TALN 2007, Toulouse, 5–8 juin 2007

Caractérisation des discours scientifiques et vulgarisés en français, japonais et russe

Lorraine GOEURIOT[1], Natalia GRABAR[2,3], Béatrice DAILLE[1]
[1] LINA/Nantes
[2] INSERM, UMR_S 872, Eq. 20, Paris, F-75006
Université René Descartes, Paris, F-75006
[3] Health on the Net Foundation, SIM/HUG, Genève, Suisse
natalia.grabar@biomath.jussieu.fr,
{lorraine.goeuriot,beatrice.daille}@univ-nantes.fr

Résumé. L'objectif principal de notre travail consiste à étudier la notion de comparabilité des corpus, et nous abordons cette question dans un contexte monolingue en cherchant à distinguer les documents scientifiques et vulgarisés. Nous travaillons séparément sur des corpus composés de documents du domaine médical dans trois langues à forte distance linguistique (le français, le japonais et le russe). Dans notre approche, les documents sont caractérisés dans chaque langue selon leur thématique et une typologie discursive qui se situe à trois niveaux de l'analyse des documents : structurel, modal et lexical. Le typage des documents est implémenté avec deux algorithmes d'apprentissage (SVMlight et C4.5). L'évaluation des résultats montre que la typologie discursive proposée est portable d'une langue à l'autre car elle permet en effet de distinguer les deux discours. Nous constatons néanmoins des performances très variées selon les langues, les algorithmes et les types de caractéristiques discursives.

Abstract. The main objective of our study consists to characterise the comparability of corpora, and we address this issue in the monolingual context through the disctinction of expert and non expert documents. We work separately with corpora composed of medical area documents in three languages, which show an important linguistic distance between them (French, Japanese and Russian). In our approach, documents are characterised in each language through their thematic topic and through a discursive typology positioned at three levels of document analysis : structural, modal and lexical. The document typology is implemented with two learning algorithms (SVMlight and C4.5). Evaluation of results shows that the proposed discursive typology can be transposed from one language to another, as it indeed allows to distinguish the two aimed discourses. However, we observe that performances vary a lot according to languages, algorithms and types of discursive characteristics.

Mots-clés : linguistique des corpus, corpus comparable, algorithmes d'apprentissage, analyse stylistique, degré de comparabilité.

Keywords: corpus linguistics, comparable corpora, learning algorithms, stylistic analysis, degree of comparability.

1 Introduction

Un corpus comparable est un ensemble de textes qui partagent entre eux un certain nombre de caractéristiques. La première de ces caractéristiques est de rassembler des documents qui portent sur des sujets proches. Par exemple, dans le contexte multilingue, les corpus comparables vont réunir les documents qui une thématique commune mais qui ne sont des traductions (Bowker & Pearson, 2002). Dans un contexte monolingue, les corpus comparables peuvent contenir les documents qui portent sur le même sujet mais relèvent de discours différents. Le terme *comparable* est donc employé afin d'indiquer que les corpus ont un certain nombre de caractéristiques en commun. Celles-ci peuvent concerner le contexte de création du texte (ex. : la période, l'auteur), mais aussi le texte lui-même (ex. : le thème, le genre). Le choix de ces caractéristiques dépend des objectifs fixés. Il influe sur le *degré de comparabilité* du corpus, notion permettant de quantifier la comparabilité d'un corpus (aux valeurs maximales dans un corpus parallèle par exemple).

Dans notre travail, nous étudions un corpus comparable portant sur le domaine médical dans trois langues à forte distance linguistique : le français, le japonais et le russe. Afin de garantir une meilleure comparabilité des documents de notre corpus, nous prenons en compte une caractéristique classique, relative à la thématique du corpus, mais aussi une caractéristique discursive : distinction entre les documents selon qu'ils relèvent du discours scientifique ou vulgarisé. Nous nous basons ici sur la notion de discours telle qu'elle est définie par (Ducrot & Schaeffer, 1999) : *"Tout ensemble d'énoncés d'un énonciateur caractérisé par une unité globale de thème"*. Pour l'implémentation de cette caractérisation, nous effectuons une analyse stylistique contrastive, inspirée des travaux de (Karlgren, 1998), et proposons une typologie pour la distinction automatique des documents scientifiques et vulgarisés en différentes langues. Cette typologie reprend trois niveaux d'analyse des documents : structurel, modal et lexical. L'objectif principal de notre travail consiste à vérifier si cette typologie, basée sur ces trois niveaux, permet effectivement de caractériser le discours d'un document du Web et d'affiner la notion de comparabilité des documents. En travaillant sur un corpus trilingue, nous pourrons aussi observer le comportement de cette typologie selon les langues. Finalement, cette typologie est utilisée afin d'adapter à notre corpus des algorithmes d'apprentissage.

