# Reconnaissance du type de discours dans des corpus comparables spécialisés

**Lorraine Goeuriot, Emmanuel Morin et Béatrice Daille**

*Université de Nantes, LINA - UMR CNRS 6241*
*{lorraine.goeuriot,emmanuel.morin,beatrice.daille}@univ-nantes.fr*

RÉSUMÉ. *Notre objectif est d'automatiser la construction de corpus comparables spécialisés à partir du Web. La comparabilité se base sur trois niveaux : le domaine, le thème et le type de discours. Le domaine et le thème peuvent être filtrés grâce aux mots-clés utilisés lors de la recherche. Nous présentons dans cet article la reconnaissance automatique du type de discours dans des documents spécialisés français et japonais, qui nécessite une analyse linguistique poussée. Une analyse contrastive des documents nous permet de déterminer quelles informations paraissent discriminantes. En s'inspirant des travaux classiques de recherche d'information, nous créons une typologie robuste et linguistiquement motivée basée sur trois niveaux d'analyse : structurel, modal et lexical. Cette typologie nous permet d'apprendre des modèles de classification qui donnent de bons résultats, ce qui montre l'efficacité de cette typologie.*

ABSTRACT. *Our goal is to automate the compilation of smart specialized comparable corpora. The comparability is based on three levels: domain, topic and type of discourse. Domain and topic can be filtered with the keywords used through web search. We present in this paper the automatic detection of the type of discourse in French and Japanese documents, which needs a wide linguistic analysis. A contrastive analysis of the documents leads us to specify which information is relevant to distinguish them. Referring to classical studies on information retrieval, we create a robust and linguistically motivated typology based on three analysis levels: structural, modal and lexical. This typology is used to learn classification models using shallow parsing. We obtain good results, that demonstrates the efficiency of this typology.*

MOTS-CLÉS : *classification automatique, type de discours, typologie multilingue, corpus comparables*

KEYWORDS: *automatic classification, type of discourse, multilingual typology, comparable corpora*

Lorraine Goeuriot, Emmanuel Morin et Béatrice Daille

## 1. Introduction

L'exploitation de corpus comparables est un domaine de recherche récent, qui vise à suppléer les inconvénients liés à l'utilisation de corpus parallèles notamment lorsqu'il s'agit de travailler avec un couple de langues ne faisant par intervenir l'anglais. Les corpus comparables sont principalement utilisés pour extraire des terminologies multilingues (Déjean *et al.*, 2002, Morin *et al.*, 2007) ou des lexiques multilingues (Fung *et al.*, 1998, Rapp, 1999). Ils représentent aussi une ressource précieuse dans le cadre d'études contrastives multilingues (Peters *et al.*, 1997) et permettent aux traducteurs (Laviosa, 1998) et enseignants d'observer la langue dans son usage.

La profusion de documents accessibles dans des langues variées sur le web incite à puiser dans ce réservoir pour constituer des corpus comparables. Néanmoins, cette tâche ne saurait se réduire à la simple collecte de documents partageant un vocabulaire commun. Il est nécessaire de respecter des caractéristiques communes telles que le thème et le domaine (Bowker *et al.*, 2002) qui sont fixées avant la construction du corpus et qui sont fonction de sa finalité (McEnery *et al.*, 2007). De nombreux travaux traitent de la construction de corpus à partir du Web (Baroni *et al.*, 2006, Chakrabarti *et al.*, 1999) mais aucun, à notre connaissance, n'est consacré à celle des corpus comparables, qui doit répondre à différentes contraintes. Nous fixons ainsi la comparabilité à trois niveaux : le domaine, le thème et le type de discours.

L'objectif que nous poursuivons dans cette étude vise la constitution automatique de corpus comparables spécialisés à partir de documents issus du web pour des couples de langues à grande distance linguistique. Plus précisément, nous cherchons à rendre opérationnelle la précédente notion de comparabilité. Le domaine et thème d'un document pouvant être filtrés grâce aux mots-clés lors de la recherche (Chakrabarti *et al.*, 1999), nous nous concentrons ici sur la reconnaissance automatique des types de discours des domaines de spécialité : scientifique et vulgarisé. Pour ce faire, nous mettons en évidence un ensemble de critères, linguistiquement motivés, discriminants et opératoires pour caractériser les types de discours scientifique et vulgarisé. Ces critères implémentés au sein d'un système de classification automatique permettent de créer un corpus comparable français/japonais spécialisé dont la qualité avoisine celle obtenue manuellement.

La suite de cette étude est structurée comme suit. Après une introduction des travaux relatifs à l'exploitation de corpus comparables dans la section 2, l'analyse stylistique effectuée sur notre corpus d'étude est présentée dans la section 3. Elle nous permet de créer une typologie des types de discours scientifique et vulgarisé dans des domaines de spécialité. L'application d'algorithmes d'apprentissage à celle-ci est décrite dans la section 4 et ses résultats dans la section 5.

