Catégorisation de textes en domaines et genres

Complémentarité des indexations lexicale et morphosyntaxique

Céline Poudat, Guillaume Cleuziou, Viviane Clavier

Dans Document numérique 2006/1 Vol. 9 , pages 61 à 76
Éditions JLE

ISSN 1279-5127

Catégorisation de textes en domaines et genres

Complémentarité des indexations lexicale et morphosyntaxique

Céline Poudat* — **Guillaume Cleuziou**** — **Viviane Clavier*****

* *Centre Orléanais de Recherche en Anthropologie et Linguistique (CORAL)*
Université d'Orléans - F-45072 Orléans cedex 2

** *Laboratoire d'Informatique Fondamentale d'Orléans (LIFO)*
Université d'Orléans - F-45067 Orléans cedex 2

*** *Groupe de Recherche sur les Enjeux de la Communication (GRESEC)*
Institut de la Communication et des Médias, 11, av. du 8 mai 1945
Université Stendhal Grenoble III - F-38130 Echirolles

{celine.poudat,guillaume.cleuziou}@univ-orleans.fr , viviane.clavier@u-grenoble3.fr

RÉSUMÉ. *Cet article traite du choix de descripteurs linguistiques appropriés pour caractériser et classifier les textes. On considère généralement que les domaines sont corrélés au niveau du contenu (mots, termes, etc.) tandis que les genres sont discriminés au niveau morphosyntaxique. Malgré les bons résultats obtenus par ces choix méthodologiques, peu de travaux ont cherché à mesurer l'impact et la complémentarité des deux niveaux de description pour la classification. Cette étude vise ainsi à évaluer l'intérêt discriminant des descripteurs morphosyntaxiques et thématiques pour classer les genres et les domaines. Des résultats encourageants sont obtenus sur un corpus pilote de textes scientifiques français.*

ABSTRACT. *This paper deals with the selection of appropriate descriptors to characterize and classify texts. In most classification tasks, domains are generally correlated to the content level (words, terms, bags of words, etc.) and genres to the morphosyntactic one. However, few studies have assessed the impact and the complementarity of the two description levels on classification. The present study aims at evaluating the discriminant interest of the lexical and morphosyntactic linguistic levels in the field of genre and domain classification. Encouraging results are obtained on a French scientific corpus, which has been built in that perspective.*

MOTS-CLÉS : *recherche d'information, genre, domaine, classification, lexique, morphosyntaxe.*

KEYWORDS: *information retrieval, genre, domain, classification, morphosyntax.*

Document numérique. Volume 9 – n° 1/2006, pages 61 à 76

1. Introduction

Toute entreprise de classification nécessite un ensemble approprié de descripteurs. Il en va ainsi en matière de classification textuelle : au même titre qu'il serait peu pertinent de proposer des descripteurs tels que « taille » ou « couleur des yeux » pour caractériser les profils financiers d'utilisateurs de comptes bancaires, il serait inapproprié de décrire les textes scientifiques à partir de variables certes discriminantes en matière de typologie textuelle littéraire mais fondamentalement absentes du discours scientifique, *e.g.* « nombre de marques de dialogue », ou « nombre de verbes conjugués au passé simple ».

Les classifications textuelles en domaines et en genres, qui représentent un enjeu pour la Recherche d'Information (RI), nécessitent de même un ensemble de descripteurs linguistiques adéquats. Dans les faits, domaines et genres sont associés à des niveaux linguistiques différents. Quand il s'agit de classification thématique ou domaniale, les textes sont souvent réduits à l'état de « sacs de mots ». Chaque document est alors décrit par le vocabulaire présent dans le corpus. Étant donné la taille de ce vocabulaire, une étape de réduction de l'espace de description est généralement effectuée[1] : sélection d'attributs par des mesures d'intérêt, reparamétrage de l'espace ou regroupement d'attributs. Ces formalismes d'indexation permettent d'obtenir des classifieurs performants, atteignant jusqu'à 90 % de précision sur grands corpus (Hofmann, 1999 ; Dhillon *et al.,* 2003). De la même manière, les classifications en genres à partir d'un jeu de variables morphosyntaxiques robuste sont à même d'obtenir de très bons résultats en matière de validation de typologies textuelles (Karlgren *et al.,* 1994 ; Kessler *et al.,* 1997 ; Malrieu *et al.,* 2001).

On notera toutefois que la plupart des travaux recensés effectuent de la classification domaniale sur corpus génériquement homogènes (*e.g.* Reuters ou Newsgroup), et de la classification générique sur corpus discursivement[2] hétérogènes (*e.g.* (Karlgren *et al.,* 1994 ; Kessler *et al.,* 1997 ; Malrieu *et al.,* 2001)), ce qui augmente le pouvoir classificatoire des variables employées mais limite l'utilisation conjointe et l'évaluation de la portée des deux niveaux descriptifs. Bien que de nombreuses applications de RI partent de données génériquement hétérogènes mais de même domaine, ce type de classification demeure problématique et est rarement mené étant donné la robustesse des jeux de variables utilisés.

Nous nous proposons d'évaluer l'impact des variables thématiques et morphosyntaxiques sur les classifications génériques et domaniales à partir d'un corpus pilote de taille restreinte développé à cet effet. Nous déterminerons ainsi les variables les plus discriminantes selon les typologies et apprécierons l'intérêt d'une utilisation conjointe des deux niveaux descriptifs.

1. Pour davantage de détails sur cette étape de réduction, se reporter à (Sebastiani, 2002).
2. Discours littéraire, juridique, scientifique, journalistique, etc. Les types de discours sont reliés à des pratiques sociales distinctes et organisent en leur sein les typologies génériques et domaniales. Le discours juridique inclut ainsi les genres de l'arrêt, du décret, de la loi, etc.

