ASp
la revue du GERAS

37-38 | 2002
Rédactologie - Situations d'apprentissage

La notion de « concept » dans les textes spécialisés : une étude comparative entre la progression thématique et la texture des concepts

Lise Fontaine et Yves Kodratoff

Édition électronique
URL : http://journals.openedition.org/asp/1490
DOI : 10.4000/asp.1490
ISBN : 978-2-8218-0390-9
ISSN : 2108-6354

Éditeur
Groupe d'étude et de recherche en anglais de spécialité

Édition imprimée
Date de publication : 30 décembre 2002
Pagination : 59-83
ISSN : 1246-8185

Référence électronique
Lise Fontaine et Yves Kodratoff, « La notion de « concept » dans les textes spécialisés : une étude comparative entre la progression thématique et la texture des concepts », *ASp* [En ligne], 37-38 | 2002, mis en ligne le 23 juillet 2010, consulté le 01 mai 2019. URL : http://journals.openedition.org/asp/1490 ; DOI : 10.4000/asp.1490

La notion de « concept » dans les textes spécialisés : une étude comparative entre la progression thématique et la texture des concepts

Lise Fontaine et Yves Kodratoff

Introduction

1 Cette étude se consacre à un réel problème pour les scientifiques non anglophones qui rédigent leurs articles de recherche en anglais, et plus précisément pour les scientifiques français rédigeant en anglais. Dans un monde universitaire où la règle « publier ou périr » règne, les problèmes linguistiques des scientifiques qui sont obligés de rédiger un article de recherche en langue anglaise étrangère ne sont pas simples et méritent l'attention qui leur a été récemment consacrée par les chercheurs en linguistique appliquée et en anglais de spécialité.

2 Souvent, même après avoir corrigé les erreurs de surface (orthographiques, lexicales et grammaticales), il reste des problèmes textuels qui rendent le texte incohérent ou incompréhensible au lecteur.

> There are also problems which only arise when a text is viewed as continuous discourse, because in some texts each sentence taken separately seems to be acceptable, even if the whole does not seem to be satisfactory. (Mauranen 1996)

3 La texture est donc aussi importante que la grammaticalité des phrases. La texture peut être définie « comme les fibres du texte – c'est elle qui supporte la cohérence » (Lemire 1996). Selon Halliday et Hasan le texte ne consiste pas en phrases, il est réalisé par ses phrases. De ce point de vue, il est nécessaire de dépasser la proposition afin de mieux comprendre cette texture.

4 La structure thématique d'un texte fait partie de sa texture; un bon tissage à ce niveau aide à rendre le texte cohérent et normalement plus facile à lire et à comprendre. Le tissage d'un texte cohérent n'a pas de modèle universel; chaque langue (et chaque culture) a sa propre façon de tisser un texte (Mauranen 1996 ; Smejrková 1996 ; Connor & Mayberry 1996 ; Ventola 1994 ; Weil 1991). Le rédacteur qui écrit dans une langue étrangère se trouve face à des problèmes textuels qui ne sont pas évidents.

5 Le domaine scientifique concerné est celui de la fouille de données (« *data mining* »). Le premier congrès international de KDD a eu lieu en 1995 et depuis 1997, deux autres congrès internationaux en KDD ont été créés[1] : PAKDD, *the Pacific-Asia conference in Knowledge Discovery and Data Mining* ; et PKDD, *the Principles and Practice of Knowledge Discovery in Databases*. Le développement de KDD en France est un des exemples de la propagation d'un domaine scientifique en langue étrangère avant de l'être dans la langue du pays. Le premier congrès francophone en KDD a eu lieu en France en janvier 2001[2] et c'est pour cette raison que nous n'avons pas travaillé sur un corpus de textes de KDD écrits en français.

6 Le taux de participation des scientifiques français à ces congrès montre des variations importantes. Hors du congrès européen, PKDD, leur taux d'acceptation est très faible. Par exemple, les taux de publication des Français en 2000 : 0 % d'articles à KDD-2000 (congrès américain), 2 % à PAKDD-2000 (congrès Pacifique-Asie) et 23 % à PKDD-2000 (congrès européen) (Fontaine 2001). Il faut noter que le comité de programme au PKDD-2000 comportait 36 % de Français comparé à 16 % provenant des pays anglophones. Au congrès KDD-2000, il était composé de 85 % d'Américains. Ceci montre bien que les scientifiques français ont une présence forte dans le domaine, mais ils n'arrivent pas à publier en anglais hors d'Europe. Le rôle du comité de programme est certainement très important. La différence entre les comités de PKDD-2000 et KDD-2000 est frappante : la probabilité qu'un article soit lu par un scientifique provenant d'un pays anglophone est de l'ordre de 15 % à PKDD et de 85 % à KDD.

7 Notre travail s'appuie sur deux corpus (voir ci-dessous), en deux étapes, constitués d'introductions d'articles rédigés en anglais par des anglophones et des francophones. Alors que les décomptes de mots sont très semblables dans les deux corpus, la façon dont les thèmes sont organisés et la façon dont les concepts sont évoqués diffèrent profondément.

8 Notre étude est consacrée aux sections d'introduction des articles du corpus. Les sections d'introduction sont souvent les plus difficiles à écrire (Swales 1990). En outre, la structure des Introductions est très variable. Paltridge note que le potentiel de structure générique (« *generic structure potential* ») des Introductions de son corpus est trop variable pour pouvoir en tirer des conclusions quant aux patrons (« *patterning* ») que peuvent présenter les éléments structuraux particuliers (1997 : 70). Il est donc raisonnable de supposer que l'Introduction contient plus de variations que le reste du texte, et qu'elle pose plus de difficultés au rédacteur non anglophone.

9 Il est bien connu que l'article de recherche a un modèle conventionnel d'organisation : ce qu'on appelle l'IMRD ou l'IMRED (Introduction, Méthodologie, Résultats, et Discussion). Selon Swales, les sections n'ont pas les mêmes distributions de caractéristiques linguistiques et rhétoriques :

> The evidence thus suggests a differential distribution of linguistic and rhetorical features across the four standard sections of the research article. (1990 : 136).

> On sait que les sections Introduction et Discussion sont en général les plus retravaillées avant la publication, tandis que les deux autres restent quasiment inchangées (Knorr-Cetina cité par Swales 1990 : 137).

10 Cette étude consiste à décrire la progression thématique de ces textes, à détecter la présence des concepts, puis à comparer leurs liens dans les deux corpus. La progression thématique est étudiée par un linguiste au moyen de la linguistique systémique fonctionnelle. Les concepts sont identifiés, par un informaticien spécialiste de fouille de données, par la position grammaticale des termes présents dans les textes. Ces deux structures nous donnent deux moyens différents d'étudier en profondeur la façon dont les scientifiques bâtissent leur argumentation. Leur étude et leur comparaison permettent de mettre à jour des stratégies de tissage textuel extrêmement différentes dans les deux corpus.

11 Nous donnerons d'abord une vue générale des progressions thématiques trouvées dans les textes, puis nous montrerons comment nous avons déterminé, automatiquement, la présence de concepts et, manuellement, la texture de ces concepts. Ceci nous conduira à une comparaison entre le rôle du thème et celui du concept dans le tissage de la texture textuelle. Enfin, nous soulignerons les résultats principaux de ce travail, ainsi que les travaux futurs qu'il suggère.

La Progression Thématique

12 Dans un premier temps, nous voulons décrire brièvement les résultats d'une étude comparant la progression thématique des textes scientifiques écrits en anglais par des auteurs francophones et anglophones (Fontaine 2001).

13 Le corpus est divisé en deux parties[3] : la partie « anglophone », celle des locuteurs de langue maternelle anglaise, contient les Introductions de seize articles de recherche écrits en anglais ; la partie « francophone », celles des locuteurs de langue maternelle française, contient celles de seize articles de recherche écrits en anglais. Les articles ont été publiés dans les actes du congrès KDD, PKDD, et PAKDD[4]. L'analyse a été faite sur les sections d'Introduction uniquement. La taille du corpus entier est de 14 415 mots : 6 895 mots dans le corpus « anglophone », avec une moyenne de 431 mots et une distribution entre 163 et 723 ; 7 520 mots dans le corpus « francophone » avec une moyenne de 470 mots et une distribution entre 203 et 1 066. Nous avons mis un exemple d'une section d'Introduction de chaque corpus dans l'annexe.

14 Afin de déterminer la catégorie linguistique de chaque article de recherche, nous avons vérifié sa langue maternelle auprès de chaque auteur, en anglais pour les auteurs présumés anglophones et en français pour les auteurs présumés francophones, afin de prendre en compte la langue maternelle de chaque auteur ainsi que son pays de travail.

15 Une analyse de la progression thématique permet de parcourir le texte en suivant d'une phrase à l'autre le déroulement du texte. Les relations entre deux phrases successives sont donc les plus importantes car une partie du texte qui est difficile à suivre s'explique au travers d'une analyse de la progression thématique et on constate une incohérence thématique dans le paragraphe (et le texte globalement) (Mauranen 1996). Cette analyse consiste d'abord en une division de la proposition en Thème et Rhème[5]. En Linguistique Systémique Fonctionnelle (LSF), le Thème est considéré comme le point de départ de la proposition en tant que message (Martin, Matthiessen & Painter 1997) et en anglais, il se

trouve en première position dans la proposition[6]. Le Thème est défini dans cette étude par ce qui précède l'élément fini dans la proposition, le Rhème étant ce qui reste[7]. En LSF, le Thème contient trois composants : le Thème textuel qui consiste en les éléments lexicaux qui permettent la connexion entre les propositions et qui servent à orienter ou à structurer le texte (p. ex. « ensuite »); le Thème interpersonnel comprend les éléments qui ont un aspect de modalité (p. ex. « peut-être ») ; le Thème expérientiel[8] qui contient une réalisation de la représentation expérientielle – un participant, un procès ou une circonstance (p. ex. « La voiture »).Toute proposition contient au moins un Thème expérientiel. L'exemple 1 montre la structure textuelle d'une proposition contenant un thème textuel, qui sert à orienter l'énoncé, et un thème expérientiel.

