Découverte et caractérisation des corpus comparables spécialisés

Lorraine Goeuriot

École Centrale de Nantes Université de Nantes École des Mines de Nantes

ÉCOLE DOCTORALE STIM
« SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE MATHÉMATIQUES »

Année 2009

Découverte et caractérisation des corpus comparables spécialisés

---

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE NANTES

Discipline : INFORMATIQUE

présentée et soutenue publiquement par

Lorraine GOEURIOT

le 30 Janvier 2009

au LINA

devant le jury ci-dessous

Président	:	Alexandre DIKOVSKY	Université de Nantes
Rapporteurs	:	Monique SLODZIAN, Professeur	INALCO
		Pierre François MARTEAU, Professeur	Université de Bretagne Sud
Examinateurs:		Béatrice DAILLE, Professeur	Université de Nantes
		Alexandre DIKOVSKY, Professeur	Université de Nantes
		Kyo KAGEURA, Professeur	Université de Tokyo
		Emmanuel MORIN, Maître de conférence	Université de Nantes

Directeur de thèse : Béatrice DAILLE
Encadrant de thèse : Emmanuel MORIN
Laboratoire : LINA (Laboratoire d'Informatique de Nantes Atlantique)

N° ED 0366-XXX

DÉCOUVERTE ET CARACTÉRISATION DES CORPUS COMPARABLES SPÉCIALISÉS

Specialized Comparable Corpora Discovery and Characterization

Lorraine GOEURIOT

favet neptunus eunti

Université de Nantes

Lorraine GOEURIOT

Découverte et caractérisation des corpus comparables spécialisés

xi+152 p.

Ce document a été préparé avec LaTeX2$_\varepsilon$ et la classe `these-IRIN` version 0.92 de l'association de jeunes chercheurs en informatique LOGIN, Université de Nantes. La classe `these-IRIN` est disponible à l'adresse :

`http://login.lina.sciences.univ-nantes.fr/`

Impression : these-lorraine.tex – 1/12/2008 – 16:47

Révision pour la classe : $Id: these-IRIN.cls,v 1.3 2000/11/19 18:30:42 fred Exp

Sommaire

Introduction

Contexte

Il existe autant de langues de spécialité que de domaines (Bowker et Pearson, 2002, p. 25). Chaque langue de spécialité possède des caractéristiques propres : syntaxe, terminologie... Les problématiques et concepts d'un domaine spécialisé sont internationaux mais les caractéristiques linguistiques ne le sont pas. De plus, le vocabulaire technique et scientifique n'est pas toujours présent dans les dictionnaires de langue générale et cela constitue un véritable problème lors de la traduction. Comment permettre aux scientifiques d'un domaine de communiquer d'une langue à une autre ? L'anglais, *lingua franca* des communications scientifiques, semble apporter une solution à ce problème. Cependant, il est toujours nécessaire de traduire la terminologie du domaine en anglais et les scientifiques et traducteurs se trouvent confrontés à ce problème. L'élaboration manuelle de dictionnaires bilingues pour les langues de spécialité semble impossible. En effet, les domaines sont très nombreux et les langues de spécialité sont en constante évolution et sont généralement propices aux néologismes. Traducteurs, terminologues, chercheurs en TAL... se sont alors intéressés à ce problème. L'une des solutions envisagées est alors d'observer les terminologies en situation, c'est-à-dire dans un corpus composé de documents issus du domaine de spécialité dans les langues étudiées. Les traducteurs et terminologues y analysent manuellement les terminologies d'une langue à l'autre, leur contexte d'usage... Quant aux informaticiens, ils extraient de manière automatique des informations multilingues de ces corpus. À partir d'une collection de textes issus d'un domaine de spécialité et sans autre connaissance sur ce domaine il est possible d'extraire des lexiques multilingues et ce pour la plupart des langues.

Objet de l'étude

Les corpus multilingues sont des corpus composés de textes en deux ou plusieurs langues. Ces corpus sont utilisés par différentes communautés, avec différents objectifs. Les traducteurs les utilisent comme aide à la traduction car ils permettent d'observer les termes en usage dans la langue, les collocations, etc. Les linguistes les utilisent afin de mener des études comparatives de phénomènes linguistiques d'une langue à une autre. Enfin, l'intérêt de ces corpus est d'en extraire des informations multilingues : terminologies, lexiques... Il existe deux principaux types de corpus multilingues : les corpus parallèles, composés de textes et de leurs traductions, et les corpus comparables, composés de textes n'étant pas des traductions mais partageant un certain nombre de caractéristiques communes. Les corpus parallèles, bien que très utilisés, présentent quelques inconvénients : les ressources sont peu nombreuses, notamment lorsque l'anglais n'est pas impliqué et ils ne sont pas pleinement représentatifs de chaque langue puisque le texte source influence souvent la traduction. Les corpus comparables permettent de pallier ces inconvénients : les ressources permettant de les construire sont beaucoup plus nombreuses et ils sont plus représentatifs des caractéristiques linguistiques et culturelles de chaque langue. Ainsi, de nombreux travaux font maintenant appel aux corpus comparables. Dans les domaines spécialisés, le besoin en ressources et en informations multilingues se fait particulièrement ressentir, en particulier le besoin de lexiques et de terminologies multilingues ainsi que leur mise à jour. En effet, les terminologies des domaines de spécialité évoluent constamment : les thématiques changent et le lexique associé s'en trouve

modifié. Des communautés scientifiques actives travaillent au sein des domaines de spécialité, réalisant régulièrement de nouvelles publications. Le besoin de communiquer, diffuser les connaissances à toute la communauté est grandissant. Cette affluence de nouvelles ressources et le caractère mouvant des domaines de spécialité impliquent un réel besoin de recenser les connaissances pointues et très spécialisées de ces domaines de façon régulière. Lexiques et terminologies multilingues peuvent être extraites et actualisées grâce aux corpus comparables. Ces corpus, contrairement aux corpus parallèles, peuvent être construits sur des domaines restreints. En effet, les contraintes de constitution des corpus comparables sont moins fortes : les documents issus de différentes langues ne doivent pas être des traductions mais partager des caractéristiques telles que le domaine, le thème, le genre... Cependant, la construction de corpus comparables pose plusieurs problèmes. Leur définition est assez vague, les caractéristiques communes varient selon l'usage du corpus. De plus, la construction des corpus comparables – comme de tout corpus – est une tâche coûteuse en temps et se trouve confrontée au manque de précision de leur définition. Cette thèse s'intéresse à ces deux problèmes que nous présentons plus en détails dans la section suivante.

Problématique

La définition des corpus comparables est assez floue. De la majorité des travaux portant sur ces corpus (Déjean et Gaussier, 2002; Fung et McKeown, 1997; Teubert, 1996; Zanettin, 1998), nous retenons que ceux-ci partagent certaines caractéristiques telles que le thème, le genre, la période... Le choix de ces caractéristiques communes dépend de l'utilisation du corpus. Ce choix varie selon les études et selon les constructeurs du corpus. Un corpus construit afin d'en extraire des terminologies multilingues ne sera pas forcément adapté à une utilisation pédagogique pour des étudiants en langues. Existe-t-il alors plusieurs définitions des corpus comparables ? Le degré de comparabilité est une notion permettant de quantifier dans quelle mesure les textes d'un corpus sont comparables. Un corpus contenant des textes n'ayant aucun rapport les uns avec les autres a un degré de comparabilité minimal et un corpus dont les textes ont de nombreuses caractéristiques communes (période, thème, genre, média par exemple) aura un degré de comparabilité élevé. Il existe donc un lien étroit entre les caractéristiques communes des textes d'un corpus comparable et son degré de comparabilité, mais comment le quantifier ? Une solution permettant de le calculer a été proposée dans quelques travaux et se base sur les similarités lexicales entre textes. Cette solution, proche de la vision en *sac de mots* des corpus (Habert et al., 1997), ne semble satisfaisante que dans certains contextes d'usage (principalement de traitement automatique des textes). Nous analysons ici la définition de ces corpus, de la comparabilité et des caractéristiques communes afin de proposer une définition plus précise et adaptée à tous les contextes d'utilisation. Une définition plus claire et détaillant les différents choix faits selon le contexte d'utilisation des corpus permettront de faciliter leur construction.

La tâche de construction de corpus est très coûteuse et nous souhaitons ici proposer un système d'assistance permettant de réaliser automatiquement certaines étapes. Les domaines de spécialité sont nombreux et les quantités de ressources nécessaires pour couvrir l'ensemble de ces domaines le sont aussi. Afin qu'une communauté scientifique puisse mieux communiquer, il faut au moins disposer de la terminologie dans les différentes langues et de lexiques multilingues. Pour que cela soit possible, la construction de corpus doit être simplifiée. Nous nous intéressons dans cette thèse à la construction de corpus comparables spécialisés dans les langues française et japonaise. Nous proposons une définition des corpus comparables spécialisés avec pour caractéristiques communes un domaine de spécialité, un thème et un type de discours (scientifique ou vulgarisé). Le thème permet de circonscrire un champ

scientifique et le type de discours permet de filtrer le niveau de communication des documents du corpus, garantissant un niveau de langue, un lexique et une syntaxe communs. La constitution d'un corpus comparable spécialisé se compose de différentes étapes. La première étape est cruciale, il s'agit de déterminer le domaine de spécialité dont le corpus sera représentatif, ainsi que de fixer les critères de choix des documents (selon les caractéristiques communes des documents fixées notamment). Cette étape ne peut être réalisée automatiquement mais nous en détaillons les principes. La seconde étape consiste à rechercher et sélectionner les documents. Il est important de sélectionner une source de données adaptée : le Web constitue la ressource principale, mais des portails scientifiques rassemblent des documents d'un domaine (par exemple Cismef pour le domaine médical) en proposant des méta-informations très utiles pour la constitution du corpus (genre, thème, mots-clés…). Les documents sont ensuite sélectionnés depuis la source en respectant les critères de choix. Dans notre cas, les documents sont sélectionnés selon leur thème et leur type de discours. Le thème pouvant facilement être filtré à l'aide de recherches par mots clés, nous nous concentrons sur la reconnaissance automatique du type de discours. Une analyse contrastive stylistique nous permet de créer une typologie de critères caractérisant le type de discours. Cette typologie est utilisée afin de générer une représentation vectorielle des documents qui nous permet d'apprendre des modèles de classification. Ces modèles sont ensuite insérés dans une chaîne logicielle d'aide à la construction de corpus comparables. Cette chaîne traite une collection de documents préalablement sélectionnés et collectés par l'utilisateur qu'elle classe, annote et documente afin d'en constituer un corpus comparable.

Plan du document

Dans le **chapitre 1** nous présentons un état de l'art nous permettant de définir les corpus comparables spécialisés, leurs caractéristiques et de lister leurs différentes utilisations. Nous définissons les corpus comparables en plusieurs étapes. Dans un premier temps, nous donnons une définition des corpus de façon générale puis les corpus électroniques. Nous abordons ensuite la notion de représentativité des corpus et effectuons une analyse des différentes typologies de corpus utilisées. Dans un second temps, nous définissons les corpus spécialisés en analysant comment les domaines ou langues de spécialité peuvent être circonscrits. Enfin, nous présentons les deux principaux types de corpus multilingues : les corpus parallèles et les corpus comparables. Nous justifions ainsi notre choix portant sur les corpus comparables en comparant ces deux types de corpus. Ce chapitre se termine par une partie portant sur les utilisations des corpus multilingues dans différents domaines.

Le **chapitre 2** est consacré aux corpus comparables. Nous analysons dans une première partie la comparabilité. Les travaux détaillant cette notion étant peu nombreux, nous tentons de la confronter avec des notions proches telles que la similarité ou la représentativité afin de dégager une définition précise. Dans une seconde partie nous listons les critères de compa rabilités choisis dans la littérature : domaine, genre, période… Cet état de l'art nous permet ainsi de présenter notre choix de critères pour des corpus comparables spécialisés. La troisième partie de ce chapitre est consacrée à la construction de ces corpus. Nous nous appuyons sur différents ouvrages traitant de la construction de corpus électroniques afin de dégager les étapes de la construction de corpus comparables spécialisés. La dernière partie de ce chapitre présente le corpus d'étude que nous avons construit, portant sur la thématique du diabète et de l'alimentation en français et japonais.

Dans le **chapitre 3** nous présentons l'analyse stylistique du corpus d'étude et la typologie des types de discours créée. Un état de l'art sur le sujet nous permet de dégager les méthodologies existantes (inductive, déductive, contrastive…) et de cibler les méthodes les plus adaptées selon les objectifs. Dans

le cas de la reconnaissance des types de discours scientifique et vulgarisé dans les langues française et japonaise, la méthode déductive et contrastive paraît la plus adaptée : partant d'un ensemble de textes préclassés, la comparaison de couples de documents appartenant à des classes différentes permet de dégager des caractéristiques discriminantes pour chaque classe. Le fruit de cette analyse, un ensemble de critères, est organisé dans une typologie composée de trois niveaux d'analyse : structurel, modal et lexical. Les critères de la modalité s'appuient sur deux théories : cella de Charaudeau (1992) appelée la modalité *locutive* et celle de Givón (1994) appelée la modalité *irrealis*.

Le **chapitre 4** est consacré à l'élaboration d'un système de classification automatique. Il se compose de deux parties : la première présente la théorie et les méthodes, la seconde est la mise en application de ces méthodes. La première phase consiste à indexer les documents du corpus d'apprentissage, c'est-à-dire générer, pour chaque document, une représentation de celui-ci compréhensible par les systèmes d'apprentissage automatique. Cette représentation s'appuie sur la valeur de chaque critère de la typologie sur le document. Une étape d'implémentation de la typologie est donc nécessaire. Dans la seconde phase, ces représentations des documents sont transmises à un système d'apprentissage automatique qui génère alors un modèle de classification reproduisant la classification fournie. La troisième phase consiste à évaluer le modèle de classification sur un second corpus. Nous présentons pour chacune de ces phases le travail que nous avons fourni et justifions nos choix.

Dans le **chapitre 5** figurent les résultats de l'évaluation des modèles de classification générés. Le corpus d'évaluation est présenté dans une première partie, ce corpus comparable porte sur le thème du cancer du sein en français et japonais. Nous présentons ensuite les résultats obtenus sur ce corpus avec la typologie complète. Nous essayons ensuite d'améliorer ces résultats en évaluant chaque catégorie de critères de la typologie, et nous comparons notamment des critères correspondant à deux théories de la modalité. La typologie composée des critères structurels, lexicaux et modaux de la théorie de Charaudeau (1992) donnent en moyenne de meilleurs résultats (pour les deux langues et les classifieurs). Nous analysons ensuite plus en détails les résultats obtenus pour les critères de la typologie en observant la fréquence de certains sur le corpus. Ce chapitre se termine sur une discussion sur l'aspect binaire de notre classification. Ce travail nous a permis de constater qu'il existe un continuum entre les deux types de discours et nous justifions cette observation avec quelques expériences.

