On the Impact of Lexical and Linguistic features in Genre- and Domain-Based Categorization

Guillaume Cleuziou[1] and Céline Poudat[2]

[1] LIFO, Université d'Orléans, FRANCE
guillaume.cleuziou@univ-orleans.fr
[2] ERTIM, INALCO, Paris, FRANCE
celine.poudat@enst.fr

Abstract. Classification in genres and domains is a major field of research for Information Retrieval (scientific and technical watch, data-mining, etc.) and the selection of appropriate descriptors to characterize and classify texts is particularly crucial to that effect.

Most of practical experiments consider that domains are correlated to the content level (words, tokens, lemmas, etc.) and genres to the morphosyntactic or linguistic one (function words, POS, etc.). However, currently used variables are generally not accurate enough to be applied to the categorization task.

The present study assesses the impact of the lexical and linguistic levels in the field of genre and domain categorization. The empirical results we obtained demonstrate how important it is to select an appropriate tagset that meets the requirement of the task. The results also assess the efficiency of the linguistic level for both genre- and domain-based categorization.

1 Introduction

Text categorization (or classification), as any classification task, requires an appropriate set of descriptors. In the same way as it would be irrelevant to characterize the financial profiles of bank account users according to variables such as "size" or "eye color", it would be inappropriate to describe scientific texts thanks to variables such as "number of dialogue marks" as far as they are absent from scientific discourse.

Genre and domain classifications are today widely used in Information Retrieval (IR) systems and they also require appropriate descriptors. It is worth emphasizing that genres and domains are generally associated with distinct linguistic levels. On the one hand, domains, or subjects, are rather related to lexical features in practice: texts are often reduced to "bags of words" and each document is described on the basis of the whole corpus lexicon. The size of the latter calls for a necessary step of reduction of the description area: selection of the attributes thanks to statistical measures (number of occurrences in the corpus), interest measures (Mutual Information, Information Gain, chi-square measure, etc.), re-parameterisation of the space with methods like *Latent Semantique Indexing* (LSI) or feature clustering. These formalisms allow us to obtain efficient classifiers which can reach a precision of 90% on large corpora [Hof99,DMK03].

Genres are on the other hand generally classified thanks to morphosyntactic (or linguistic) variables which have proved to be quite efficient to validate of text typologies [KC94,KNS97,MR01].

Nevertheless, domain-based categorization is generally conducted on genre-homogene-ous corpora (*e.g.* Reuters[1] or Newsgroup[2]) whereas genres are most often classified on discourse-homogeneous ones (*e.g.* [KC94,KNS97,MR01]): this increases the classificatory power of the variables but prevents the joint use, and the evaluation of the scopes of the two levels.

The aim of the present study is to assess the impact of thematic and morphosyntactic variables on genre and domain classifications. The experiment will be conducted on a pilot-corpus that will allow us to determine the interest of a joint use of the two levels of description.

After a brief overview of the use of the notions of genres and domains in IR, we will discuss about the relation between the two concepts in Section 2. Section 3 presents the corpus and the methodology we adopted to evaluate the complementarity between linguistic and lexical features. The experimental aspects of this assessment and the obtained results are detailed respectively in Sections 4 and 5.

2 Genres and Domains

Although the notions of genres and domains are more and more common in IR, they are scarcely used conjointly as far as they are traditionally associated with variables or cues belonging to distinct linguistic levels. Indeed, domains are generally related in practice to lexical features whereas the notion of genre is rather connected to morphosyntactic variables.

Domains, or subjects, are indeed supposed to reflect particular fields of knowledge and are often described in terms of lexical relations, as in ontologies for instance. Different methods have been developed to characterize and classify texts in domains according to their contents. The most commonly used measures are computed from the basis of words, word clusters (unequally called topics, themes, etc.) or word stems frequencies which have turned out to be quite efficient in various applications. Word-based classification is still besides the most widespread because of its lower cost.