Dans la suite de cet article, nous présentons d'abord le corpus étudié (sec. 2), la typologie établie pour la caractérisation des discours scientifique et vulgarisé (sec. 3), et les méthodes utilisées pour la détection automatique de ces discours (sec. 4). Nous présentons ensuite les résultats (sec. 5) et concluons (sec. 6).

2 Collecte et caractéristiques du corpus comparable trilingue

Notre corpus de travail est un corpus comparable comportant des documents en trois langues à grande distance linguistique : le français, le japonais et le russe. Dans le cadre de notre étude, relative à la recherche d'information multi- et translangue, nous souhaitons disposer d'un corpus au degré de comparabilité assez élevé, sans pour autant effacer les diversités culturelles et linguistiques propres à chacune des langues étudiées. Nous situons la comparabilité à deux niveaux :

– comme c'est souvent le cas en recherche d'information, nous assurons un premier niveau de comparabilité grâce à la thématique commune partagée par les documents en trois langues.

Nous avons choisi le domaine médical et, plus précisémment, la thématique “diabète et alimentation” : ce thème touche un large public et présente une garantie potentielle de collecter une diversité de documents sur le web. Par ailleurs, du point de vue applicatif, les corpus comparables trilingues étant en partie dédiés à l’extraction d’informations multilingues, un thème commun renforce la garantie de trouver un vocabulaire et des caractéristiques linguistiques communes dans les langues traitées.

– de plus, afin d’augmenter la comparabilité des textes, nous cherchons à assurer un deuxième niveau de comparabilité et distinguons les textes du corpus selon leur type de discours, scientifique et vulgarisé.

Dans la suite de cette section, nous abordons la méthodologie de constitution du corpus trilingue et donnons ses principales caractéristiques.

2.1 Méthodologie de constitution des corpus

Le corpus de cette étude est un corpus comparable dans les langues française, japonaise et russe. Les documents sont extraits du Web. La démarche de constitution du corpus repose sur trois étapes principales :

1. Recherche de pages web correspondant à la thématique visée,
2. Sélection des pages pertinentes,
3. Classement de ces pages selon leur type de discours.

Ainsi, lors de la première étape de recherche des pages web, nous avons utilisé trois approches : (1) Recherche sur le web à l’aide de moteurs de recherche généraux ; (2) Recherche interne sur des portails (médicaux) en utilisant le cas échéant les moteurs de recherche propres aux sites ; (3) Exploitation des liens entre les pages. Les deux premières approches nécessitent l’utilisation de mots clés. Afin d’obtenir un large spectre de documents, les requêtes utilisées sont formées avec des combinaisons variées de mots clés tels que *alimentation*, *diabète* et *obésité* étendus avec i) leurs synonymes relevés dans les dictionnaires, et ii) aux termes équivalents extraits des pages visitées. Notons aussi que dans le cas d’utilisation d’un moteur de recherche spécifique à un portail, les mots clés sont également spécifiques à ce portail.

Parmi ces documents, nous avons sélectionné manuellement les documents pertinents pour la thématique visée. Et enfin, les pages sélectionnées ont ensuite été classées selon le type de discours émanant. Lors de la classification manuelle, nous utilisons les heuristiques suivantes :

– un document scientifique est rédigé par des spécialistes à destination de spécialistes.
– en ce qui concerne la vulgarisation scientifique, nous distinguons deux degrés de vulgarisation : les textes écrits par “le grand public” à destination de tous, et les textes écrits par des spécialistes à destination du “grand public”.

Nous ne distinguerons pas par la suite ces deux niveaux mais accorderons cependant une plus grande place aux documents écrits par des spécialistes au détriment des discussions sur des forums par exemple. Les documents écrits par les spécialistes s’avèrent en effet être plus riches en vocabulaire et plus complets en contenu. La classification manuelle est donc basée sur ces heuristiques, elle est appuyée par des éléments supplémentaires : la nature du site contenant le document, le vocabulaire utilisé dans le document, etc. Il faut noter cependant que la tâche de classification manuelle reste assez empirique. Cela nous a conduit à ne pas inclure certains documents “ambigus” (documents inclassables ou sur lesquels les avis divergeaient) dans les corpus d’apprentissage.