Après un bref rappel des notions de domaines et de genres en RI, nous reviendrons sur les relations entre les deux concepts en section 2. La partie 3 présente la méthodologie que nous avons développée pour réaliser l'expérience, de même que le corpus test utilisé. Enfin, les sections 4 et 5 sont dédiées aux aspects expérimentaux de cette évaluation et à l'analyse des résultats obtenus.

2. Genres et domaines

Bien que les notions de genres et de domaines soient de plus en plus exploitées en RI, elles sont rarement utilisées conjointement, dans la mesure où elles sont généralement associées à des variables ou traits appartenant à des niveaux linguistiques différents. Les domaines se situeraient sur le plan lexical, tandis que les genres, ou les styles, seraient déterminés au niveau morphosyntaxique.

Ainsi, les domaines sont souvent décrits en termes de relations lexicales, dans la mesure où ils sont supposés être le reflet de champs de connaissance particuliers. Ils se positionnent donc sur le plan du contenu, que différentes techniques de classification de documents ont tenté d'appréhender. Les mesures les plus fréquentes sont calculées sur les mots, les clusters de mots - inégalement appelés « thèmes », « sujets », « topics », etc. - ou encore les racines (ou word stems) (Porter, 1980), et se sont avérées plutôt efficaces dans diverses entreprises. De manière générale, on demeure au niveau du mot en raison de son faible coût de traitement.

La notion de genre[3], philologique et littéraire au départ, est de plus en plus employée en RI et catégorisation textuelle (Prime-Claverie *et al.,* 2002 ; Crowston *et al.,* 2004). En effet, le genre possède des propriétés linguistiques formelles qui permettent de l'identifier et de le différencier : certains marqueurs sont ainsi absents de certains genres, comme les points d'exclamation dans les textes juridiques. De manière générale, on recourt aux parties du discours, de même qu'à des catégories fonctionnelles pour discriminer et décrire les genres. A la suite de Biber (Biber, 1988), c'est l'utilisation de variables morphosyntaxiques qui a été privilégiée pour valider des typologies textuelles et identifier les genres.

On considère généralement que les genres et les domaines sont des notions orthogonales. Il est souvent souligné qu'on peut retrouver des domaines distincts à l'intérieur de genres différents, et inversement, ce qui laisserait penser que les deux dimensions sont indépendantes. Les deux niveaux de caractérisation des notions sont par conséquent rarement utilisés de manière conjointe ; certaines études (*e.g.* (Poudat *et al.,* 2003 ; Lee *et al.,* 2002)) ont pourtant corrélé des variables lexicales aux genres et ont obtenu des résultats tout à fait encourageants. La classification des domaines à partir du niveau morphosyntaxique reste encore, à notre connaissance, en suspens. Pourtant, il semble qu'à l'instar des genres, les domaines sont susceptibles d'entraîner des régularités stylistiques.

3. Ou « style », « registre », voire « type de textes ».

Prenons par exemple le cas du discours scientifique : la pratique sociale de la « communication scientifique » a donné lieu à la création d'un ensemble de genres tant écrits qu'oraux (articles, actes, présentations de conférence, etc.), dans laquelle on retrouve des « domaines » correspondant aux différentes aires de l'activité scientifique (médecine, économie, recherche d'information, informatique, etc.). L'ensemble des productions de cette pratique communicative, qui partagent des propriétés linguistiques communes, forme ce que l'on appelle le « discours scientifique ». Si les genres ont développé au sein de cette pratique une structure et un style propre qui permettent de les identifier par-delà les domaines - on reconnaîtra un article scientifique, qu'il porte sur le domaine médical, biologique ou informatique -, il paraît raisonnable d'émettre l'hypothèse que les domaines peuvent être discriminés au moyen de variables morphosyntaxiques.

Notre objectif étant d'évaluer l'intérêt des niveaux morphosyntaxiques et thématiques en matière de classifications en genre et en domaine, il nous semble toutefois primordial d'initier cette entreprise sur un corpus textuel discursivement homogène[4], quitte à étendre l'étude à un corpus plus large et plus hétérogène dans une étape ultérieure.

3. Méthodologie

Les notions de domaines et de genres intéressant spécifiquement le discours scientifique et les applications qui s'y attachent (veille scientifique et technique, recherches documentaires, etc.), c'est sur un corpus de textes scientifiques français que nous avons travaillé. Comme les textes scientifiques sont soumis à de fortes contraintes rédactionnelles qui limitent leur(s) variation(s), ils possèdent des propriétés de genre plus stables qui conviennent particulièrement à notre entreprise.

3.1. *Sélection de descripteurs adéquats*

Parmi les variables lexicales envisageables, ce sont les substantifs que nous avons sélectionnés. En effet, les noms sont des parties du discours non vides davantage susceptibles de pointer sur des concepts scientifiques, que les adverbes, verbes ou adjectifs. Ils sont donc potentiellement plus discriminants et peuvent aisément être extraits. Le poids des substantifs au singulier et au pluriel (dans la mesure où ils peuvent renvoyer à des concepts différents, *e.g.* « la langue » en linguistique ne renvoie pas à la même notion que « les langues ») a également été pris en compte.