Exemple 1

For example,	the data streams being monitored	may be streams of numbers, mid type feature vectors, or free text documents.
Thème textuel	Thème expérientiel	
Thème		**Rhème**

16 La progression thématique, étudiée à l'origine par Danes (source : Duszak 1994, voir aussi Mauranen 1996), cherche à trouver la source de chaque thème, c'est-à-dire d'où vient le thème, à quoi fait-il référence ? Ainsi, on met en évidence la structure de la construction du discours. « *Since the processes of text understanding are knowledge-based as well as context-and-task sensitive, thematic configurations can also be seen as structures of expectation in discourse* » (Duszak 1994). Danes a identifié trois modèles principaux de la progression thématique[9] :

17 • *La progression à Thème Constant* : le thème d'une proposition est dérivé du thème de la proposition précédente, comme dans l'exemple 2.

Exemple 2

$T_1 \rightarrow R_1$ ↑	**We**	begin by discussing briefly cellular phone fraud detection as a domain to illustrate source of the issues in activity monitoring.
$T_2 \rightarrow R_2$	**We**	define the problem formally and present an evaluation methodology.

18 *La progression Linéaire* : le Thème d'une proposition provient du Rhème de la proposition précédente (voir exemple 3).

Exemple 3

$T_1 \rightarrow R_1$ ↑	We	present **a framework** within which these tasks have a natural expression.
$T_2 \rightarrow R_2$	**This framework**	modifies similarities of the tasks and highlights significant differences.

19 • *La progression à Thème Dérivé* : le Thème d'une proposition provient d'un Hyper-Thème qui n'est pas explicitement dans le contexte immédiat du Thème – ceci peut être le titre de l'article, ou le thème du paragraphe, etc. Par exemple, le Thème de la phrase en (4) est dérivé du titre.

Exemple 4

[HT] <titre> ↑ $T_1 \rightarrow R_1$	**The goal of activity monitoring**	is to issue alarms accurately and in a timely fashion.

20 Quand le Thème d'une proposition n'est pas motivé par le contenant d'un Thème ou Rhème précédant, il s'agit d'une rupture dans la progression thématique. Le Thème 13.1 dans l'exemple 5 est effectivement une rupture[10] de la progression thématique car il n'y a pas de lien dans le texte qui puisse motiver ce thème.

Exemple 5

La progression	#	Thème	Rhème
linéaire Th12 → Rh11	12	A window of size w	can then provide patterns of size 1 up to w.
Rupture	13	For temporal transactions,	this restricts the events to be bounded within a maximal time interval.
linéaire décalé Th14 →Rh12	14	An ambiguous pattern	contains at least one event that present a given degree of polymorphism.

21 On trouve également des progressions douteuses (notées par « ?? ») où il n'y a rien dans le texte pour aider le lecteur à voir clairement le lien éventuel avec la phrase précédente. Cependant, ce type de progression a été identifié uniquement dans le corpus francophone. L'absence de lien explicite ne permet pas l'identification d'une progression thématique ; la progression reste donc douteuse, comme dans l'exemple 6. On ne peut pas être sûr que le Thème 10.2 soit dérivé du Thème 10.1.

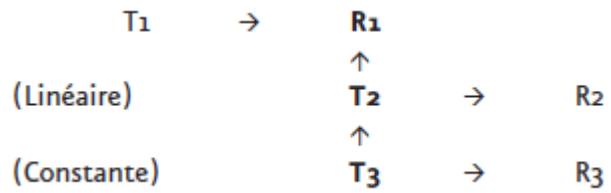
Exemple 6

la progression	#	Thème	Rhème
linéaire Th10.1 → ThRh10.1	10.1	Such a human-centered approach	has known an increasing development in Decision Aid during the last decade -from decision making in selection tasks to manufacturing and process control [17],
??constant Th10.2 →Th10.1	10.2	and some models	start being developed in KDD [5].

22 Dans les deux corpus, ce sont les progressions thématiques linéaires qui sont les plus fréquentes. La progression à thème constant est moins fréquente chez les auteurs francophones que les auteurs anglophones (21 % comparé à 32 %). Les thèmes dérivés sont à peu près employés de la même façon par les auteurs des deux corpus. La différence la plus claire se trouve dans le choix des thèmes qui causent une rupture dans le texte : 12 % des progressions provoquent une rupture dans le corpus francophone tandis qu'on n'en trouve que 3 % dans le corpus anglophone.

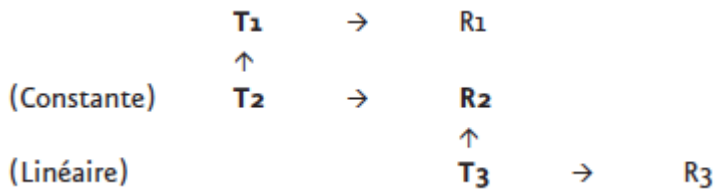
23 En prenant en compte les combinaisons de progressions thématiques, on trouve que la combinaison la plus fréquente dans les deux corpus est une progression d'un thème à progression linéaire vers un thème à progression constante (LC). (Voir l'annexe pour des exemples du corpus de chaque type de combinaison.) Cette combinaison est beaucoup plus fréquente chez les auteurs anglophones que chez les auteurs francophones, comme nous voyons dans le tableau 1.

La combinaison LC



24 La combinaison Constant - Linéaire (CL) est employée presque aussi fréquemment que la combinaison LC chez les auteurs francophones, mais les auteurs anglophones l'utilisent beaucoup moins souvent que la combinaison LC ; chez ces derniers, nous voyons que l'emploi de CL est de 42 % inférieur à celui de l'emploi de LC.

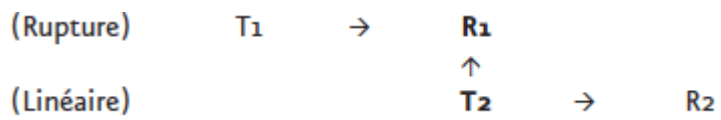
La combinaison CL



25 Aucune de ces deux combinaisons n'est préférable par rapport à l'autre ; elles sont toutes les deux acceptables. Cependant, le fait que les francophones l'utilisent aussi souvent que la progression LC peut indiquer un emploi aléatoire dans le choix du Thème, et donc l'absence d'un schéma préalable pour structurer le texte.

26 Les ruptures qui sont plus fréquentes dans le corpus francophone nous montrent deux différences intéressantes entre les deux corpus. Premièrement, on ne trouve presque jamais une rupture suivie d'une autre rupture dans le corpus anglophone, tandis que dans le corpus francophone, on trouve neuf exemples de ruptures suivies d'une autre rupture. Deuxièmement, dans le corpus francophone, une rupture a la même chance d'être suivie par une progression à thème constant que par une progression linéaire. En revanche, dans le corpus anglophone, une rupture est le plus souvent suivie d'une progression linéaire, ce qui suggère que la rupture puisse servir aux auteurs anglophones de technique d'introduction d'une idée dans le rhème qui sera ensuite reprise dans le thème de la proposition suivante.

La combinaison RL



27 Nous avons calculé le nombre de paires adjacentes possibles dans le corpus, ceci nous permet de comparer les combinaisons CC, CL, LL et LC.

Combinaisons CC, CL, LL et LC

	corpus anglophone, $n_a = 390$	corpus francophone, $n_f = 427$
Nombre de paires adjacentes, $N_x = (n - 1)$	$N_a = 389$	$N_f = 426$

28 Il y a donc 389 paires adjacentes dans le corpus anglophone et 426 paires adjacentes dans le corpus francophone. Nous pouvons calculer ensuite la fréquence de ces combinaisons comme suit : Fréquence = [(nx)/ (N)]*100], où nx = nombre d'occurrences de la combinaison, et N = Nombre de paires adjacentes. Le Tableau 1 montre les fréquences des combinaisons, suivies par le nombre d'occurrences, n, en parenthèse.

Tableau 1. Fréquences des combinaisons de progressions thématiques

	Fréquence de la combinaison dans le corpus anglophone, (n_a)	Fréquence de la combinaison dans le corpus francophone, (n_f)
LC	16,5 (64)	9,1 (39)
CC	10,3 (40)	4,9 (21)
CL	8,5 (33)	8,0 (34)
LL	4,6 (18)	7,0 (30)
LC + CC	26,8	14,0
CL + LL	13,1	15,0

29 Les auteurs anglophones ont plus tendance à développer les progressions dans le thème, vers la gauche. Les francophones ne montrent pas de combinaison de préférence et une progression est en règle générale suivie par n'importe quelle autre progression. Il faut noter que la combinaison CC présente chez les auteurs francophones une chute sévère de fréquence par rapport aux autres paires.

30 Nous n'avons pas trouvé les mêmes structures thématiques dans les deux corpus. Les anglophones ont, en général, tendance à négocier leurs informations vers la gauche de la proposition, dans le thème, tandis que les francophones n'ont pas montré de véritable structure thématique. Les auteurs francophones semblent avoir tendance à éviter la gauche : ils préfèrent faire progresser leurs informations vers la droite de la proposition, ce qui est le contraire des auteurs anglophones. Les difficultés qu'ont les auteurs francophones à gérer les fils thématiques dans leurs textes écrits en anglais rendent ces textes difficiles à suivre.

31 Dans ses recherches sur la cohésion textuelle, Carter-Thomas a montré que les apprenants en anglais langue étrangère ont des difficultés importantes à structurer leurs écrits sur le plan thématique. Ceci nous permet de cerner de plus près certains aspects essentiels à la réussite d'un texte (2000). Elle souligne l'importance de prendre en compte la structure thématique dans l'enseignement de langue étrangère : « en choisissant de ne pas traiter de cette composante textuelle, nous ne fournissons pas aux étudiants tous les outils nécessaires à l'amélioration de leur écrits » (2000). Elle a certainement raison. Mais une telle démarche aiderait seulement les jeunes chercheurs en formation.