Le **chapitre 6** est consacré à la mise en commun des travaux présentés dans les cinq chapitres précédents. Nous y présentons l'élaboration d'un système d'aide à la construction de corpus comparables spécialisés en français et japonais. À l'aide de la plateforme *UIMA* (Unstructured Information Management Architecture), que nous présentons dans une première partie, nous créons un système permettant, à partir d'une collection de documents en français et japonais relevant d'une même thématique spécialisée, de les classer et de générer un corpus annoté et documenté. Nous suivons donc les différentes étapes de construction présentées dans le chapitre 2 et intégrons un classifieur présenté dans le chapitre 4. Cet outil, encore évolutif, permet à l'utilisateur de créer des corpus en ne se souciant plus que de la sélection des documents selon une thématique, tout en gardant un contrôle sur les différentes étapes et le corpus ainsi généré.

CHAPITRE 1

Des collections de textes aux corpus comparables spécialisés

1.1 Introduction

L'évolution du Web produit depuis quelques années une augmentation très importante des ressources textuelles accessibles. Cette brusque augmentation de la quantité de données textuelles a eu un impact sur la perception et la constitution de corpus. Une collection de textes rassemblant des centaines de millions de mots telle que le British National Corpus et une collection de plusieurs centaines de milliers de mots seront appelés de la même façon *corpus*. Nous nous intéressons dans cette thèse à la constitution de corpus comparables spécialisés et il est nécessaire avant toute chose de bien les définir. Dans ce premier chapitre, nous présentons l'objet de notre étude, les corpus comparables spécialisés. Ces corpus ont de plus quelques caractéristiques : ils doivent être adaptés à un traitement automatique et ils sont construits à partir du Web. En tenant compte de ces propriétés nous définissons dans ce chapitre chaque élément constituant les corpus comparables spécialisés : les corpus, les corpus spécialisés et les corpus comparables. Le terme corpus désigne généralement une collection de documents, prenant différentes formes selon la discipline. Nous nous intéressons ici aux corpus dans le cadre d'études multilingues (en TAL, linguistique ou traduction). Nous partons d'une définition très générale des corpus que nous centrons ensuite sur le TAL et les études multilingues. Nous abordons ensuite le problème de l'adéquation entre les corpus et les objectifs de leur étude, c'est-à-dire la représentativité. La constitution d'un corpus est généralement liée à un besoin particulier, qui influence fortement sa composition. Dans la suite, nous nous appuyons sur différentes théories des langues de spécialité afin de donner une définition et de caractériser les corpus spécialisés. La dernière partie de ce chapitre est consacrée aux corpus multilingues et à leur exploitation.

1.2 Les corpus

1.2.1 Définition

Dans les disciplines philosophiques, un corpus est un « *recueil réunissant ou se proposant de réunir, en vue de leur étude scientifique (linguistique, socio-linguistique, etc.), la totalité des documents disponibles d'un genre donné, par exemple épigraphiques, littéraires, etc.* » (TLFi, 1960). Cette définition semble limitée pour plusieurs raisons. Tout d'abord, il paraît difficile de collecter la totalité des textes disponibles d'un genre donné (par exemple pour les genres du Web, comment collecter la totalité des

articles de forum ?). De plus, les travaux en linguistique des corpus se penchent généralement sur des phénomènes particuliers de la langue et nécessitent un corpus représentant ces phénomènes. Il faut donc se pencher du côté de la linguistique et du TALN pour obtenir des définitions plus opérationnelles des corpus. Les définitions suivantes correspondent en fait à la notion de *corpus informatisé* ou *corpus électronique*[1], souvent appelés *corpus* par abus de langage. La plupart des disciplines, en particulier la linguistique de corpus se basent sur des corpus électroniques. Sinclair (1996a, p. 5) en donne la définition suivante :

> « *A computer corpus is a corpus which is encoded in a standardised and homogenous way for open-ended retrieval tasks.* »

Deux conceptions des corpus sont distinguées par Rastier (2002) : les sacs de mots et les archives de textes. Selon lui, un ensemble de mots ou de phrases ne peut être considéré comme un corpus, ils nécessitent d'être observés dans un contexte qui est le texte :

> « *Si le mot [...] est l'unité élémentaire, le texte est pour une linguistique évoluée l'unité minimale, et le corpus l'ensemble dans lequel cette unité prend son sens.* »

L'importance de l'unité *texte* est aussi présente dans Péry-Woodley (1995, p. 8), pour qui « *un corpus se compose par définition de discours, de langue « concrète », et c'est immanquablement sous la forme de textes [...] que la langue se réalise en discours.* »

Habert (2000) apporte à cela une précision en distinguant les réservoirs à corpus (ou bases de données textuelles) aux corpus eux-mêmes. En effet, on trouve de nombreuses bases de données textuelles, telles que le BNC (British National Corpus) ou Frantext (textes littéraires datant du XVIème siècle à nos jours rassemblés par l'INaLF). Ces bases de données sont très souvent utilisées afin de générer des corpus, mais elles ne peuvent selon lui pas être considérées comme des corpus. C'est, « *l'opération de choix raisonné parmi les composants disponibles qui crée un corpus* » (Habert, 2000, p. 4).

La définition de Sinclair (1996a) résume bien les précisions précédentes : « *A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language* ». Le terme « données langagières » (« pieces of language ») est utilisé afin d'indiquer qu'un corpus ne contient pas nécessairement uniquement des œuvres complètes mais aussi des extraits. Les critères linguistiques évoqués correspondent à tous les éléments linguistiques déterminants lors de la construction du corpus, et qui permettront d'effectuer des analyses sur un corpus *représentatif* du(es) phénomène(s) linguistiques visés. Ces critères correspondent généralement à la récurrence de traits linguistiques dans les données langagières.

L'expression « échantillon du langage » indique que le corpus doit être représentatif du langage. Cependant, représenter un langage dans sa globalité grâce à un échantillon est un projet qui paraît irréalisable. Un sous-ensemble de base de données textuelles peut au mieux représenter un phénomène linguistique d'un langage (Habert, 2000) ou d'un sous-langage (Bowker et Pearson, 2002). Habert (2000, p. 1) reprend ainsi cette définition en la restreignant :

> « *un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques et* ***extra-linguistiques*** *explicites pour servir d'échantillon* ***d'emplois déterminés*** *d'une langue* »

Teubert (1996) affirme que « *les corpus sont les ressources les plus adaptées pour les études empiriques sur le langage, plus fiable que l'introspection* » (Teubert, 1996, p.240). Pour lui, la conception d'un corpus approprié dépend des connaissances que l'on souhaite en extraire. Un corpus est toujours

[1] « *Computer corpus* », appelé « *corpus informatisé* » chez Dubreil (2006) par exemple, et « *corpus électronique* » chez Habert et al. (1997).

construit dans le cadre d'une étude, afin d'observer un ou des phénomènes linguistiques précis, ou « emplois déterminés d'une langue ». L'objectif ainsi fixé du corpus détermine les critères de construction du corpus. Dubreil (2006) résume ces différentes définitions des corpus en les caractérisant selon trois axes :

la nature : le corpus est composé de données langagières.

la structure : les données du corpus sont sélectionnées, mises en formes et enrichies. Leur sélection se fait selon des critères de choix, de façon à ce que le corpus soit représentatif des objectifs visés. Les critères appliqués sont d'ordre linguistiques ou extra-linguistiques. Le corpus ainsi constitué est ensuite mis en forme (normalisation) et enrichi (documentation).

la finalité : obtenir un corpus représentatif d'un langage, d'un sous-langage ou de certains phénomènes linguistiques.

Cette définition est très générale, Bowker et Pearson (2002, p. 11) affirment qu'il existe autant de types de corpus qu'il existe d'études. Parmi les types les plus généraux, elles citent :

les corpus de références (*general reference corpora*) : corpus très larges, représentatifs d'un langage dans son ensemble (par exemple le British National Corpus (Aston et Burnard, 1998)) ; et *les corpus d'études* (*special purpose corpora*), corpus créés afin d'observer un aspect particulier du langage (par exemple le corpus créé par Beauvisage (2001) dans le cadre de son étude sur des sous-genres du roman policier ;

les corpus écrits : corpus composés de textes écrits et *les corpus oraux*, corpus composé de transcription de matériel oral (par exemple le corpus *Switchboard*, composé de conversations téléphoniques (Godfrey et al., 1992) ;

les corpus monolingues et **les corpus multilingues** (voir section 1.4) ;

les corpus synchrones : contenant des documents datant d'une période restreinte, permettant d'étudier le langage sur une période précise et *les corpus diachroniques*, rassemblants des écrits de périodes différentes, afin d'observer l'évolution du langage (voir (Kyto et al., 1994) pour plus de détails) ;

les corpus ouverts (« monitor corpus ») : corpus auxquels des textes sont constamment ajoutés et *les corpus fermés*.

Quel que soit le type de corpus, il sert généralement « d'échantillon d'emplois déterminés d'une langue » Habert (2000). Il est donc adapté à la finalité du cadre dans lequel il est construit. Mais qu'est-ce exactement qu'un corpus représentatif de phénomènes langagiers ? Comment déterminer qu'un corpus contient suffisamment d'informations pour généraliser des observations ? Nous présentons dans la section suivante une définition de la représentativité.

1.2.2 Représentativité des corpus

Comme nous l'avons mentionné dans la section 1.2.1, un corpus fait office d'échantillon du langage, et se doit d'en être représentatif. Biber (1993, p. 1) définit la représentativité en ces termes :

> « *Representativeness refers to the extent to which a sample includes the full range of variability in a population.* »

Par population, Biber désigne une langue, un langage, un sous-langage ou un ensemble de phénomènes langagiers. Au sein d'une population ou d'un corpus, deux types de variations apparaissent : les variations de situation et les variations linguistiques. Les variations de situation portent sur tous les éléments « extérieurs » au texte, tandis que les variations linguistiques concernent uniquement le texte. En travaillant

sur un corpus composé d'articles scientifiques, les variations de situation porteront sur le domaine, le thème, l'auteur, la période, etc. Les variations linguistiques seront plutôt d'ordre stylistique : syntaxe, lexique, etc.
Un corpus représentatif d'une population langagière doit inclure (Biber, 1993, p. 1) :

- un ensemble des types de textes de la population langagière ;
- un ensemble de distributions linguistiques de la population langagière.

Il paraît cependant irréaliste de prétendre pouvoir construire un corpus représentatif de n'importe quelle population langagière. En effet, la langue générale semble par exemple impossible à cerner dans sa globalité. Il en va de même pour des études plus circonscrites, la constante évolution de la langue rend difficile la tâche de trouver un échantillon pleinement représentatif d'un phénomène linguistique. Il est donc nécessaire de prendre en considération ces difficultés rencontrées lors de la création de corpus.

Les premières considérations lors de la construction d'un corpus sont souvent le type des textes, le nombre de textes, leur longueur, etc (Biber, 1993). Les documents sont souvent collectés parce qu'ils sont faciles d'accès, et l*a priori*té est donnée à la disponibilité des données et leur quantité plutôt qu'à la population qui doit être représentée (Habert, 2000). Selon Biber (1993), la représentativité d'un corpus dépend de plusieurs éléments. Elle dépend en premier lieu de la définition de la population ciblée, c'est-à-dire de la population que l'échantillon (le corpus) doit représenter. Cette population se définit à deux niveaux, que nous illustrons grâce au corpus LOB (Johansson et al., 1978)[2] :

- par ses bornes : quels textes sont inclus dans le corpus, quels textes en sont exclus (ex. : tous les textes publiés en anglais en 1961 au Royaume Uni) ;
- par l'organisation hiérarchique au sein de la population : quelles catégories de textes sont inclues, quelles sont leurs définition (ex. : 15 catégories principales et de nombreux sous-genres au sein de ces catégories dans le corpus LOB).

La représentativité dépend en second lieu de l'éventail des distributions linguistiques au sein du corpus, qui doit être équivalent à celui de la population ciblée. Cet éventail des distributions linguistiques correspond à l'ensemble des traits linguistiques, leurs variations et leurs distributions au sein d'un même texte, entre textes et entre types de textes. Le corpus, échantillon de la population, doit rendre possible des analyses de ces distributions, ainsi qu'une généralisation de leurs résultats à la population. Cet aspect de la représentativité dépend fortement du premier aspect : si un corpus ne représente pas l'ensemble des types de textes d'une population, il ne représentera pas l'ensemble des distributions linguistiques.

En dernier lieu, la représentativité dépend de l'échantillonnage des textes de la population dans le corpus. Pour cela, des choix de répartitions doivent être faits, par exemple sélectionner au hasard x documents de chacun des types de textes de la population ciblée. Cet échantillonnage induit un certain nombre de choix techniques, parmi lesquels figurent le nombre de textes par catégories, la taille des textes, etc.

Habert (2000) aborde la représentativité sous un angle plus pratique. Selon lui, un corpus peut être considéré, du point de vue purement statistique, comme l'échantillon d'une population. Deux problèmes apparaissent alors avec la notion d'échantillon : l'*incertitude*, qui survient quand un échantillon est trop petit pour représenter la population, et la *déformation*, survenant quand les caractéristiques d'un échantillon sont systématiquement différentes de celles de la population cible.

De ces problèmes découlent deux approches chez les créateurs de corpus. La première, qu'Habert nomme « gros, c'est beau », part du postulat que l'augmentation des données produit des corpus plus représentatifs de la langue (en englobant le maximum d'énoncés possibles). Cette méthode fait écho aux réservoirs à corpus, tels que le BNC, conçus pour englober l'ensemble des phénomènes d'un langage.

[2] Ce corpus sert aussi d'exemple dans Biber (1993).

Cependant, sur des populations plus restreintes, cette approche peut apporter du bruit dans le corpus, c'est ce qu'Habert appelle « l'insécurité des grands ensembles ». Cette seconde approche privilégie les conditions de production et réception des documents, en corrélation avec les caractéristiques langagières de la population cible.

Cette deuxième approche nous amène alors à nous intéresser de plus près au processus de construction d'un corpus. La représentativité d'un corpus se caractérise par une dimension interne (détermination des différents emplois du langage ciblé) et/ou une dimension externe (conditions de production et réception des documents). Les documents sont catégorisés avant d'être introduits dans un corpus, et chaque étude donne lieu à une nouvelle catégorisation. « Malaise dans la classification ? », Habert (2000) observe un « éparpillement typologique », dû aux nombreuses études sur corpus, amenant inévitablement de nombreux besoins et de nouvelles classifications. Nous allons tenter dans la section suivante de faire un bref tour d'horizon des travaux typologiques.

1.2.3 Typologies de corpus

Pour Adam (1992, p. 6), « *la catégorisation des textes fait partie des activités cognitives spontanées des sujets. [...] Sans l'existence de catégories, notre appréhension des énoncés produits serait probablement impossible* ». Habert (2000) l'a bien remarqué, il existe de nombreuses classifications de textes dans les corpus, et elles sont dues à la diversité des études sur corpus. Les classifications les plus fréquentes sont les suivantes : genres, registres, discours, domaines, thèmes... Néanmoins, les textes d'un corpus peuvent aussi être classés selon les cooccurrences de certains traits linguistiques.