The notion of genre[3], which is traditionally philological and literary, is more and more common in IR and text categorization. Indeed genres can be identified and contrasted thanks to their specific linguistic properties: for instance, legal

[1] http://www.research.att.com/lewis/reuters21578.html

[2] http://people.csail.mit.edu/jrennie/20Newsgroups/

[3] Or “style”, “register” or even “text type”.

texts do not contain exclamation marks. Genre analysis and characterization are generally conducted according to a set of methods inherited from quantitative stylistics, *i.e.* using part-of-speech (POS) or function words categories. Besides, the morphosyntactic method, initiated by [Bib88], has been successful in various text and genre classification studies.

Since different domains can been retrieved inside different genres and vice versa, we are tempted to consider that domains and genre are not correlated. The associated descriptive levels (resp. lexical and linguistic) are then rarely used together in practice. Although some encouraging recent studies tried to use lexical features to improve genre-based categorization [WK99,LM02,PC03], the characterization of domains thanks to morphosyntactic variables is still undone, as far as we know. However, in the same way, domains might be properly classified thanks to linguistic variables or at least, this additional descriptive level may improve a word-based domain classification.

3 Methodology

3.1 Development of a pilot corpus

For our study, traditional benchmarks such as *Reuters* or *Newgroups* were excluded as they are generically homogeneous. Furthermore, since our goal is to evaluate the interests of two descriptive levels for genre- and domain-based categorization, we decide (for this study) to eliminate the discursive variability[4].

As genres and domains are key-notions for scientific discourse description and applications (scientific watch, document retrieval, etc.), we conducted the following experiments on scientific texts. As they are subjected to an important bureaucracy (peer reviewing, anonymity policy), scientific texts have to meet linguistic and structural constraints that might reduce variation.

We use a pilot corpus especially developed for this study : it is composed of 371 French scientific texts published about 2000, that is three different genres (articles, journal presentations[5] and reviews) and two scientific domains (linguistics and mechanics), described in Table 1.

	Linguistics	Mechanics
Articles	224	49
Journal presentations	45	
Reviews	53	

Table 1. Presentation of the pilot corpus.

[4] Indeed, types of discourses seem to appear in first (before genres, domains or personal styles) with morphosyntactic characterizations [MR01].

[5] or introductive articles, describing and presenting the topic of the journal issue, and the scientific articles it contains. Because of their specific purpose and design, journal presentations are clearly distinct from scientific articles.

The relative small size of text collection is a common problem to all studies which requires such a specific corpus (*e.g.* [WK99]). Although the significance of the results is then limited, this first stage of experiments gives a crucial starting point for further experiments in wide corpus.

With regards to the following experiments, we will use the two following subcorpora (in addition to the global corpus):

- *ART*-corpus refers to the text collection composed of *articles* only (first line in Table 1),
- *LING*-corpus refers to the collection which contains only the texts about *linguistics* (first column in Table 1).

Furthermore we will differentiate *local* and *global* corpora: *global* corresponds to the whole corpus whereas *local* refers to a subcorpus, homogeneous in genre (*ART*-corpus) or in domain (*LING*-corpus).

3.2 Feature selection for scientific texts

Among the possible lexical variables, the choice of the most frequent substantives, or noun descriptors seems to be appropriate and quite economical, as they are potential scientific concepts rather than verbs, adverbs or adjectives. In that respect and as far as scientific domains are concerned, they are more discriminatory and have the advantage to be easily extracted. As singular and plural nouns might relate back to different concepts[6], the singular and plural forms of the nouns have been taken into account. About 10,000 singular nouns and 4,000 plural nouns are then extracted from the global corpus.

As far as they represent our generic descriptive hypothesis, the selection of morphosyntactic variables has been subjected to a precise linguistic expertise; indeed, it would be quite inappropriate to describe scientific texts according to features they do not possess, or with too general variables that would not include scientific texts properties. In addition to the traditional POS (nouns, verbs, adverbs, adjectives, prepositions, etc.), we selected a set of cues gathering the general descriptive hypothesis put forward in the literature focusing on scientific discourse. Table 2 describes these additional tags.

Finaly, a set of 136 variables is selected to describe the morphosyntax of the scientific texts. The tagging has been performed by learning with the tagger TnT (Trigrams'n'Tags) [Bra00] on the selected feature set.

3.3 Classifiers used

Document classification (or categorization) has led to numerous works requering to machine learning technics. In this field of research, the most commonly used classifiers are : *Naïve Bayes* [LR94], SVM[7] [Joa98] and *Decision Trees* [CH98].