Dans un deuxième temps, nous avons sélectionné 136 variables morphosyntaxiques dédiées au discours scientifique : il serait en effet peu pertinent de décrire

4. Il semblerait en effet que les types de discours sont les premiers à émerger au niveau morphosyntaxique, bien avant les genres, les domaines ou les styles personnels (Malrieu *et al.,* 2001). Étant donné que nous nous intéressons aux notions de genres et de domaines dans la présente étude, il semble pertinent d'évacuer momentanément le problème des discours.

les textes scientifiques à partir de descripteurs trop généraux ou non caractéristiques qui n'incluent pas ses traits spécifiques. Outre les parties du discours traditionnelles (noms, adjectifs, verbes, adverbes prépositions, etc.), nous avons donc retenu un ensemble de traits « caractéristiques » du discours scientifique dans la littérature existante (tableau 1).

Variable	**Description**
ABR	Abréviations
CON (+ attributs)	Connecteurs : addition, cause, conséquence, conclusion, exemplification, disjunction, opposition, reformulation, espace, temps, etc.
FGW	Eléments étrangers (non français)
NUM (+ attributs)	Numéraux : dates, cardinaux, ordinaux + références dans le text (*e.g.* « Voir en 1.2 »)
LS	Indices de structuration (titres et listes)
PON (+ attributs)	Ponctuation : deux points, crochets, guillemets, parenthèses, slashs, etc.
VER :mod :[temps]	Modaux
SIG	Acronymes
SYM	Symboles

Tableau 1. *Principaux traits caractéristiques du discours scientifique*

3.2. *Développement et prétraitement du corpus pilote*

Nous avons été contraints d'exclure les corpus de référence traditionnels comme Reuters ou Newsgroup en raison de leur homogénéité générique et avons été développé un corpus pilote adapté à notre problématique.

Le corpus est de taille restreinte : il contient au total 371 textes scientifiques français publiés autour de 2000. Trois genres (articles, présentations de revue et comptes rendus) et deux domaines différents (linguistique et mécanique) y sont représentés. La répartition des documents de ce corpus est présentée dans le tableau 2.

Les spécificités des expérimentations présentées *infra* nous ont amené à effectuer différentes partitions du corpus correspondant à des tâches de classification distinctes :

– *ART-corpus* correspond au sous-corpus constitué uniquement des textes de genre « article » (1ère ligne du tableau 2),

– *LING-corpus* correspond au sous-corpus constitué uniquement des textes de domaine « linguistique » (1ère colonne du tableau 2).

L'étiquetage a été réalisé à partir des textes bruts *via* un processus incrémental d'apprentissage avec le tagger TnT (Trigrams'n'Tags) (Brants, 2000) sur le jeu d'étiquettes sélectionné.

	Linguistique	Mécanique
Articles	224	49
Présentations de revues	45	
Comptes rendus	53	

Tableau 2. *Présentation du corpus utilisé*

3.3. *Classifieurs utilisés*

La classification (ou catégorisation) automatique de documents a donné lieu à de nombreux travaux recourant aux méthodes d'apprentissage automatique. Les techniques les plus utilisées dans ce domaine d'application sont : le classifieur naïf de Bayes (Lewis *et al.,* 1994), les machines à support vectoriel (SVM) (Joachims, 1998) ou encore les arbres de décisions (Cohen *et al.,* 1998).

Les expérimentations que nous proposons par la suite visent à (1) évaluer l'influence de chaque type de description sur la classification (précision du classifieur) et (2) observer l'articulation des deux ensembles d'attributs combinés dans un même classifieur. Dans cette perspective, nous utilisons deux méthodes très différentes mais complémentaires de ce point de vue, à savoir la classification par SVM et par arbres de décision.

Les SVMs sont reconnus pour leurs performances inégalées dans l'application à la catégorisation de textes (Dumais *et al.,* 1998). De manière simplifiée, cette méthode consiste à apprendre un classifieur dans un nouvel espace d'attributs de dimension plus importante que l'espace initial. Ce nouvel espace peut être obtenu par différents types de fonctions noyaux (*e.g.* linéaire, polynomial, RBF, etc.)[5]. Plusieurs études empiriques (*e.g* (Dumais, 1998)) ayant montré que les meilleures performances en classification de textes sont obtenues avec des SVMs linéaires, c'est ce type de noyau que nous avons retenu dans nos expérimentations. La classification par SVMs permettra alors d'appréhender quantitativement l'importance de chaque ensemble d'attributs : lexical, morphosyntaxique et combiné, notés respectivement $\mathcal{L}$, $\mathcal{M}$ et $\mathcal{L} \oplus \mathcal{M}$.

Les Arbres de Décision (ADs), contrairement aux SVMs, procèdent par apprentissage symbolique. Bien que moins performants sur cette application, les arbres générés par cette méthode permettent l'analyse et l'interprétation du rôle joué par chaque attribut. La présence et la position d'un attribut dans l'arbre indique son importance dans le processus de classification ainsi que la classe favorisée par ce dernier. De l'arbe peut être extrait un ensemble de règles explicatives « caractérisant » les classes ciblées. Dans nos expérimentations, nous utiliserons l'algorithme C4.5 (Quinlan, 1993).

5. Pour plus de précisions sur la technique d'apprentissage par SVM, se reporter à (Vapnik, 1995).

3.4. *L'évaluation*

Afin de mesurer l'impact des différents jeux de variables sur les classifications en genre et esn domaine, il est nécessaire d'observer l'influence de chacun des trois ensembles d'attributs ($\mathcal{L}$, $\mathcal{M}$ et $\mathcal{L} \oplus \mathcal{M}$) sur le corpus global et les corpus locaux[6].