Habert (2000) distingue deux types de classifications :

Les classifications *a priori* : le classement est basé sur des catégories ou caractéristiques des documents ;

Les classifications *a posteriori* : dans lesquelles le classement repose sur les caractéristiques linguistiques des textes uniquement, à partir desquelles sont générées les catégories.

Les classifications *a priori* sont généralement basées sur les classes fréquemment utilisées, tandis que les classifications *a posteriori* se basent uniquement sur les caractéristiques des textes, qui ne correspondent généralement pas à des classes prédéfinies. Les classes créées à partir de cooccurrences de traits linguistiques sont appelées « type de texte » par Biber.

1.2.3.1 Les classifications a priori

Les classifications *a priori* peuvent reposer sur (Habert, 2000, p. 14):

- les conditions de production des textes ;
- les buts visés par les textes ;
- leur inscription dans des genres ou autres catégories (sous-genres, types de discours...) ;
- l'emploi ou non de certaines marques linguistiques.

Les conditions de production des textes correspondent à la définition de la situation de communication dans laquelle s'inscrit le document. Le locuteur ou destinateur, ainsi que le destinataire doivent donc être connus. Cette identification constitue une première difficulté, puisqu'ils peuvent être totalement inconnus (lors d'une recherche de textes sur le Web par exemple). De plus, ces informations doivent être renforcées par la représentation que l'auteur a de lui-même et de son destinataire. Les buts visés par les textes correspondent à la fonction visée par le texte. Adam (1992) distingue par exemple sept fonctions : narrative, descriptive, explicative, argumentative, prédictive, conversationnelle et rhétorique. Le

troisième critère sur lequel peut reposer la classification est le choix d'une catégorie textuelle. Les plus fréquentes sont les domaines, les thèmes et les genres. Nous les présentons succinctement dans cette section, elles seront détaillées dans le chapitre 2.

Les catégorisations en domaine ou en thème Vite limitées, selon Habert, par la constante évolution des thèmes et domaines, ainsi que des parutions (notamment sur le Web), ces classifications permettent d'assurer une certaine représentativité et forment un filtre imparfait mais efficace. Sinclair (1996b) dresse une liste des différents thèmes et domaines utilisés dans de nombreux corpus : religion, droit, sciences, histoire... Habert y note cependant un certain nombre d'inconvénients. Tout d'abord les corpus traitant un thème le recouvrent rarement. De plus, l'ensemble des thèmes / domaines représentés ne présentent aucune structure : ils sont plus ou moins généraux, certains thèmes en regroupent d'autres, etc.

Les catégorisations en genres Les genres, selon Biber (1989), sont « *les catégories de textes distinguées spontanément par les locuteurs confirmés d'une langue ; par exemple les genres de l'anglais incluent les romans, les articles de journaux, les éditoriaux, les articles de recherche, les discours en public, les nouvelles radiophoniques et la conversation de tous les jours* »[3]. Les genres sont des « *dispositifs de communication socio-historiquement définis* » (Maingueneau, 1996, p. 43). Ainsi, les textes d'un même genre partagent généralement une même fonction communicative, et un certain nombre de caractéristiques linguistiques. Une classification par genre assure donc une certaine représentativité dans un corpus. Karlgren et Cutting (1994) font par exemple appel au corpus Brown, dans lequel les documents sont classés selon leur genre. Dubreil (2006) utilise un corpus composé d'articles scientifiques.

Le quatrième critère est l'emploi de certaines marques linguistiques. La sélection des documents d'un corpus peut se faire sur la présence (ou l'absence) de certains traits linguistiques. Benveniste (1966)[4] s'est par exemple intéressé aux *embrayeurs*, unités linguistiques dont la valeur référentielle nécessite de connaître les conditions de leur énonciation (lieu, moment, identité des locuteurs et interlocuteurs…) (Habert, 2000, p. 19). L'étude de son corpus lui permet alors de distinguer deux catégories : l'histoire (énonciation historique) et le discours (énonciation personnelle).

Les classifications des documents au sein des corpus peuvent également être basées sur des combinaisons de ces quatre critères.

1.2.3.2 Les classifications a posteriori

Autrement appelées typologies inductives, ce mode de classification repose sur les caractéristiques des textes uniquement. À partir de ces catégories sont générés des groupements de textes ayant la même tendance à recourir à un ensemble de traits linguistiques et à en éviter d'autres. Un corpus est utilisé afin d'observer la répartition des traits, et repérer les plus discriminants. Bronckart (1996)[5] réalise une typologie discursive de textes en croisant deux paramètres :

- le rapport de l'auteur à la situation d'énonciaton : présence d'embrayeurs dans les textes notifiant un ancrage du texte dans un certain contexte, ou absence d'embrayeurs, donc autonomie du texte par rapport à tout contexte ;
- le rapport de l'auteur au monde : distance entre la représentation du monde faite dans le texte et l'interaction sociale en cours. Si celle-ci est grande, les faits présentés dans le texte paraissent passés, sinon les faits énoncés semblent proches du lecteur et peuvent se produire dans son monde.

[3] La traduction vient de Habert (2000, p. 16).
[4] Cité dans (Habert, 2000).
[5] Cité dans Habert (2000, p. 23).

Une analyse inductive des variations de ces deux paramètres lui permettent d'aboutir à 4 catégories, appelées « architypes discursifs » (Habert, 2000). Ces catégories sont présentées dans le tableau 1.1.

		Rapport au monde	
		Conjonction	Disjonction
Rapport interactif à la situation	Implication	Discours interactif	Récit
	Autonomie	Discours théorique	Narration

Table 1.1 – Architypes discursifs de Bronckart (1996)

Ces travaux inductifs font bien évidemment écho aux travaux de Biber, qui sont présentés dans le chapitre 2.

1.2.3.3 Synthèse

Cette liste de typologies, qui ne se veut pas exhaustive nous permet d'entrevoir le fossé entre la représentativité théorique des corpus et le côté pratique lors de la construction de ceux-ci. Pour Biber, un corpus (considéré comme échantillon d'une population) est représentatif s'il inclut l'ensemble des variations caractéristiques de la population. Faisons l'analogie avec un sondage, comment s'assurer que les sondés soient représentatifs de l'ensemble de la population ? Classer les sondés par âge, ville ou profession peut permettre d'avoir une idée plus précise de l'échantillon qu'ils représentent et ainsi viser les catégories manquantes pour que l'ensemble soit plus représentatif. Les textes des corpus sont donc répartis suivant plusieurs classes, à l'intérieur desquelles chaque « sous-classe » est représentée jusqu'à obtenir un corpus suffisamment important et varié. Le plus important semble donc de définir rigoureusement la « population ciblée » avant la construction du corpus. Dans le cas des corpus dits « de spécialité », la définition de la population ciblée se restreint souvent à la simple sélection d'un domaine de spécialité. Un domaine est-il suffisant pour circonscrire une langue de spécialité ? Nous présentons dans la section suivante les langues de spécialité et communautés de discours et proposons une définition des corpus spécialisés.

1.3 Les corpus spécialisés

Un corpus spécialisé porte sur un domaine de connaissance ou une situation de communication particuliers. Il doit être représentatif de la langue de spécialité en usage dans le domaine. Les textes d'un corpus spécialisé sont sélectionnés selon une *configuration domaine-genre* (Péry-Woodley, 2000), par exemple des articles de recherche en TALN pour Dubreil (2006, p. 66). Les corpus spécialisés se doivent d'être représentatifs d'un domaine ou d'une situation de communication, ils doivent contenir des documents d'une certaine qualité (traitant réellement et rigoureusement du domaine) et représenter l'ensemble des variétés du domaine auquel il a trait.

Dubreil (2006, p. 67) en donne la définition suivante : « tout regroupement de données langagières créé à des fins spécifiques et représentatif d'une situation de communication ou d'un domaine dans la pratique. » Elle précise que la constitution d'un corpus spécialisé implique donc que « le domaine des textes inclus dans le corpus soit défini et délimité » et que « les textes soient représentatifs de ce domaine pour garantir l'authenticité des conclusions qu'on en tire ».

Il est donc primordial, lors de la construction, de savoir définir et délimiter un domaine. Plusieurs modèles théoriques permettent de mener à bien cette tâche : les langues de spécialité, les sous-langages, les communautés de discours.

1.3.1 Délimiter un domaine

Bowker et Pearson (2002, p. 25) opposent la langue générale ou LGP (Language for General Purpose) à la langue de spécialité ou LSP (Language for Specialized Purpose). La LGP correspond au langage que nous utilisons tous les jours, pour parler de sujets ordinaires dans des situations communes. La LSP correspond aux langages utilisés pour parler de domaines de connaissance spécialisés, par exemple le domaine de la chimie. On parle en réalité de LSP au pluriel, car il existe autant de langues de spécialité que de domaines. À chaque langue correspond une langue générale et des langues de spécialité. Un locuteur natif d'une langue est généralement un expert de la langue générale, il en connaît le vocabulaire et la syntaxe. La connaissance d'une langue de spécialité est par contre réservée aux spécialistes d'un domaine scientifique ou technique.

Lerat (1995) présente trois notions constitutives des langues de spécialité[6] :

Une origine : une langue de spécialité appartient à un domaine particulier ;

Une nature : une langue de spécialité est une variété de la langue générale, elle possède les mêmes caractéristiques, mais celles-ci sont restreintes ;

Une fonction : une langue de spécialité sert à communiquer, à transmettre des informations.

De là peuvent apparaître différentes caractéristiques propres aux langues de spécialité (Bowker et Pearson, 2002, p. 26) :

- un vocabulaire spécialisé ;
- des combinaisons particulières de mots et des méthodes de présentation des informations[7] ;
- des caractéristiques stylistiques particulières.

À titre d'exemple, la rédaction des résultats d'une expérimentation scientifique se fait en utilisant le vocabulaire spécialisé propre au domaine, en utilisant un style et une syntaxe particuliers. Nous illustrons ceci avec un extrait tiré d'un article scientifique écrit par Fung et Yee (1998), correspondant à la figure 1.1. Nous trouvons dans cet exemple un vocabulaire spécialisé au TAL ou plus généralement à l'informatique : « algorithm », « corpus », « ranking candidates », etc. , des méthodes de présentation particulière : les deux items « Evaluation I: unknown words » et « Evaluation II: known words » par exemple. Quant au style, nous pouvons observer des marqueurs de glose dans les phrases, ainsi que des quantités numériques et des références à différents tableaux et figures.

La notion de langue de spécialité est largement débattue chez les linguistes. Selon Dubreil (2006, p. 68), tant du point de vue lexical que grammatical, les langues de spécialité ne se distinguent pas clairement de la langue générale et peuvent y être inclues ou se chevaucher.

Les langues de spécialité sont souvent comparées à la théorie des sous-langages, bien que le terme sous-langage soit souvent utilisé comme « fourre-tout » pour tous les langages scientifiques et spécialisés (Williams, 1999, p. 40). Un sous-langage (SL) correspond à : « *the language used by a particular community of speakers, say, those concerned by a particular subject matter or those engaged in a specialized occupation* » (Sager (1986) cité par Péry-Woodley (1995)). Les SL semblent donc s'ancrer au niveau de la communauté concernée par le domaine de spécialité plus que sur le domaine lui-même. Notons que le

[6] Cité dans (Dubreil, 2006).

[7] « Special ways of combining words and arranging information. »

3.9 Experimental Results

Evaluation I: unknown words In order to apply the above algorithm to find the translation for 流感/*liougan* from the newspaper corpus, we first use a script to select the 118 English content words which are not in the lexicon as possible candidates. The highest ranking candidates of 流感 are *flu, Lei, Beijing, poultry* respectively. We also apply the algorithm to the frequent Chinese unknown words and the 118 English unknown words from the English newspaper. The output is ranked by the similarity scores. The highest ranking translated pairs are shown in Table 6.

Evaluation II: known words A second evaluation is carried out on randomly selected 40 known English words from the English newspaper against 900 known Chinese words from the Chinese newspaper. This evaluation is more automatic because a dictionary can be used to find correct translations. We have added *flu*/流感 in the dictionary.

The five highest ranking candidates for *flu, shop, virus* are shown in Table 7.

For the test set of 40 English words against 900 Chinese candidates, translation accuracy ranges from 30% when only the top candidate is counted, to 76% when top 20 candidates are considered, and up to 88% when top 40 are counted. We suggest that it is not unreasonable for the system to to give 20+ translation candidates for each word when the system is used as translator-aid.

Figure 1.1 – Exemple : évaluation d'expérimentations extrait de (Fung et Yee, 1998)

terme sous-langage est un faux ami, les sous-langages n'étant pas forcément des sous-ensembles de la langue générale (Habert et al., 1997, p. 149). Cette théorie se base sur l'hypothèse que les SL utilisent un lexique fini, il est donc en théorie possible de délimiter un SL. Dubreil (2006, p. 70) examine les différents critères permettant la délimitation. Les deux premiers critères, linguistiques, sont le lexique et la grammaire. Pour le lexique, nous avons vu qu'il est en théorie fini. En pratique, il faut prendre en compte le « principe dynamique des langues vivantes » : de nouveaux mots sont perpétuellement intégrés aux langages ou sous-langages. Quant à la grammaire, elle est en théorie un sous-ensemble fini de la grammaire de la langue générale (Harris, 1988). Cependant, cette hypothèse théorique ne tient pas compte du fait que « *la réalité est plus expressive que la grammaire simplifiée de Harris* » (Williams, 1999, p.50). Les critères extra-linguistiques sont le thème et la communauté des locuteurs. Circonscrire un SL par son thème revient à se baser sur une classification pré-établie. Cependant, il faut en théorie disposer d'une telle classification ou de suffisamment de connaissances pour pouvoir statuer sur l'appartenance d'un texte à une catégorie. Le dernier critère, la communauté des locuteurs de la SL, se base sur le fait que les locuteurs appartenant à un même domaine partagent certaines habitudes. La théorie des sous-langages ne définit cependant pas l'appartenance d'un locuteur à un domaine. Ce modèle théorique est très utilisé en TAL afin de construire des corpus, mais il est plus difficile théoriquement d'en définir

ses contours. Williams (1999) met en exergue le fait que les sous-langages, selon certains chercheurs, seraient des langues artificielles, construites à des fins scientifiques uniquement. Circonscrire un sous-langage consisterait en effet à délimiter le domaine sur lequel il porte. La délimitation d'un domaine pose une fois de plus problème. Williams (1999) et Dubreil (2006) font donc appel à la théorie des communautés de discours, qui semble être un compromis entre la définition théorique et la pratique : la constitution de corpus.

Cette fois-ci, la définition de la théorie fait en partie abstraction de la notion de langage, pour s'attarder sur une communauté particulière et le discours qui lui est propre.

Knowles et Roe (1994, p. 138)[8], une communauté correspond à :

> « *any group of individuals who are defined by a shared global purpose to which all publicly subscribe, and who have evolved or adopted mechanisms and procedures for achieving their shared objectives* ».