[6] "la langue" - language - and "les langues" - languages - are for instance different linguistic notions.

[7] Support Vector Machine.

Tag	Description
ABR	Abbreviations
CON (+ attributes)	Connectives: addition, cause, consequence, conclusion, exemplification, disjunction, opposition, rephrasing, space, time, etc.
FGW	Foreign (non-French) elements
NUM (+ attributes)	Numerals: date, cardinal, ordinal + references in the text (e.g. "See in 12")
LS	Title cues and list marks
PON (+ attributes)	Punctuation marks : colon, square brackets, quotation marks, braces, slashs, etc.
VER:mod:[tense]	Modals
SIG	Acronyms
SYM	Symbols

Table 2. Description of the morphosyntactic descriptors.

Because goals of the following experiments are two fold, we chose to use two methods very different in nature : texts will be classified with SVM in order to evaluate the accuracy rate obtained from various initial descriptions (lexical, linguistic or combined) and decision trees (DT) will help us to explain how lexical and linguistic features may be combined within the classifier.

The SVM method is acknowledged to outperform other methods in text categorization [DPHS98]. To simplify matters, it consists in learning a classifier in a new feature space, far more dimensioned than the original one. The new space is obtained from different kernel functions (*e.g.* linear, polynomial, rbf, etc.). As several studies showed that best accuracies were obtained with a linear SVM [Dum98], we decided to use this type of kernel in our experiments. For each classification task (genre or domain-based), it will then be possible to measure the relevance of each set of features: lexical features only ($\mathcal{L}$), morphosyntactic ones only ($\mathcal{M}$) and combined features ($\mathcal{L} \oplus \mathcal{M}$).

In contrast with the SVM numerical approach, DT proceeds in a more symbolic way. Although it usually provides less accurate results in text classification, the learned trees are easier to analyse and to interpret and the study of the trees enables us to bring out the role played by each of the features. In our experiments, we will use the well-known C4.5 method [Qui93].

3.4 Evaluation framework

In this section we first give formal details about the feature vectors construction before describing the set of experiments.

Let $\mathcal{D}$ be a set of texts and $\mathcal{C}$ be a set of classes such that a unique class $c(d_i) \in \mathcal{C}$ is associated to each text $d_i \in \mathcal{D}$ (genre or domain). $\mathcal{D}$ is divided into a training set $\mathcal{D}_{train}$ and a test set $\mathcal{D}_{test}$.

$\mathcal{L}_{\mathcal{D}} = \{l_1, \ldots, l_{|\mathcal{L}|}\}$ denotes the ordered set of substantives (singular and plural) which occur within the texts from $\mathcal{D}_{train}$ (lexical description). In $\mathcal{L}_{\mathcal{D}}$, substantives are ordered by decreasing relevance for the given classification task $\mathcal{C}$ using the Mutual Information (MI) :

$$\forall l_i \in \mathcal{L},\ MI(l_i, \mathcal{C}) = \sum_{c_j \in \mathcal{C}} P(c_j).\log \frac{P(l_i|c_j)}{P(l_i)}$$

$\mathcal{M} = \{m_1, \ldots, m_{136}\}$ denotes the ordered set of 136 morphosyntactic (or linguistic) features described in section 3.2. We use the Information Gain (IG) coefficient to measure the interest of each descriptor according to the target classification function. Since features in $\mathcal{M}$ are continuous, a discretization step is necessary (cf. [Mit97]).

$\mathcal{L} \oplus \mathcal{M}$ corresponds to a mixture of the two feature sets $\mathcal{L}$ and $\mathcal{M}$ in the following order: $\mathcal{L} \oplus \mathcal{M} = \{l_1, m_1, l_2, m_2, \ldots, l_{136}, m_{136}, l_{137}, l_{138}, \ldots, l_{|\mathcal{L}|}\}$.

In order to determine the impact of the variables on genre and domain classification, it is necessary to observe the influence of each of the three feature sets ($\mathcal{L}$, $\mathcal{M}$ and $\mathcal{L} \oplus \mathcal{M}$) on local and global corpora. It is also interesting to observe the influence of the size of the feature vector ; in this way we will report results for different sizes : from 1 to 500.