Soient $\mathcal{D}$ un ensemble de textes scientifiques et $\mathcal{C}$ un ensemble de classes (genre ou domaine selon l'étude) tels que chaque texte $d_i \in \mathcal{D}$ est associé à une unique classe $c(d_i) \in \mathcal{C}$. $\mathcal{D}$ est divisé en deux sous-ensembles, d'entraînement et de test, notés respectivement $\mathcal{D}_{train}$ et $\mathcal{D}_{test}$.

On note $\mathcal{L}_{\mathcal{D}} = \{l_1, \ldots, l_{|\mathcal{L}|}\}$ l'ensemble ordonné des substantifs (singuliers et pluriels) apparaissant dans les textes de $\mathcal{D}_{train}$ (description lexicale). Les substantifs sont ordonnés par Information Mutuelle (IM) décroissante. Soit $\mathcal{C}$ la variable de classe et l_i une variable lexicale de $\mathcal{L}$:

$$IM(l_i, \mathcal{C}) = \sum_{c_j \in \mathcal{C}} P(c_j).\log \frac{P(l_i|c_j)}{P(l_i)} \quad [1]$$

On note $\mathcal{M} = \{m_1, \ldots, m_{136}\}$ l'ensemble ordonné des 136 attributs morpho-syntaxiques décrit en section 3.1. On utilise le Gain d'Information (GI) pour mesurer l'intérêt de chaque attribut pour la classification cible et ainsi ordonner $\mathcal{M}$:

$$\forall m_i \in \mathcal{M},\ GI(m_i, \mathcal{C}) = \max_s \{E_{\mathcal{C}}(\mathcal{D}_{train}) - P(\{m_i < s\}).E_{\mathcal{C}}(\mathcal{D}_{train}^{m_i<s}) - P(\{m_i \geq s\}).E_{\mathcal{C}}(\mathcal{D}_{train}^{m_i \geq s})\} \quad [2]$$

Les attributs dans $\mathcal{M}$ sont continus (*e.g. % prépositions*) ; ils sont alors discrétisés de façon analogue à l'algorithme C4.5 (Quinlan, 1993). Ainsi dans [2], les valeurs de s correspondent aux différents seuils de discrétisation possibles pour les valeurs de m_i, $\mathcal{D}_{train}^{m_i<s}$ et $\mathcal{D}_{train}^{m_i \geq s}$ aux sous-ensembles de documents induits par cette discrétisation. Enfin, E désigne la fonction « entropie », $E_{\mathcal{C}}(X)$ mesurant la pureté d'un ensemble X étant donné une schéma de classification attendu $\mathcal{C}$:

$$E_{\mathcal{C}}(X) = -\sum_{c_j \in \mathcal{C}} P_X(c_j).\log_2 P_X(c_j) \ \ avec\ \ P_X(c_j) = \frac{\#\{x_i \in X | c(x_i) = c_j\}}{\#X} \quad [3]$$

Rappelons que dans [3], $c(x_i)$ indique la classe associée à l'élément x_i.

Enfin, $\mathcal{L} \oplus \mathcal{M}$ correspond à une fusion ordonnée des deux ensembles d'attributs $\mathcal{L}$ et $\mathcal{M}$, suivant l'ordre d'alternance suivant : $\mathcal{L} \oplus \mathcal{M} = \{l_1, m_1, l_2, m_2, \ldots, l_{136}, m_{136}, l_{137}, l_{138}, \ldots, l_{|\mathcal{L}|}\}$.

6. Corpus homogènes en genre (*ART-corpus*) ou en domaine (*LING-corpus*).

Les expérimentations présentées en section 4 correspondent à des résultats moyens obtenus sur 5 validations croisées à 2 blocs (*2-fold cross-validations*) : $\mathcal{D}$ est divisé en deux sous-ensembles de tailles équivalentes, chaque sous-ensemble étant utilisé à son tour comme corpus d'entraînement et de test. Les valeurs reportées correspondent à des micro-précisions[7].

Concernant l'apprentissage par SVM, dans le cas de problèmes multiclasses plusieurs SVMs sont appris (un par classe) puis combinés.

4. Expérimentations

Nous considérerons, dans ce qui suit, plusieurs sous-corpus correspondant chacun à une tâche différente de classification. En premier lieu, les expérimentations présentées porteront sur une classification en domaines.

Sur le corpus local (*ART-corpus*), la classification consistera à distinguer les deux domaines « linguistique » et « mécanique » pour un ensemble de documents homogène en genre (uniquement des articles). Le corpus « global » permettra en revanche d'appréhender l'introduction d'un paramètre de variation générique (articles, présentations et compte rendus).

De façon analogue, dans un second temps, la classification en genre sera expérimentée sur un corpus « local » homogène en domaine (*LING-corpus*) puis sur le corpus « global » faisant intervenir une variation générique au sein des domaines.

4.1. *Classification en domaines*

Les résultats obtenus avec la méthode SVM (figures 1 et 2) montrent clairement et contre toute attente que les variables morphosyntaxiques sont plus discriminantes que les variables lexicales. De plus, on note qu'une utilisation conjointe des deux types de variables est globalement plus efficace que chacun des deux ensembles choisi séparément.

On obtient donc l'ordre de précédence suivant, avec ou sans variations génériques :

$$\textbf{\{Indexation-}\mathcal{L} \oplus \mathcal{M}\textbf{\}} > \textbf{\{Indexation-}\mathcal{M}\textbf{\}} > \textbf{\{Indexation-}\mathcal{L}\textbf{\}}$$

D'autres tests, effectués avec C4.5 indiquent les mêmes tendances, bien que les taux de précision obtenus par les ADs soient moins bons qu'avec la méthode SVM. L'indexation par le lexique semble également moins pertinente qu'une indexation morphosyntaxique ou mixte. Il semble donc que les domaines scientifiques se distinguent davantage par des traits stylométriques que par des informations lexicales,

7. La *micro-précision* mesure la proportion de textes classés correctement, quelle que soit la classe. *A contrario*, la *macro-précision* mesure pour chaque classe séparément la proportion de textes bien classés avant d'effectuer la moyenne.

constat surprenant si l'on considère que les deux domaines à discriminer (linguistique et mécanique) sont conceptuellement très éloignés.