À une communauté correspondent des moyens du discours et un lexique. L'appartenance à une communauté nécessite d'adopter ces règles. Une communauté de discours se distingue d'un domaine ou d'un sous-domaine par six caractéristiques Williams (1999, p. 52) :

1. Un but commun accepté par tous les membres de la communauté ;
2. Des mécanismes d'interaction entre les membres ;
3. Des mécanismes de participation garantissant l'information et le retour d'information ;
4. L'utilisation et la possession d'un ou plusieurs genres ;
5. L'acquisition d'un lexique spécifique ;
6. Un seuil de membres.

Ce modèle semble rendre possible la définition et la délimitation de la communauté et sa mise en pratique lors de la constitution d'un corpus.

1.3.2 La théorie face à la pratique

Nous venons de lister et comparer trois modèles théoriques permettant de délimiter un domaine en vue de constituer un corpus spécialisé. Les linguistes Williams (1999); Dubreil (2006) ont analysé ces modèles et montrent les failles que peuvent avoir ces modèles théoriques. Néanmoins, chacun d'entre eux est utilisé par les chercheurs lors de la création de leurs corpus. Compte tenu du principe dynamique de la langues et de ses constantes évolutions et ambiguïtés, certaines théories comme les LSP ou les SL omettent des éléments bancals afin de se concentrer sur les aspects pratiques.

Les communautés de discours semblent alors fournir un compromis entre la théorie et la pratique. En se concentrant sur la caractérisation de la communauté de discours, les problèmes de limites floues sont écartées. Cependant, baser la définition d'un corpus spécialisé uniquement sur la communauté de discours dont sont issus les documents pose aussi quelques problèmes pratiques. Cette définition des corpus spécialisé implique que l'auteur de chaque document inséré dans un corpus soit identifié. Nombreux sont les corpus construits à partir du Web, qui constitue une ressource intarissable de données textuelles. Cependant, la provenance des documents du Web et les informations sur l'auteur n'étant pas toujours fournies, il est parfois difficile d'associer un document à une communauté de discours.

Dans cette thèse, nous travaillons sur des corpus spécialisés dont les documents sont extraits du Web. Bien que certains portails permettent d'accéder à de nombreuses informations sur les documents, il est

[8] Cités dans Williams (1999, p. 51).

assez difficile pour la majorité des ressources de disposer d'informations sur la publication. Il est ainsi difficile de cibler une communauté de discours. De plus, la théorie des communautés de discours est utilisée dans le cadre d'études très ciblées avec un groupe d'auteurs très restreint et souvent un genre particulier (par exemple le corpus de Dubreil (2006) composé d'articles de la conférence TALN et de la revue TAL). Les contraintes sont moins fortes pour nos corpus et le manque potentiel d'informations sur les documents du Web nous poussent à utiliser la théorie de Bowker et Pearson (2002) des langues de spécialité. Le terme corpus spécialisé désignera donc ici un corpus composé de documents issus de langues de spécialité.

1.4 Du corpus monolingue au corpus multilingue

Depuis les années 80, les travaux en linguistique de corpus, auparavant en majorité sur la langue anglaise, se sont ouverts sur les langues européennes et asiatiques (McEnery et Xiao, 2007). De là sont apparus les corpus multilingues, corpus composés de textes en plusieurs langues[9]. Les corpus multilingues représentent des ressources très utiles dans de nombreux domaines : traduction automatique, aide à la traduction, extraction d'informations multilingues, étude comparatives…

McEnery et Xiao (2007, p. 2) listent trois types de corpus multilingues :

- les corpus de type A : composés de textes sources accompagnés de leurs traductions ;
- les corpus de type B : composés de corpus monolingues partageant des caractéristiques (sampling frame) ;
- les corpus de type C : combinaisons de A et de B.

Selon les travaux et les périodes, ces corpus sont appelés comparables ou parallèles (voir McEnery et Xiao (2007, p. 2)). Nous appellerons ici parallèles les corpus de type A et comparables les corpus de type B et C. Il existe dans la littérature d'autres appellations pour les corpus multilingues. Fung et McKeown (1997) parlent de corpus non-parallèles ou de corpus parallèles bruités, et Rapp (1995) de corpus non-liés. Si l'ensemble de ces types de corpus devait être classé selon un degré de similarité des textes les composant, nous pourrions obtenir le classement de la figure 1.2.

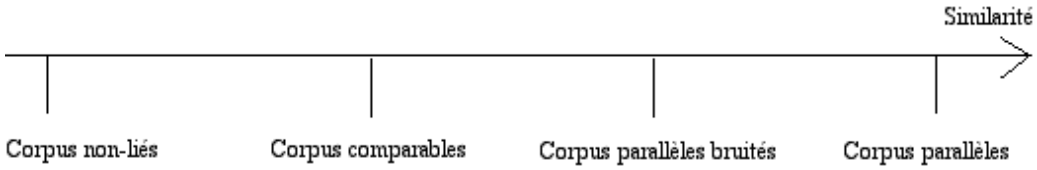


Figure 1.2 – Classification des types de corpus multilingues

Nous présentons dans la suite de cette section les deux principaux types de corpus : les corpus parallèles et comparables.

[9] Dans le cas où seulement deux langues sont représentées dans un corpus, on parle de corpus bilingue. Par soucis de simplicité, nous ne distinguerons pas les deux et parlerons dans tous les cas de corpus multilingue.

1.4.1 Les corpus parallèles

Un *corpus parallèle*[10] est un ensemble de textes accompagnés de leurs traductions dans une ou plusieurs langues (Bowker et Pearson, 2002, p. 92). L'ancêtre le plus connu des textes parallèles est la Pierre de Rosette. Découverte au 18ème siècle, elle permit à Champollion en 1822 de découvrir la clé du déchiffrement de l'écriture hiéroglyphique (Véronis, 2000). Même si le terme parallèle est utilisé afin d'indiquer qu'un corpus contient des textes et leurs traductions, les paires de textes dans un corpus parallèle ne sont pas forcément des traductions directes, elle peuvent être des traductions d'un troisième texte. Cette situation apparaît souvent dans certains *environnements multilingues* (l'Union Européenne par exemple) dans lesquels il peut être impossible de savoir dans quelle langue a été rédigé le texte source.

Parmi les corpus parallèles de référence, on compte :

Le corpus Hansard : créé dans les années 80, ce corpus est composé de texte anglais et français tirés des transcriptions des débats du parlement canadien de 1970 à 1988. Ce corpus contient plusieurs dizaines de millions de mots (Véronis, 2000) ;

Le corpus Europarl : corpus rassemblant des textes du Parlement Européen dans 11 langues, avec plus de 20 millions de mots par langue (Koehn, 2004) ;

Le corpus Hong-Kong Hansard : créé par le LDC (Linguistic Data Consortium), ce corpus rassemble les textes en anglais et français issus des discussions, rapports, etc. du parlement de Hong Kong ;

Le corpus de l'UBS (Union des banques suisses), utilisé par exemple par Gale et Church (1993), organisme publiant dans plusieurs langues (français, anglais, italien, allemand) des rapports sur le développement de l'économie suisse ;

Certains ouvrages lus dans le monde entier, comme la Bible, sont traduits dans la plupart des langues et peuvent constituer un corpus parallèle.

1.4.2 Les corpus comparables

Pour Teubert (1996), les corpus comparables sont des corpus en deux ou plusieurs langues ayant une composition ou une structure[11] similaire (ou quasi-similaire). Les textes du corpus sont sélectionnés selon des critères linguistiques ou extra-linguistiques (le domaine par exemple), ce qui permet de garantir aux textes un ensemble de caractéristiques communes (ou composition). Nous avons vu dans la section 1.2 que les éléments d'un corpus étaient sélectionnés en fonction de certains critères. La définition générale de Teubert ne nous permet pas de déterminer ce qu'est une composition similaire et quels critères permettent de l'obtenir.

Ainsi, nous nous ramenons à la définition de Bowker et Pearson (2002), selon laquelle les corpus comparables sont composés de documents en plusieurs langues, qui ne sont pas des traductions, mais qui partagent certaines caractéristiques (Bowker et Pearson, 2002, p. 93).

Il existe quelques travaux abordant la notion de corpus comparables monolingues, constitués de deux ensembles de textes, l'un composé de textes écrits dans une langue et l'autre composé de textes traduits dans cette même langue (Zanettin, 1998; Culo et al., 2008). Le but de ces corpus étant d'étudier le processus de traduction, un certain nombre de contraintes de composition du corpus sont identiques au contexte multilingue : thématique, vocabulaire, syntaxe communs. Nous ne nous intéresserons dans cette thèse qu'aux corpus comparables multilingues.

[10] L'Homme (2004) parle de *corpus alignés*.
[11] « *Composition* » dans l'article anglais.

Nous ne nous intéressons ici qu'au contexte multilingue et ne développerons pas cet aspect de la comparabilité.

En rassemblant les différents critères présents dans la littérature, on voit que ceux-ci vont de la période de rédaction des documents, à leur thème, au média utilisé, etc. Déjean rassemble ceux-ci en deux catégories :

- les critères qualitatifs : critères utilisés en stylistique tels que le genre, l'auteur, la période, le média, etc. ;
- les critères quantitatifs : critères basés sur les mesures de fréquences de certains traits linguistiques (ex. : fréquence de certains termes).

L'ensemble des critères de comparabilité choisi pour un corpus fait varier son *degré de comparabilité*.

Il n'existe pas, à notre connaissance, de corpus comparable de référence. Teubert (1996) note toutefois que le projet NERC (Network of European textual Reference Corpora) constitue une bonne base pour la construction de corpus comparables (Calzolari, 1993).

Il existe deux principales catégories de corpus comparables :

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 McKeown (1997), par exemple, utilisent un corpus anglais/japonais composé d'articles tirés 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 tirés de grands journaux nationaux allemands 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 Gaussier (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.

1.4.3 Synthèse

Au premier abord, les corpus parallèles peuvent paraître plus adaptés à tout type de tâche d'extraction d'informations multilingues puisque l'alignement de phrases y est facilité. De nombreuses bases de données textuelles ou des corpus de références servent de base à la création de corpus parallèles, par exemple le corpus *Hansard*, composé de débats du parlement canadien publiés dans les langues nationales, français et anglais ; ou *l'European Corpus Initiative* (ECI), composé de textes parallèles dans différentes langues européennes. Néanmoins, ce type de corpus comporte un certain nombre d'inconvénients. Tout d'abord, ces ressources sont limitées, voire rares pour certaines langues peu répandues, il est en effet plus facile de construire un corpus comparable dans un domaine donné qu'un corpus parallèle de bonne qualité (Fung et Yee, 1998). De plus, la traduction est généralement influencée par les ressources à traduire. En effet, les tournures de phrases et le vocabulaire du texte traduit sont fortement liés au texte source. Cet écart faible entre texte source et texte cible n'est pas très dommageable pour les langues proches (langues romanes/anglais), mais peut l'être pour de nombreuses langues à grande distance typologique et culturelle. Les corpus comparables permettent de plus de passer outre les restrictions de langues et autres contraintes imposées par les corpus parallèles. Enfin, ces corpus sont beaucoup plus représentatifs des particularités linguistiques de chaque langue.

Revenons à la figure 1.2. Il est possible de considérer l'axe comme celui de la comparabilité. Ainsi, un corpus parallèle serait un corpus à forte comparabilité, tandis qu'un corpus non-lié aurait une comparabilité minimale. Un continuum peut être perçu entre ces différents types de corpus. Cependant, nous

avons vu en définissant la comparabilité que celle-ci s'instancie sur les caractéristiques propres aux textes. Peut-on considérer qu'un corpus parallèle portant sur une thématique particulière par exemple a un degré de comparabilité plus élevé que celui d'un corpus comparable sur cette même thématique ? Si les textes d'un corpus parallèle ne sont pas considérés par paires, celui-ci peut alors être utilisé comme un corpus comparable. Mais est-il plus comparable qu'un corpus ne contenant aucune traduction ? Le fait est qu'un tel corpus provoquerait une certaine perte d'informations. En effet, un corpus comparable, s'il ne contient aucune traduction, sera composé uniquement de textes originaux, ayant tous un contenu, un style et des informations propres, ce qui ne sera pas le cas dans un corpus parallèle. Ce continuum peut exister si l'on ne considère que le point de vue informatique et applicatif des corpus, mais ce n'est plus le cas dès lors que l'on élargit le cadre. Nous considérerons donc que ces deux types de corpus sont bien distincts et qu'il n'y a pas lieu d'évaluer la comparabilité d'un corpus parallèle.

1.5 Exploitation des corpus multilingues

Les corpus multilingues sont utilisés dans de nombreux types de travaux. Nous allons dans cette section présenter succinctement quelques travaux majeurs traitant de l'exploitation des corpus. Ces travaux relèvent de différentes disciplines, informatiques ou linguistiques principalement.

1.5.1 Aide à la traduction et enseignement

Les dictionnaires et lexiques multilingues constituent des ressources insuffisantes pour les traducteurs. Il leur est nécessaire d'observer la langue dans son usage et les corpus multilingues sont un bon moyen d'y arriver.

> « *Any work of translation must [...] be not only linguistically correct but also appropriate to the frame of reference of the source, i.e. it must also achieve equivalence at the level of style, register, cultural and social context, etc.* » (Peters et al., 1996, p. 68).

Les travaux de plus en plus nombreux en linguistique de corpus et les progrès effectués sur la création et le traitement des corpus ont permis de « démocratiser » ces ressources et ainsi de lier la communauté TAL à celle des traducteurs (Zanettin, 2002).

Laviosa (1998)[12] en fait le constat :

> « *The corpus-based approach is evolving, through theorical elaboration and empirical realisation, into a coherent, composite and rich paradigm that addresses a variety of issues pertaining to theory, description and the practice of translation studies.* »

Les travaux de traduction basée sur corpus (*corpus-based translation*) se partagent en deux parties : l'une théorique et l'autre pratique. Les travaux théoriques se basent sur des corpus multilingues afin d'étudier le processus de traduction, c'est-à-dire comment une idée est exprimée d'une langue à une autre, ou encore étudier certains traits linguistiques et leur fréquences. Par exemple, Xiao et McEnery (2002) utilisent un corpus parallèle anglais-chinois afin d'observer comment les expressions temporelles et aspectuelles anglaises sont exprimées en chinois. Certains travaux se basent sur des corpus comparables monolingues pour analyser le processus de traduction au sein d'une langue : comparaison entre textes traduits et textes écrits dans la langue, comparaison entre traductions faites par des professionnels et par des étudiants, etc. (Zanettin, 1998; Kübler, 2008; Culo et al., 2008).

[12] Citée dans McEnery et Xiao (2007, p. 5).