The experimentations proposed in section 4 are the result of 2-fold cross-validations: $\mathcal{D}$ is splitted into two equal subcorpora, each of them being by turn used as test and training set. The reported values correspond to micro-averaging precisions[8] on 5 cross-validations.

For the SVM classifier, in case of multi-class problems, several binary classifiers are learned and combined.

4 Experimentations

The first experimentations are devoted to domain classification. They are based on "local" (*ART*-corpus) and "global" (whole corpus) corpora. The first set will be the basis of the discrimination of the two domains within the same genre whereas the second one will enable us to introduce a generic variation parameter.

Genre classification will then be conducted in the same way: first on the "local" corpus (*LING*-corpus), and next on the same "global" corpus.

4.1 Domain classification

The results we obtained with the SVM method (figures 1 and 2) clearly show, against all expectations, that morphosyntactic variables are more discriminatory than lexical ones. Moreover, it is worth emphasizing that, for the same number

[8] Micro-averaging measures the proportion of well classified texts whatever the class. It differs from the macro-averaging which measures the average of the accuracies for each class separately.

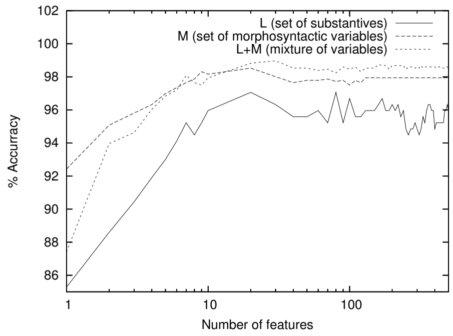

Fig. 1. Domain-based categorization with SVM on the ART-Corpus.

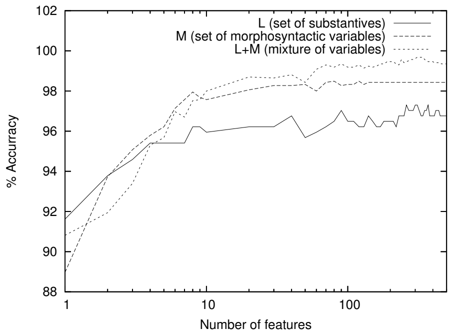

Fig. 2. Domain-based categorization with SVM on the global Corpus.

of features, a combination of the two types of variables is on the whole more efficient than each of the two sets on their own.

The following precedence order is obtained (with or without generic variation):

$$\{\mathcal{L} \oplus \mathcal{M} - indexing\} > \{\mathcal{M} - indexing\} > \{\mathcal{L} - indexing\}$$

The same trends are noted with a decision tree classifier, although the accuracy rates are weaker than with SVM. The lexical indexation is also less efficient than the morphosyntactic and mixed ones.

The result is quite surprising as mechanics and linguistics are conceptually and lexically very different, or even opposed and scientific domains might be better discriminated thanks to morphosyntactic variables than with lexical features.

4.2 Genre classification

The results we obtained with the SVM method (figures 3 and 4) confirm morphosyntactic variables are relevant to capture the genre dimension. The accuracy rate is higher using the feature sets containing morphosyntactic information than the lexical one. It must be emphasized that the domain differences do not disrupt this conclusion:

$$\{\mathcal{L} \oplus \mathcal{M} - indexing\} \approx \{\mathcal{M} - indexing\} \gg \{\mathcal{L} - indexing\}$$

Figures 5 and 6 report the results obtained with the decision tree classifier. The accuracy rates obtained are once again noticeably weaker than with SVM: 84% best rate *vs.* 88% with SVM. However, the precedence order obtained with C4.5 is rather different. Lexical cues are efficient on the global corpus (from 100 features) and this seems to corroborate the results obtained by [LM02] :

$$\{\mathcal{L} \oplus \mathcal{M} - indexing\} > \{\mathcal{M} - indexing\} \gg \{\mathcal{L} - indexing\}$$

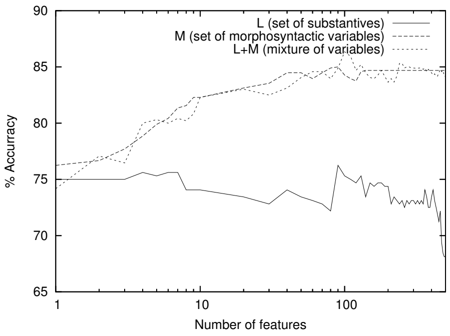

Fig. 3. Genre-based categorization with SVM on the LING-Corpus.