2.2 Caractéristiques des corpus

Le tableau 1 présente les principales caractéristiques du corpus ainsi constitué : le nombre de documents et le nombre de mots dans chacune des langues et pour chaque type de discours.

	Français		Japonais		Russe	
	SC	VU	SC	VU	SC	VU
Nb. documents	65	183	119	419	45	150
Nb. mots	425 800	267 900			318 596	175126
Nb. caractéres	2 668 783	2 845 114	493 587	1 154 773	2298306	2 165 768

TAB. 1 – Caractéristiques du corpus

Ce corpus rassemble ainsi plus de 1 500 000 mots dans trois langues. Les chiffres donnés pour la langue japonaise correspondent au nombre de caractères, le nombre de mots étant difficilement estimable. L'ensemble de nos documents utilise plus de 3 alphabets et un grand nombre d'encodages différents. C'est pourquoi les textes ont tous été transcodés en Unicode UTF-8, seul codage permettant de traiter les alphabets latin et cyrillique, ainsi que les caractères kanjis japonais. Les documents du corpus appartiennent à différents formats, parmi lesquels on compte les formats usuels du web (html, xhtml, php, etc.), mais aussi d'autres formats (pdf, ps, doc, etc.). Toutes les pages ont été conservées dans leur format original, mais aussi converties en texte brut. Les genres du Web (Bretan *et al.*, 1998) ne sont pas tous représentés dans le corpus français, dans lequel on trouve en majorité des rapports et articles (de presse ou scientifiques), contrairement au corpus japonais dans lequel on trouve une grande diversité (allant du rapport scientifique à l'offre d'emploi). Le corpus russe montre également une variabilité de genres (articles, ouvrages, recettes de cuisine, guides de bonne pratique, discussions sur des forums spécialisés, ...).

3 Une typologie pour l'analyse stylistique du corpus

L'analyse stylistique est une discipline linguistique trouvant son application informatique dans le domaine de la catégorisation textuelle. Ces travaux se basent sur des méthodes relevant de l'apprentissage automatique. Les algorithmes d'apprentissage les plus connus sont les séparateurs à vastes marges (SVM), les réseaux de neurones, les classifieurs de Bayes et les arbres de décision (Sebastiani, 2002). Deux grands courants se distinguent dans l'analyse stylistique. Les travaux adoptant la *démarche inductive* permettent de faire émerger d'un corpus des corrélations déterminant des classes de similarités (pouvant varier selon le type de caractéristiques utilisé). Cette méthode permet de créer des typologies dites inductives. Dans le cadre de l'apprentissage automatique, cette méthode s'apparente à l'apprentissage non-supervisé, ou clustering (Biber, 1989). La seconde démarche, appelée *démarche déductive*, consiste à analyser un ensemble de documents préclassés afin de caractériser l'appartenance d'un élément à une classe, sous la forme d'une typologie. Cette démarche s'apparente à l'apprentissage supervisé (Bretan *et al.*, 1998). Avec la démarche déductive, deux techniques permettent d'arriver à une typologie : l'analyse des documents un à un, ou l'analyse contrastive de documents appartenant à deux classes distinctes. Dans notre travail, nous avons choisi d'utiliser l'approche contrastive.

Les algorithmes de catégorisation textuelle s'appliquent à de nombreuses typologies. Les plus fréquentes sont les typologies thématiques ou de genres (Bretan *et al.*, 1998). Les travaux por-

tant sur les typologies de discours étant en revanche moins nombreux, nous allons adapter les algorithmes de la démarche déductive et contrastive à la typologie des discours.

Les documents de notre corpus, collectés sur le Web, présentent une structure propre que nous ne pouvons pas négliger. Ainsi, contrairement à un grand nombre de travaux traitant de l'analyse stylistique de textes, par exemple (Malrieu & Rastier, 2002 ; Biber, 1989), nous prenons en compte aussi bien le contenu textuel que la structure des documents. Nous nous basons sur ces deux types d'informations afin de dégager une typologie propre aux discours ciblés.