Les utilisations pratiques des corpus multilingues dans le cadre de la traduction sont plus nombreuses. Ils représentent une base d'entraînement et d'apprentissage pour la traduction, ainsi qu'une base de développement d'applications telles que la traduction automatique (« machine translation »), et la traduction assistée par ordinateur (« computer-assisted translation ») (McEnery et Xiao, 2007, p. 9) :

> « *As corpora can be used to raise linguistic and cultural awareness in general, they provide a useful and effective workbench for translators and trainee* ».

Bowker (1998) affirme que la traduction assistée par ordinateur permet d'obtenir des textes traduits de meilleure qualité tant au niveau de la compréhension du sujet traité qu'à celui du choix des termes et expressions idiomatiques. Les corpus parallèles sont très utiles pour la traduction puisqu'ils apportent des équivalents de traduction de mots ou d'expressions, par le biais des concordances dans le corpus. S'ils contiennent plusieurs traductions pour un même texte, ils sont alors encore plus riches et permettent par exemple d'analyser la structure et les stratégies d'une traduction. Les corpus parallèles étant des ressources rares et présentant l'effet de *translationese* (influence de la source sur la traduction produite), les corpus comparables sont aussi utilisés pour l'aide à la traduction et pour l'apprentissage des langues. Sharoff et al. (2006) présente un outil faisant appel à des corpus comparables afin de trouver des équivalents aux expressions de la langue générale difficiles à traduire. Partant d'une phrase ou d'une expression dans la langue source, le système identifie dans un corpus comparable un ensemble d'expressions équivalentes utilisées dans un même contexte dans la langue cible. Ce système nécessite bien entendu des corpus de taille conséquente afin d'assurer la résolution du plus grand nombre de problèmes mais cela semble plus simple à atteindre avec des corpus comparables.

Zanettin (1998) fait quant à lui appel aux corpus comparables afin de former les traducteurs ou de favoriser l'apprentissage d'une langue de façon plus générale. Selon lui, les corpus comparables peuvent être utilisés dans différentes tâches :

- Aide à la traduction : « by looking for homographs [...], cognates and perceived equivalents, learners were able to evaluate the respective behaviour in the two languages of similar discourse units and to draw from a selection of citations in the target language suitable candidates [...], in adherence to the linguistic and genre convention of the receiving culture » ;
- Apprentissage de domaines et de leur terminologie : ils permettent de parcourir différents domaines, spécialement les domaines de spécialité quelle que soit la langue, et d'en découvrir leurs terminologies ;
- Fouille textuelle : ils permettent enfin d'analyser des genres et thématiques particulières (ce qui peut être une tâche pré- ou post-traduction), et d'étudier leurs caractéristiques linguistiques communes, leurs similarités.

Dans le cadre de l'aide à la traduction, McEnery et Xiao (2007) ajoutent que les corpus comparables sont principalement utiles dans les domaines de spécialité pour valider et éviter les erreurs lors d'une traduction vers la langue maternelle, et pour chercher des suggestions de traduction et confirmer les choix lors d'une traduction depuis la langue maternelle.

1.5.2 Lexicographie et terminologie

Dans les corpus parallèles peuvent être alignées les phrases équivalentes d'une langue à l'autre. Cette méthode permet d'obtenir pour chaque phrase, expression ou terme dans une première langue, un ou des équivalents de traduction dans la seconde langue. Ces séquences alignées permettent aux lexicographes d'accéder à l'ensemble des équivalents de traduction d'une expression, ce qui peut être utile pour des expressions particulières ou des collocations (Véronis, 2000).

De très nombreux travaux portent sur l'extraction de terminologies multilingues et la création automatique de lexiques à partir de corpus multilingues. Les premiers portent sur les corpus parallèles, Catizone et al. (1989) furent parmi les premiers à publier un article sur l'extraction de lexique multilingue grâce à des méthodes statistiques et à un dictionnaire électronique, sur un corpus parallèle anglais-allemand. À l'époque, les lexiques extraits de corpus parallèles portaient généralement sur des termes simples, mais les recherches se sont vite penchées sur des phénomènes plus complexes, tels que les collocations, expressions, phrases, etc. (Daille et al. (1994), Dagan et Church (1994) par exemple). C'est en 1995 que Fung et Rapp cherchent à pallier le manque de ressources parallèles en créant des méthodes permettant d'aligner des corpus non-parallèles (corpus parallèles bruités puis corpus comparables). S'en suivent alors un grand nombre de travaux portant d'abord sur les termes simples, puis les termes composés, collocations, etc., notamment Chiao (2004), Déjean et Gaussier (2002), Morin et Daille (2004)...

Les travaux de désambiguisation du sens des mots (word sense disambiguation) utilisent aussi des corpus multilingues. En effet, en supposant que la plupart des ambiguïtés proviennent du niveau lexical, le recours à un corpus de même thématique dans une autre langue peut permettre de lever l'ambiguïté (Brown et al., 1991).

1.5.3 Autres travaux

Les travaux en CLIR (Cross Language Information Retrieval) visent à rechercher des documents dans une langue à l'aide de requêtes dans une autre langue. Les outils de CLIR sont très utilisés afin d'effectuer des recherches multilingues sur le Web (Véronis, 2000). Ces travaux requièrent des outils de traduction terminologiques et font donc appel aux corpus multilingues (Oard et Diekema, 1998). Certains utilisent les corpus afin d'en extraire directement des lexiques bilingues utilisés ensuite pour traduire des requêtes, tandis que d'autres utilisent les corpus multilingues comme bases documentaires (une recherche dans une langue obtiendra comme résultats les traductions des documents correspondants, pour des corpus parallèles) (*ibid.*).

1.6 Synthèse

Dans ce chapitre nous avons défini notre sujet d'étude, les corpus comparables spécialisés, en les décomposant. Dans un premier temps nous avons défini les corpus. En partant d'une définition très générale couvrant plusieurs domaines nous avons donné une définition des corpus s'appliquant au domaine du TAL. Un corpus se caractérise par : une nature (des données langagières), une structure (la sélection, mise en forme et documentation des données) et une finalité (être représentatif d'un phénomène langagier). Nous avons ensuite défini les corpus spécialisés. Nous avons comparé trois théories permettant de circonscrire des domaines de spécialité lors de la création de corpus spécialisés. Compte tenu des contraintes imposées lors de la construction de nos corpus (adaptation à un traitement automatique et extraction des documents depuis le Web), la théorie la plus adaptée à notre travail est celle de Bowker et Pearson (2002), des LSP (*language for special purpose*). Un corpus spécialisé est alors composé de documents relevant d'une langue de spécialité. Celles-ci sont caractérisées par une origine (un domaine particulier), une nature (variété de la langue générale) et une fonction (transmission d'informations). Cette théorie permet de construire des corpus spécialisé malgré les inconvénients du Web (origine des documents souvent inconnue) puisqu'elle est moins restrictive que les théorie des sous-langages ou des communautés de discours. Nous avons enfin passé en revue les différents types de corpus multilingues et leur exploitation. Cet état de l'art nous permet de faire un comparatif entre les deux types de corpus

multilingues les plus utilisés, les corpus parallèles et les corpus comparables, ainsi que d'envisager leurs différents usages. Les corpus comparables, auxquels nous nous intéressons, sont des ensembles de textes dans plusieurs langues qui ne sont pas des traductions mais partagent un certain nombre de caractéristiques. Le choix de ces caractéristiques et la comparabilité des corpus étant des concepts un peu flous, nous nous y intéressons dans le chapitre suivant.

CHAPITRE 2

Corpus comparables

2.1 Introduction

Nous avons présenté dans le chapitre précédent l'objet de notre étude : les corpus comparables spécialisés. Nous en avons donné une définition générale : ils sont composés de textes dans plusieurs langues partageant certaines caractéristiques. Le choix de ces caractéristiques dépend de l'utilisation des corpus. Un état de l'art nous a permis de dégager des groupements récurrents de caractéristiques : pour les corpus comparables de langue générale, le choix se porte souvent vers le genre, la période, la médium, tandis que pour les corpus spécialisés, le choix se porte plus souvent sur le thème, le genre, le type de discours... Nous avons également introduit la notion de comparabilité : dans quelle mesure les textes d'un corpus sont-ils comparables ? La comparabilité dépend des caractéristiques communes aux textes : plus ils ont de caractéristiques communes, plus ils sont comparables. Cependant, le choix des caractéristiques dépendant des objectifs applicatifs du corpus, la comparabilité en dépend-elle aussi ?

Le but de ce travail est d'explorer la notion de comparabilité des corpus dans le cadre des langues de spécialité. Nous abordons dans un premier temps la notion de comparabilité, sa définition ainsi que ses ambiguïtés. Nous présentons dans un second temps les différents critères de comparabilité utilisés dans la littérature puis nous terminons sur la construction des corpus comparables et l'impact de ces critères sur celle-ci. Enfin, nous présentons notre corpus d'étude, sa construction et ses caractéristiques.

2.2 La comparabilité des corpus

Nous nous basons ici sur la définition des corpus comparables de Bowker et Pearson (2002, p. 93) présentée dans le chapitre précédent : les corpus comparables sont des ensembles de textes en plusieurs langues qui ne sont pas des traductions mais partagent un certain nombre de caractéristiques.

Les caractéristiques communes aux textes, ou critères de comparabilité, permettent de garantir une certaine homogénéité dans le corpus. Déjean et Gaussier (2002) rassemble ceux-ci en deux catégories :

- les critères qualitatifs : critères utilisés en stylistique tels que le genre, l'auteur, la période, le médium, etc. ;
- les critères quantitatifs : critères basés sur les mesures de fréquence de certains traits linguistiques.

La notion de comparabilité formalisée grâce à des traits communs aux textes est floue. Le choix de ces traits communs dépend des objectifs fixés lors de la construction. Des textes écrits sur la même période et tirés d'un même type de média permettront, par exemple, de comparer la diversité des thèmes abordés selon la langue et la culture. Dans le cadre de l'extraction automatique d'informations multilingues, des textes portant sur un même thème et émanant d'un même genre seront plus adaptés. La restriction d'un corpus permet en effet de garantir dans ses textes des structures syntaxiques et un lexique communs (Déjean et Gaussier, 2002).

Dans le domaine de l'aide à la traduction, Zanettin (1998) utilise différents corpus comparables afin d'analyser leur utilité dans le processus de traduction. Le premier, composé de documents d'un même genre (journalistique), sur un même domaine (le sport), issus de journaux anglais et italiens datant de 1992 est utilisé afin d'assister et de former des traducteurs. Le second, composé d'articles médicaux anglais et italiens traitant de l'hépatite C, sert à étudier la terminologie d'un domaine. Le troisième permet d'analyser comment un même thème est traité selon les sources et les langues, il est composé de documents issus de journaux nationaux anglais et italien traitant de la France. Ces corpus sont relativement petits (moins d'un million de mots). Lewis (2005) fait appel à un corpus comparable de deux millions de mots contenant des discours politiques prononcés entre 1995 et 2002 en anglais, français et irlandais, afin de faire une analyse contrastive des connecteurs adversatifs. Morin et Daille (2004) utilisent, quant à eux, un corpus comparable spécialisé de cinq millions de mots français et anglais tirés d'une revue spécialisée du domaine de la foresterie afin d'en extraire des terminologies bilingues.

Le choix de ces traits communs influe sur le *degré de comparabilité* des corpus, notion permettant de quantifier la comparabilité des textes d'un corpus. Cette donnée reste théorique, puisqu'elle paraît difficile à calculer sans tenir compte des objectifs applicatifs du corpus. Il existe à l'heure actuelle peu de travaux abordant le calcul de ce degré. Si les caractères communs de corpus comparables peuvent être si variés, comment calculer, mais aussi formaliser le degré de comparabilité de ces corpus ?

C'est en comparant deux notions *proches* de la comparabilité que nous tentons de répondre à cette question.

2.2.1 Comparabilité et similarité

Kilgarriff (2001) introduit la notion de *similarité entre corpus*. Il cherche à déterminer comment calculer la similarité (ou la distance) entre deux corpus. Cette notion paraît tout d'abord utile dans un cadre purement théorique, mais elle fait aussi référence à différents problèmes plus concrets. On peut par exemple se demander lors de la construction d'un corpus s'il n'existe pas d'autres corpus disponibles et similaires ; un automate valide sur un corpus pourra être valide sur un corpus similaire ; quel coût peut avoir l'application d'une méthode d'un corpus émanant d'un domaine sur un corpus d'un autre domaine... Cependant, la quantification d'une telle mesure fait face à de nombreuses difficultés. En effet, juger de la similarité entre des éléments complexes et multidimensionnels est une tâche subjective, des éléments peuvent être similaires sous un certain angle, mais totalement opposés sous un autre. On peut alors assumer qu'il existe autant de mesures de similarités qu'il existe de contextes d'étude. Il part cependant dans son étude de l'hypothèse qu'il existe une telle mesure, permettant d'évaluer la similarité entre deux corpus. Celle-ci est intimement liée à la notion d'homogénéité des deux corpus. Kilgarriff appelle homogène un corpus composé de documents du même type, au sens de Biber (1989) : un type de texte correspondant à un ensemble de textes au sein duquel certains traits linguistiques sont fortement corrélés. Cette similarité étant principalement basée sur des fréquences de termes, elle ne peut être appliquée au contexte multilingue. Mais la démarche de Kilgarriff permettant d'évaluer différentes mesures, peut être intéressante. Pour quantifier une telle mesure, il faut disposer d'un ensemble de corpus sur lequel les similarités sont connues (*Known-similarity corpus*). Il propose par exemple de construire N corpus composés de textes de type A ou B. Le corpus 1 est alors composé de 100 % de textes de type A, le corpus 2 de 90 % de textes de type A et 10 % de textes de type B, etc. Il est ainsi possible d'évaluer les différentes mesures de similarité sur des corpus sur lesquels le résultat attendu est connu. En adaptant cette méthode à des mesures de similarité particulières, ainsi qu'une définition de l'homogénéité adaptée aux corpus comparables, il est possible d'évaluer différentes mesures du degré de comparabilité.

2.2.2 Comparabilité et représentativité

Nous avons présenté dans le chapitre précédent (section 1.2) la notion de représentativité dans un corpus. Rappelons que, selon Habert (2000), un corpus se caractérise par : la définition de la population ciblée, l'éventail des distributions linguistiques au sein de cette population et le nombre de mots par texte et le nombre de textes par type. Construire un corpus représentatif consiste donc, dans un premier temps, à délimiter la population cible. En réalité, cela revient bien souvent à définir des classes d'appartenance des documents (cf. chapitre 2 section 1.3). Ces classes peuvent être définies *a priori*, auquel cas elles correspondent à des catégories définies manuellement, comme le domaine, le thème, le genre… Ces classes peuvent aussi être déterminées *a posteriori*, sur la base de caractéristiques linguistiques préalablement observées dans un corpus. La classification qui en résulte se base sur des corrélations de traits linguistiques entre les textes. Biber (1989) nomme ces classes *types de textes*, en opposition aux catégories définies par des caractéristiques externes aux textes. Ces types, pouvant ne correspondre à aucune classification cohérente pour l'homme, garantissent cependant une homogénéité linguistique dans le corpus.