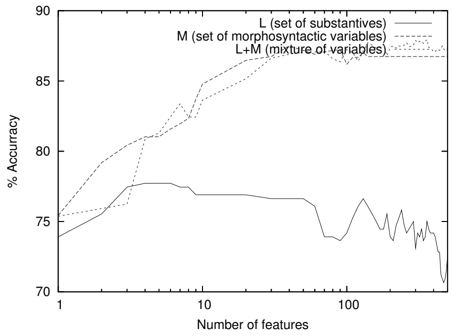

Fig. 4. Genre-based categorization with SVM on the global Corpus.

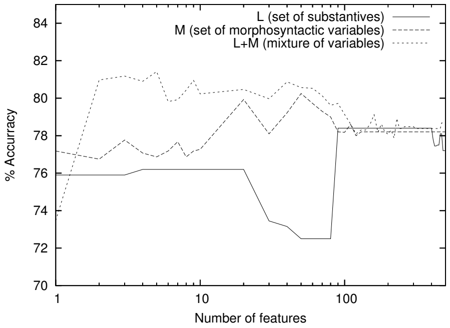

Fig. 5. Genre-based categorization with DT on the ART-Corpus.

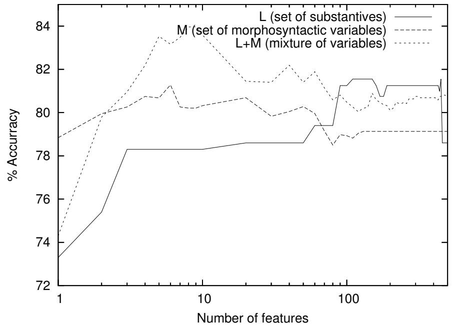

Fig. 6. Genre-based categorization with DT on the global Corpus.

Nevertheless the precedence order we obtain on the local corpus is quite similar to the one obtained with SVM.

From a technical point of view, the differences obtained between the two classifiers may be due to the different methods they are implemented on. Indeed, the SVM approach considers a new space of representation of the documents, with a high dimensionality and of which dimensions are defined by - linear - combinations of the initial descriptors. The method calls for the whole of the variables whereas the construction of a decision tree generally calls for a small set of precisely selected cues.

4.3 Further analysis : micro *vs.* macro-precision

Before detailing the preceding results with the study of the decision trees, let us consider an intermediate synthesis of the experimentations we conducted so far.

Table 3 reports the macro and micro-precisions inducted by the decision trees learned from the global corpus for a defined number of descriptors. This is quite important because of the large size variations of the classes:

Type of classification	Type of precision	Nature and size of the feature set		
		$\mathcal{M}_{136}$	$\mathcal{L}_{500}$	$\{\mathcal{M} \oplus \mathcal{L}\}_{500}$
Domain	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%

Table 3. Micro and macro-precisions on the global corpus with C4.5.

The macro-precision analysis brings out phenomena that were hidden by the influence of the linguistic articles class (60% documents of the global corpus). Thereby, we can observe a clear emphasis of the relevance of the combined set for domain classification (+4.5%). A larger number of documents belonging to the mechanics domain is misclassified with the $\mathcal{M}$ or $\mathcal{L}$ descriptions than with a combined one. This observation reinforces once again the complementarity of the two levels in domain clustering.

5 Analysis of the discriminatory descriptors

5.1 Domain descriptors

Features		
Morphosyntactic	**Lexical**	**Combined**
References	*équation*	*équation*
Personal pronouns	*écoulement*	*vitesse*
Symbols, acronyms, abbreviations	*vitesse*	*écoulement*
Modal past participles	*coefficient*	*vitesses*
Adverbs and connectives	*déformation*	*laboratoire*
Reflexive pronouns	*amélioration*	Reflexive adjectives
	augmentation	Adverbial phrase
	courbes	Adverbs and connectives
	essais	Concessive connectives
	laboratoire	Nomber of ”JE” (”I”)
	mécanique	Prepositions
	vitesses	Punctuation (points)

Table 4. Features retrieved from domain decision trees.