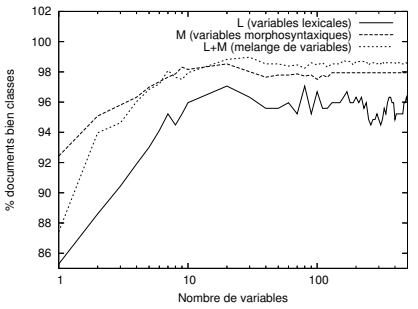

Figure 1. *Classification en domaines avec SVM sur corpus local (ART-Corpus)*

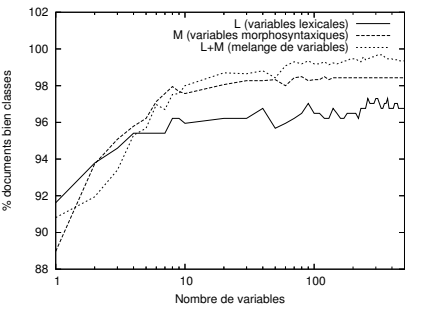

Figure 2. *Classification en domaines avec SVM sur corpus global*

4.2. *Classification en genres*

Les résultats obtenus avec le classifieur SVM (figures 3 et 4) confirmeraient l'hypothèse selon laquelle les genres sont effectivement corrélés au niveau morphosyntaxique : le taux de précision obtenu est plus élevé avec les jeux de variables comprenant des attributs morphosyntaxiques qu'avec les variables lexicales uniquement. Notons que les différences de domaines ne perturbent pas cet ordre.

$$\textbf{\{Indexation-}\mathcal{L} \oplus \mathcal{M}\textbf{\}} \approx \textbf{\{Indexation-}\mathcal{M}\textbf{\}} \gg \textbf{\{Indexation-}\mathcal{L}\textbf{\}}$$

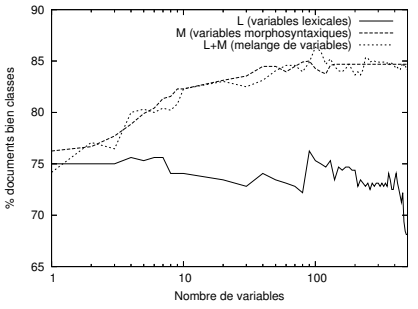

Figure 3. *Classification en genres avec SVM sur corpus local (LING-Corpus)*

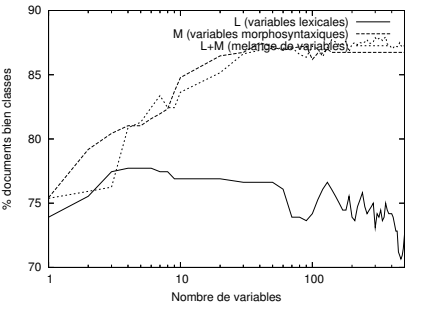

Figure 4. *Classification en genres avec SVM sur corpus global*

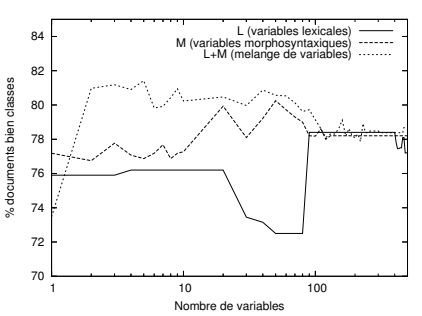

Figure 5. *Classification en genre avec AD sur corpus local (LING-Corpus)*

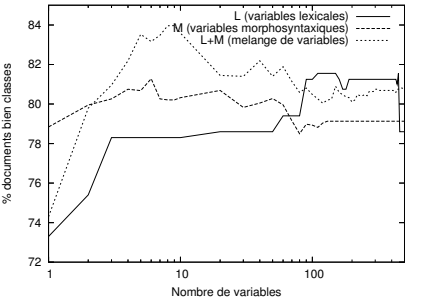

Figure 6. *Classification en genre avec AD sur corpus global*

Nous présentons en figures 5 et 6 les résultats obtenus avec C4.5. On observe en premier lieu que les taux de précision obtenus avec cette méthode sont encore une fois sensiblement inférieurs aux résultats obtenus avec la méthode SVM : 84 % au mieux avec C4.5 contre 88 % avec SVM. De plus, l'ordre de établi précédemment ($\mathcal{L} \oplus \mathcal{M} \approx \mathcal{M} \gg \mathcal{L}$) diffère avec cette nouvelle approche : les variables lexicales, combinées aux attributs morphosyntaxiques forment un jeu de descripteurs plus efficace au niveau global, ce qui confirmerait l'existence d'une possible corrélation des genres avec le niveau lexical, hypothèse soutenue par Lee et Myaeng (Lee *et al.,* 2002), qui associent des traits lexicaux au genre de la « homepage » :

$$\textbf{\{Indexation-}\mathcal{L} \oplus \mathcal{M}\textbf{\}} > \textbf{\{Indexation-}\mathcal{M}\textbf{\}} \gg \textbf{\{Indexation-}\mathcal{L}\textbf{\}}$$

D'un point de vue plus technique, ces différences obtenues entre les deux classifieurs[8] peuvent en partie s'expliquer par les méthodes très différentes auxquelles ces deux classifieurs font appel. Notamment, l'approche SVM considère un nouvel espace de représentation des documents, à forte dimensionalité, et dont les dimensions sont définies par combinaisons - ici linéaires - des descripteurs initiaux. Cette méthode fait donc intervenir de façon plus ou moins marquée l'ensemble des descripteurs considérés tandis que la construction d'un arbre de décision nécessite généralement très peu de descripteurs mais bien choisis.