Ces méthodes, permettant de garantir la représentativité au sein d'un corpus, font écho à la comparabilité des corpus. En effet, dans les deux cas, les textes sont minutieusement classés dans différentes catégories de façon à homogénéiser le corpus. La finalité semble pourtant différente. Dans le cadre de la représentativité des corpus, le but est de créer un corpus représentatif d'une population langagière, alors que dans le cadre de la comparabilité, le but est de garantir que les textes soient suffisamment comparables pour pouvoir exploiter le corpus. Un corpus comparable est-il représentatif ? Un corpus représentatif peut-il être qualifié de comparable ?

Il apparaît dans un premier temps que ces deux notions n'interviennent pas au même niveau. La question de la représentativité d'un corpus intervient au moment de la définition de l'étude portant sur le corpus, lorsque la population langagière cible est identifiée. La comparabilité intervient au moment de la construction du corpus. Si la population langagière peut être représentée et est suffisamment restreinte et délimitée pour chaque langue, alors le corpus multilingue qui en résulte pourra être considéré comme comparable. À l'inverse, un corpus comparable dont les critères sont rigoureusement définis peut être représentatif.

2.2.3 Calcul de la comparabilité

Les quelques travaux s'attelant à la tâche du calcul de la comparabilité au sein d'un corpus le font dans le cadre de l'extraction de lexiques multilingues. Déjean et Gaussier (2002) proposent ainsi un *critère minimal*, permettant de déterminer si deux corpus peuvent être qualifiés de *comparables*[1]. Ce critère est le suivant :

> « *Deux corpus de langues l_1 et l_2 sont dits comparables s'il existe une sous-partie non négligeable du vocabulaire du corpus de langue l_1, respectivement l_2, dont la traduction se trouve dans le corpus de langue l_2, respectivement l_1.* »

Déjean et Gaussier (2002) se basent ici sur le résultat de l'exploitation de leurs corpus afin de statuer sur leur comparabilité, ce qui devrait être déterminé avant l'exploitation. Cependant, cette définition peut être utilisée afin de déterminer la comparabilité d'un corpus avant même de l'exploiter dans le cadre d'extractions terminologiques multilingues.

[1] Ce critère minimal servira de base afin d'affirmer qu'un corpus est utilisable dans le cadre de l'extraction terminologique.

Saralegi et al. (2008) utilisent quant à eux la notion de similarité afin d'évaluer la comparabilité dans un corpus comparable anglais - basque composé de documents de vulgarisation scientifique destiné à l'extraction de terminologies bilingues basée sur la comparaison des contextes des mots. Selon eux, cette similarité s'instancie au niveau sémantique et ils basent leur travail sur l'hypothèse suivante : plus les documents sont similaires, plus le contexte des mots doit l'être[2]. La comparabilité dans leurs corpus correspond donc à la quantité de contextes de mots similaires d'une langue à l'autre, garantissant une certaine efficacité lors de l'extraction de traductions. Ils considèrent pour cela chaque document comme le vecteur contenant les mots les plus représentatifs. Pour cela, ils étiquettent et lemmatisent leurs documents et en extraient les termes les plus représentatifs : noms propres, entités nommées et termes composés principalement. En traduisant les termes de chaque vecteur à l'aide d'un dictionnaire, ils peuvent ainsi comparer des documents issus de deux corpus de langues différentes. À l'aide de la métrique cosinus, ils mesurent la distance entre chaque paire de textes de langues différentes (Saralegi et Alegría, 2007). Partant d'un corpus C_1, composé de documents eu_i ($i \in [0 \dots n]$) en langue L_1 et C_2, composé de documents en_j ($j \in [0 \dots m]$) en langue L_2, ils créent une matrice des distances d_{ij} entre chaque couple bilingue de documents (eu_i, en_j) issus des corpus C_1 et C_2 :

$$\begin{bmatrix} d_{11} & \dots & d_{1j} & \dots & d_{1m} \\ \vdots & & \vdots & & \vdots \\ d_{i1} & \dots & d_{ij} & \dots & d_{im} \\ \vdots & & \vdots & & \vdots \\ d_{n1} & \dots & d_{nj} & \dots & d_{nm} \end{bmatrix}$$

À l'aide de cette matrice, ils calculent ensuite la similarité entre les deux corpus à l'aide de la *Earth Movers Distance* (EMD). Cette méthode, uniquement basée sur le lexique, est directement héritée de la notion de similarité de Kilgarriff. Elle permet de mesurer si les contextes des mots d'un corpus dans une langue sont similaires à ceux d'un corpus dans une autre langue. Cette adaptation d'une technique à la base monolingue à un environnement bilingue leur permet d'évaluer l'homogénéité lexicale de deux corpus. Leur hypothèse de départ étant que si deux corpus sont comparables, les contextes des mots des textes de ces corpus le sont aussi, cette méthode peut être considérée comme une mesure de la comparabilité. Elle semble efficace sur un nombre limité de documents, mais les résultats de son extension à la mesure de la comparabilité au sein d'un corpus sont moins bons. Le principe de cette mesure est proche de celle de Déjean et Gaussier (2002) : les résultats de l'exploitation du corpus dans le cas de Déjean et Gaussier (2002) (un vocabulaire commun), le traitement effectué sur le corpus dans le cas de Saralegi et al. (2008) (contexte des mots similaires). La comparabilité paraît donc directement liée aux objectifs du corpus comme le confirment ces deux applications.

2.2.4 Bilan et définition

Nous avons vu que la communauté utilisant les corpus comparables afin d'en extraire des lexiques multilingues s'est intéressée à la notion de comparabilité et en a proposé des définitions ainsi que des méthodes de calcul. Pascale Fung (Fung et Yee, 1998; Fung, 2000) utilise des corpus comparables afin d'en extraire des lexiques multilingues. Elle caractérise la comparabilité selon différentes hypothèses :

- Sur un même thème, les mots ont des contextes comparables quelque soit la langue ;

[2] « The more similar the documents are, the more similar the context of words should be [...] ».

- Dans une période ou un domaine communs, les mots sont utilisés selon la même forme d'usage (« usage pattern »).

Le choix de ces critères communs est dicté par les objectifs d'exploitation du corpus fixés au préalable. Les corpus comparables sont exploités dans de nombreux travaux à des fins variées. Ainsi, le degré de comparabilité et les critères de comparabilité doivent dépendre de l'objectif applicatif du corpus.

Plus nombreux sont les critères communs et plus comparables sont les textes du corpus. Il en va de même pour la granularité de ces critères : plus ces derniers sont précis et plus les textes sont comparables. Cependant, plus ils sont nombreux et précis, plus il est difficile de récolter de textes correspondants. Un compromis doit être trouvé entre la quantité, la précision des critères et la taille du corpus.

La comparabilité semble donc, dans certains cas, calculable, et dans d'autres plus empirique. Dans le cadre de la traduction automatique, la comparabilité équivaut à la garantie de trouver dans les textes bilingues (ou multilingues) du corpus le même vocabulaire et des contextes similaires. Les corpus n'étant exploités qu'automatiquement, une tendance à la construction de « gros » corpus peut être observée, afin de garantir cette couverture lexicale. De plus, cette comparabilité opératoire peut parfois restreindre les corpus comparables à un « sac de mots », privilégiant des correspondances au niveau lexical, au détriment des textes, qui sont pourtant l'unité minimale au sein d'un corpus (Péry-Woodley, 1995). La comparabilité équivaut ainsi à la garantie de la réussite du traitement automatique du corpus. Elle ne se base donc pas réellement sur des critères linguistiques ou des classifications pré-établies, mais certaines combinaisons de critères peuvent toutefois approcher la comparabilité souhaitée. La comparabilité dans ce cadre peut-être assimilée à la notion de similarité lexicale comme le font Saralegi et al. (2008).

Les linguistes, traducteurs et enseignants, faisant essentiellement une utilisation manuelle des corpus comparables, cherchent à affiner manuellement et empiriquement cette notion de comparabilité, en sélectionnant et classant rigoureusement les documents. L'unité texte prend tout son sens dans ces travaux, et les différents critères de comparabilité portent donc essentiellement sur les textes composant les différents sous-corpus. La comparabilité équivaut alors plutôt à la représentativité des corpus, l'objectif étant de construire un corpus de grande qualité, pleinement représentatif de certaines communautés de discours, thématiques, genres, etc.

Des mesures, telles que la similarité multilingue de Saralegi et al. (2008) semblent donc adaptées dans un contexte d'extraction automatique d'informations multilingues, mais peut paraître obsolète dans d'autres cas. Pour conclure cette étude, rassemblons l'ensemble des éléments collectés permettant de définir au mieux et de façon la plus générale possible les corpus comparables et la comparabilité :

> **Un corpus comparable est un ensemble de textes dans plusieurs langues (deux ou plus) qui ne sont pas des traductions et partagent des caractéristiques.** Ces caractéristiques peuvent être :
> - qualitatives : caractéristiques extra-linguistiques (auteur, période, thème…) ou catégories pré-établies (genre, type de discours…) ;
> - quantitatives : basées sur les mesures de fréquences de certains traits linguistiques (ex. : types de textes de Biber).
>
> Le choix de ces critères dépend directement des objectifs applicatifs du corpus et des méthodes employées pour atteindre les objectifs. Les choix récurrents de critères sont les suivants (Morin, 2007, p. 29, 30):
>
> **Pour les corpus comparables spécialisés** : domaine, thème, type de discours, genre, etc. ;
>
> **Pour les corpus comparables généraux** : thème, médium, période, genre…

Examinons maintenant les principaux critères de comparabilité utilisés dans la littérature.

2.3 Critères de comparabilité

Nous l'avons vu dans la partie 2.2.2, la comparabilité est assez proche de la représentativité. En effet, elle repose sur une catégorisation des textes afin de délimiter les variations de textes au sein du corpus. Nous exposions dans le chapitre précédent les différentes typologies de corpus recensées par Habert (2000). Ces différentes catégorisations sont aussi effectuées sur les documents des corpus comparables, nous y voyons des critères de comparabilité. Nous listons dans cette section les critères de comparabilité les plus fréquemment utilisés.

2.3.1 Domaine et Thématique

Comme c'est souvent le cas dans les travaux sur les corpus comparables, un premier niveau de comparabilité est assuré grâce à un domaine, voire une thématique commune. Rapp (1995) fut l'un des premiers à travailler sur l'extraction de lexiques à partir de corpus non-parallèles. Sans pour autant utiliser le terme « corpus comparable », les textes composant ses corpus portaient sur une thématique commune. Rapp supposait en effet que la qualité des lexiques extraits d'un corpus dépendait en grande partie de la *comparabilité thématique* des textes le composant.

De nombreux travaux s'intéressent à la construction de lexiques et de terminologies multilingues grâce aux corpus comparables spécialisés. Il existe peu de ressources pour les domaines spécialisés pourtant propices à l'apparition régulière de néologismes (Fung et McKeown, 1997). Selon Harris (1991), les langues de spécialité par opposition aux langues générales, sont caractérisées par un lexique et une syntaxe restreints. Ces particularités linguistiques permettent d'atténuer les phénomènes polysémiques (Fung et McKeown, 1997) et problématiques dans le cadre de l'extraction d'information multilingues.

2.3.2 Discours

Malrieu et Rastier (2002) donnent une représentation hiérarchique des différents niveaux de classification textuelle (voir figure 2.1) dans laquelle le discours se trouve être le niveau le plus général. Ducrot et Schaeffer (1999) définissent le discours comme « *tout ensemble d'énoncés d'un énonciateur caractérisé par une unité globale de thème (topic)* ».

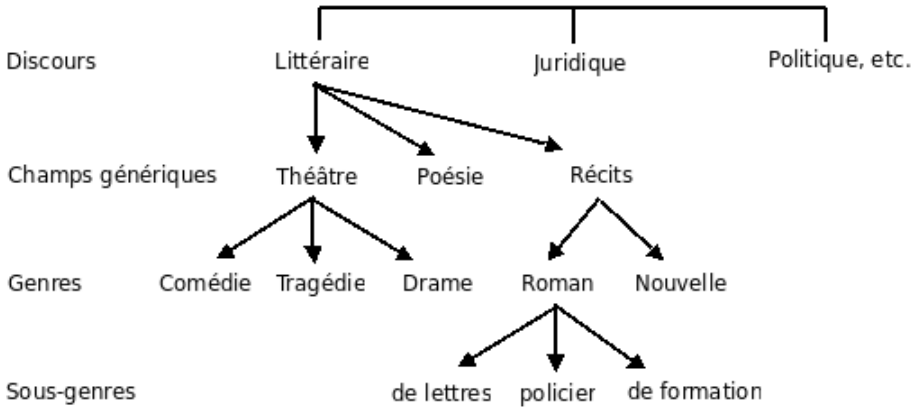


Figure 2.1 – Niveaux de classification selon Malrieu et Rastier

Bowker et Pearson (2002) présentent une notion similaire au discours : dans un corpus de langue de spécialité, différents niveaux communicationnels (*communicative settings*, (Bowker et Pearson, 2002, p. 36)) apparaissent : la communication d'expert à expert, d'expert à initié, d'expert à non initié et de professeur à élève. Dans des textes relevant de domaines spécialisés, deux niveaux communicationnels appelés *discours* en référence à Malrieu et Rastier (2002) apparaissent :

- le discours scientifique, regroupant tous les écrits d'experts ou d'initiés à destination d'experts ou d'initiés ;
- le discours de vulgarisation scientifique, dans lequel on trouve différents degrés de vulgarisation : les écrits d'experts ou d'initiés à des non-initiés, ainsi que les écrits de non-initiés à non-initiés, très fréquents maintenant sur le Web pour certaines thématiques du domaine scientifique et en particulier médical.

Cette catégorisation, moins détaillée que celle de Bowker, nous permet de conserver une certaine homogénéité dans le corpus, du point de vue lexical, mais aussi syntaxique.

2.3.3 Genres

Le terme *genre* est généralement utilisé pour classer différents types de textes littéraires (roman, comédie, etc.). En vue des récents développements sur l'analyse du discours, les *genres* font maintenant référence à des notions plus générales. Les travaux de traitement automatique du langage, et plus particulièrement l'analyse des genres, sont à l'origine de l'élargissement de cette notion, maintenant différente de celle utilisée dans le champs littéraire (Beauvisage, 2001).

Selon Biber (1989), les genres sont des catégories de textes distinguées spontanément par les locuteurs confirmés (matures) d'une langue ; par exemple, les genres de l'anglais incluent les romans, les articles de journaux, les éditoriaux, les articles de recherche… Établir une typologie des genres est une tâche subjective, les avis divergent sur ce qui constitue un genre et sur les critères d'appartenance à ce genre (Finn et Kushmerick, 2005). Ainsi, selon ces derniers, le genre serait une classe de documents qui émanent de l'étude du style de langage et du texte utilisé dans l'ensemble de ces documents, procurant aux utilisateurs des fonctions discriminatoires efficaces. Dans leur classification hiérarchique, Malrieu et Rastier (2002) (figure 2.1) placent les genres en dessous des discours.