Sinclair (1996) dans ses travaux typologiques introduit une notion de niveaux dans les typologies textuelles. En effet, il est selon lui plus pertinent de distinguer deux catégories de critères : les critères externes, caractéristiques du contexte de création du texte ; et les critères internes, caractéristiques linguistiques du texte. Notre corpus étant construit à partir de documents issus du Web, nous considérons les critères externes comme étant les critères relatifs à la création du document et à sa structure (caractéristiques "non-linguistiques"). La partie interne concerne les caractéristiques linguistiques du document. Cependant, l'analyse stylistique met en évidence différents niveaux de granularité dans les critères. La distinction entre les documents scientifiques et vulgarisés induit une prise en compte du locuteur dans son discours, c'est-à-dire de la modalité. De plus, le discours scientifique peut se caractériser par le vocabulaire employé, la longueur des mots, et autres critères relevant du lexique. La typologie que nous adoptons distingue donc trois niveaux d'analyse des documents :

Caractéristiques structurelles : éléments de la structure graphique et textuelle du texte ;

Caractéristiques modales : éléments caractérisant la modalité dans le texte ;

Caractéristiques lexicales : éléments relatifs au lexique employé dans le texte.

Critère	Français	Japonais	Russe
Format d'URL	×		
Format de document	×	×	×
Méta-informations (présence)	×	×	×
Titre de la page (présence)	×	×	×
Techniques de mise en page	×	×	×
Fonds	×	×	×
Images	×	×	×
Paragraphes	×	×	×
Listes	×	×	×
Nombre de phrases	×	×	×
Typographie	×	×	×
Longueur du document	×	×	×

TAB. 2 – Caractéristiques structurelles

Les caractéristiques structurelles, présentées dans le tableau 2, concernent en majeure partie l'aspect graphique du document (format, images, fonds, ...) ainsi que les éléments de sa structure pris en compte ici grâce aux balises `HTML` (paragraphes, listes, titre, ...). L'ensemble de critères structurels sont détectables dans les trois langues.

Dans le tableau 3 sont présentées les caractéristiques modales, qui correspondent à la modalité dans les textes, c'est-à-dire à la position du locuteur dans son propre discours. Ces critères sont directement inspirés des théories de Charaudeau (1992), et ont été adapté à notre corpus (Krivine *et al.*, 2006; Nakao, 2006). Parmi les actes locutifs énoncés par Charaudeau (1992), nous

	Critère	Français	Japonais	Russe
Modalité allocutive	Pronoms personnels sujets allocutifs	×	×	
	Modalité de l'injonction	×	×	×
	Modalité de l'autorisation	×		×
	Modalité du jugement	×		
	Modalité de la suggestion	×	×	×
	Modalité de l'interrogation	×	×	×
	Modalité de l'interpellation	×		×
	Modalité de la requête	×	×	×
Modalité élocutive	Pronoms personnels sujets élocutifs	×	×	
	Modalité du constat	×	×	×
	Modalité du savoir	×	×	×
	Modalité de l'opinion	×	×	×
	Modalité de la volonté	×	×	×
	Modalité de la promesse	×	×	×
	Modalité de la déclaration		×	×
	Modalité de l'appréciation	×		×
	Modalité de l'obligation	×		×
	Modalité de la possibilité	×		×
	Modalité de l'interdiction			×

TAB. 3 – Caractéristiques modales

Critère	Français	Japonais	Russe
Vocabulaire spécialisé	×	×	×
Caractères numériques	×	×	×
Unités de mesure	×	×	×
Longueur des mots	×		×
Bibliographie	×	×	×
Citations bibliographiques	×	×	×
Ponctuation	×	×	×
Fins de phrases[1]		×	
Parenthèses	×	×	×
Autres alphabets (latin, hiragana, katakana)		×	×
Symboles[2]		×	

TAB. 4 – Caractéristiques lexicales

n'avons conservé que ceux qui s'avérent opératoires dans les documents analysés. Par exemple, la modalité de l'opinion peut être détectée en français grâce aux verbes comme *penser, paraître, sembler*. Si, dans leur majorité, ces critères se retrouvent dans les trois langues, quelques critères (jugement, interdiction) s'avèrent spécifiques à une ou deux des langues traitées, essentiellement parce que l'instantiation de la typologie était basée sur des études assez isolées des corpus dans chaque langue.

Enfin, dans le tableau 4, nous présentons les critères lexicaux. Plus que dans les deux types précédents, ces critères montrent une dépendance selon les langues, comme l'usage de caractères hiragana ou katakana pour la langue japonaise et celui de l'alphabet latin en russe. Notons aussi que certains de ces critères sont spécifiques des documents scientifiques, comme les bibliographies et citations bibliographiques, le vocabulaire spécialisé ou les unités de mesure.