4.3. *Analyse complémentaire : micro* vs *macro-précision*

Avant de fournir une explication plus précise des résultats précédents nous proposons un résultat intermédiaire synthétisant l'ensemble des expérimentations présentées ci-dessus. Pour un nombre fixé de descripteurs, nous étudions dans le tableau 3, les macro et micro-précisions induites par les arbres de décisions appris sur le corpus glo-

8. Cette observation confirme l'importance d'utiliser plusieurs méthodes de classification, utilisant des approches différentes d'apprentissage.

bal. Cette étude a son importance compte tenu des grandes variations de tailles entre les classes, aussi bien pour la classification en domaines que pour la classification en genres.

Type de classification	Type de précision	Nature et taille de l'ensemble de descripteurs		
		$\mathcal{M}_{136}$	$\mathcal{L}_{500}$	$\{\mathcal{M} \oplus \mathcal{L}\}_{500}$
Domaine	micro	92.2 %	93.3 %	94.1 %
	macro	**80.3 %**	**80.4 %**	**84.8 %**
Genre	micro	79.9 %	80.1 %	81.1 %
	macro	59.3 %	61.9 %	61.4 %

Tableau 3. *Micro et macro-précisions sur le corpus global avec C4.5*

L'analyse en terme de macro-précision révèle certains phénomènes masqués par l'influence d'une classe fortement majoritaire (60 % des documents du corpus global sont des articles de linguistiques). Notamment pour la tâche de classification en domaine, la macro-précision permet de mettre en évidence un écart plus important entre les ensembles d'attributs pris séparément (80.3 % et 80.4 %) et la combinaison des deux ensembles (84.8 %). En effet, on note beaucoup plus de documents du domaine de la mécanique classés en linguistique avec les niveaux de descriptions $\mathcal{M}$ ou $\mathcal{L}$ qu'avec une description combinée $\mathcal{L} \oplus \mathcal{M}$. Cette remarque confirme à nouveau la complémentarité entre les deux niveaux de description pour la classification en domaines.

5. Analyse des descripteurs discriminants

5.1. *Les descripteurs de domaine*

On reporte dans le tableau 4 les variables apparaissant dans au moins 2 des 10 arbres de décision obtenus (5 validations croisées à 2 blocs)[9].

Les variables lexicales discriminantes sont toutes caractéristiques du domaine scientifique mécanique. Par exemple, on observe sur un échantillon d'entraînement que si le terme « écoulement » apparaît au moins deux fois, il permet de discriminer la moitié des textes de mécanique. Les textes de linguistique sont donc différenciés de manière négative : dans le même échantillon 90 % des textes linguistiques sont bien classés s'ils contiennent au plus une fois le terme « écoulement » et ne contiennent ni « mécanique », ni « vitesse » et ni « essais ». Cette discrimination par des termes de mécanique s'explique par : la taille plus importante des textes de linguistique qui augmente le nombre et la diversité des descripteurs, et les textes de mécanique qui semblent plus homogènes au niveau lexical.

9. L'ordre d'apparition des variables dans les colonnes du tableau est totalement arbitraire.

Variables		
Morphosyntaxiques	**Lexicales**	**Mixtes**
Indices de renvois (*e.g.* « voir en **1.1** »)	*équation*	*équation*
Pronoms personnels	*écoulement*	*vitesse*
Prépositions	*vitesse*	*écoulement*
Symboles, sigles, abréviations	*coefficient*	*vitesses*
Participes passés modaux	*déformation*	*laboratoire*
Adverbes et connecteurs	*amélioration*	Adjectifs réflexifs
Pronoms clitiques	*augmentation*	Locutions adverbiales
	courbes	Adverbes et connecteurs
	essais	Connecteurs de concession
	laboratoire	Nombre de « JE »
	mécanique	Prépositions
	vitesses	Ponctuation (points)

Tableau 4. *Descripteurs morphosyntaxiques et lexicaux discriminants en matière de classification en domaines*

Les descripteurs morphosyntaxiques les plus discriminants semblent par contre plus caractéristiques du domaine linguistique : par exemple, on observe sur un échantillon que la variable « préposition », lorsqu'elle dépasse un certain seuil, permet de différencier jusqu'à 90 % des textes de linguistique. De même, un nombre élevé de pronoms personnels et de marques de renvois discrimine les textes de linguistique des textes de mécanique.

En ce qui concerne les classifications mixtes, notons qu'elles recourent davantage aux variables morphosyntaxiques qu'aux variables lexicales malgré la prépondérance des traits lexicaux dans l'espace de description ($|\mathcal{L}| = 364 > 136 = |\mathcal{M}|$). Pourtant, les variables lexicales interviennent toujours en premier dans l'arbre de classification (cf. figure 7), les traits morphosyntaxiques permettant de raffiner la classification. Elles sont donc les plus discriminantes, mais ne suffisent pas à classer les documents de manière satisfaisante. Le rôle du niveau morphosyntaxique est donc loin d'être négligeable en matière de classification en domaines.