32 Mauranen arrive à des conclusions similaires à la suite de ses recherches sur la structure thématique dans les textes institutionnels. Elle a comparé les textes écrits en anglais par des auteurs anglophones et finnois, et en finnois par des auteurs finnois. Son étude a pu identifier les difficultés qu'éprouvent les rédacteurs en anglais langue étrangère au niveau textuel. Quant à l'apprentissage d'une langue étrangère, elle dit qu'il faut aller encore plus loin :

> while it is useful for writers to widen their foreign-language competence to include text linguistic skills, it is also at the same time necessary to increase general awareness of global discourse structures as outcomes of specific and culturally shaped strategies. (Mauranen 1996)

33 Pour les chercheurs expérimentés, il n'existe toujours pas de véritable solution. C'est pourquoi nous cherchons à trouver un outil automatique qui puisse servir comme un outil de rédaction en anglais langue étrangère.

La Texture Conceptuelle

34 Avant de pouvoir aborder l'étude des relations entre concepts dans les phrases successives d'un texte, il est nécessaire d'expliquer comment nous avons déterminé la présence de concepts. Pour cette phase de notre recherche, nous avons augmenté la taille des deux corpus. Le corpus « anglophone » contient désormais les sections d'Introduction de 100 articles et le corpus « francophone » en contient 31 ; tous les textes sont en anglais.

35 Nous sommes en train de construire un système de reconnaissance de concepts dans les textes (voir, par exemple, Kodratoff 1999), appelé Rowan. Bien entendu, ce système n'est pas encore terminé et nous présentons ici une sorte de vue « instantanée » de Rowan, tel qu'il fonctionne en mai 2002. Il est en cours d'application sur trois domaines, les introductions d'articles de fouille de textes (dont nous parlons dans le présent article) écrits par des anglophones et par des francophones, des textes de psychologie en ressources humaines appartenant à la société PerformanSe, et des CV appartenant à la société Vediorbis.

36 La première étape de Rowan est une mise en forme « canonique » des textes qui comprend deux sous-étapes.

37 **Sous-étape n° 1** : nettoyage des textes. Cette étape est spécifique à chaque domaine. Par exemple, les références entre parenthèses sont éliminées des introductions d'articles scientifiques, ou encore les CV redondants sont éliminés de la base de CV.

38 **Sous-étape n° 2** : terminologie. Les termes utilisés par les auteurs sont détectés et une remise en forme canonique est effectuée. L'essentiel de cette mise en forme consiste à utiliser, un programme inductif proposant la création des termes qui sont introduits ensuite systématiquement dans les textes. Par exemple, « *data mining* » est partout remplacé par « *data-mining* ». Cette étape est très importante puisque le nombre de termes proposés est de l'ordre du millier dans les corpus que nous utilisons. Le programme de création de termes repose sur plusieurs heuristiques. D'une part, suivant les travaux de Jacquemin (1997), nous calculons une mesure d'association entre mots voisins, sur la base de leur présence exclusive ou non dans cette combinaison. Par exemple, « *data* » se trouve dans de nombreuses combinaisons avec d'autres mots que « *mining* » et donc son association à « *mining* » est faible. Inversement, « *mining* » se trouve presque seulement à la suite de « *data* », et son association avec « *data* » est très élevée, ce

qui favorise la formation du terme *data-mining*. D'autre part, nous élaguons les termes qui n'apparaissent pas suffisamment souvent, et ce en fonction de leur rang d'association et du nombre total de termes possibles. Enfin, notre détermination est itérative : après un premier passage, nous en effectuons un second, au cours duquel la mesure d'association est modifiée dans le sens que plus d'importance est donnée aux mots qui ont participé à la formation de termes au cours du premier passage. L'expérience nous a montré qu'un troisième passage est inutile.

39 Les termes sont construits après un étiquetage effectué par l'étiqueteur de Brill (1994 ; en anglais, il est en libre-service sur la toile, en français, on l'obtient auprès du CNRS), ce qui nous permet de déterminer des termes de la forme « nom-nom » comme « *association-rule* », de la forme « adjectif-nom » comme « *aggregate-function* », de la forme « nom-adjectif » (en pratique, ce cas ne se trouve qu'en français), de la forme « nom-préposition-nom » comme « *order-of-magnitude* », de la forme « nom-verbe » comme « *data-mining* ». L'itération permet de découvrir de nouveaux termes négligés au premier passage et de construire des termes contenant plus de deux mots, comme « *look-ahead-itemset* », mais au plus de quatre mots comme « *decision-treeinduction-algorithm* ».

40 La deuxième étape de Rowan est une analyse syntaxique superficielle (« *shallow parsing* »). L'intérêt de cette étape découle de l'hypothèse développée et confirmée par Claire Nédellec (Faure & Nédellec 1998 ; Nédellec 2000) que les relations syntactiques permettent de supprimer la polysémie intrinsèque au langage naturel, même de spécialité. Ainsi, le mot « *algorithm* » en tant que sujet du verbe « *to develop* » est une instance du concept « description générale de l'algorithme », alors que le mot « *algorithm* » en tant que sujet du verbe « *to discover* » est une instance du concept « description des sorties de l'algorithme », tout au moins dans les textes que nous avons analysés. Nous avons pu constater que cette hypothèse est exacte à environ 80 %, ce qui à la fois la rend intéressante, et exige des améliorations futures. Notons que cette hypothèse est statistiquement vraisemblable : alors qu'un texte contient de l'ordre de quelques milliers de mots significatifs, ces mots se retrouvent dans une centaine de milliers de relations syntaxiques. Ceci explique une forme de désambiguïsation des sens des mots. D'autre part, cela montre l'énormité de la tâche à effectuer au cours de l'étape 3, ci-dessous, et l'importance de l'étape 4.

41 Nous avons ajouté à cette hypothèse celle que les termes désignent un concept en dehors de toute autre relation syntaxique que le fait d'être un terme. Cette hypothèse se révèle exacte à environ 70 % dans nos corpus.

42 Pour exécuter l'analyse syntaxique superficielle, nous avons utilisé une version de l'analyseur syntaxique superficiel développé par Xerox qui a été prêtée gracieusement au LRI. Les entrées de Rowan ont la forme des sorties de cet analyseur, et donc l'utilisation d'un autre analyseur ne demanderait qu'une adaptation des entrées de Rowan.

43 La troisième étape de Rowan est entièrement dans les mains de l'expert du domaine. Il ou elle décide d'abord des concepts intéressants à déterminer dans les textes. Pour les introductions des articles de fouille de données, le spécialiste est l'un de nous.

44 Partant de principes bien connus depuis longtemps[11], et dus à Arnault et Nicole (1662) :

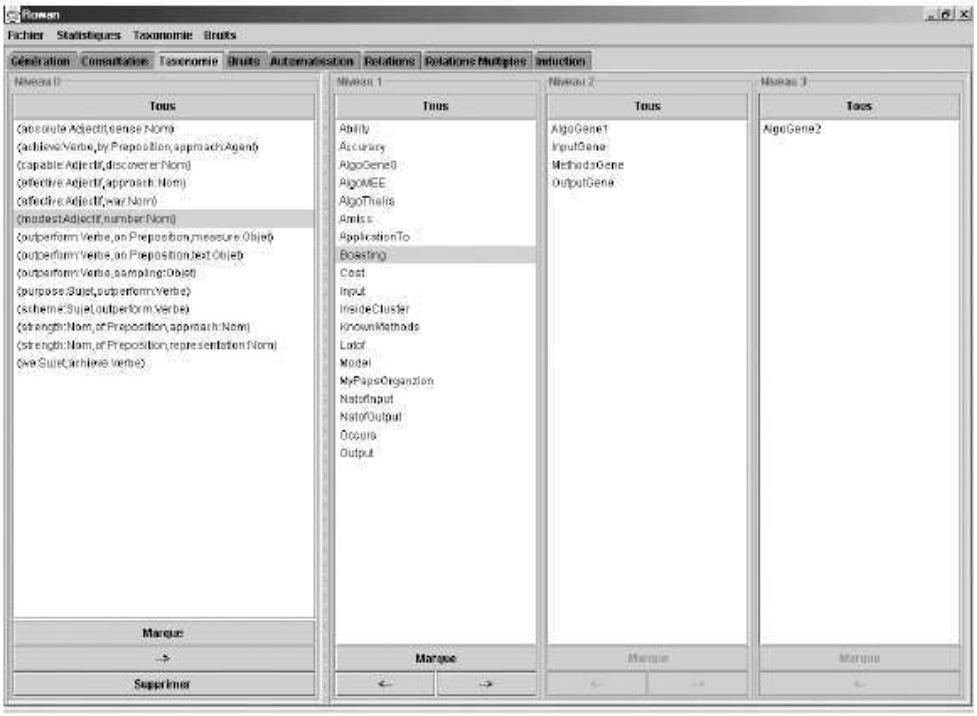
> J'appelle compréhension de l'idée les attributs qu'elle enferme en soi et qu'on ne peut leur ôter sans la détruire ... J'appelle étendue de l'idée, les sujets à qui cette idée convient,

nous pouvons affirmer que l'expert possède des « définitions en intention » des concepts, éventuellement inconscientes. Face à un texte, nous demandons à l'expert de fournir des

« définitions en extension » de la façon de décrire un concept dans un texte. Il est notable que notre approche ne définit jamais ce qu'est un concept, mais se contente de caractériser la présence d'un concept dans le texte.