4 Classification automatique selon les discours

La typologie tripartite (caractéristiques structurelles, modales et lexicales) sert de base aux algorithmes d'apprentissage automatique. Nous présentons dans cette section la méthode : les algorithmes d'apprentissage utilisés et les principes d'évaluation des résultats.

4.1 Apprentissage

Comme nous l'avons annoncé, l'objectif de notre travail consiste à adapter les algorithmes d'apprentissage à la catégorisation des documents en fonction de deux discours, scientifique et vulgarisé. Dans notre travail, l'analyse stylistique, et la catégorisation des documents, est effectuée à travers la typologie de critères proposée (sec. 3). En effet, les algorithmes d'apprentissage perçoivent les documents sous formes de vecteurs, où chaque élément du vecteur représente la valeur d'un critère pour le document correspondant. La longueur d'un vecteur correspond à la fréquence (brute ou pondérée) de chaque critère dans le document traité. En partant d'un corpus d'apprentissage, où les documents sont répartis en deux échantillons (scientifique et vulgarisé), et d'une liste de critères, les méthodes d'apprentissage génèrent une procédure de classification. Cette procédure est ensuite appliquée à de nouvelles données, pour effectuer de nouvelles classifications.

Il existe différentes techniques de classification automatique de textes (réseaux de neurones, classifieurs de Bayes, séparateurs à vastes marges, etc.) pour lesquelles Sebastiani (2002) a effectué un travail de rassemblement et comparaison. Appliquées sur un corpus de dépêches Reuters, ces techniques ont ainsi montré des performances variables en fonction de l'utilisation des approches supervisée ou non-supervisée, de la taille du corpus, du nombre de catégories, etc. Nous avons choisi d'utiliser les séparateurs à vastes marges SVMlight (Joachims, 2002) ainsi que les arbres de décision C4.5 (Quinlan, 1993). Visant un système de classification rapide, nous avons privilégié, pour l'implémentation des critères, des techniques simples basées sur des patrons lexicaux ou lexicaux-syntaxiques et qui analysent superficiellement les documents et leur contenu.

Le fonctionnement du système de catégorisation repose sur la reconnaissance de formes lexicales dans les documents, que ce soit pour les critères graphiques (reconnus à travers les balises), modaux (reconnus à travers les marqueurs de modalité) ou lexicaux.

4.2 Évaluation

Le corpus doit être échantillonné en deux parties : un échantillon étant réservé à l'apprentissage, l'autre au test. Pour cela, nous avons utilisé la méthode dite *par validation croisée* (*N-fold cross validation*)(Cornuéjols & Miclet, 2002).

Cette méthode consiste à diviser le corpus d'apprentissage en n sous-échantillons de tailles égales. On retient ensuite un des n échantillons (celui de numéro i) qui sera utilisé pour la phase de test, les autres servant à l'apprentissage. Une fois les résultats collectés, on réitère cette opération en faisant varier i de 1 à n. Nous avons choisi de poser $n = 5$. Nous avons donc 80% de nos documents (en terme de caractères) dans l'échantillon d'apprentissage, les autres 20% dans l'échantillon de test. Les résultats présentés dans la section 5 sont les moyennes sur

les 5 partitionnements.

Par ailleurs, nous utilisons les métriques de précision et de rappel pour évaluer nos résultats :
– le rappel, correspondant au nombre de documents correctement classés dans une classe C sur le nombre de documents appartenant à cette classe ;
– la précision, correspondant au nombre de documents correctement classés dans la classe C sur le nombre de documents classés dans la classe C.

5 Analyse et discussion des résultats

Nous avons donc appliqué les algorithmes *SVMlight* et *C4.5* à notre corpus. Les résultats de la classification pour ces deux algorithmes figurent dans les tableaux 5 et 6.