## 2. Contexte

« A comparable corpus can be defined as a corpus containing components that are collected using the same sampling frame and similar balance and representative-

ness » (McEnery *et al.*, 2007, p. 20). La comparabilité est garantie grâce à des caractéristiques pouvant référer au contexte de création des documents (période, auteur...) ou aux documents eux-mêmes (thème, genre...). Le choix des caractéristiques communes, qui définissent le contenu du corpus, influent sur le *degré de comparabilité*, notion permettant de quantifier dans quelle mesure deux corpus sont comparables. Ce choix dépend des objectifs applicatifs du corpus, nous distinguons dans les travaux sur ces corpus deux types :

**Les corpus comparables généralistes** : composés généralement d'articles de journaux. Les documents sont souvent extraits de journaux nationaux, et portent sur une même période, voire une même thématique. Fung *et al.* (1997), par exemple, utilisent un corpus anglais/japonais composé d'articles extraits du Wall Street Journal et du Nikkei Financial News (journaux traitant du domaine financier) sur une même période. Rapp (1999) utilise lui aussi des articles extraits de grands journaux nationaux allemand et anglais sur une même période, mais sans cibler de domaine particulier.

**Les corpus comparables spécialisés** : composés de documents émanant d'un domaine spécialisé, souvent scientifique, faisant appel à un langage spécialisé. Déjean *et al.* (2002) utilisent par exemple un corpus composé de documents médicaux tirés de la base de données médicales MEDLINE, ainsi que Chiao (2004) , utilisant les bases CISMEF, CLINIWEB et OSHUMED.

Dans les travaux sur la langue générale, les documents partagent souvent des caractéristiques telles que le domaine et le thème. Étant souvent extraits de journaux périodiques, il est important de les limiter à une certaine période afin de garantir la comparabilité. Dans les travaux sur les langues de spécialité, un premier niveau de comparabilité peut être assuré grâce au domaine ou au thème. De plus, différentes situations de communication peuvent être observées dans les langues de spécialité (Pearson, 1998, p. 36) : la communication d'expert à expert, d'expert à initié, de semi-expert à non-initié, d'enseignant à élève... Malrieu *et al.* (2002) relèvent différents niveaux de classification de textes, chaque niveau correspondant à une certaine granularité. Le premier niveau est le *discours*, défini comme un ensemble d'énoncés d'un énonciateur caractérisé par une unité globale de thème (Ducrot *et al.*, 1999). Le second niveau est le *genre*, défini comme les catégories de textes distinguées spontanément par les locuteurs d'une langue. Par exemple, au discours littéraire correspondent les genres : théâtre, poésie, récit... En s'inspirant de ces situations de communication et ces niveaux de classification de textes, nous avons choisi de distinguer deux situations de communication, que nous appelons *types de discours*, dans les domaines de spécialité : scientifique (textes écrits par des spécialistes à destination de spécialistes) et vulgarisé (textes écrits pour des non-spécialistes par des non-spécialistes, semi-spécialistes ou spécialistes). Ce niveau de comparabilité, le type de discours, reflète le « contexte de production ou d'usage » des documents (Habert *et al.*, 2001) et garantit une homogénéité lexicale dans le corpus (Bowker *et al.*, 2002, p. 27). De plus, Morin *et al.* (2007) montrent qu'un corpus comparable dont les documents partagent un thème et un type de discours est très adapté à l'extraction de terminologies multilingues.

Dans cette étude, nous nous intéressons à la catégorisation automatique de documents selon leur type de discours. Elle est basée sur une typologie composée de critères caractérisant ce type de discours, élaborée grâce à une analyse stylistique contrastive (Karlgren, 1998). Son but est de trouver des critères linguistiquement motivés, correspondant à différents niveaux d'analyse, dont la combinaison caractérise un type de discours.

## 3. Analyse des types de discours

La première étape de cette analyse des types de discours est une analyse stylistique manuelle, basée sur les méthodes déductives et contrastives, dont le but est de mettre en évidence des critères discriminants et linguistiquement motivés caractérisant les types de discours scientifique et vulgarisé. La principale difficulté de cette tâche réside dans la recherche de critères pertinents adaptés à chaque langue. Ces critères sont ensuite rassemblés dans une typologie qui sera utilisée afin d'apprendre des modèles de classification. Celle-ci se devra d'être robuste, générique et extensible à d'autres langues. La généricité sera garantie par une typologie couvrant une grande variété de caractéristiques textuelles, la robustesse par des critères opératoires et un traitement aisément adaptable aux textes comme aux documents du Web.