45 La figure 1 donne une idée des concepts que nous avons introduits.

Figure 1

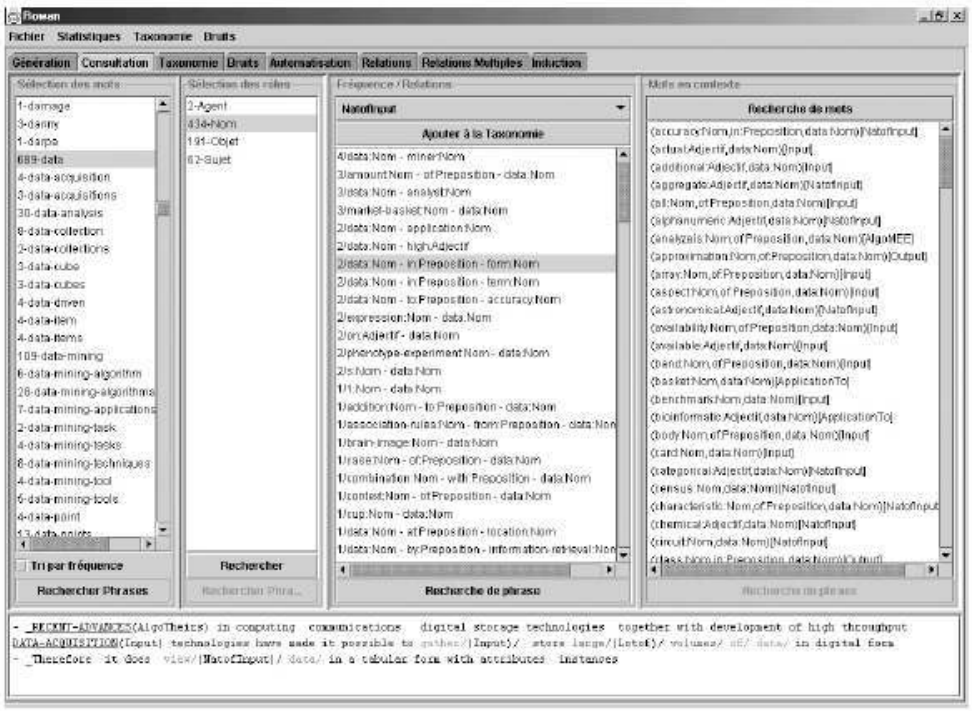


46 La plupart de ces concepts sont « de bon sens » comme « *AlgoMEE* » qui signifie que l'auteur parle de son algorithme. Certains ont été introduits en examinant les textes, comme « *Boasting* » (« autosatisfaction », surtout présent dans les textes des francophones) et « *Amiss* » (où l'auteur met en doute l'efficacité de son algorithme, surtout présent dans les textes des anglophones).

47 Le spécialiste doit ensuite parcourir les formes syntaxiques et les termes, et attribuer un concept à certains d'entre eux. Par exemple, la phrase contenant « *modest number* » est en fait une phrase d'autosatisfaction d'un auteur anglophone, d'où sa place dans la taxonomie, comme dans la figure 1.

48 Rowan fournit de nombreuses indications au spécialiste pour l'aider dans la tâche de détermination des concepts par les formes syntaxiques. Sur la figure 2, on voit l'écran de base de Rowan, où l'expert voit que le mot « *data* » apparaît dans 699 relations syntaxiques et, pour 434 d'entre elles, comme un nom. Il apparaît deux fois dans la relation syntaxique « *data in form* », et les phrases contenant cette association se trouvent dans la fenêtre du bas.

Figure 2



49 Les concepts déjà découverts dans ces phrases sont indiqués en couleur pour les relations syntaxiques et en majuscules soulignées pour les termes. La liste des concepts déjà découverts et contenant le mot « *data* » se trouve dans la colonne de droite appelée « recherche de mots ». Si l'utilisateur désire ajouter l'expression « *data in form* » aux instances du concept « *NatofInput* », il clique simplement sur « Ajouter à la taxonomie ».

50 Nous évaluons à un millier d'heures de travail intensif la constitution de la version manuelle de la taxonomie des introductions des articles anglophones de fouille de données. Créer une taxonomie de toutes pièces est donc très difficile. Rowan est surtout utile pour construire une taxonomie à partir d'une autre existant déjà. Par exemple, nous avons créé une taxonomie spéciale pour les textes francophones en une centaine d'heures, malgré les très grandes différences constatées (et qui ne font pas l'objet de cet article).

51 La quatrième étape de Rowan, et celle qui va demander le plus de développements dans l'avenir, en est la phase inductive au cours de laquelle une taxonomie existante est complétée automatiquement. Présentement, environ 80 % de la taxonomie des introductions des articles anglophones de fouille de données ont été obtenus manuellement, et 20 % par induction automatique. L'algorithme d'induction actuel se fonde sur la détermination de quelques milliers de groupes, appelés noyaux[12], contenant des mots ayant le plus grand nombre possible de relations syntaxiques en commun. On fait croître ces noyaux en leur ajoutant de nouvelles relations syntaxiques qui n'entraînent pas une trop grande diminution du nombre de relations en commun. La notion de « diminution » est fixée par des paramètres choisis par l'utilisateur. Nous avons expérimentalement fixé les paramètres de sorte que l'algorithme propose environ 5 000 concepts. Ceux-ci sont comparés aux concepts obtenus manuellement et lorsque les concepts induits contiennent au moins une relation en commun avec les concepts

obtenus manuellement et qu'aucune de leurs relations en commun n'est différente de celles des concepts obtenus manuellement, alors les relations non en commun et leurs concepts associés sont ajoutés à la taxonomie.

52 Une fois la taxonomie construite, les concepts d'un nouveau texte du même registre sont facilement identifiés et codés dans le texte. Jusqu'à présent, le système identifie les concepts une phrase à la fois, comme dans les exemples (7) et (8) :

Exemple 7

> In this paper the desired knowledge is major topics in a collection; DATA-MINING [Known-Methods] is used to discover [Output] patterns that disclose those topics[13].

Exemple 8

> The discovered [Output] substructures allow [Ability] abstraction over detailed structure in the original [Input] data[14]

53 Dans cette étude, les liens entre les concepts des phrases successives ont été identifiés à la main pour une analyse fouillée. Ensuite les liens entre concepts ont été analysés afin d'identifier la texture conceptuelle. Nous pouvons maintenant comparer l'analyse de la progression et la texture des concepts.

Une comparaison de la texture thématique et conceptuelle

54 Nous avons démontré que les auteurs français ne maîtrisent pas la texture des textes en anglais scientifique. Une meilleure compréhension de ces difficultés textuelles aidera certainement les auteurs écrivant dans une langue étrangère. Pour les scientifiques expérimentés, il est trop tard pour parler de la pédagogie, même si une telle approche serait très utile dans les cours d'anglais de spécialité. Ce problème ne se résout pas facilement car il y a peu de ressources disponibles pour le rectifier. C'est pourquoi nous voulons comparer les techniques des linguistes et ceux des scientifiques du domaine pour voir si une collaboration ne serait pas possible. Maintenant que la possibilité d'identifier automatiquement un « concept » dans un texte de spécialité n'est plus une utopie (même si cela reste encore peu commode), il nous reste à déterminer à quel point la présentation des concepts dans un texte peut influencer la cohérence structurelle pour arriver à décerner la possibilité de créer un outil automatique d'aide à la rédaction scientifique.

55 L'analyse linguistique en LSF respecte la relation dynamique entre langue et contexte. Le contexte social est réalisé par un registre à trois variables : le mode discursif, le rôle discursif et le champ discursif. Dans notre corpus, le mode discursif est assez complexe ; nous allons simplement dire que l'article de recherche se trouve en mode de texte imprimé dans les Actes d'un colloque, mais il paraît aussi en mode électronique[15]. Le rôle discursif comprend les relations sociales entre locuteurs. L'article de recherche se trouve dans une situation multilogue où plusieurs auteurs ont écrit un seul article en collaboration et il y a certainement plusieurs récepteurs. Les relations dans ce réseau sont forcément variables. Le champ discursif concerne l'activité de discours, le but du discours

et les participants. Dans notre corpus, le but est de transmettre des connaissances nouvelles et intéressantes aux collègues scientifiques dans la communauté KDD.

56 Ces trois variables contextuelles définissent le registre de la communauté KDD internationale. La relation entre registre et genre est complexe. Le texte produit dans un genre réalise donc le but communicatif de ce genre grâce au registre : « *ideology is realised by genre, which is in turn realised by register* » (Martin 1993). Le fait de partager le même champ discursif, et en conséquence, le même registre, permet l'interprétation sémiotique du discours dans un domaine. Le champ discursif est plus important que les autres variables dans cette étude car c'est ici qu'on trouve la représentation des connaissances partagées parmi les membres de cette communauté scientifique ; c'est-à-dire ce qui est important ou non. Un concept tel que nous l'entendons est non seulement une représentation de ces connaissances, mais l'analyse de sa représentation textuelle permet de comprendre, d'une part, comment les locuteurs se servent de ces concepts et, d'autre part, comment ils les expriment dans le texte. À titre d'hypothèse de travail et suivant un principe de simplicité, nous admettrons ici que les concepts sont organisés dans l'esprit des scientifiques du domaine spécialisé en forme de taxinomie[16]. Nous pouvons nous appuyer sur cette taxinomie afin d'étudier un nouveau type de cohérence textuelle.

57 Ce que nous avons vu quant à l'analyse textuelle est que la progression thématique concerne explicitement le rôle du Thème et donc le rôle du Rhème n'apparaît pas sauf lié à un Thème. Cependant le Rhème est porteur d'informations importantes et nous allons voir que les concepts que nous avons identifiés se trouvent le plus souvent dans le Rhème. Le Thème sert à organiser la proposition du point de vue du locuteur (ce dont il parle) et le Rhème porte l'information nouvelle pour le co-locuteur (ici il s'agit du lecteur) (Martin 1992).

> Le Rhème est très intéressant car il contient toujours l'élément qui complète le développement de la communication [...] les éléments rhématiques portent une information qui est plus importante que les éléments thématiques.[17] (Firbas 1995)

58 Il ne sera donc pas étonnant que les concepts se trouvent dans le Rhème le plus souvent, et rarement dans le Thème. L'étude de ces éléments normalement rhématiques est très importante pour complémenter notre compréhension de la texture. La texture des concepts n'est donc pas vraiment comparable à la progression thématique car les deux apportent différentes contributions au tissage et à la compréhension du texte. Nous allons néanmoins les comparer afin de commencer à comprendre ce nouveau réseau de liens textuels.

