanguage Information Structures », in Edward H. Bendix, *The Uses of Linguistics,* New York, New York Academy of Sciences, 1990, p. 161-179.

   Naomi Sager, « Syntactic Formatting of Science Information », in Richard Kittredge and John Lehrberger, *Sublanguage : Studies Language in Restricted Semantic Domains,* Berlin, de Gruyter, 1982, p. 9-26.

7. Chantal Contant, *Génération automatique de texte : application au sous-langage boursier français,* Université de Montréal, Département de linguistique, mémoire de maîtrise, décembre 1985, p. 11.

8. Ralph Grishman and Richard Kittredge, *Analyzing Language in Restricted Domains : Sublanguage Description and Processing,* Hillsdale, Lawrence Erlbaum Associates, 1986, 246 p.

   Richard Kittredge and John Lehrberger, *Sublanguage : Studies Language in Restricted Semantic Domains,* Berlin, de Gruyter, 1982, 240 p.

   Richard Kittredge, « Sublanguages », *American Journal of Computational Linguistics,* vol. 8, no. 2 (April-June 1982), 79-84.

Figure 3
Extrait du fichier-maître pré-traité

* doc = hydro1
* page = hydro1/1
* subd = tit * txt = st Poste \Mauricie à 315 - 230 Kv,

Raccordement au Poste des lignes 315 et 230 Kv

* subd = descri * txt = pt DESCRIPTION DU PROJET

* txt = pt Les aménagements prévus par \Canards \Illimités \Canada se situent sur l'extrémité est de l'île \Dupas, elle-même localisée dans les îles de \Sorel à l'ouest du Lac \St-Pierre.

Ces aménagements comprennent 3 ouvrages principaux:

1- À l'extrémité ouest de la \Commune, l'on a aménagé en alternant des planches agricoles arrondies et des rigoles se déversant dans un fossé collecteur muni d'un bassin piscicole.

[...]

Description du projet

Les recherches sur les langages de spécialité intéressent particulièrement trois champs d'activités. Les linguistes les étudient pour leurs propriétés uniques qui en font des sous-systèmes linguistiques largement autonomes et très structurés par rapport au langage naturel commun. Les spécialistes en intelligence artificielle y trouvent une matière plus cohérente et plus aisément assimilable pour un système informatique, notamment dans la conception des systèmes-experts. Enfin, pour les spécialistes en sciences de l'information, les langages de spécialité permettent d'entrevoir une amélioration significative du formatage, de l'analyse et du repérage de l'information textuelle. Ainsi que le déclarait Yves Hudon,

> *Le professionnel du texte (documentaliste, linguiste, etc.) s'attaque maintenant à la comparaison de textes, à leur enrichissement (annotations), à leur analyse, voire à la recherche sur le contenu des textes de l'organisation*[9].

Cette troisième préoccupation a guidé notre recherche.

Considérant que la description en bonne et due forme d'un langage de spécialité était une tâche trop lourde pour les moyens et le temps dont nous disposions, notre projet s'est arrêté à la première étape de toute étude sur les langages de spécialité, celle de la constitution de classes de mots représentatifs du type de textes sélectionné.

Nous avons défini notre objectif de recherche en quatre phases successives: observer les noms communs, les noms propres, les verbes et les adjectifs distinctifs d'un type de textes; établir des classes de termes en attribuant diverses propriétés relatives à leur degré de pertinence, leur portée sémantique et le temps des verbes; faire des recoupements entre diverses propriétés; examiner les cooccurrences des termes les plus représentatifs.

Pour débuter, deux éléments étaient indispensables. D'une part, il nous fallait un corpus de textes présentant des affinités avec nos propositions théoriques. Il nous est apparu que les rapports d'analyse environnementale du ministère de l'Environnement du Québec correspondaient bien au portrait-robot du langage de spécialité. Ces rapports s'inscrivent dans la procédure d'examen et d'évaluation des impacts sur l'environnement et contiennent les jugements des chargés de projets sur les estimations des promoteurs. Notre choix s'est arrêté sur sept textes dont les auteurs, les longueurs et les dates de publication différaient. Ces textes couvraient une variété de projets allant des aménagements fauniques aux constructions de routes, et du transport de l'énergie électrique à l'épuration des eaux usées. Ils avaient tous été rédigés dans les années 1980 et leur longueur variait entre 10 et 20 pages. Il faut garder à l'esprit que c'est la parenté des textes à la fois dans le type de document et dans leur cadre de production qui favorise l'éclosion d'un langage de spécialité.

D'autre part, il fallait compter sur un logiciel suffisamment souple et performant pour manipuler les textes à la lumière des nombreuses stratégies de fouille désirées. Le logiciel SATO (Système d'analyse de textes par ordinateur) produit par le Centre d'analyse des textes par ordinateur de l'UQAM (ATO) a constitué un précieux auxiliaire, en dépit du fait que ses applications en sciences de l'information commencent seulement à être exploitées.