Sinclair (1996) distingue deux niveaux d'analyse dans son rapport sur les typologies textuelles : un niveau externe, concernant le contexte de création des textes et un niveau interne, correspondant aux caractéristiques linguistiques des textes. Nos corpus étant composés de documents extraits du Web, nous considérons les critères du niveau externe comme tous les critères liés à la création des documents et à leur structure (critères non-linguistiques), nous les appelons *critères structurels*. Ke *et al.* (2009) utilisent les fréquences des mots et caractères ainsi que certaines informations lexicales et structurelles afin de distinguer les types de discours en chinois. Notre analyse stylistique a permis de mettre en évidence différents niveaux de granularité dans les critères linguistiques, le niveau d'analyse interne est donc composé de deux catégories. Pour distinguer les deux niveaux de communication que sont nos types de discours, nous devons tout d'abord considérer le locuteur dans son discours : les critères modaux (Nakao, 2008). De plus, le discours scientifique peut être caractérisé par son vocabulaire, la longueur des mots et autres critères lexicaux. Notre typologie est donc composée de trois niveaux d'analyse : structurel, modal et lexical.

### 3.1. *Les critères structurels*

Nos documents étant extraits du Web, nous devons considérer leur structure et le contexte de leur création. Dans le cadre de la classification de documents du Web, différents éléments apportent des informations pertinentes : les images, les vidéos et d'autres contenus multimédia (Asirvatham *et al.*, 2001) ; les méta-informations, le titre et la structure `HTML` (Riboni, 2002). Les critères structurels de notre typologie sont : le patron d'URL, le format du document, les balises `META`, la balise `TITLE`, la mise en

page (utilisation de CSS, cadres, tableaux…), le fond des pages, les images, les liens, les paragraphes, les listes, le nombre de phrases, la typographie (italique, gras…) et la longueur des documents (nombre de caractères).

### 3.2. *Les critères modaux*

Le degré de spécialisation requis par le lecteur ou l'interlocuteur est caractérisé par la relation établie dans l'énoncé entre le locuteur ou l'auteur et l'interlocuteur ou le lecteur[1]. Cette relation est caractérisée par le ton du locuteur et par l'emploi de certains traits linguistiques. La modalisation est une interprétation de l'attitude du locuteur vis-à-vis du contenu de son discours (Querler, 1996), elle est caractérisée par différents marqueurs textuels : les verbes, les adverbes, les formules de politesse… La plupart des théories de la modalisation sont dépendantes de la langue et font appel à une description des phénomènes spécifiques à chaque langue. Pourtant, la théorie de Charaudeau (1992) semble indépendante de la langue et opérationelle pour les langues française et japonaise. Selon lui (Charaudeau, 1992, p. 572), la modalisation permet d'expliciter au sein d'un énoncé la position du locuteur par rapport à l'interlocuteur, à lui-même et à son discours. Elle est composée d'actes locutifs, qui sont des positions particulières du locuteur dans son discours. Chacun d'entre eux est caractérisé par différentes modalités. Nous en observons deux dans cette théorie :

**L'acte allocutif** : le locuteur implique l'interlocuteur dans son discours (ex. : « *Tu dois le faire* ») ;

**L'acte élocutif** : le locuteur est impliqué dans son discours, il révèle sa propre position (ex. : « *J'aimerais le faire* »).

Les modalités sont présentées dans le tableau 1 avec des exemples en français. Certaines d'entre elles ne sont pas utilisées dans un langage ou l'autre si elles sont trop peu fréquentes ou trop ambiguës.

### 3.3. *Les critères lexicaux*

Biber (1988, 1989) utilise des informations lexicales afin d'observer les variations entre des textes et plus particulièrement entre leurs genres et leurs types. Karlgren *et al.* font appel aux critères lexicaux afin de caractériser les genres textuels et observer les variations stylistiques entre textes. Nous considérons ainsi que les informations lexicales peuvent être pertinentes dans la distinction des types de discours scientifique et vulgarisé. La première raison est que le vocabulaire spécialisé est l'une des principales caractéristiques des textes issus de domaines de spécialité (Bowker *et al.*, 2002, p. 26). De plus, les documents scientifiques contiennent davantage d'unités lexicales complexes, groupes nominaux ou phrases nominales que les documents vulgarisés

1. Comme nous travaillons sur des domaines de spécialité, nous considérons le locuteur comme auteur des textes et l'interlocuteur comme lecteur.

| **Modalité** | **Exemple** | **Français** | **Japonais** |
|---|---|---|---|
| **Acte allocutifs** | | | |
| Pronoms personnels allocutifs | *Tu, vous* | × | |
| Injonction | *Ne fais pas ça* | × | × |
| Autorisation | *Tu peux le faire* | × | |
| Jugement | *Bravo, tu as réussi !* | × | |
| Suggestion | *Tu devrais le faire* | × | × |
| Interrogation | *Quand arrives-tu ?* | × | × |
| Interpellation | *Comment allez-vous, monsieur ?* | × | |
| Requête | *S'il vous plaît, faites-le* | × | × |
| **Actes élocutifs** | | | |
| Pronoms personnels élocutifs | *Je, nous, on* | × | × |
| Constat | *Je remarque qu'il est parti* | × | × |
| Connaissance | *Nous savons qu'il est parti* | × | × |
| Opinion | *Je pense qu'il est parti* | × | × |
| Volonté | *Je voudrais qu'il parte* | × | × |
| Promesse | *Je te promets qu'il sera là* | × | × |
| Déclaration | *Je t'assure qu'il est parti* | | × |
| Appréciation | *Je l'aime bien* | × | |
| Obligation | *Nous devons le faire* | × | |
| Possibilité | *Je peux le leur dire* | × | |