59 Notre méthodologie d'analyse comparative suit le travail de Mauranen (1996). Elle avait émis une hypothèse pour tester son analyse. Les notions de base sont les suivantes : « Puisque la structure Thème-Rhème sert à organiser la phrase, elle doit obéir aux contraintes du contexte immédiat afin de faire progresser le texte de façon cohérente »[18] (Mauranen 1996). Cette position se trouve aussi dans les travaux faits dans le domaine du Traitement du Langage Naturel : « *In discourse, the preceding major clause has a substantial effect on the interpretation of the next major clause* » (Allen 1995). Les relations entre les phrases sont donc très importantes pour une texture réussie.

60 Nous supposons que chaque phrase est liée à la phrase précédente et, par conséquent, il y a un lien identifiable entre deux phrases successives (Mauranen 1996).

61 Pour l'analyse de la progression thématique, ceci veut dire que chaque thème expérientiel est lié au thème ou au rhème (ou aux deux) de la proposition précédente. Pour l'analyse de la texture conceptuelle, ceci veut dire que chaque phrase contient au moins un

concept qui est lié à la phrase précédente par un concept identique, ou un concept qui est en relation de co-hyponymie ou hyponymie (p. ex. le concept [NATURE-OFINPUT] est lié au concept [INPUT] par une relation d'hyponomie).

62 Si un lien ne figure pas dans le texte, il sera considéré comme une déviation du tissage « standard » du texte.

63 Dans l'exemple (9), extrait du texte [E3] du corpus anglophone, nous voyons que tous les Thèmes sauf le premier sont liés. Le Thème 6 n'est pas lié à la phrase précédente mais il est explicitement marqué dans le texte comme étant lié au Thème 1. Nous voyons aussi que ce fait est noté par la texture des concepts identifiés. Également dans cet extrait, chaque phrase est liée à la phrase précédente par un concept identique ou par une relation de co-hyponomie.

64 Parmi les quinze concepts qui se trouvent dans cet extrait, treize sont contenus dans le Rhème et deux se trouvent dans le Thème. Il est intéressant de noter qu'ici les liens vers les concepts en position de Thème (marqué par "*th") sont identiques à ceux trouvés dans la progression thématique.

Clé des abréviations de concepts

KM = [KNOWN-METHODS]
Acc = [ACCURACY]
AG = [ALGORITHM-GENERAL-0] ;
AM = [ALGORITHM-MINE]
NI = [NATURE-OF-INPUT]
Occ = [occurs]
LotOf = [LotOf]
NO = [NATURE-OF-OUTPUT]
O = [OUTPUT]
PapOrg = [paper-organisation]

65 Une ligne solide indique un lien entre deux concepts identiques ; une ligne en pointillés signifie un lien entre un concept et son co-hyponyme.

Exemple 9

[E3]	Progression thématique	Texture des concepts
Titre	*Explicitly Representing Expected Cost: An Alternative to ROC Representation*	KM
S1	Thème 1 [dérivé] → hyper thème = général	Acc – AG
S2	Thème 2 [linéaire] → Rhème 1	AG - NI – Occ
S3	Thème 3 [constant] → Thème 2	NI – LotOf
+S4	Thème 4 [linéaire] → Rhème 2 & 3	NI*th
S5	Thème 5 [constant] → Thème 4	AG - AG – NI
S6	Thème 6 [linéaire (-5)] → Rhème 1	Acc*th - KM - AG – NO
S7	Thème 7 [linéaire] → Rhème 6	O – NO

66 On peut suivre effectivement une progression des concepts présentés dans le texte (voir l'annexe pour le texte complet de cet exemple). Nous trouvons deux fils : le lien entre les instances du concept [NATURE-OF-INPUT] et celui entre les instances du concept [

ALGORITHM-GENERAL-0]. Ils progressent ensemble dans le texte et mènent au concept [NATURE-OF-OUTPUT]. Cette section du texte va certainement parler d'un algorithme de façon générale avec plus de développement sur la communication de la nature des données d'entrée et les implications avec les données de sortie.

67 L'exemple (10), extrait du texte [F3] du corpus francophone (voir l'annexe pour le texte complet de cet exemple), ne montre pas autant de connexions entre les phrases successives.

Exemple 10

[F3]	Progression Thématique	Texture des concepts
Titre	*Approximation of Frequency Queries by Means of Free-Sets*	
S1	Thème 1 [dérivé] → hyper thème = général	KM*th - O - NI
S2	Thème 2 [dérivé] → hyper thème = auteurs	AM - KM
S3	Thème 3 [linéaire] → Rhème 2	LotOf -O
S4	Thème 4 [dérivé] → hyper thème = article	AM
S5	Thème 5 [linéaire] → Rhème 4	NI
S6	Thème 6 [linéaire] → Rhème 5	PapOrg

68 Dans cet extrait, les concepts se trouvent partagés dans le texte de façon presque égale et donc on n'arrive pas facilement à saisir ce dont il parle. Ces problèmes sont confirmés dans l'analyse de la progression thématique où l'on voit qu'il y a aussi plus de déviations et les phrases ne se sont pas liées l'une à l'autre en séquence.

Résultats

69 Nous avons tiré trois textes de chaque corpus (auteurs anglophones : [E1], [E2], [E3]; auteurs francophones : [F1], [F2], [F3]) pour une analyse détaillée de la progression thématique et la texture des concepts dans les deux groupes. Les deux groupes sont semblables en ce qui concerne le nombre de mots, de phrases et de propositions dans les textes. Nos résultats quantitatifs montrent des différences plus importantes entre les deux groupes en ce qui concerne la progression des concepts. Le tableau 2 montre les déviations dans les textes écrits par des auteurs anglophones. Ce qui ressort clairement est que le taux de déviations de la progression thématique est très proche de ceux de la texture conceptuelle. Ceci veut dire que nous trouvons un taux stable qui rend compte d'un niveau 'acceptable'[19] de déviations dans les deux types de texture chez les auteurs anglophones. Les taux relativement bas des déviations dans le corpus anglophone confirme notre hypothèse de départ : deux phrases d'un texte donné seront liées au travers du Thème et aussi dans le Rhème par deux concepts liés.

Tableau 2. Répartitions des textes écrits par les auteurs anglophones

Textes	[E1]	[E2]	[E3]	somme
# déviations de la progression thématique taux***	4 (0.2)	3 (0.2)	6 (0.2)	13 (0.2)
# déviations de la texture conceptuelle taux	8* (0.4)	2 (0**) (0.1)	7 (0**) (0.3)	17 (0.3)
# phrases	20	15	24	59
# propositions	25	18	33	76
# mots	369	356	542	1267
				moyennes
rapport concepts/mots	1 : 10.9	1 : 7.1	1 : 11.1	1 : 9.7
rapport liens/phrase	1 : 1	1 : 0.35	1 : 0.60	1 : 0.65

* trois de ces phrases sont des exemples et aucun concept n'a été trouvé.

** nombre de déviations, si on accepte un lien à une distance de (n-2), c'est-à-dire la phrase avant la précédente.

*** le taux est calculé en divisant le nombre de déviations par le nombre de (1) propositions moins 1 pour les progressions thématiques et (2) phrases moins 1 pour la texture conceptuelle.

70 Les résultats du corpus francophone ne permettent pas une telle confirmation de notre hypothèse (voir Tableau 3). Nous constatons d'abord que dans le corpus francophone, les textes ont un peu plus de mots mais moins de propositions. Cependant, il y a deux fois plus de déviations de la progression thématique dans ce corpus que chez les auteurs anglophones et cette proportion augmente encore quand on considère les déviations de la texture conceptuelle. Ce fait ne met pas en défaut notre hypothèse mais signale plutôt une absence, chez les auteurs francophones écrivant en anglais, d'une représentation syntaxique et textuelle des concepts scientifiques en anglais.

Tableau 3. Répartitions des textes écrits par les auteurs francophones

textes	[F1]	[F2]	[F3]	Somme
# déviations de la progression thématique taux***	6 (0.4)	11 (0.3)	8 (0.4)	25 (0.3)
# déviations de la texture conceptuelle taux	10 (8**) (0.8)	16 (13**) (0.8)	16 (10**) (0.8)	41 (0.7)
# phrases	14	22	22	58
# propositions	16	34	23	73
# mots	393	561	458	1412
				Moyennes
rapport concepts/mots	1 : 39.3	1 : 16.0	1 : 16.4	1 : 23.9
rapport liens/phrase	1 : 3.5	1 : 1.2	1 : 1.4	1 : 2.0

71 Il est clair que les auteurs francophones manquent de connexions conceptuelles entre les phrases. Même si on ne considère pas seulement les phrases successives et que l'on

compte le nombre de liens entre tous les concepts dans un texte, les francophones ont toujours moins de concepts par rapport au nombre de mots ainsi que par rapport au nombre de liens entre toutes les phrases. Les auteurs anglophones ont tissé des textes conceptuellement plus denses que leurs collègues francophones : la répartition des concepts est de 1 concept tous les 10 mots en moyenne chez les anglophones et de 1 concept tous les 24 mots chez les francophones. Les concepts représentés dans les textes des anglophones sont aussi plus connectés : dans les textes du corpus anglophone, on trouve en moyenne presque deux liens pour chaque phrase tandis que, dans les textes des francophones, on trouve un lien entre deux concepts toutes les deux phrases. Le nombre de concepts et les liens qui les connectent font que les textes des auteurs anglophones sont d'un tissage plus serré que les textes écrits en anglais par les scientifiques francophones. On peut voir ceci comme des « trous » dans le texte, des endroits vides qui rendent le texte plus difficile à lire. Ces résultats sont cohérents avec ceux de Fontaine (2001) qui a montré que la texture des auteurs francophones écrivant en anglais est moins explicite et moins structurée, car ayant moins de liens inter-phrastiques.

72 Ce manque relatif de liens explicites s'explique au moins partiellement par le travail de Bachschmidt (1999). Dans une étude comparative de rhétorique anglophone et francophone, Bachschmidt explique que :

> Les auteurs francophones se fondent sur des faits connus pour résoudre le problème constitué par leur hypothèse ou répondre à la question qui a motivé leur recherche. [...] Bien différente est la rhétorique de l'anglophone [...] il s'agit pour l'auteur anglophone de justifier de la manière la plus explicite possible l'interprétation des données scientifiques. [...] Le co-énonciateur anglophone n'est pas véritablement appelé à inférer quoi que ce soit. (200-202).