On notera que ces résultats contiennent des indices descriptifs susceptibles d'intéresser la caractérisation des domaines.

5.2. *Les descripteurs de genre*

On reporte dans le tableau 5 les variables apparaissant dans au moins trois arbres de décision sur l'ensemble des arbres appris.

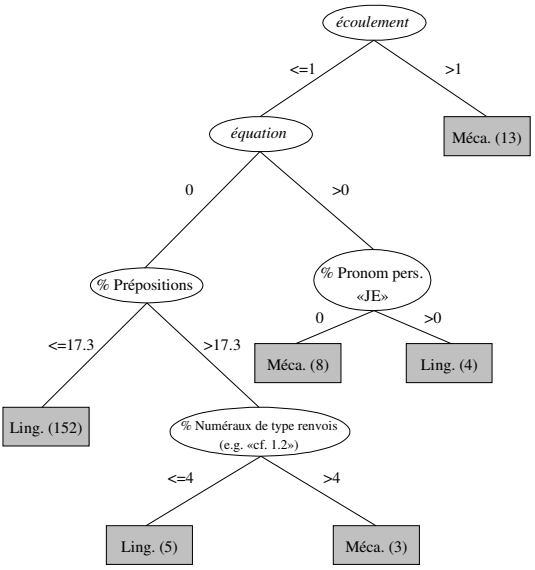

Figure 7. *Arbre représentatif pour la classification en domaines avec l'ensemble de descripteurs* $\mathcal{L} \oplus \mathcal{M}$

On notera que les arbres de décision font intervenir plus de variables lexicales pour classifier les genres que pour la classification des domaines[10], ce qui ne semble pas surprenant. Les substantifs présentés dans le tableau 5 sont caractéristiques des comptes rendus et des présentations de revues. Les articles sont donc classés relativement à l'absence de marqueurs caractéristiques des deux autres genres : ainsi, la quasi totalité des articles est correctement classée si les textes ne contiennent ni « contributions », ni « chapitres » et au plus une occurrence de « chapitre », les contributions étant aussi bien caractéristiques des comptes rendus que des présentations de revues. « chapitres » permettrait par contre de discriminer les comptes rendus. Certains indices lexicaux semblent donc caractéristiques du genre, conformément à ce que soutiennent (Lee *et al.,* 2002). Toutefois, les éléments lexicaux ne sont pas aussi efficaces pour distinguer les genres que pour la classification des domaines, les genres n'étant pas discriminés de manière aussi claire que les domaines.

Les variables morphosyntaxiques semblent caractéristiques des articles scientifiques : ainsi, les indices de structuration textuelle (LS) sont particulièrement discriminants et interviennent d'ailleurs en premier dans la plupart des arbres de classification.

10. Ce phénomène n'est pas visbible dans le tableau qui ne présente que les substantifs fréquents.

En effet, les comptes rendus ne sont jamais structurés, à l'inverse des articles et des présentations de revues. Notons que si les articles sont caractérisés par un niveau élevé de structuration, il n'en va pas de même des présentations, qui peuvent être structurées sans que cela soit pour autant caractéristique du genre.

Variables		
Morphosyntaxiques	**Lexicales**	**Mixtes**
Indices de structuration textuelle (LS)	*chapitres*	LS
Noms propres	*contributions*	*articles*
Passifs/passés composés	*articles*	*chapitres*
Symboles	*presses*	*contributions*
Ponctuation (deux points)	*chapitre*	Passifs/passés composés
Ponctuation (points)	*bibliographie*	Connecteurs de concession
Connecteurs de conséquence	*journées*	Connecteurs spatiaux
Éléments de langue étrangère	*linguistique*	Éléments de langue étrangère
Indices de renvois	*numéro*	Indices de renvois
Pronom personne « NOUS » clitique	*politique*	Pronom personne « NOUS » clitique

Tableau 5. *Descripteurs morphosyntaxiques et lexicaux discriminants en matière de classification en genres*

Enfin, en ce qui concerne la classification mixte, on note que seuls trois items lexicaux participent à la classification de manière significative : les substantifs « articles », « chapitres » et « contributions », qui ne sont pas caractéristiques des articles. De la même manière que pour la classification à partir du plan morphosyntaxique seul, les indices de structuration interviennent en premier dans l'arbre de classification (cf. figure 8).

6. Conclusion

Nous avons cherché à évaluer de manière expérimentale l'incidence des niveaux morphosyntaxique et lexical sur la classification en domaines et en genres dans le cas particulier des textes scientifiques.

Dans cette perspective, un ensemble de descripteurs morphosyntaxiques adapté aux caractéristiques du discours scientifique a été développé. Nous avons parallèlement opté pour le choix des substantifs au singulier et au pluriel au niveau lexical, dans la mesure où ils pointent potentiellement sur des concepts.