**Tableau 1.** *Critères modaux*

(Sager, 1990). Nous présentons dans le tableau 2 les critères lexicaux. Il est à noter que ceux-ci sont plus dépendants de la langue que les critères présentés précédemment.

| **Critère** | **Français** | **Japonais** |
|---|---|---|
| Vocabulaire spécialisé | × | × |
| Caractères numériques | × | × |
| Unités de mesure | × | × |
| Longueur des mots | × | |
| Bibliographie | × | × |
| Citations bibliographiques | × | × |
| Ponctuation | × | × |
| Fins de phrase | | × |
| Parenthèses | × | × |
| Autres alphabets (latin, hiragana, katakana) | | × |
| Symboles | | × |

**Tableau 2.** *Critères lexicaux*

## 4. Reconnaissance automatique du type de discours

L'élaboration d'un système de classification automatique est réalisée en trois étapes : l'indexation des documents, l'apprentissage du classifieur et son évaluation (Sebastiani, 2005, p. 112, 113). L'indexation des documents consiste à générer une représentation compacte des documents pouvant être interprétée par un classifieur. Dans notre cas, chaque document $d_i$ est représenté par un vecteur de poids des critères :

$$\vec{d_i} = \{w_{1i}, \ldots, w_{ni}\}$$

où $n$ représente le nombre de critères de la typologie et $w_{ji}$ représente le poids du $j^{eme}$ critère dans le $i^{eme}$ document. Chaque poids de critère est normalisé, en le divisant par le total. L'indexation des documents est ici effectuée grâce à la typologie (cf. section 3) et à l'implémentation de ces critères.

L'implémentation des critères de notre typologie se fait au moyen de patrons lexico-syntaxiques (*i.e.* des expressions régulières).

### 4.1. *Critères structurels*

La majorité des critères structurels (présentée en section 3.1) est implémentée par des opérations de recherche de motifs. Par exemple, le patron d'URL permet de déterminer si un document est issu d'un site hospitalier (`http://www.chu-***.fr`) ou d'un site universitaire (`http://www.univ-***.fr`)... Quant aux images, paragraphes, liens, etc., une simple recherche de balises a été effectuée.

### 4.2. *Critères modaux*

Les marqueurs de présence du locuteur dans un texte peuvent être implicites ou ambigus. Nous avons préféré utiliser des marqueurs simples afin d'éviter d'introduire trop de bruit dans notre système (précision forte). Nous y introduisons toutefois du silence (rappel fort) : toutes les occurrences d'une modalité ne sont pas détectées mais celles qui le sont sont correctes. Certains pronoms sont spécifiques à l'acte locutif : par exemple, les pronoms français *je* et *nous*, et les japonais 私 (je), 私達 (nous) et 我々 (nous) sont caractéristiques de l'acte élocutif. Nous utilisons de plus des marqueurs lexicaux : par exemple, la modalité du savoir peut être détectée en français grâce aux verbes *savoir, connaître* et en japonais avec le verbe 知る (savoir), dans la forme polie 知っています et dans la forme neutre 知っている.

### 4.3. *Critères lexicaux*

Nous présentons ici l'implémentation des onze critères lexicaux introduits dans le tableau 2. Certains d'entre eux sont spécifiques aux documents scientifiques, comme

les bibliographies, citations bibliographiques ou vocabulaire spécialisé. Pour mesurer la densité terminologique en français (proportion de vocabulaire spécialisé dans un texte), nous recherchons des affixes gréco-latins (Namer *et al.*, 2007) et des adjectifs relationnels particulièrement fréquents dans les domaines scientifiques (Daille, 2000). Nous avons dénombré près de 50 affixes tels que *inter-*, *auto-* ou *nano-* et 10 suffixes relationnels tels que *-ique* ou *-al*. Ces affixes peuvent être présents dans les deux types de discours, mais dans des proportions différentes. Par exemple, le terme *ovariectomie* peut être fréquent dans un document scientifique tandis qu'il sera très rarement employé dans un document vulgarisé, et ce au profit du terme *ablation des ovaires*. Les fins de phrases sont des particules de terminaison spécifiques, par exemple la particule か qui est souvent utilisée à la fin des phrases interrogatives.