À partir de 1995, Biber utilise le terme *registre* pour cette notion de genre plus large. Ces notions floues posent des problèmes d'ordre terminologique. Selon Péry-Woodley (2000), « *les frontières entre registre et genre sont inévitablement floues et il serait vain de chercher à en établir une liste exhaustive* ».

Le genre est un critère de comparabilité fréquent. Dans les travaux sur la langue générale, ceux-ci sont souvent couplés à une période et à un médium, par exemple des articles journalistiques dans (Fung et McKeown, 1997). Dans les travaux sur les langues de spécialité, le genre est souvent associé au type de discours et au domaine, par exemple des articles de recherche dans (Déjean et Gaussier, 2002). Un nombre restreint de genres apparaît dans les corpus comparables puisque ce critère est souvent mêlé à d'autres critères.

La notion de genres peut poser quelques problèmes, parmi lesquels la limitation des ressources pour un genre donné, mais aussi l'évolution incessante des genres sur le Web, illustrée par l'apparition récente des blogs, des forums… Cependant, au sein d'un même genre, les textes peuvent partager un style particulier (terminologie, syntaxe…)[3]. Par exemple, les articles scientifiques adoptent généralement une même structure (résumé, bibliographie…). Des textes partageant un même genre ont généralement une structure et un style propre.

[3] Nous détaillerons cette notion de style dans le chapitre suivant.

2.3.4 Autres critères

Les corpus comparables de langue générale contiennent, la plupart du temps, des textes tirés d'une même *période*. Lorsque le corpus contient uniquement des articles journalistiques tirés d'un domaine particulier (ex. : articles du Monde et du Times tirés de la rubrique économique), il est important de délimiter leur période de parution, d'une part pour filtrer une partie des articles et d'autre part pour garantir leur comparabilité. En effet, sans délimitation de la période, les thèmes peuvent être très différents sur plusieurs mois, et d'autant plus sur plusieurs années. Le style des auteurs change au cours des années, le vocabulaire et la syntaxe peuvent être différents.

Morin et Daille (2004) traitent un corpus dont les documents sont tirés d'un même médium : un magazine publié en plusieurs langues consacré aux forêts et industrie forestière. Néanmoins, la comparabilité dans ce cas revient celle d'un corpus ayant un un thème et un genre communs. De plus, les articles ne sont pas issus d'une période particulière.

Selon Déjean et Gaussier (2002), les critères peuvent être d'ordre qualitatif (comme ceux exposés dans cette section) ou quantitatifs, c'est-à-dire relatifs à des fréquences de traits linguistiques dans le corpus. Ces critères quantitatifs correspondent aux « typologies inductives » présentées dans le chapitre précédent. N'ayant pas rencontré de travaux faisant appel à ce type de critère, nous ne pouvons en discuter l'efficacité. Cependant, nous pouvons supposer qu'ils sont utiles dans le cadre de l'étude contrastive multilingue de traits linguistiques par exemple. De plus, les travaux en classification automatique par genres, discours, etc., se basant sur des caractéristiques externes et internes aux textes, nous laissent penser qu'une classe basée sur les caractéristiques propres aux textes peut être pertinente et homogène.

2.4 Construction des corpus comparables

Nous nous intéressons ici à la construction des corpus, et plus particulièrement des corpus comparables spécialisés. Nous considérons le document comme unité minimale dans les corpus. La constitution d'un corpus peut se décomposer en quatre phases essentielles : la délimitation du cadre de l'étude permettant de cibler la population langagière visée et les caractéristiques textuelles correspondantes ; la sélection et la collecte des données textuelles ; la normalisation et l'annotation des données sélectionnées, puis la documentation du corpus. Selon Biber (1993), avant de considérer la collecte des documents terminée (et passer à la phase de normalisation et annotation), l'ajustement des critères de sélection des documents peut être fait de façon cyclique (voir figure 2.2).

Nous détaillons dans les sections suivantes les quatre étapes de la construction d'un corpus, en tentant de faire un résumé des différents travaux sur la construction des corpus tout en ciblant celle-ci sur les corpus comparables et spécialisés.

2.4.1 Délimitation du cadre

2.4.1.1 Définition de la population

La délimitation du cadre de l'étude est l'étape essentielle de la construction d'un corpus. Comme nous l'avons vu dans la partie 1.3 du chapitre 1, pour pouvoir généraliser des phénomènes langagiers observés sur un corpus à un langage, celui-ci doit être représentatif de ce langage. Une attention toute particulière doit être portée à la définition de la population et de ses caractéristiques. Dans notre cas, la construction d'un corpus comparable spécialisé, il est nécessaire de définir la langue de spécialité. Les langues de spécialité se caractérisent par :

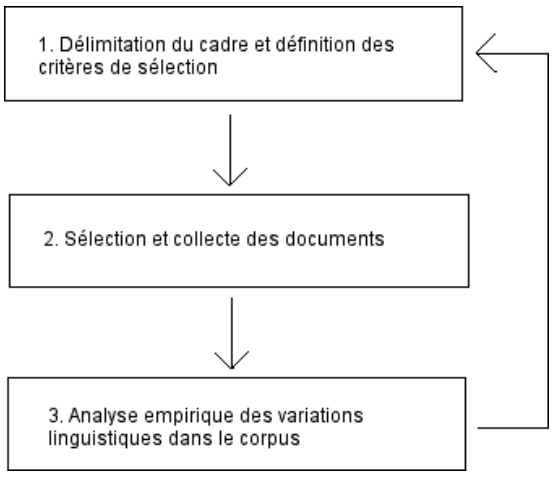


Figure 2.2 – Processus cyclique d'ajustement du corpus de Biber

une origine : une langue de spécialité appartient à un domaine de spécialité ;

une nature : une langue de spécialité est une variété de la langue générale ;

une fonction : une langue de spécialité sert à communiquer, à transmettre des informations.

Un domaine de spécialité doit donc être sélectionné. Il est important de vérifier qu'il soit actif et que suffisamment de documents soient publiés dans chaque langue du corpus.

Biber (1993, p. 380-385)[4] permet d'ajouter un niveau de restrictions supplémentaire. Ses *paramètres situationnels* permettent de restreindre le type de texte souhaité :

1. Canal : écrit / parlé / écrit lu ;
2. Format : publié / non publié ;
3. Cadre : institutionnel / autre cadre public / privé - interpersonnel ;
4. Destinataire :
   a. Pluralité : non compté / pluriel / individuel / soi-même ;
   b. Présence : présent / absent ;
   c. Interaction : aucune / peu / beaucoup ;
   d. Connaissances partagées : générales / spécialisées / personnelles ;
5. Destinateur :
   a. Variation démographique : sexe, âge, profession, etc. ;
   b. Statut : individu / institution dont l'identité est connue ;
6. Factualité : informatif-factuel / intermédiaire / imaginaire ;
7. Objectifs : persuader, amuser, édifier, informer, expliquer, donner des consignes, raconter, décrire, enregistrer, se révéler, améliorer les relations interpersonnelles, etc. ;
8. Thèmes : . . .

[4] Cité dans Habert et al. (1997, p. 152).

Utiliser et conserver ce type d'information lors de la constitution d'un corpus permet de garder un lien entre un objet d'analyse purement linguistique (les textes du corpus) et les paramètres situationnels de chacun des composants du corpus (Habert et al., 1997, p. 153).

Dans le cadre de la construction de corpus comparables, nous distinguons deux types de critères : les critères de constitution et les critères de comparabilité. Nous avons vu précédemment (section 2.2.2) que ces deux notions n'intervenaient pas au même niveau mais pouvaient être redondantes. La plupart[5] des travaux portant sur les corpus comparables utilisent les critères de comparabilité afin de construire le corpus. En effet, lorsque ces critères portent sur la thématique, la période ou les genres, ils conditionnent la recherche et la sélection des documents. Cependant, des critères de comparabilité quantitatifs peuvent permettre de mesurer la comparabilité *a posteriori*.

2.4.1.2 Taille du corpus

Quelle est la taille idéale d'un corpus ? Cette question est souvent débattue et les chiffres évoluent au gré des progrès technologiques. Dans les années 60, le Brown Corpus rassemblait 1 million de mots, ce qui était énorme à l'époque (Williams, 1999, p. 29). Le corpus BNC, datant des années 90, rassemble quant à lui 100 millions de mots à l'heure actuelle et il existe de plus gros corpus.

Cette course au « gigantisme » (*ibid.*) n'est pas sans rappeler le penchant qu'Habert nommait « gros, c'est beau ». Cependant, cette course n'a lieu d'être que dans les travaux sur la langue générale. Les travaux sur les langues spécialisées se penchent maintenant de plus en plus sur des corpus de taille moindre dont les documents sont rigoureusement sélectionnés. Il faut tout de même rester prudent sur la taille réduite, elle doit toujours être adaptée aux besoins de l'étude et les textes doivent être suffisamment nombreux et représentatifs pour pouvoir en tirer des généralités. Selon Williams (1999), la taille généralement acceptée d'un corpus spécialisé s'approche des 500 000 mots. Bowker et Pearson (2002) estiment qu'il n'existe pas de taille idéale pour un corpus, mais que celle-ci doit être un compromis entre les besoins de l'étude et le temps disposé pour la collecte.

Un corpus spécialisé peut avoir une taille restreinte car les langues de spécialité sont des sous-ensembles d'une langue plus restreints que la langue générale. Elles considèrent qu'un corpus spécialisé peut aller d'une dizaine de milliers de mots à plusieurs centaines de milliers. Enfin, un équilibre doit être trouvé entre longueur et largeur dans le corpus (Williams, 1999) : le nombre de mots est important, mais le nombre de textes l'est aussi.

2.4.1.3 Échantillonnage

Péry-Woodley (1995, p. 221) présente trois critères de constitution d'un corpus :

> « *(1) C'est d'abord au niveau de l'unité texte, unité fonctionnelle, que vont se faire les choix de constitution de corpus ; (2) Les textes devront comporter des indications permettant de les situer en termes de domaine et de registre ; (3) Pour que l'unité texte puisse être prise en compte dans l'analyse ou le traitement, le corpus doit être constitué de textes entiers et non d'extraits* ».

Travaillant sur des domaines spécialisés, il est peu fréquent de trouver des textes très longs et il est généralement plus pertinent de choisir des textes entiers.

Biber (1993) s'est penché sur les méthodes d'échantillonnage dans un corpus : quelles catégories de documents choisir et dans quelles proportions ? Selon lui, la définition de la population et des méthodes

[5] Tous ne détaillent malheureusement pas la construction des corpus.

d'échantillonnage priment sur la taille du corpus. La méthode d'échantillonnage choisie dépend des caractéristiques non linguistiques des textes. Si le corpus doit contenir des textes appartenant à différentes classes, il est important que le nombre de documents par classes soit homogène, et que le contenu même de ces classes soit représentatif. Pour cela, Biber propose entre autres d'utiliser la méthode de *stratified sampling* consistant à sélectionner de façon aléatoire plusieurs échantillons ou textes par classe.

2.4.2 Sélection et collecte des données textuelles

2.4.2.1 Sélection des données

Une fois les caractéristiques du corpus correctement définies et les documents recherchés ciblés, la recherche et la sélection de ces documents peut commencer. Nous l'avons vu dans le chapitre 1, les corpus construits de nos jours sont en grande majorité des corpus électroniques, afin de pouvoir leur appliquer des traitements informatiques. Bowker et Pearson (2002, p.61) listent deux principales ressources de textes électroniques :

1. Les CD-ROMs et bases de données en ligne ;
2. Le Web.

Les CD-ROMs et bases de données en ligne sont des collections de documents relevant d'une institution ou d'un organisme mettant à disposition ses textes sous format électronique. Le journal *Le Monde* met par exemple à disposition ses archives sous la forme de CD-ROM. L'avantage de ces ressources réside dans la garantie de leur qualité. En effet, elles sont généralement soumises à des charges éditoriales lourdes et la qualité de leur contenu est fiable. Cependant, ces collections sont le plus souvent payantes (pour les raisons énoncées précédemment), ce qui explique en partie pourquoi nombre de chercheurs et utilisateurs de corpus se tournent vers le Web.

Kilgarriff et Grefenstette (2003) s'interrogent sur le fait que le Web puisse être considéré comme un corpus. En effet, les linguistes et chercheurs en TAL s'y penchent car il constitue une incontournable ressource langagière. En 2003, ils estimaient à plusieurs centaines de milliards les documents sur le Web qui ont l'avantage d'être disponibles instantanément et gratuitement. De plus, il contient des documents dans toutes les langues et leur quantité laisse supposer la grande diversité de thème, de genre, etc. Sans forcément le considérer comme un corpus, nous le considérons toutefois comme l'une des ressources textuelle les plus importantes. Cependant, la profusion des données et l'absence totale de contrôle sur celles-ci oblige les constructeurs de corpus à une grande vigilance et à un soin particulier quant à la sélection des documents. Bowker et Pearson (2002, p. 61-62) présentent deux outils de recherche de documents sur le Web : les moteurs de recherche et les classifications thématiques[6]. Les moteurs de recherche sont des outils permettant de chercher des documents sur le Web grâce à des mots clés recherchés dans les textes et dans leurs méta informations. L'avantage de ce type d'outil est qu'il cherche dans l'intégralité des pages indexées par le moteur (plusieurs dizaines de milliards pour les moteurs les plus connus). La probabilité de trouver dans cette gigantesque collection des textes correspondant à des critères très précis n'est donc pas négligeable. De plus, l'indexation se faisant sur les termes des textes eux-mêmes, la recherche peut donc être très précise. L'un des inconvénients de ces moteurs de recherche est le temps nécessaire pour trier de telles quantités de données. De plus, l'indexation étant faite automatiquement, les résultats d'une requête peuvent contenir des textes ne correspondant pas du tout à la recherche effectuée. Si les mots clés sont peu précis, les résultats peuvent traiter de thématiques ou de

[6] « *Subject directories* » dans le texte.

domaines différents. Les moteurs de recherche nécessitent des mots clés. Si les mots clés de départ sont peu nombreux, il est possible de les étendre en utilisant :

- des dictionnaires de synonymes ;
- des mots sémantiquement liés trouvés dans les pages visitées.

Les classifications thématiques sont des arborescences thématiques créées manuellement dans lesquelles sont classées des documents. L'avantage de ces classifications est l'aspect manuel, qui garantit la pertinence des pages classées. Cependant, les documents disponibles sont beaucoup moins nombreux et les arborescences souvent peu précises (au regard des nombreux critères de sélection d'un corpus).

Pour certains domaines de spécialité, il existe un troisième type d'outils : les portails thématiques, sites contenant des documents portant tous sur un domaine particulier. Habert et al. (2001) utilisent par exemple des portails médicaux pour construire un corpus représentatif du domaine médical.