		Français		Japonais		Russe	
		Préc.	Rapp.	Préc.	Rapp.	Préc.	Rapp.
svml	Scientifique	1,00	0,36	0,20	0,41	1,00	0,52
	Vulgarisé	0,80	1,00	0,72	0,80	0,75	1,00
c4.5	Scientifique	0,89	0,80	0,13	0,12	0,50	0,38
	Vulgarisé	0,91	0,94	0,84	0,86	0,74	0,82

TAB. 5 – Précision et rappel pour chaque catégorie de critères obtenus pour les algorithmes SVM light et C4.5 sur les trois langues

		Français		Japonais		Russe	
		Préc.	Rapp.	Préc.	Rapp.	Préc.	Rapp.
svml	Caractéristiques structurelles	0,90	0,67	0,59	0,71	0,85	0,74
	Caractéristiques modales	0,60	0,50	0,50	0,49	0,28	0,50
	Caractéristiques lexicales	0,91	0,75	0,58	0,53	0,98	0,97
c4.5	Caractéristiques structurelles	0,85	0,85	0,41	0,44	0.62	0,68
	Caractéristiques modales	0,89	0,91	0,39	0,44	0,34	0,68
	Caractéristiques lexicales	0,85	0,85	0,47	0,45	0,45	0,52

TAB. 6 – Résultats pour chaque catégorie de critères

Selon le tableau 5, quels que soient la langue et l'algorithme, les documents du discours vulgarisé sont toujours mieux catégorisés. On remarque ainsi que les résultats obtenus avec les documents en français sont dans l'ensemble assez satisfaisants, avec un rappel moyen de 87%, et une précision moyenne de 90% toutes catégories confondues pour le classifieur C4.5 (soit plus de 200 documents bien classés parmi les 250 du corpus). Les résultats de la classification en japonais sont bons pour les documents du discours vulgarisé, mais assez médiocres pour les documents scientifiques. Finalement, avec les documents russes, on obtient de meilleurs résultats avec le système SVMlight, avec un rappel supérieur à 75 % et une précision de 87%. En ce qui concerne les performances plus faibles de la catégorisation des documents scientifiques en japonais et en russe, cela peut être expliqué par la plus forte proportion de documents vulgarisés pour ces langues, au détriment des documents scientifiques (voir le tableau 1). La normalisation des vecteurs d'apprentissage permet de pallier ces trop grandes variations mais il est évident

qu'un corpus d'apprentissage plus grand contient des données plus variées et permet de générer un modèle de la langue plus complet.

Par contre, ce résultat est assez surprenant pour le français : le nombre d'occurrences dans les documents scientifiques y est presque deux fois plus important, même si le nombre de documents scientifiques reste inférieur. Quant aux résultats globalement moins performants en japonais, ils peuvent s'expliquer par une forte disparité dans les genres des documents. Par conséquent, il est plus difficile de caractériser ce type de discours par un ensemble de caractéristiques "stable". Les critères utilisés lors de la classification manuelle (voir sec. 2.1) ne sont peut être pas suffisants dans cette langue et devraient être affinés.

Dans le tableau 6 figurent les résultats obtenus par chacun des deux algorithmes en fonction des catégories de la typologie. Chacune de ces catégories montre son importance dans la typologie. En effet, quel que soit le type de classifieur, les résultats obtenus dans chaque langue indiquent qu'il est possible de classer correctement plus de la moitié de nos documents en tenant compte des critères n'appartenant qu'à une seule catégorie. Cependant, aucune des catégories ne se distingue de façon significative dans ces tests. Par ailleurs, les meilleures catégories ne sont pas les mêmes selon le classifieur utilisé. Ainsi, avec *SVMlight*, ce sont les caractéristiques structurelles et lexicales qui se montrent les plus efficaces dans les trois langues. Par contre, avec *C4.5*, chaque langue privilégie des caractéristiques différentes : la modalité en français, le lexique en japonais et la structure des documents en russe. De manière plus détaillée, nous avons pu également constater que, parmi les critères les plus discriminants, on trouve le type d'URL (format de l'URL : site hospitalier, universitaire, etc.), les pronoms délocutifs et le nombre de phrases narratives pour le corpus français ; la longueur des documents, les pronoms élocutifs et les formules de politesse en fin de phrases pour le japonais ; le titre des documents, les images et le vocabulaire spécialisé pour le corpus russe.