Table 4 reports the variables found in at least two decision trees out of the 10 obtained (five 2-fold cross-validations).

The discriminatory lexical variables are all specific to mechanics. For instance, we observe in a sample that the term “écoulement” (*flow*) enables us to discriminate half of the texts of the training corpus belonging to mechanics. Linguistics texts are thus negatively differentiated: in the same sample, 90% of the linguistics corpus is correctly classified if the texts do not contain the term “écoulement” more than once and if they contain neither “mécanique” (*mechanics*), neither “vitesse” (*speed*) and nor “essais” (*test*). This discrimination is due to two reasons: (1) the more important size of the texts belonging to linguistics increases the number and the diversity of the descriptors and (2) mechanics articles seem to be more homogeneous in terms of lexicon.

On the contrary, the discriminatory morphosyntactic descriptors are more specific to the linguistic field: for instance, the number of prepositions enable us to differentiate up to 90% of the training corpus. In the same way, linguistics texts contain more personal pronouns and reference marks than mechanics ones.

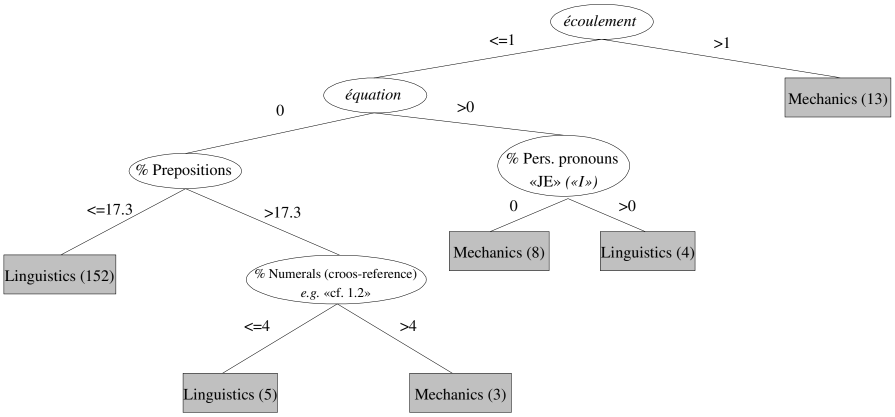

Fig. 7. Representative tree obtained with the combined feature set for domain-based categorization.

Joint classifications use a higher number of morphosyntactic variables than lexical ones, in spite of the predominance of the lexical cues in the description space ($|\mathcal{L}| = 364 > |\mathcal{M}| = 136$). However, lexical cues are always the first used in the classification tree (cf. 7), morphosyntactic variables enable us to refine the classes. The morphosyntactic level seems to be discriminatory although it does not enable us to classify the documents in an acceptable way.

5.2 Genre descriptors

Table 5 reports the variables found in at least three decision trees out of the whole of the trees. It is first to notice that the decision trees use more lexical variables to classify genres than domains. The substantives given in table 5 are characteristic of the reports and journal presentations. Most of the articles are correctly classified if the texts contain neither “contributions” (*contri-*

Features		
Morphosyntactic	**Lexical**	**Combined**
Title cues (LS)	*chapitres*	Title cues (LS)
Proper nouns	*contributions*	*articles*
Passives/present perfect	*articles*	*chapitres*
Symbols	*presses*	*contributions*
Punctuation (colon)	*chapitre*	Passives/present perfect
Punctuation (points)	*bibliographie*	Concessive connectives
Consequence connectives	*journées*	Space connectives
Foreign elements	*linguistique*	Foreign elements
References	*numéro*	References
Reflexive ”NOUS” (”WE”)	*politique*	Reflexive ”NOUS” (”WE”)

Table 5. Features retrieved from genre decision trees.

butions), neither “chapitres” (*chapters*) and not more than one occurrence of “chapitre” (*chapter*). Lexical items are thus efficient to characterize genres, as [LM02] pointed out.

Morphosyntactic variables are particularly efficient to distinguish articles: title cues (LS) are indeed very discriminatory, as reviews are never structured and journal presentations far less than articles.