Bien qu'ils aient été obtenus sur un corpus de taille restreinte, les résultats de l'expérience sont particulièrement encourageants parce qu'ils soulignent l'intérêt d'une complémentarité des deux niveaux pour la classification en domaines et confirment celui des variables morphosyntaxiques en matière de classification en genres. En effet, la discrimination des deux domaines observés est nettement plus précise si l'on

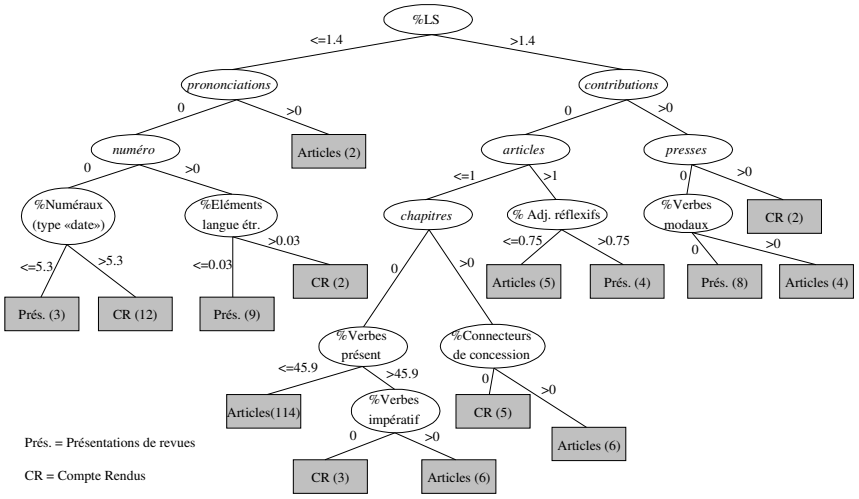

Figure 8. *Arbre représentatif pour la classification en genres avec l'ensemble de descripteurs* $\mathcal{L} \oplus \mathcal{M}$

utilise les deux jeux de variables conjointement avec les deux types de classifieurs employés, dans la mesure où les variables morphosyntaxiques permettent d'affiner considérablement les partitions obtenues avec le lexique. Enfin, nous avons pu apprécier la performance de la méthode SVM par rapport à la méthode C4.5 en matière de classification générique morphosyntaxique. Nous n'écartons pas toutefois l'intérêt de l'utilisation de variables lexicales pour discriminer les genres, l'étude des descripteurs s'étant révélée encourageante.

Nous envisageons d'approfondir et de préciser les résultats obtenus sur d'autres types de domaines et de genres. La pertinence des descripteurs utilisés sera également évaluée plus exactement : le jeu de variables morphosyntaxiques employé sera ainsi comparé aux jeu d'étiquettes du Penn Treebank Corpus utilisé par des taggers comme Brill ou TreeTagger par exemple, et d'autres types de descripteurs lexicaux seront extraits afin d'évaluer la pertinence de l'approche substantivale que nous avons adoptée.

7. Bibliographie

Biber D., *Variation across Speech and Writing*, University Press, Cambridge, 1988.

Brants T., « TnT - A Statistical Part-of-Speech Tagger », *Proceedings of the Sixth Applied Natural Language Processing Conference (ANLP'00)*, Seattle, WA, 2000.

Cohen W., Hirsh H., « Joins that generalize : text classification using WHIRL », *in* , R. Agrawal, , P. E. Stolorz, , G. Piatetsky-Shapiro (eds), *Proceedings of KDD-98, 4th International Conference on Knowledge Discovery and Data Mining*, AAAI Press, Menlo Park, US, New York, US, 1998, p. 169-173.

Crowston K., Kwasnik B., « A Framework for Creating a Facetted Classification for Genres : Addresssing Issues of Multidimensionality », *37th Hawaii International Conference on System Sciences, (IEEE Computer Society)*, Hawaii, 2004.

Dhillon I. S., Mallela S., Kumar R., « A divisive information theoretic feature clustering algorithm for text classification », *Journal of Machine Learning Researches*, 2003, vol. 3, p. 1265-1287.

Dumais S.*IEEE Intell. Systems*, , « Using SVMs for text categorization », 1998.

Dumais S., Platt J., Heckerman D., Sahami M., « Inductive learning algorithms and representations for text categorization », *CIKM '98 : Proceedings of the seventh international conference on Information and knowledge management*, ACM Press, 1998, p. 148-155.

Hofmann T., « Probabilistic Latent Semantic Indexing », *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*, Berkeley, California, August, 1999, p. 50-57.

Joachims T., « Text categorization with support vector machines : learning with many relevant features », *in* , Claire Nédellec and Céline Rouveirol (ed.), *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398, Springer Verlag, Heidelberg, DE, Chemnitz, DE, 1998, p. 137-142.

Karlgren J., Cutting D., « Recognizing text genres with simple metrics using discriminant analysis », *Proceedings of COLING 94*, Kyoto, 1994.

Kessler B., Nunberg G., Schültze H., « Automatic detection of text genre », *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL'97)*, 1997, p. 32-38.

Lee Y.-B., Myaeng S. H., « Text genre classification with genre-revealing and subject-revealing features », *SIGIR '02 : Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM Press, 2002, p. 145-150.

Lewis D. D., Ringuette M., « A comparison of two learning algorithms for text categorization », *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, US, 1994, p. 81-93.

Malrieu D., Rastier F., « Genres et variations morphosyntaxiques », *Traitement Automatique des langues*, 2001, vol. 42, n° 2, p. 548-577.

Porter M. F., « An algorithm for suffix stripping », *Program 14 :*, 1980p. 130-137.

Poudat C., Cleuziou G., « Genre and Domain Processing in an Information Retrieval Perspective », *in* , LNCS (ed.), *Third International Conference on Web Engineering*, Springer, Oviedo, Spain, 2003, p. 399-402.

Prime-Claverie C., Beigbeder M., Lafouge T., « Clusterisation du Web en vue d'extraction de corpus homogènes », *INFORSID*, 2002, p. 229-242.

Quinlan J. R., *C4.5 : Programs for Machine Learning*, Morgan Kaufmann, 1993.

Sebastiani F., « Machine learning in automated text categorization », *ACM Comput. Surv.*, 2002, vol. 34, n° 1, p. 1-47.

Vapnik V., *The nature of statistical theory*, Springer Verlag, 1995.