6 Conclusion et perspectives

En partant d'un corpus comparable composé de documents issus du Web en français, japonais et russe, nous avons mené une analyse stylistique et contrastive et avons élaboré une typologie pour la caractérisation des discours scientifique et vulgarisé sur le Web. Cette typologie se base sur trois aspects des documents du Web : l'aspect structurel, l'aspect modal et l'aspect lexical. Notre typologie, implémentée grâce aux algorithmes d'apprentissage des séparateurs à vastes marges (SVMlight) ainsi qu'aux arbres de décision (C4.5) donne des résultats de classification satisfaisants. Nous pouvons ainsi "calculer" le type de discours d'un document du Web. Ces résultats montrent en outre que chacune des catégories de la typologie est discriminante pour le système de catégorisation. En effet, chacune d'elles permet d'obtenir des résultats acceptables, tandis que leur combinaison permet de les améliorer. Le discours d'un document du Web peut donc être caractérisé selon ces trois aspects. Nous remarquons par ailleurs que certains des critères de la typologie sont présents quelle que soit la langue, ce qui permet de statuer sur leur caractère "universel". Il semble donc qu'il existe des éléments communs dans les documents en langues différentes qui permettent d'indiquer leur type de discours et que, de manière générale, cette typologie est portable d'une langue à une autre. Une des limites principales, observable dans notre travail, semble provenir de la taille insuffisante des corpus. Par exemple, pour la catégorisation des documents scientifiques en japonais et en russe. Par ailleurs, les faibles résultats de la classification des documents scientifiques japonais sont dus également à la diversité des genres dans ce corpus. Cette constatation nous amène à nous interroger sur la composition et

la catégorisation manuelle du corpus japonais. Ainsi, suite à la classification hiérarchique des notions de genres et de discours de (Malrieu & Rastier, 2002), on peut se demander si cette distinction ne permettrait pas d'affiner nos résultats. En respectant la distinction des genres, nous pouvons rendre les catégories d'apprentissage plus homogènes et garantir ainsi un plus fort degré de comparabilité dans les corpus. De plus, suite à notre travail, nous nous interrogeons sur la légitimité de la catégorisation binaire effectuée. Nous pensons qu'il pourrait être intéressant de considérer les catégories scientifiques et vulgarisées des documents comme un continuum. Ceci conduirait à attribuer un "degré de vulgarisation" à chaque document plutôt qu'un type de discours.

Remerciements

Ce travail a été mené dans le cadre du projet DECO, programme CNRS-TCAN 2004-2006 en partenariat avec le NII et l'INaLCO. Nous remercions Estelle Dubreil, Sonia Krivine et Masaru Tomimitsu pour leur participation à la construction du corpus comparable.

Références

BIBER D. (1989). A typology of english texts. *Linguistics*, **27**, 3–43.

BOWKER L. & PEARSON J. (2002). *Working with Specialized Language : A Practical Guide to Using Corpora*. London/New York, Routeledge.

BRETAN I., DEWE J., HALLBERG A., WOLKERT N. & KARLGREN J. (1998). Web-specific genre visualisation. In *Proceedings of the 3rd World Conference on the WWW and Internet*.

CHARAUDEAU P. (1992). *Grammaire du sens et de l'expression*. Hachette.

CORNUÉJOLS A. & MICLET L. (2002). *Apprentissage artificiel : Concepts et algorithmes*. Eyrolles.

DUCROT O. & SCHAEFFER J.-M. (1999). *Nouveau dictionnaire encyclopédique des sciences du langage*. Seuil.

JOACHIMS T. (2002). *Learning to Classify Text using Support Vector Machines*. Kluwer.

KARLGREN J. (1998). *Natural Language Information Retrieval*, chapter Stylistic Experiments in Information Retrieval. Tomek, Kluwer.

KRIVINE S., TOMIMITSU M., GRABAR N. & SLODZIAN M. (2006). Relever des critères pour la distinction automatique entre les documents médicaux scientifiques et vulgarisés en russe et en japonais. In P. MERTENS, C. FAIRON, A. DISTER & P. WATRIN, Eds., *Actes de la 13ème conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, volume vol.1, p. 522–531, Leuven.

MALRIEU D. & RASTIER F. (2002). Genres et variations morphosyntaxiques. *TAL*, **42**(2), 548–577.

NAKAO Y. (2006). étude sémantico-discursive contrastive d'un corpus comparable français-japonais. Master's thesis, Université de Nantes.

QUINLAN J. R. (1993). *C4.5 : Programs for Machine Learning*. San Francisco, CA, USA : Morgan Kaufmann Publishers.

SEBASTIANI F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, **34**(1), 1–47.

SINCLAIR J. (1996a). *Preliminary recommendations on Corpus Typology*. Rapport interne, EAGLES (Expert Advisory Group on Language Engineering Standards).

SINCLAIR J. (1996b). *Preliminary recommendations on Text Typology*. Rapport interne, EAGLES (Expert Advisory Group on Language Engineering Standards).