73 Ce dernier point suggère que le rôle *attendu* du co-énonciateur ne soit pas le même pour un francophone et un anglophone. La relation entre rédacteur et lecteur peut être considérée comme un contrat de compréhension mutuelle, où chacun comprend sa tâche, et ce contrat est défini dans la communauté culturelle des co-énonciateurs.

74 Le fait qu'un texte puisse paraître étrange et même incohérent n'est pas forcément le résultat d'une pensée fautive, mais plutôt dû à des difficultés à réaliser les stratégies discursives d'une langue étrangère (Mauranen 1996) et, par conséquent, le texte ne présentera ni le sens voulu par le rédacteur ni la texture attendue par le lecteur. Carter-Thomas (2000) a montré l'importance d'un certain niveau de « dextérité syntaxique » pour manipuler la structure thématique et par conséquent la cohérence textuelle. Ceci est aussi vrai pour améliorer la représentation des concepts dans un texte écrit dans une langue étrangère. Les textes écrits par des francophones présentent beaucoup moins de concepts que les auteurs anglophones. Bien entendu, les auteurs francophones ne sont pas moins spécialistes ni intelligents, donc comment expliquer cette absence ?

75 Dans l'exemple (11) nous trouvons une phrase provenant du corpus francophone où le concept [OUTPUT] a été identifié en rencontrant les mots « *knowledge discovery* ». Nous savons que cette identification est une erreur car l'auteur essaie de parler de processus et donc du concept [KNOWNMETHODS]. Le problème est particulièrement typique puisqu'il s'agit d'une citation « presque exacte » d'une phrase célèbre dans le domaine, mais la citation complète contient « *knowledge discovery in databases* » qui est devenu le nom universitaire du « *data mining* », donc une instance de [KNOWNMETHODS]. Dans ce domaine, « *knowledge discovery* » a deux sens en anglais : soit pour parler du processus de la

découverte de connaissances, soit pour parler de ce qui a été découvert (dans ce cas, les connaissances sont effectivement une sortie du processus).

Exemple 11

> KNOWLEDGE-DISCOVERY [OUTPUT] has been defined as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

76 Puisque cette phrase parle du processus, et donc du concept [KNOWNMETHODS], le lien avec les phrases environnantes est perdu et le lecteur ne peut pas être certain en première lecture de ce dont l'auteur parle.

77 Une deuxième explication se trouve dans l'exemple (12). L'auteur parle effectivement d'un concept quand il écrit « *continuous attributes* » (celui de [NATURE-OF-INPUT]), mais les auteurs anglophones n'utilisent pas cette expression, ils disent : « *continuous features* » qui indique en effet le concept [NATURE-OF-INPUT].

Exemple 12

> However, considering that rules using intervals on continuous attributes [*vide*] are often difficult to interpret and that strict thresholds are often too abrupt, we have decided to use fuzzy logics.

78 Les endroits vides dans les textes, c'est-à-dire sans concept identifié, ne sont pas véritablement vides de concept mais le concept n'est pas exprimé de façon facilement reconnaissable. C'est la combinaison des mots, termes et structure syntaxique qui permet d'exprimer, de façon mutuellement compréhensible, un concept. Pour pouvoir régler ce problème de représentation des concepts, deux capacités sont nécessaires : la connaissance de la façon ritualisée dont les concepts s'expriment dans un domaine spécialisé, et un niveau suffisant de dextérité syntaxique pour pouvoir manipuler l'expression d'un concept.

Conclusions

79 Notre tâche dans cette étude était de dégager la structure thématique, de détecter la présence des concepts dans les textes, puis de comparer leurs liens dans deux corpus de textes écrits en anglais – ceux des auteurs anglophones et ceux des auteurs francophones. Nos résultats montrent que les auteurs francophones maîtrisent moins efficacement la construction textuelle autant du point de vue de la présence des concepts que de la structure thématique, laissant plus de ruptures et de progressions ambiguës difficiles à suivre, et plus de vides conceptuels. Il ne faut pas oublier que ces textes ont été sélectionnés par un comité de sélection et ensuite publiés dans un livre des actes du congrès. Ainsi, ce sont des textes approuvés par cette communauté scientifique. On ne peut que supposer que ces problèmes textuels sont encore plus marquants dans les textes rejetés et les textes des jeunes chercheurs (moins expérimentés).

80 Jusqu'à présent, il y a peu de solutions. Le chercheur non-anglophone qui rédige en anglais peut éventuellement essayer de trouver un cours enseignant comment mieux gérer l'ensemble du texte. Il y a peut-être dans son département un service d'aide à la

rédaction, mais les problèmes textuels ne sont pas souvent abordés (Carter-Thomas 2000 ; Mauranen 1996). Il peut essayer de trouver quelqu'un pour éditer son travail, ce qui pose un autre problème car, en effet, une analyste de texte peut certainement comprendre tout ce qui est lié à la texture, mais comprendra mal le travail dans le domaine spécialisé et, en conséquence, ceci deviendra vite un projet énorme qui prendra beaucoup de temps. Enfin, il peut essayer de trouver un collègue dans la langue ciblée pour relire et corriger le texte, mais encore une fois ce n'est pas simple et le problème devient que personne n'a le temps ! Il nous semble donc utile d'explorer ce que les technologies de l'information peuvent apporter à la rédaction en langue étrangère, notamment en anglais.

81 L'identification des concepts et la structure de leurs représentations dans le texte font partie de la texture de ces textes, créant des liens textuels jusqu'à présent ignorés. Contrairement à la progression thématique qui permet au rédacteur de positionner ce qu'il veut dire, la texture des concepts trace le mouvement des éléments essentiels. Nous n'avons fait qu'aborder ce sujet et, en un sens, cette étude se termine par encore plus de questions. La plus importante reste de savoir si, selon notre hypothèse, la correction des déviations de la texture conceptuelle suffit à améliorer le texte, et par conséquent, à aider le rédacteur.

82 En appliquant une approche issue de l'informatique, nous avons voulu commencer une recherche en vue de la création d'un outil d'aide à la rédaction scientifique. Si la réponse à la question que nous venons de poser ci-dessus est l'affirmative, nous voyons immédiatement deux possibilités quant à l'aide informatique à la rédaction de textes dans le domaine de KDD.

83 La première, une aide à rédaction : pour trouver une structure syntaxique et lexicale pour un concept donné, sélectionner le concept dans une liste pour voir les structures associées. Ceci constituerait une sorte de dictionnaire des formes syntaxiques associées à un concept.

84 La seconde, un Vérificateur de Texture : un système identifiant les concepts dans un texte donné et signalant toute phrase sans lien avec la phrase précédente; il signalera ainsi tout espace vide où un concept devrait normalement (statistiquement) se trouver. En réalité, nous ne sommes pas loin de la réalisation d'un tel outil, dans le domaine de KDD au moins. Pour d'autres domaines, la réponse est moins positive car, sans la taxonomie de concepts au départ, l'outil est impossible. Nous avons vu à quel point la construction d'une telle taxonomie est laborieuse ; il est aussi nécessaire de trouver un spécialiste du domaine pour la construire. La seule motivation linguistique n'est évidemment pas suffisante pour effectuer un tel travail. Les autres domaines auxquels nous appliquons notre méthode sont motivés par l'apport de notre méthode aux techniques de découverte automatique de patrons (« extraction de l'information ») dans les textes de spécialité. Chaque fois qu'une telle motivation économique existera, les retombées linguistiques seront immédiatement exploitables. Ceci étant, les progrès, essentiellement dans les méthodes inductives de construction automatique de taxonomies de concepts, détermineront la possibilité d'élargir les recherches sur la texture des concepts à d'autres domaines.

85 Le but ne doit pas être de forcer les auteurs écrivant dans une langue étrangère à se conformer à un style anglo-américain, mais de les aider à rendre un texte cohérent. Nous ne devons pas oublier le rôle important que jouent les lecteurs dans le discours scientifique : les collègues et les éditeurs. Il est vrai que les rédacteurs en langue anglaise étrangère doivent faire un effort, et nous avons consacré cet article à une partie du

processus de la rédaction. Cependant, il faut aussi que les lecteurs se rendent compte de ces problèmes textuels afin d'éviter de rejeter un travail pour des raisons superficielles. Selon Irena Vassileva (1998) :

> Members of the (international) academic discourse community should be made aware of the existence of other, different cultures rhetorics and learn to be tolerant towards such styles. [...] Speakers of other languages who use English for international communication should be taught how to do it in a way acceptable for the intended audience, while at the same time preserving their cultural identity.

BIBLIOGRAPHIE

Allen, James. 1995. *Natural Language Understanding*, 2nd ed. Californie : The Benjamin Cummings Publishing Company.

Arnauld, Antoine & Pierre Nicole. 1965 [1662]. *La Logique, ou l'art de penser*. Paris : PUF.

Bachschmidt, Patrick. 1997. « Procédure de constitution d'un corpus attesté d'articles de recherche scientifique en vue d'une étude contrastive ». *ASp* 15-18, 133-138.

Bloor, Thomas & Meriel Bloor. 1995. *The Functional Analysis of English: A Hallidayan Approach*. Londres : Arnold.

Brill, E. 1994. « Some advances in transformation-based part of speech tagging ». Proc. AAAI 1, 722-727.

Carter-Thomas, S. 2000. *La Cohérence textuelle : pour une nouvelle pédagogie de l'écrit*. Paris : L'Harmattan.

Carter-Thomas, Shirley & Elizabeth Rowley-Jolivet. 2001. « Structure informationnelle et genre : le cas de la communication scientifique de congrès ». Journée d'étude, juin 2001, ILPGA, Paris.