En nous interrogeant sur les intentions qui motivent la production de ces rapports, nous sont apparus les deux centres d'intérêt primordiaux de ces textes: les impacts écologiques attribués aux projets, et leurs contreparties destinées à les contenir, les mesures d'atténuation (également appelées mesures de mitigation). Nous avons formulé l'hypothèse que les descriptions d'impacts et de mesures d'atténuation étaient toutes désignées pour être des « incubateurs » de langages de spécialité. Le projet pouvait alors se résumer à ceci: comment sont décrits les impacts et les mesures d'atténuation dans les rapports d'analyse environnementale? Ce type d'interrogation se pose forcément durant les opérations d'indexation et de repérage.

Avant d'entamer le coeur du projet, toute une série de pré-traitements fut nécessaire pour préparer les textes aux opérations d'analyse. Se sont succédé les retouches orthographiques et typographiques exigées par le mode de fonctionnement du logiciel; l'attribution de propriétés textuelles pour distinguer les textes et leurs composantes: 7 documents (doc) pouvant contenir jusqu'à 12 subdivisions

9. Yves Hudon, « La recherche textuelle: un choix technique et administratif », *Documentation et bibliothèques*, vol. 37, no 2 (avril-juin 1991), 73.

(subd)[10] partagées en sous-titres (st) et en plein-texte (pt); le blocage des noms composés dans le lexique; et enfin la catégorisation grammaticale de tous les termes. Les textes ont été alors mis bout à bout pour ne former qu'un seul corpus. (figure 3)

Bien que longue et ardue, cette étape préliminaire de la recherche ne devait pas être escamotée, sous peine de conduire à des résultats invalides.

Les éléments étaient alors en place pour interroger le texte avec SATO. Le plan d'action était le suivant: 1- constitution de classes de termes représentatifs avec attribution de propriétés; 2- exploitation des termes sélectionnés selon leurs localisations; 3- mise en lumière de leurs co-occurrences représentatives.

1) En premier lieu, notre classement des termes s'appuyait sur trois variables:

A - Leur proximité dans une même phrase des termes *impact(s), mesure(s) de mitigation et mesure(s) d'atténuation* à l'extérieur des subdivisions consacrées respectivement aux impacts et aux mesures d'atténuation (introduction, historique, description, recommandations, etc.); B - Leur fréquence d'occurrences à l'intérieur de ces subdivisions; C - Le nombre de documents dans lesquels ils apparaissent.

Une série de commandes traitées par SATO a donc produit cinq indices pour chacun des noms, verbes et adjectifs du texte. Au cours d'une première opération effectuée par sélection manuelle, tous les mots ou familles de mots dont la moitié ou plus des fréquences totales se produisait à proximité d'un des mots témoins (A), ou dans une des deux subdivisions sous observation (B), ont pu être retracés. Déjà, un texte de départ de 80 pages était ramené à une liste de près de 350 termes soupçonnés de représenter avec le plus d'acuité et de régularité les impacts et les mesures d'atténuation. Cette première liste a été raffinée en deuxième étape pour démarquer les mots ou familles de mots apparaissant majoritairement dans la subdivision d'un des concepts (valeur sémantique supposément supérieure) de ceux qui surviennent majoritairement à côté d'un des mots témoins (valeur sémantique réduite).

Figure 4
Exemple des données nécessaires au classement

terme	freq	docu	imp	subi	mdm	mda	subm
ABORD*	2	hydro1 epu1	0	1	0	0	0
ABORDS*	1	faun2	0	1	0	0	0
ACCEPTABILITÉ*	4	hydro1 hydro2 route1	3	0	0	0	0
ACCEPTABILITÉ ENVIRONNEMENTALE	10	faun1 faun2 route2 epu1	3	0	0	0	0
ACCEPTABLES	13	hydro1 hydro2 faun1 faun2 route1 route2 epu1	3	1	1	1	0
ACCROISSEMENT*	1	epu1	0	1	0	0	0
ACCRUES*	1	route2	0	1	0	0	0
ACHATS*	1	hydro2	0	1	0	0	0
ACHETANT*	1	route1	0	1	0	0	0

Légende freq = nom du (des) texte(s) dans le(s)quel(s) apparaît le terme
docu = nom du (des) texte(s) dans le(s)quel(s) apparaît le terme
imp = proximité à cinq mots ou moins de *impact(s)*
subi = présence dans la subdivision consacrée aux impacts
mdm = proximité à cinq mots ou moins de *mesure(s) de mitigation*
mda = proximité à cinq mots ou moins de *mesure(s) d'atténuation*
subm = présence dans la subdivision consacrée aux mesures d'atténuation (ou mesures de mitigation)
* = terme retenu