With regards to joint classification, it is worth emphasizing that three lexical items only are discriminatory: the substantives “articles”, “chapitres” and “contributions”, which are besides specific to articles. In the same way as in morphosyntactic classification, title cues are the first variables used in the classification tree.

6 Conclusion

In this paper we have presented an experimental assessment of the impact of the morphosyntactic and lexical variables to classify scientific genres and domains.

Although they were conducted on a small corpus, the results we obtained are quite encouraging as they not only corroborate the interest of linguistic features to classify genres, but illustrate the strong complementarity of the two levels in domain classification. Indeed, the joint use of the two sets of descriptors seems to be more efficient to discriminate domains, as morphosyntactic variables enable us to refine the partitions obtained with the lexicon. Moreover, it is worth emphasizing that genre classification are far better with the morphosyntactic level and the SVM classifier.

Further experiments will take into account additional genres and domains and will specify the impact of the two description levels. We also plan to assess the relevance of the descriptors we used: the morphosyntactic tagset we developed will be contrasted to the Penn TreeBank one [MSM94], and other lexical sets will be extracted to compare the relevance of the substantive-based approach we adopted.

References

[Bib88] D. Biber. *Variation across Speech and Writing.* University Press, Cambridge, 1988.

[Bra00] T. Brants. TnT - A Statistical Part-of-Speech Tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference (ANLP'00)*, Seattle, WA, 2000.

[CH98] W.W. Cohen and H. Hirsh. Joins that generalize: text classification using Whirl. In Rakesh Agrawal, Paul E. Stolorz, and Gregory Piatetsky-Shapiro, editors, *Proceedings of KDD-98,*, pages 169–173, New York, US, 1998. AAAI Press, Menlo Park, US.

[DMK03] I. S. Dhillon, S. Mallela, and R. Kumar. A divisive information theoretic feature clustering algorithm for text classification. *Journal of Machine Learning Researches*, 3:1265–1287, 2003.

[DPHS98] Susan Dumais, John Platt, David Heckerman, and Mehran Sahami. Inductive learning algorithms and representations for text categorization. In *Proceedings of CIKM'98*, pages 148–155. ACM Press, 1998.

[Dum98] S. Dumais. Using svms for text categorization. *IEEE Intell. Systems*, 13(4), 1998.

[Hof99] Thomas Hofmann. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*, pages 50–57, Berkeley, California, August 1999.

[Joa98] T. Joachims. Text categorization with support vector machines: learning with many relevant features. In Claire Nédellec and Céline Rouveirol, editor, *Proceedings of ECML-98*, number 1398, pages 137–142, Chemnitz, DE, 1998. Springer Verlag, Heidelberg, DE.

[KC94] J. Karlgren and D. Cutting. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of COLING 94*, Kyoto, 1994.

[KNS97] B. Kessler, G. Nunberg, and H. Schültze. Automatic detection of text genre. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL'97)*, pages 32–38, 1997.

[LM02] Yong-Bae Lee and Sung Hyon Myaeng. Text genre classification with genre-revealing and subject-revealing features. In *Proceedings of the 25th annual international ACM SIGIR*, pages 145–150. ACM Press, 2002.

[LR94] David D. Lewis and Marc Ringuette. A comparison of two learning algorithms for text categorization. In *Proceedings of SDAIR-94*, pages 81–93, Las Vegas, US, 1994.

[Mit97] T. Mitchell. *Machine Learning.* McGraw Hill, New York, 1997.

[MR01] D. Malrieu and F. Rastier. Genres et variations morphosyntaxiques. *Traitement Automatique des langues*, 42(2):548–577, 2001.

[MSM94] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330, 1994.

[PC03] C. Poudat and G. Cleuziou. Genre and Domain Processing in an Information Retrieval Perspective. In LNCS, editor, *Third International Conference on Web Engineering*, pages 399–402, Oviedo, Spain, 2003. Springer.

[Qui93] J. R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, 1993.

[WK99] Maria Wolters and Mathias Kirsten. Exploring the use of linguistic features in domain and genre classification. In *Proceedings of the ninth conference EACL'99*, pages 142–149, 1999.