Connor, Ulla & Susan Mayberry. 1996. « Learning discipline-specific academic writing: A case study of a Finnish graduate student in the United States ». In Ventola, Eija & Anna Mauranen (eds.), *Academic Writing. Intercultural and Textual Issues*. Amsterdam/Philadelphie : John Benjamins, 231-253.

Crosnier, Elisabeth. 1994. « Enquête sur l'évaluation de l'anglais par les anglophones dans les publications des scientifiques français ». *ASp* 3, 39-55.

Diday E., J. Lemaire, J. Poujet & F. Testu. 1982. *Eléments d'analyse des données*. Paris : Dunod.

Duszak, Anna. 1994. « On Thematic Configurations in Texts: Orientation and goals ». In Mejrková, S. & F. Sticha (eds.), *The Syntax of Sentence and Text*. Amsterdam/Philadelphie : John Benjamins, 105-119.

Faure, D. & C. Nédellec. 1998. « A Corpus-based conceptual clustering method for verb frames and ontology acquisition ». LREC workshop on Adapting lexical and corpus ressources to sublanguages and applications. Granada, Espagne, 5-12.

Firbas, J. 1995. « A contribution on a panel discussion on Rheme ». In Fontaine, L. 2001. Note de Recherche de DEA, Université Victor Segalen Bordeaux 2.

Fontaine, L. 2001. « Une étude comparative des articles de recherche écrits en anglais par des auteurs anglophones et francophones : une analyse textuelle des sections d'introduction en KDD (*Knowledge Discovery and Datamining*). ». Note de Recherche de DEA, Université de Victor Segalen Bordeaux 2.

Fries, Peter H. 1995. « Themes, methods of development, and texts ». In Ruqaiya Hasan & Peter H Fries (eds.), *On Subject and Theme: A Discourse Functional Perspective*. Amsterdam/Philadelphie : John Benjamins, 317-359.

Fries, Peter H. 1994. « On theme, rheme and discourse goals » In M. Coulthard (ed.), *Advances in Written Text Analysis*. Londres : Routledge, 229-249.

Ghadessy, Mohsen (ed.) 1995. *Thematic Development in English Text*. Londres : Pinter.

Halliday, M.A.K. 1994. *An Introduction to Functional Grammar*. 2nd Edition. Londres : Arnold.

Halliday, M.A.K. & R. Hasan. 1976. *Cohesion in English*. Malaisie : Pearson Education Limited.

Jacquemin, C. 1997. « Variation terminologique : reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus ». Mémoire d'Habilitation à diriger des recherches, Université de Nantes.

Kodratoff Y. 1999. « Knowledge discovery in texts: A definition and applications ». In Ras & Skowron (eds.), *Foundation of Intelligent Systems*, LNAI 1609. Springer.

Kodratoff, Yves & Edwin Diday. 1991. *Induction symbolique et numérique à partir de données*. Toulouse : Cépaduès-Édition.

Lemire, Gilles. 1996. *Grammaire française : vision systémique en linguistique*. <http://www.fse.ulaval.ca/fac/Grammaire-BEPP>.

McCabe, Anne. 1999. « Theme and thematic patterns in Spanish and English history Texts ». Doctoral Thesis, Aston University.

Martin, Jacky. 1997. « Du bon usage des corpus dans la recherche sur le discours spécifique ». *ASp* 15-18, 75-84.

Martin, J.R. 1992. *English Text: System and Structure*. Amsterdam/Philadelphie : John Benjamins.

Martin, J.R., & Guenter Plum. 1996. « A systemic functional perspective on genre: modelling genre ». The Texlinguistic Approach to Genre Colloquium - AAAL 1996, 1-18.

Martin, J.R., Christian Matthiessen, & Clare Painter. 1997. *Working with Functional Grammar*. Sydney : Arnold.

Matthiessen, Christian. 1995. « THEME as an enabling resource in ideational 'Knowledge' construction ». In Mohsen Ghadessy (ed.), *Thematic Development in English Texts*. Londres : Pinter Publishers, 20-54.

Mauranen, Anna. 1996. « Discourse competence – Evidence from thematic development in native and non-native texts ». In Ventola, Eija & Anna Mauranen (eds.), *Academic Writing. Intercultural and Textual Issues*. Amsterdam/Philadelphie : John Benjamins, 195-231.

Michalski, R.S. & R.E. Stepp. 1983. « Learning from observation: Conceptual clustering ». In Michalski, R.S., J.G. Carbonell & T.M. Mitchell (eds.), *Machine Learning: An artificial intelligence approach*. Morgan Kaufmann, 331-363.

Nédellec, C. 2000. « Corpus-based learning of semantic relations by the ILP system Asium ». In Cussens, James & Saso Dzerovski (eds.), *Learning Language in Logic*. Springer Verlag.

Paltridge, B. 1997. *Genre, Frames and Writing in Research Settings*. Amsterdam/Philadelphie : John Benjamins.

Smejrková, S. 1996. « Academic writing in Czech and English » In Ventola, Eija & Anna Mauranen (eds.), *Academic Writing. Intercultural and Textual Issues*. Amsterdam/Philadelphie : John Benjamins, 137-153.

Vassileva, I. 1998. « Who am I/who are we in academic writing ». *International Journal of Applied Linguistics* 8/2, 163-190.

Ventola, E. 1994. « From syntax to text: problems in producing scientific abstracts in L2 ». In Smejrková S. & F. Sticha (eds.), *The Syntax of Sentence and Text*. Amsterdam/Philadelphie : John Benjamins, 283-303.

Ventola, E. & A. Mauranen (eds.). 1996. *Academic Writing: Intercultural and Textual Issues*. Amsterdam/Philadelphie : John Benjamins.

Ventola, E. & A. Mauranen. 1991. « Non-native writing and native revising of scientific articles. » In Functional and systemic linguistics: approaches and uses, E. Ventola (ed), 457-490. Berlin : Mouton de Gruyter.

Weil, Henri. 1991. *De l'ordre des mots dans les langues anciennes comparées aux langues modernes*. Paris : Didier Érudition.

Yakhontova, Yatyana. 2001. « Textbooks, contexts and learners ». *English for Specific Purposes* 20/1, 397-415.

ANNEXES

Annexe 1 - Deux exemples des sections Introductions

[Corpus Francophone - F3] Basis of a Fuzzy Knowledge Discovery System

1 Introduction

Knowledge discovery has been defined as The non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [9].To realize a set of knowledge discovery tools, one has some major choices to achieve: the discovery process may be supervised or not, knowledge representation may use decision trees, association rules, neural networks, ... Having to design such tools, we have chosen, for simplicity reasons, a supervised process and a knowledge representation by rules. However, considering that rules using intervals on continuous attributes are often difficult to interpret and that strict thresholds are often too abrupt, we have decided to use fuzzy logics.

Fuzzy logics may be considered as extensions of multivalued logics, allowing usage of intermediate truth-values between false and true [18]. They allow expression of knowledge in a more natural way that classical binary logics, using graduated attributes as in X is rather high (for X is high is rather true). Fuzzy logics offer many logical operators [12], which permits a good expressiveness of various knowledge forms. In this paper, we detail the primary operations needed to extract fuzzy knowledge from a database.

First, usage of fuzzy logics in knowledge discovery needs to convert numerical attributes to their fuzzy representations; for this, it is necessary to define for each classical attribute, a mapping from its possible values to a set of truth-values for each fuzzy attribute. This mapping is often realized by a fuzzy partition, or rather by a fuzzy pseudo partition, and it is then possible to translate classical attributes by valuations on their fuzzy correspondents. These operations are called fuzzification.

To extract rules from a database, one needs to evaluate each possible rule in order to establish which rules must be kept; for this purpose, many indexes are available, of which we have only retained three indexes: the confidence of a rule, its support and a less usual index, called the intensity of implication. After recalling the principles of these indexes, we expose how they can be evaluated in fuzzy logics.

It is then possible to use a knowledge extraction algorithm using the same principles as in classical logics. Our algorithm is an exploratory search in a tree of possible rules, with evaluation of each rule. Fuzzy logics also allow specific methods, based on genetic algorithms, to search the most representative set of weights for a general set of fuzzy rules, but we will not consider this possibility here.

Several evaluations of the fuzzy logical operators are possible. If one only wants to extract rules for a human expert, the nature of the operators does not matter, but if these rules are to be processed by an expert system, a choice of fuzzy operators is necessary.To find the most adequate set of fuzzy operators, we expose a justified restriction to only four possible sets of fuzzy operators and we give a method to find amongst them, the more consistent with a database.

To reduce the huge set of rules that can then be extracted, one may want to use classical reduction schemes. We show that to be valid in a fuzzy logic, classical reduction schemes need specific choices of fuzzy operators. We conclude by recalling the interest to use fuzzy attributes instead of numerical intervals for continuous attributes in the database, and by considering some possible improvements of the systems we have described.

[Corpus Anglophone - E14] Learning Limited Dependence Bayesian Classifiers

Introduction

Recently, work in Bayesian methods for classification has grown enormously (Cooper & Herskovits 1992) (Buntine 1994). Bayesian networks (Pearl 1988) have long been a popular medium for graphically representing the probabilistic dependencies which exist in a domain. It has only been in the past few years, however, that this framework has been employed with the goal of automatically learning the graphical structure of such a network from a store of data (Cooper & Herskovits 1992) (Heckerman, Geiger, & Chickering 1995). In this latter incarnation, such models lend themselves to better understanding of the domain in which they are employed by helping identify dependencies that exist between features in a database as well as being useful for classification tasks. A particularly restrictive model, the Naive Bayesian classifier (Good 1965), has had a longer history as a simple, yet powerful classification technique. The computational efficiency of this classifier has made it the benefactor of a number of research efforts (Kononenko 1991).

Although general Bayesian network learning as well as the Naive Bayesian classifier have both shown success in different domains, each has it shortcomings. Learning in the

domain of unrestricted Bayesian net-works is often very time consuming and quickly becomes intractable as the number of features in a domain grows. Moreover, inference in such unrestricted models has been shown to be NP-hard (Cooper 1987). Alternatively, the Naive Bayesian classifier, while very efficient for inference, makes very strong independence assumptions that are often violated in practice and can lead to poor predictive generalization. In this work, we seek to identify the limitations of each of these methods, and show how they represent two extremes along a spectrum of data classification algorithms.