Parallèlement, les termes n'apparaissant que dans un des textes étaient écartés, faute d'être représentatifs (C). La liste se voyait alors amputée de plus de 200 termes, pour un total de 178 lexèmes. (figure 4)

Pour compléter les résultats du tableau, ce classement purement statistique a été doublé d'un classement sémantique composé de six catégories (lieu, période, action, qualificatif, cause, objet) de façon à isoler les enjeux des analyses environnementales. L'analyse des termes a permis d'observer que lorsque le texte traite des impacts, une action est posée (*perturbation, dérangement, modification,* etc.), elle engendre (*cause, entraîne,* etc.) des conséquences (*accroissement, émettre, subir,* etc.) soigneusement évaluées (*majeur, moyen,* etc.) sur un objet (*habitat, résidence,* etc.) et dans un espace-temps précis (*automne, printemps, abord,* etc.). Lorsqu'il se rapporte aux mesures d'atténuation qui lui sont associées, le texte mentionne des ressources pour *limiter* ou *atténuer* les dommages en démarrant des *plantations* ou en disposant des *écrans*.

10. Il fallait pouvoir distinguer l'introduction de la conclusion, la partie traitant des impacts de celle traitant des mesures d'atténuation, etc.

Figure 5
Extrait de la liste des termes apparaissant à titre de cooccurrents dans plus de 75 % des cas avec le terme *impact(s)*

Terme	Fréquence des co-occurrences
Acceptabilité	6/8
Ajoute(ent,er)	3/4
Anticipé(ée,s)	3/3
Appréhendés	2/2
Attendu(s)	2/2
Atténuant(es) Atténue(er,eront,és)	9/9
Cause(ée,er)	3/5
Diminuer(eraient,nués)	4/5
Dus(û)	4/5
Négatif	8/9
Suscite(és)	9/9

Par ailleurs, nous avons essayé un classement selon le temps des verbes. Cette tentative s'est révélée infructueuse et elle fut abandonnée.

Toutes ces manipulations ont donné lieu à l'attribution de propriétés lexicales aux termes qui composent les différentes classes, de manière à les catégoriser. La composition de telles listes est l'une des plus intéressantes propriétés de SATO.

2) À partir de ces listes de termes et des propriétés textuelles divisant les parties des textes, il a été possible de faire des recoupements selon les types d'aménagement (hydroélectriques, fauniques, etc.). En reprenant les termes n'apparaissant que dans un type d'aménagement, nous pouvions dégager les interventions et les éléments naturels particuliers aux différents travaux: *ferme* ou *scarification* pour la construction de routes, *qualité de l'eau* ou *faune ichtyenne* pour l'érection d'un émissaire d'épuration, etc.

D'autres recoupements ont été effectués selon la répartition entre le plein-texte et les sous-titres afin d'éliminer les termes à fréquence élevée provenant des sous-titres, tels que *principaux, conclusion,* etc.

Enfin, nous avons pu repérer les termes communs aux deux subdivisions étudiées. La liste révèle une majorité de termes contextuels désignant l'environnement physique *(berges, boisés, fossés, rigoles,* etc.).

Ces nouvelles listes jetaient un autre éclairage sur la composition du vocabulaire.

3) L'analyse des cooccurrences représentatives est essentielle à l'étude des langages de spécialité. C'est ainsi que nous avons pu analyser les termes qui gravitent le plus souvent autour des termes *impact(s), mesure(s) d'atténuation, mesure(s) de mitigation* et de certains verbes jugés particulièrement représentatifs. Les résultats ont été les plus probants avec le mot impact(s). (figure 5)

De la sorte, nous ouvrons la voie à une représentation du modèle informatif de ce type de texte. Pour le cerner, il resterait à procéder à une analyse syntaxique qui conduirait à définir les structures de phrases types mettant en relation les classes de termes signalées.

Conclusion

Nos premières observations nous avaient amené à croire que les rapports d'analyse environnementale se révéleraient facilement en langage de spécialité. Les données recueillies ne suffisent pas à corroborer cette hypothèse. La taille modeste de l'échantillonnage et le caractère volontairement incomplet de l'étude ne permettent pas d'en juger. Un fait est à signaler: une très forte disparité dans la récolte de termes existe entre les classes abondantes associées aux impacts et celles, maigrelettes, associées aux mesures d'atténuation. La subdivision et les termes associés aux impacts se sont avérés nettement plus intéressants. Ceci dit, nos espérances nous portaient à exagérer la richesse de ce filon. Ainsi, à la lumière de nos données partielles, le vocabulaire repéré reste assez volatile et les temps de verbes n'ont démontré aucune régularité significative. Peut-être s'agit-il d'un langage de spécialité en devenir? Somme toute, des textes tirés d'un domaine technique bien délimité ne sont pas nécessairement le gage de la présence d'un langage de spécialité. Cela illustre la terrible complexité de l'objet textuel et la difficulté d'en représenter le contenu. De multiples facteurs contextuels (caractère de convention attribué au type de texte, auteurs, dates, directives de rédaction, etc.) contribuent à relativiser toute tentative de généralisation.