### 4.4. *Algorithmes de classification automatique*

La classification automatique est un processus qui, partant d'un ensemble de vecteurs dans une classe $c$ ou $\overline{c}$, détermine quelles caractéristiques doit avoir un nouveau document pour être classé dans l'une de ces classes[2]. À partir d'une indexation de documents, il existe plusieurs algorithmes permettant de réaliser ce processus (les réseaux de neurones, les classifieurs bayesiens, les machines à vecteurs de support…) dont (Sebastiani, 2002) a mené une comparaison. Appliquées à des corpus de dépêches Reuters, ces méthodes donnent des résultats variables selon le nombre de classes, de critères… Dans cette étude, les systèmes *SVMlight* (Joachims, 2002) et *C4.5* (Quinlan, 1993) donnent de très bons résultats dans un contexte similaire au notre : petits corpus, classification binaire, moins de 100 critères.

## 5. Expérimentations

Nous décrivons dans cette section les deux corpus comparables que nous avons utilisés et présentons les expériences menées sur ceux-ci. Le premier corpus est utilisé afin d'apprendre un modèle de classification basé sur notre typologie (la phase d'apprentissage), tandis que le second corpus sert à évaluer ce modèle de classification sur de nouveaux documents (la phase d'évaluation).

### 5.1. *Corpus comparables*

Les corpus utilisés dans nos expériences sont composés de documents français et japonais extraits du Web. Ils sont issus du domaine médical, sur les thématiques *diabète et alimentation* pour la phase d'apprentissage et *cancer du sein* pour la phase d'évaluation. La collecte des documents a été menée manuellement. Leur domaine et leur thématique ont été filtrés grâce aux mots-clés : par exemple, *alimentation, diabète*

2. Dans le cas binaire ; voir (Sebastiani, 2005) pour les autres cas.

et *obésité* pour la partie française et 糖尿病 (*diabète*) et 肥満 (*surpoids*) pour la partie japonaise du corpus d'apprentissage. Les documents ont ensuite été manuellement sélectionnés puis classés par des locuteurs natifs de chaque langue, qui ne sont pas des spécialistes du domaine médical, selon leur type de discours : scientifique (SC) ou vulgarisé (VU). La classification manuelle se base sur les heuristiques suivantes :

– Un document scientifique est écrit par des spécialistes, à destination de spécialistes.

– En ce qui concerne les documents vulgarisés, nous distinguons deux niveaux de vulgarisation : les documents écrits par des spécialistes pour le grand public et les documents écrits par le grand public pour le grand public. Nous ne distinguons pas ici ces deux niveaux mais accordons toutefois plus d'importance aux documents écrits par des spécialistes, potentiellement plus riches en contenu et en vocabulaire (les conseils d'un médecin à ses patients peuvent être plus riches qu'une discussion de forum).

Notre classification manuelle des documents se base donc sur ces deux heuristiques, ainsi que sur différents éléments empiriques : l'origine du site Web, le vocabulaire employé... Pour quelques documents, il a été difficile de déterminer le type de discours (par exemple des documents écrits par des personnes dont le degré de spécialisation n'était pas clair). Ils n'ont pas été conservés dans le corpus.

Nous avons donc créé deux corpus comparables :

– [DIABÈTE] portant sur le thème *diabète et alimentation* et utilisé lors de la phase d'apprentissage.

– [CANCER] portant sur le thème *cancer du sein* et utilisé lors de la phase d'évaluation.

Le tableau 3 présente les principales caractéristiques de chaque corpus : le nombre de documents et le nombre de mots[3] pour chaque langue et chaque type de discours.

| | | | # doc. | # mots |
|---|---|---|---|---|
| [DIABÈTE] | FR | SC | 65 | 425 781 |
| | | VU | 183 | 267 885 |
| | JP | SC | 119 | 234 857 |
| | | VU | 419 | 572 430 |
| [CANCER] | FR | SC | 50 | 443 741 |
| | | VU | 42 | 71 980 |
| | JP | SC | 48 | 211 122 |
| | | VU | 51 | 123 277 |

**Tableau 3.** *Principales caractéristiques des deux corpus comparables*

3. Pour le japonais, le nombre de mots est le nombre d'occurrences reconnues par ChaSen (Matsumoto *et al.*, 1999)

### 5.2. *Résultats de la phase d'apprentissage*

Dans cette première expérience, nous entraînons et testons nos classifieurs sur le corpus [DIABÈTE]. Nous utilisons la méthode par validation croisée (*N-fold cross validation method*) qui consiste à diviser le corpus en $n$ partitions de même taille. Si nous fixons $n = 5$, à chaque itération, le sous-corpus d'apprentissage compte 80 % des documents du corpus initial (en terme de caractères) et les 20 % restants (correspondant à la $i^{eme}$ partition) sont utilisés pour l'évaluation. Les résultats que nous donnons sont des moyennes sur ces 5 partitions et nous utilisons les métriques de précision et de rappel pour évaluer l'efficacité des classifieurs :

$$\text{Précision} = \frac{\text{\# doc. correctement classés dans } c}{\text{\# doc. classés dans c}}$$

$$\text{Rappel} = \frac{\text{\# doc. correctement classés dans } c}{\text{\# doc. appartenant à c}}$$