Annexe 2 - Exemples du corpus des quatre types de progressions thématiques

• La combinaison CL

Constant	We	present a framework within which these tasks have a natural expression.
linear	This framework	modifies similarities of the tasks and highlights significant differences.

	T1	→	R1	
	↓			
(Constante)	**T2**	→	**R2**	
			↓	
(Linéaire)			**T3**	→ R3

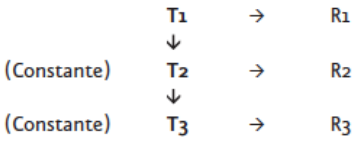
• La combinaison LC

linear	These considerations	led to our adoption of ROC analysis.
Constant	Asymmetric misclassification costs and highly imbalanced classes	often arise in Knowledge Discovery and Data Mining (KDD) and Machine Learning (ML)

	T1	→	**R1**	
			↓	
(Linéaire)			**T2**	→ R2
			↓	
(Constante)			**T3**	→ R3

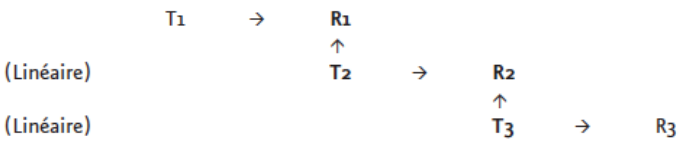
• La combinaison CC

constant	The discovered substructures	allow abstraction over detailed structure in the original data.
Constant	Iteration of the substructure discovery and replacement process	constructs a hierarchical description of the structural data in terms of the discovered substructures.



• La combinaison LL

linear	The result	is a refined set of clusters.
linear	Each cluster	is represented as a set of named entities and corresponds to an ongoing topic in the corpus.



Annexe 3 - Le texte de l'exemple [9]

1. Provost and Fawcett [9] have argued persuasively that accuracy is often not an appropriate measure of classifier performance.

2. This is certainly apparent in classification problems with heavily imbalanced classes (one class occurs much more often than the other).

3. It is also apparent when there are asymmetric misclassification costs (the cost of misclassifying an example from one class is much larger than the cost of misclassifying an example from the other class).

4. Class imbalance and asymmetric misclassification costs are related to one another.

5. One way to correct for imbalance is to train a cost sensitive classifier with the misclassification cost of the minority class greater than that of the majority class, and one way to make an algorithm cost sensitive is to intentionally imbalance the training set.

6. As an alternative to accuracy, Provost and Fawcett advocate the use of ROC analysis, which measures classifier performance over the full range of possible costs and class frequencies.

7. They also proposed the convex hull as a way to determine the best classifier for a particular combination of costs and class frequencies.

Annexe 4 - Le texte de l'exemple [10]

1. Several data mining tasks (e.g., association rule mining [1]) are based on the evaluation of frequency queries to determine how often a particular pattern occurs in a large data set.

2. We consider the problem of frequency query evaluation, when patterns are itemsets, in dense data sets1 like, for instance in the context of census data analysis [4] or log analysis [8].

3. In these important but difficult cases, there is a combinatorial explosion of the number of frequent itemsets and computing the frequency of all of them turns out to be intractable.

4. In this paper, we present an efficient technique to approximate closely the result of the frequency queries, and formalize it within the E-adequate representation framework [10].

5. Intuitively an E-adequate representation is a representation of data that can be substituted to another representation to answer the same kind of queries, but eventually with some lost of precision (bound by the E parameter).

6. First evidences of the practical interest of such representations has been given in [10,5].

NOTES

1. Un autre congrès KDD prend de plus en plus d'importance : ICDM, the IEEE International Conference on Data Mining. Ce congrès n'avait pas encore eu lieu (novembre 2001) au moment de cette étude, il n'a donc pas pu être pris en compte.

2. EGC2001 « avait pour objectif de rassembler d'une part une communauté académique multidisciplinaire (systèmes d'information et bases de données, apprentissage automatique, ingénierie des connaissances, statistiques et analyse de données...) et d'autre part des spécialistes de l'entreprise autour de la double thématique de l'Extraction des Connaissances dans les bases de Données ECD (KDD : *Knowledge Discovery in Databases*) et de la 'Gestion des Connaissances' GC (KM : *Knowledge Management*) » <http://www.sciences.univ-nantes.fr/irin/EGC2001>.

3. Voir Fontaine (2001) pour une description détaillée des deux corpus.

4. Voir Fontaine (2001) pour plus d'informations sur la construction du corpus et la validité d'une étude des textes publiés dans les actes du congrès.

5. Voir Carter-Thomas (2000) pour une présentation de divers points de vue sur la fonction textuelle et une discussion complète du Thème et du Rhème.

6. La limite entre Thème et Rhème n'est pas parfaitement standardisée en LSF, car certains linguistes préfèrent placer cette limite au point de l'élément fini dans les propositions déclaratives, comme, par exemple, le fait Mauranen (1996).

7. La présentation des notions de Thème et Rhème donnée ici est brève. Pour une très bonne introduction à l'analyse en LSF, voir Bloor et Bloor (1995), et pour une très bonne discussion des approches différentes au Thème, voir Carter-Thomas (2000).

8. Le terme en anglais est « *Topical Theme* » qui ne se traduit pas facilement en français, nous préférons donc le terme « Thème expérientiel » qui respecte le parallèle des trois fonctions propositionnelles.

9. Il existe des recherches assez récentes qui développent davantage ces trois modèles, notamment le travail de Duszak (1994).

10. Si nous avions suivi une autre division entre Thème-Rhème, comme par exemple celle de Mauranen (1996), le sujet de la proposition 13 aurait été inclus dans le Thème et on aurait eu une progression linéaire. La proposition 13 reste problématique néanmoins car, même dans le cadre de Mauranen, « *For temporal transactions* » ne sert pas de thème d'orientation (*orienting theme*).

11. Il est remarquable que ces notions soient encore valides en Analyse des Données, voir par exemple Kodratoff et Diday (1991). On dit maintenant 'intention' au lieu de 'compréhension' et 'extension' au lieu de 'étendue'.

12. En d'autres termes, nous utilisons une version parallélisée de l'algorithme des nuées dynamiques de Diday (Diday *et al.* 1982) ou de l'algorithme AQ de Michalski (Michalski & Stepp 1983).

13. 'DATA-MINING' est un terme que nous avons repéré dans les textes, c'est pourquoi il est écrit en majuscules. Nous avons nommé « *KnownMethods* » le concept dont il est l'instance. La relation grammaticale (soulignée) *to discover* (verbe) *patterns* (complément d'objet direct) est une instance du concept de « sorties de l'algorithme proposé par l'auteur », appelé « OUTPUT ».

14. De même, discovered substructures est une structure grammaticale indiquant la présence du concept « OUTPUT », *allow abstraction* indique celle du concept « ABILITY », *in original data* celle du concept « INPUT ».

15. Voir Fontaine (2001) pour la description complète du registre.

16. Nous sommes bien conscients qu'il s'agit d'une simplification, comme le montre le succès actuel des recherches en construction d'« ontologies ». La structure la plus probable est celle d'une « pyramide » c'est-à-dire une sorte de pseudo-taxinomie dans laquelle deux concepts peuvent partager certaines instances. Notre travail présent consiste à examiner en quoi une structure strictement taxinomique peut être utile, avant de considérer l'hypothèse plus complexe d'une pyramide, puis d'une ontologie.

17. Traduit de l'anglais par les auteurs.

18. Traduit de l'anglais par les auteurs.

19. Nous admettons que le fait d'être rédacteur en langue maternelle n'implique pas obligatoirement que le texte soit bien écrit et donc ces taux rendent compte aussi d'un certain nombre de liens non réussis.

RÉSUMÉS

Cet article met en évidence deux types de structures textuelles identifiables dans les articles de recherche écrits en anglais. La première, la progression thématique, est analysée dans un cadre de linguistique textuelle. Ensuite nous comparons ces résultats avec ceux d'un deuxième type de texture : celle des liens entre concepts identifiés par une méthode de « fouille de textes ». Nos conclusions soulignent la nature des difficultés textuelles qu'éprouvent les auteurs n'ayant pas l'anglais comme langue maternelle. Ce travail propose aussi un travail futur en vue de créer un outil d'aide à la rédaction pour les non anglophones s'exprimant en langue anglaise.

This paper explores two types of text structure in scientific research articles written by native and non-native writers of English. We use a text linguistics analysis to study thematic progression in these texts. We then compare these results to the study of a new type of texture found in texts: that is, the texture formed by concept identification using a method from Text Mining. Our conclusions point out some of the specific difficulties that non-native writers face in managing the structure of their texts. We also look toward developing a semi-automatic aide for non-native writers.

INDEX

Mots-clés : anglais scientifique, construction d'ontologies, identification de concepts, linguistique de texte, progression thématique, rhème, texture, thème
Keywords : concept identification, ontology building, rheme, scientific English, textlinguistics, texture, thematic progression, theme

AUTEURS

LISE FONTAINE

Lise Fontaine est lectrice d'anglais à l'Université Paris-Dauphine. Elle prépare un doctorat à l'Université de Bretagne Occidentale dans l'équipe ERLA, Université de Bretagne Occidentale à Brest, sous la direction de David Banks. lfontaine@Cardiff.ac.uk

YVES KODRATOFF

Yves Kodratoff est directeur de recherches au CNRS au Laboratoire de Recherche en Informatique de l'Université Paris-Sud Orsay (LRI), où il a créé en 1980 l'équipe de recherche « Inférence et Apprentissage » spécialisée dans la simulation du raisonnement inductif. Il a publié plus de cent articles et chapitres de livres sur ce sujet au cours de sa carrière et dirigé une cinquantaine de thèses de doctorat. Il travaille maintenant à la création d'outils inductifs pour la découverte de structures au sein de masses de textes (« fouille de textes »). Yves.Kodratoff@lri.fr