Toutefois, parmi les termes le plus clairement assimilés aux concepts étudiés, il subsiste une concentration de noms d'actions, de verbes et d'adjectifs dont le spécialiste de l'information peut tirer profit pour ses activités d'indexation ou de repérage, particulièrement à travers l'usage du terme *impact(s).*

De toute manière, notre étude visait d'abord à vérifier modestement l'intérêt de certaines perspectives. À travers l'approfondissement des langages de spécialité, le professionnel de la documentation se voit offrir la possibilité de franchir un degré supplémentaire de connaissance du texte et du bagage conceptuel qui lui est rattaché. Dans ses tâches d'analyse ou de référence, il lui est permis, en quelque sorte, de faire corps avec les paramètres de représentation textuelle de la spécialité qu'il sert. Il acquiert une

compréhension approfondie du contenu des textes qu'il traite sans être contraint à s'ériger en linguiste chevronné ou en spécialiste du domaine. La relation avec le texte devient plus intime et donne accès à une indexation et des patrons de fouille plus performants. Il y a là un outil puissant pour prendre le pouls épistémologique d'une science, d'une discipline, d'une technique, bref de tout univers conceptuel délimité, maîtrisé ... et écrit. N'est-ce pas là le lot d'un bon nombre de bibliothécaires et d'archivistes travaillant en milieu spécialisé ? Il y a fort à parier que, dans les années à venir, archivistes, gestionnaires de documents ou de bases de données, indexeurs et rédacteurs de résumés pourront y trouver leur compte. Aspect non négligeable, les applications du logiciel SATO autorisent une telle souplesse dans l'analyse textuelle que l'on peut envisager des opérations et des traitements additionnels. À peu de choses près, les combinaisons de stratégies de fouille peuvent inspecter un texte sous toutes ses coutures et en livrer les secrets les mieux tapis.

À celui ou celle qui souhaiterait révéler le modèle informatif d'un quelconque type de texte, soit les relations entre les classes lexicales et les structures de phrases types, nous ne pouvons cacher que cette forme d'inventaire exige des ressources considérables. Cependant, il faut considérer les bénéfices, confirmés ou potentiels, qui peuvent en être tirés. Une fois exécutée la description anatomique détaillée d'un type de texte, le praticien n'a plus qu'à s'en remettre à des mises à jour moins laborieuses pour suivre à la trace les modifications qui, par la suite, pourraient surgir dans les documents (nouveaux concepts, néologismes, structure remodelée, etc.). Pour cela, il faut dès maintenant que les spécialistes de l'information se mesurent aux défis des bases de données en texte intégral, des bases de connaissances et des systèmes experts. Les caractéristiques de ces nouveaux outils appellent des modes de manipulation différents de ceux employés pour les bases de données bibliographiques. Le milieu de la documentation ne peut pas se permettre de rester spectateur de ces innovations. Le formatage, l'analyse et le repérage du texte intégral doivent en bonne partie lui revenir. À cheval sur les recherches interdisciplinaires et les besoins des usagers, la place de nos professions documentaires est vitale et doit nous inciter à maintenir un dialogue constant avec les autres disciplines scientifiques. En puisant dans les concepts et les instruments mis au point à d'autres fins (traduction, analyse de la langue, etc.), nous nous rattachons à des concepts et des méthodes qui peuvent parfois paraître éloignés des nôtres. Pourtant, ces travaux recèlent des ouvertures susceptibles de faire évoluer nos outils de travail au même rythme que les besoins des usagers et les possibilités de développement. Ils rejoignent nos préoccupations professionnelles dans la mesure où nous faisons l'effort de les comprendre et de nous les approprier. L'excuse à propos du manque d'outils théoriques et informatiques tient de moins en moins ...

RÉFÉRENCES BIBLIOGRAPHIQUES SUPPLÉMENTAIRES

BONZI, Susan. « Syntactic Patterns in Scientific Sublanguages: a Study of Four Disciplines », *Journal of the American Society for Information Science,* vol. 41, no. 2 (1990), 121-131.

DAOUST, François. *SATO, système de base d'analyse de textes par ordinateur: manuel de référence,* Université du Québec à Montréal, Centre d'ATO, 1990, 134 p.

GRISHMAN, Ralph. « Natural Language Processing », *Journal of the American Society for Information Science,* vol. 35, no. 5 (1984), 291-296.

MEUNIER, Jean-Guy et al., « A Call for Enhanced Representation of Content as a Mean of Improving Online Full-Text Retrieval », *International Classification,* vol. 14, no. 1 (1987), 2-10.