Les résultats obtenus avec les systèmes *SVMlight* et *C4.5* sur le corpus [DIABÈTE] sont présentés dans le tableau 4. Notre mesure de référence est la suivante : nous considérons pour chaque classe (scientifique ou vulgarisée) que 50 % des documents lui appartenant sont correctement classés. Ainsi, le rappel est toujours de 50 % tandis que la précision varie : elle est faible pour les documents scientifiques et satisfaisante pour les documents vulgarisés. Nous pouvons constater que quels que soient la langue et le système de classification, notre méthode donne des résultats corrects sur les documents vulgarisés. Nous améliorons le rappel et la précision de la méthode de référence dans quasiment tous les cas de figure. En ce qui concerne les documents scientifiques, nos résultats sont bien meilleurs que ceux de référence pour le français comme pour le japonais avec le système *C4.5*. En revanche, les résultats obtenus avec le système *SVMlight* sont plus diffus, notamment en ce qui concerne le rappel des documents français.

Si nous ne tenons pas compte de la distinction des documents selon le type du discours, les résultats obtenus en français sont globalement satisfaisants avec un rappel moyen de 87 % et une précision moyenne de 90 % avec le système *C4.5* (plus de 215 documents sur 248 sont correctement classés). Les résultats de la classification des documents japonais sont bons avec le classifieur *C4.5* : plus de 90 % des documents sont correctement classifiés et la précision atteint en moyenne 80 %. Les résultats les plus faibles obtenus sur les documents japonais peuvent s'expliquer par la grande variété de genres dans ce corpus (articles de recherche, de journaux, recettes de cuisine, offres d'emploi, discussions de forums...).

Nous présentons dans le tableau 5 les résultats de la classification obtenus pour chaque catégorie de critères considérée indépendamment, avec les deux systèmes de classification sur le corpus [DIABÈTE]. Dans chaque cas, les représentations vectorielles des documents ne contiennent que les poids des critères de la catégorie concernée. Quel que soit le classifieur, nous n'observons pas de grande baisse des résultats

| | | Français | | Japonais | |
|---|---|---|---|---|---|
| | | Préc. | Rapp. | Préc. | Rapp. |
| *Mesure de référence* | SC | 0,26 | 0,50 | 0,22 | 0,50 |
| | VU | 0,74 | 0,50 | 0,78 | 0,50 |
| *SVMlight* | SC | 1,00 | 0,36 | 0,70 | 0,65 |
| | VU | 0,80 | 1,00 | 0,72 | 0,80 |
| *C4.5* | SC | 0,89 | 0,80 | 0,76 | 0,96 |
| | VU | 0,91 | 0,94 | 0,95 | 0,99 |

**Tableau 4.** *Précision et rappel pour chaque langue et classifieur pour le corpus* [*DIABÈTE*]

en ne conservant qu'une seule catégorie de critères. Par contre, les résultats sur les documents japonais sont inférieurs. Nous pouvons en déduire que la combinaison de chacune de ces catégories de la typologie permet d'améliorer les résultats de la classification. Cependant, aucune catégorie ne se distingue clairement dans cette expérience, les plus efficaces ne sont pas les mêmes selon le système utilisé et la langue. Avec *SVMlight*, les critères lexicaux et structuraux semblent les plus discriminants. Avec *C4.5*, les critères modaux donnent de meilleurs résultats sur les documents français, tandis que les critères lexicaux améliorent les résultats pour le japonais. Chaque catégorie semble discriminante pour une langue ou un système de classification et les expériences sur la typologie complète montrent que leur combinaison améliore les résultats.

| | | Français | | Japonais | |
|---|---|---|---|---|---|
| | | Préc. | Rapp. | Préc. | Rapp. |
| *SVMlight* | Structurel | 0.90 | 0.67 | 0.59 | 0.71 |
| | Modal | 0.60 | 0.50 | 0.50 | 0.49 |
| | Lexical | 0.91 | 0.75 | 0.58 | 0.53 |
| *C4.5* | Structurel | 0.85 | 0.85 | 0.41 | 0.44 |
| | Modal | 0.89 | 0.91 | 0.39 | 0.44 |
| | Lexical | 0.85 | 0.85 | 0.47 | 0.45 |

**Tableau 5.** *Résultats de chaque catégorie de critères sur le corpus* [*DIABÈTE*]

### 5.3. *Résultats de la phase d'évaluation*

Afin d'évaluer l'impact de l'application des modèles de classification générés sur de nouveaux documents, une nouvelle expérience a été menée : les classifieurs ap-

pris sur le corpus [DIABÈTE] sont testés sur le corpus [CANCER]. Les résultats sont présentés dans le tableau 6.

Nous notons une baisse globale des résultats de la classification sur ce corpus d'évaluation bien qu'ils restent satisfaisants. Les documents français sont classés avec une précision supérieure à 75% et un rappel de plus de 75%, ce qui représente plus de 70 documents correctement classés sur 92. La classification des documents japonais donne de bons résultats, avec une précision de 76% et un rappel de 77% en moyenne, ce qui représente 23 documents mal classés sur 99. Ces modèles de classification semblent donc efficaces pour distinguer les types de discours scientifique et vulgarisé dans des documents spécialisés français et japonais.

Selon les objectifs applicatifs du corpus, il peut être souhaitable de privilégier la précision ou le rappel. Par exemple, (Morin *et al.*, 2007) montrent qu'un corpus composé de documents scientifiques est plus adapté à l'extraction de termes complexes bilingues dans des domaines de spécialité qu'un corpus mêlant les deux types de discours. Dans ce cas, la précision doit être privilégiée au rappel et *SVMlight* le permet.

| | | Français | | Japonais | |
|---|---|---|---|---|---|
| | | Préc. | Rapp. | Préc. | Rapp. |
| *SVMlight* | SC | 0,92 | 0,53 | 0,90 | 0,61 |
| | VU | 0,64 | 0,95 | 0,66 | 0,98 |
| *C4.5* | SC | 0,70 | 0.92 | 0,76 | 0,70 |
| | VU | 0,87 | 0,56 | 0,75 | 0,80 |

**Tableau 6.** *Précision et rappel pour chaque langue et classifieur pour le corpus* [*CANCER*]

## 6. Conclusion

Dans cet article nous avons décrit une première étape de la construction automatique de corpus comparables spécialisés en français et en japonais. Une qualité proche des corpus construits manuellement est garantie par le choix des caractéristiques communes aux textes : un domaine, un thème et un type de discours. Une analyse stylistique contrastive nous a permis de créer une typologie composée de critères caractérisant les types de discours scientifique et vulgarisé dans des documents spécialisés issus du Web. Cette typologie est basée sur trois niveaux d'analyse des documents : le niveau structurel, le niveau modal et le niveau lexical. Ses critères ont été mis en œuvre et des modèles de classification ont été générés avec les systèmes *SVMlight* et *C4.5*. Ces derniers donnent de bons résultats, sur le corpus d'apprentissage ainsi que sur le corpus d'évaluation avec une précision moyenne de 80 % et un rappel moyen de 70 %.

Toutefois, l'aspect binaire de notre classification nous parait discutable. Il peut être en effet intéressant de considérer les classes scientifique et vulgarisée comme un continuum, ce qui nous mènerait à évaluer pour chaque document un degré de spécialisation plutôt qu'une appartenance à une classe. De plus, *SVMlight* attribue à chaque document un score, que nous interprétons comme l'appartenance à l'une des deux classes. Nous envisageons de considérer ces scores du point de vue du continuum. Nous pourrions ainsi distinguer un plus grand nombre de situations de communication : de spécialiste à spécialiste, de spécialiste à non-spécialiste, de non-spécialiste à non-spécialiste… (Pearson, 1998). Nous avons souhaité que notre typologie soit générique, de façon à pouvoir être adaptée à d'autres langues. Les critères étant déjà définis, il sera nécessaire pour ajouter une langue de trouver les marqueurs pour chacun des critères et de créer un corpus sur lequel un modèle sera appris pour cette nouvelle langue.

Remerciements

Ce travail a été mené dans le cadre du projet ANR C-Mantic 2007-2009. Nous remercions Yukie Nakao pour son travail sur la typologie et les marqueurs japonais.

## 7. Bibliographie

Asirvatham A. P., Ravi K. K., « Web Page Classification Based on Document Structure », *IEEE National Convention*, 2001.

Baroni M., Kilgarriff A., « Large Linguistically-Processed Web Corpora for Multiple Languages », *EACL'06*, The Association for Computer Linguistics, p. 87-90, 2006.

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

Biber D., « A typology of English texts », *Linguistics*, vol. 27, p. 3-43, 1989.

Bowker L., Pearson J., *Working with Specialized Language : A Practical Guide to Using Corpora*, London/New York, Routeledge, 2002.

Chakrabarti S., van den Berg M., Dom B., « Focused crawling : a new approach to topic-specific Web resource discovery », *Computer Networks (Amsterdam, Netherlands : 1999)*, vol. 31, n˚11–16, p. 1623-1640, 1999.

Charaudeau P., *Grammaire du sens et de l'expression*, Hachette, 1992.

Chiao Y.-C., Extraction lexicale bilingue à partir de textes médicaux comparables : application à la recherche d'information translangue, PhD thesis, Université Pierre et Marie Curie (Paris 6), juin, 2004.

Daille B., « Morphological Rule Induction for Terminology Acquisition », *COLING'00*, Sarrbrucken, Germany, p. 215-221, 2000.

Déjean H., Gaussier E., Sadat F., « An Approach Based on Multilingual Thesauri and Model Combination for Bilingual Lexicon Extraction », *COLING'02*, 2002.

Ducrot O., Todorov T., *Dictionnaire encyclopédique des sciences du langage*, Éditions du Seuil, 1999.

Fung P., McKeown K., « Finding terminology translations from non-parallel corpora », *Proceedings of the 5th annual workshop on very large corpora (VLC 97)*, Hong Kong, p. 192-202, 1997.

Fung P., Yee L. Y., « An IR Approach for Translating New Words from Nonparallel, Comparable Texts », *in* , C. Boitet, , P. Whitelock (eds), *COLING'98*, vol. 1, Montreal, Quebec, Canada, p. 414-420, 1998.

Habert B., Grabar N., Jacquemart P., Zweigenbaum P., « Building a text corpus for representing the variety of medical language », *in* , P. Rayson, , A. Wilson, , T. McEnery, , A. Hardie, , S. Khoja (eds), *Corpus Linguistics 2001*, Lancaster, p. 245-254, february, 2001.

Joachims T., *Learning to Classify Text using Support Vector Machines*, Kluwer Academic Publishers, 2002.

Karlgren J., *Natural Language Information Retrieval*, Tomek, Kluwer, chapter Stylistic Experiments in Information Retrieval, 1998.

Karlgren J., Cutting D., « Recognizing Text Genres with Simple Metrics Using Discriminant Analysis », *COLING'94*, vol. 2, Kyoto, Japan, p. 1071-1075, 1994.

Ke G., Zweigenbaum P., « Catégorisation automatique de pages web chinoises », *Actes de la 6ème Conférence en Recherche d'Informations et Applications (CORIA'09)*, 2009. À paraître.

Laviosa S., « Corpus-based Approaches to Contrastive Linguistics and Translation Studies », *Meta*, vol. 43, n˚4, p. 474-479, 1998.

Malrieu D., Rastier F., « Genres et variations morphosyntaxiques », *Traitement Automatique des Langues (TAL)*, vol. 42, n˚2, p. 548-577, 2002.

Matsumoto Y., Kitauchi A., Yamashita T., Hirano Y., Japanese Morphological Analysis System ChaSen 2.0 Users Manual, Technical report, Nara Institute of Science and Technology (NAIST), 1999.

McEnery A., Xiao Z., « Parallel and comparable corpora : What is happening ? », *in* , G. Anderman, , M. Rogers (eds), *Incorporating Corpora : The Linguist and the Translator*, Clevedon : Multilingual Matters, 2007.

Morin E., Daille B., Takeuchi K., Kageura K., « Bilingual Terminology Mining – Using Brain, not brawn comparable corpora », *ACL'07*, Prague, Czech Republic, p. 664-671, 2007.

Nakao Y., « Multilingual modalities for specialised languages », *Workshop on Multilingual and Comparative Perspectives in Specialized Language Resources (MCPSLR 2008), Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC 2008), Marrakech, Morocco, 26 May 2008*, European Language Resources Association (ELRA), 2008.

Namer F., Baud R., « Defining and relating biomedical terms : Towards a cross-language morphosemantics-based system », *International Journal of Medical Informatics*, vol. 76, n˚2-3, p. 226-233, 2007.

Pearson J., *Terms in Context*, John Benjamins publishing company, 1998.

Peters C., Picchi E., « Using Linguistic Tools and Resources in Cross-Language Retrieval », *in* , D. Hull, , D. Oard (eds), *Cross-Language Text and Speech Retrieval. Papers from the 1997 AAAI Spring Symposium, Technical Report SS-97-05*, p. 179-188, 1997.

Querler N. L., *Typologie des modalités*, Presses universitaires de Caen, Caen, 1996.

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

Rapp R., « Automatic Identification of Word Translations from Unrelated English and German Corpora », *ACL'99*, College Park, Maryland, USA, p. 519-526, 1999.

Riboni D., « Feature Selection for Web Page Classification », *in* , H. Shafazand, , A. M. Tjoa (eds), *Proceedings of the 1st EurAsian Conference on Advances in Information and Communication Technology (EURASIA-ICT)*, Springer, Shiraz, Iran, p. 473-478, 2002.

Sager J. C., *A Pratical Course in Terminology Processing*, John Benjamins, Amsterdam, 1990.

Sebastiani F., « Machine Learning in Automated Text Categorization », *ACM Computing Surveys*, vol. 34, n°1, p. 1-47, 2002.

Sebastiani F., « Text Categorization », *in* , A. Zanasi (ed.), *Text Mining and its Applications to Intelligence, CRM and Knowledge Management*, WIT Press, Southampton, UK, p. 109-129, 2005.

Sinclair J., Preliminary recommendations on Text Typology, Technical report, EAGLES (Expert Advisory Group on Language Engineering Standards), 1996.