Quel que soit l'outil utilisé lors de la recherche, quelques précautions doivent être prises en plus des critères de sélection des documents. Tout d'abord, ne pas oublier que « n'importe qui peut publier n'importe quoi » sur le Web. Il est donc important de s'assurer que les documents sont attestés par une institution ou une organisation certifiant sa qualité. De plus, certains textes peuvent être répartis sur de nombreuses pages ou s'appuyer sur un grand nombre de documents multimédias, ce qui peut se révéler problématique pour la construction d'un corpus (nous perdons alors l'unité document que nous avons fixé comme unité minimale).

2.4.2.2 Collecte des données

Considérer le Web comme la principale ressource textuelle est quelque peu erroné. Les documents en ligne sont rarement des textes. Tous les formats de fichiers sont y disponibles. Parmi ceux qui contiennent principalement du texte, le format HTML (HyperText Markup Language) est le plus fréquent. D'autres formats de fichiers contenant du texte sont assez courants, comme le format PDF, PS, Microsoft Word… L'aspiration des documents peut se faire depuis la page Web, ou grâce à des logiciels « d'aspiration », tel que `wget`. La structuration des fichiers une fois aspirés ne doit pas être négligée, il est important de conserver le plus d'informations possibles sur chacun de ces fichiers (voir partie 2.4.3.2), de classer ceux-ci si nécessaire et de trouver un système de nommage de fichiers pertinent.

2.4.3 Normalisation, annotation des textes et documentation du corpus

2.4.3.1 Nettoyage et annotation

Une phase de nettoyage intervient après la collecte des documents. En effet, cette étape ne permet que de collecter un ensemble de fichiers sous différents formats et encodages. Des outils de conversion permettent de convertir facilement certains formats en textes, notamment le HTML et le PDF, les plus fréquents. Bien qu'efficaces, ces programmes laissent parfois passer des erreurs, notamment pour les fichiers PDF, qui sont dues en général à la mauvaise qualité des fichiers source. Un nettoyage manuel peut être envisagé, bien que certaines erreurs peuvent être négligées si elles sont peu nombreuses. Il est ensuite important de s'assurer que tous les fichiers utilisent le même encodage. Tout comme le format des fichiers, tous les encodages se trouvent sur le Web. L'American Standard Code for Information Interchange (ASCII) est la norme de codage informatique la plus connue et la plus répandue. Bon nombre de normes nationales sont des extensions de celle-ci, par exemple l'ISO 8859-1 ou Latin1 est une extension de l'ASCII auquel ont été ajoutés les caractères accentués pour coder la majorité des langues occidentales dont le français. Pour la langue japonaise, de nombreux encodages sont utilisés, parmi lesquels

EUC-JP, l'ISO 2022-JP… Deux projets visent à créer une norme d'encodage universelle : Unicode[7], développé par le Consortium Unicode et ISO/CEI 10646[8], développé par l'Organisation Internationale de standardisation et la Commission Électrotechnique Internationale. Ces projets sont directement liés, les trois organisations ayant décidé d'unir leurs efforts. Leur but est de donner aux caractères de tous les alphabets un nom et une identification numérique de manière unifiée. Ce code pourrait permettre de remplacer les normes nationales, posant des problèmes lors de communications internationales (un même code peut correspondre à différents caractères d'une langue à l'autre). Contrairement à l'ASCII et à la plupart des normes nationales dont les caractères sont codés sur 8 bits, les codages de l'Unicode utilisent 16 à 32 bits pour chaque caractère. Il existe plusieurs formats Unicode : UTF-8, UTF-16 et UTF-32. UTF-8 est le plus commun, il est couramment utilisé sur Unix et sur le Web. Le codage des caractères est de taille variable, ce qui est moins coûteux en mémoire. De plus, il est compatible avec la manipulation de chaînes en ASCII dans les langages de programmation ainsi qu'à la transmission de données dans des réseaux de systèmes hétérogènes.

Il est important, surtout pour des corpus multilingues, de bien choisir l'encodage. L'UTF8 est adapté à tous les alphabets et permet de stocker n'importe quel corpus multilingue. Les outils `iconv` ou `utrac` permettent de convertir un fichier d'une langue à l'autre.

Une fois le corpus nettoyé et normalisé, différentes opérations peuvent être effectuées sur celui-ci, afin d'y ajouter des informations. Selon sa finalité, plusieurs informations peuvent être nécessaires : morphologiques, syntaxiques, sémantiques, etc.

2.4.3.2 Normalisation et documentation

« Sans une documentation jointe, un corpus est mort-né » (Habert et al., 1997, p. 156). Des documents électroniques sont maintenant instantanément disponibles et ce sans aucun coût. Cela peut conduire à négliger les conditions de production et de réception du document, en n'enregistrant pas les références. Le corpus n'est pas réutilisable si ces informations ne sont pas conservées. Habert et al. (1997, p. 153) distinguent la représentation physique de la représentation logique des documents. La représentation physique correspond aux aspects visuels d'un texte. L'entrée d'un dictionnaire par exemple fait appel à des codes visuels qu'il est nécessaire de connaître pour la comprendre. La représentation logique correspond à la structuration d'un document. Un balisage logique permet d'indiquer quelles sont ses subdivisions et leurs relations. Il s'effectue en deux étapes (*ibid.)* : l'identification des différents éléments structurant le texte, donnant lieu à la définition d'une « grammaire de texte », ou définition de type de document (DTD) ; puis l'introduction de ces informations structurelles (balises) dans le document en respectant la DTD.

La norme la plus utilisée de nos jours est la norme TEI (Text Encoding Initiative). Elle a été créée en 1994 par un consortium composé de chercheurs de l'*Association for Computers and the Humanities* (ACH)[9], l'*Association for Computational Linguistics* (ACL)[10], et l'*Association for Literary and Linguistic Computing* (ALLC)[11] souhaitant proposer une norme avec suffisamment d'éléments pour pouvoir représenter tout type de texte et leur structure. Depuis 1994, les membres du consortium TEI tiennent à jour la TEI DTD, maintenant utilisée pour normaliser de nombreux corpus (le BNC par exemple).

[7] `http://www.unicode.org`
[8] `http://anubis.dkuug.dk/JTC1/SC2/WG2/`
[9] `http://www.ach.org/`
[10] `http://www.aclweb.org/`
[11] `http://www.allc.org/`

Un corpus correctement documenté doit contenir, pour chaque document, la trace des sources utilisées et les responsabilités éditoriales de sa constitution (*ibid.*) :

- les sources primaires utilisées ;
- les références bibliographiques précises ;
- les objectifs visés par la constitution du corpus ;
- les révisions qu'a subies le corpus au fil de sa mise au point.

Ces informations constituent l'ensemble minimal d'informations documentant un corpus. Elles sont utilisées par la norme TEI, mais aussi le Dublin Core[12].

2.5 Corpus d'étude

Afin d'effectuer une première étude sur les corpus comparables spécialisés et leur construction, nous avons créé un corpus d'étude. Ce corpus donnera lieu par la suite à diverses analyses linguistiques présentées dans le chapitre 3. Ce corpus d'étude est un corpus comparable en français, russe et japonais[13].

Nos corpus doivent répondre à différentes contraintes. Tout d'abord, nous situons l'unité minimale au sein de nos corpus au niveau du document. Ainsi, les critères de comparabilité des corpus dépendent des caractéristiques des documents. Comme nous l'avons vu dans la section 2.2.4, le degré de comparabilité maximal d'un corpus spécialisé peut être atteint si les documents partagent un domaine, un thème et un type de discours.

Dans cette partie, nous présentons étape par étape la construction de ce corpus. La délimitation du cadre permet de déterminer quel domaine de spécialité notre corpus représentera et l'incidence de ce choix sur les critères de comparabilité. En fonction des choix effectués, les documents sont sélectionnés et collectés. Nous choisissons alors la ressource dont sont extrait les documents, la méthode de recherche et l'échantillonnage des documents. Nous présentons ensuite la phase de nettoyage, d'annotation et de documentation des documents et du corpus. Nous présentons dans une dernière partie les caractéristiques principales du corpus ainsi constitué.

2.5.1 Délimitation du cadre

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 précisément, la thématique « diabète et alimentation ». Le choix du domaine médical permet de garantir un certain nombre de ressources accessibles, grâce aux portails médicaux et à la quantité importantes de publications et d'informations. La restriction à la thématique permet d'augmenter la comparabilité au sein du corpus, grâce à un lexique et à des particularités linguistiques propres. Toutefois, ce thème touche un large public et présente une garantie potentielle de collecter une diversité de documents sur le Web.

Différents niveaux de communication sont observables dans le corpus : communication d'expert à expert, d'expert à non-expert, de non-expert à non-expert… Nous les rassemblons en deux types de discours : vulgarisé (communications s'adressant aux non-spécialistes du domaine) et scientifique (communications s'adressant aux spécialistes).

[12] `http://dublincore.org/`

[13] Ce corpus a été construit dans le cadre du projet DECO, programme TCAN-CNRS 2004-2006 en partenariat avec le NII et l'INaLCO.

La taille du corpus est fixée à 200 000 mots par langue et par type de discours. Nous cherchons donc des documents issus du Web portant sur le thème « diabète et alimentation », que nous classons dans les deux catégories scientifique ou vulgarisé. Comme nous construisons un corpus comparable, il est important de ne pas avoir de traduction dans le corpus, pour que les textes soient pleinement représentatifs des particularités linguistiques et culturelles de chaque langue.

2.5.2 Sélection et collecte des données

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 :

- Recherche de pages web correspondant à la thématique visée ;
- Sélection des pages pertinentes ;
- 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 de vulgarisation mais avons cependant accordé une plus grande place aux documents écrits par des spécialistes au détriment des discussions sur des forums par exemple. Ils s'avèrent en effet être plus riches en vocabulaire et plus complets en contenu. La classification manuelle est donc basée sur ces heuristiques et 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 » dans les corpus d'apprentissage : les documents dont le type de discours était très ambigu et ceux sur lesquels les avis des personnes construisant le corpus et classant les documents divergeaient.

Nous n'avons collecté que des documents HTML et PDF (les autres formats étant peu présents, et pour certains impossibles à manipuler). La collecte des documents s'est faite en utilisant l'outil `wget`.

2.5.3 Normalisation et annotations

Afin de pouvoir mener à bien diverses analyses linguistiques, nous avons choisi d'appliquer sur notre corpus un analyseur morpho-syntaxique (en utilisant le système `Brill`), puis une lemmatisation (en utilisant `Flemm`).

2.5.4 Documentation

Des informations sur les fichiers sont inclues dans une version du texte au format XML suivant la norme TEI. Les informations générales sur les documents sont indiquées dans la partie *header* du fichier, délimitée par la balise `<teiHeader>` (TEI Consortium, 2007). Pour chacun des textes sélectionnés, nous conservons :

- la source du texte : son URL (champ `fileDesc` dans la TEI) ;
- la méthode de recherche utilisée : moteur de recherche et mots clés par exemple (champ `fileDesc` dans la TEI) ;
- la date de téléchargement du document (champ `profileDesc` dans la TEI) ;
- la langue du document (champ `profileDesc` dans la TEI).

2.5.5 Caractéristiques du corpus

Le tableau B.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 (SC = scientifique, VU = vulgarisé).

	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. caractères	2 668 783	2 845 114	493 587	1 154 773	2298306	2 165 768

Table 2.1 – 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 (`HTML` et `PDF`). 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 davantage de 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…).

2.6 Conclusion

Dans ce chapitre nous avons présenté une analyse de la notion de comparabilité. Celle-ci permet de mesurer dans quelle mesure deux corpus peuvent être qualifiés de comparables. Peu étudiée, cette notion a une grande importance. En effet, la comparabilité permet de garantir la réussite de l'exploitation du corpus. En étudiant les quelques propositions de mesure de la comparabilité et en la comparant à des notions similaires, la similarité de Kilgarriff et la représentativité, nous avons pu donner une description

plus précise. Il apparaît alors que la comparabilité est fortement liée aux objectifs applicatifs du corpus : un corpus comparable utilisé pour l'extraction de traductions ne sera pas forcément adapté à une étude linguistique inter-langues. Elle dépend alors des critères de comparabilité, caractéristiques communes aux documents d'un corpus. Nous avons proposé dans ce chapitre une étude des différents critères existants et proposons une méthode de construction de corpus comparables spécialisés avec pour caractéristiques communes un domaine, un thème et un type de discours. Cette méthode est ensuite appliquée afin de construire un premier corpus d'étude portant sur le thème du diabète et de l'alimentation. L'objectif de ce travail est de proposer ensuite un système d'aide à la construction de corpus comparables, basé sur les différentes étapes présentées dans ce chapitre. Nous considérons dans les chapitres suivants la première partie de la construction : la sélection des documents. Nos documents doivent partager trois caractéristiques : un domaine, un thème et un type de discours. Ce sont ces critères de comparabilité qui permettront de les sélectionner. Lorsqu'ils sont extraits du Web ou de portails spécialisés, leur domaine et leur thème peuvent être filtrés grâce aux mots-clés utilisés lors de la recherche. La reconnaissance du type de discours nécessite quant à elle de créer un système de reconnaissance automatique. Pour cela, nous présentons dans le chapitre 3 une analyse stylistique du corpus d'étude nous permettant de créer une typologie des types de discours scientifiques et vulgarisés en français et japonais contenant des critères caractérisant l'appartenance d'un document à l'une de ces classes. Cette typologie sera ensuite utilisée afin de créer un système de reconnaissance automatique du type de discours.

CHAPITRE 3

Analyse stylistique et typologies multilingues

3.1 Introduction

Dans le chapitre précédent nous présentions une analyse de la comparabilité ainsi qu'un mode opératoire pour la construction de corpus comparables spécialisés. Notre choix de caractéristiques communes s'est porté sur un domaine, un thème et un type de discours. Ces caractéristiques permettent de garantir une comparabilité élevée dans nos corpus. Le corpus d'étude construit porte sur le domaine médical, la thématique est celle du diabète et de l'alimentation et les documents sont répartis en deux classes : les documents scientifiques et les documents vulgarisés. Afin d'automatiser une partie du processus de constitution d'un corpus comparable spécialisé, le type de discours d'un document doit être reconnu automatiquement. Pour cela, nous devons trouver un ensemble pertinent de caractéristiques permettant de distinguer ces deux classes. Elles pourront par la suite être utilisées afin d'apprendre un système de classification automatique.

L'objectif d'une analyse stylistique sur corpus est de faire émerger un ensemble de critères relevant du style dans un document. Cette méthode est par exemple utilisée afin de trouver des critères discriminants entre deux classes. Un état de l'art des différents travaux adoptant cette méthode nous permet d'envisager son adaptation à notre travail : des documents extraits du Web, un contexte multilingue, deux classes… (section 3.2) L'analyse stylistique nous permet donc de faire émerger des documents une typologie composée de critères discriminants (section 3.3). En les analysant selon différents niveaux (section 3.4), nous avons rassemblé des critères en une typologie bilingue, générique et robuste (section 3.5).

3.2 Analyse Stylistique

Nous présentons dans cette section un état de l'art sur l'analyse stylistique. Dans un premier temps nous décrivons le principe général et la finalité de cette méthode. Nous décrivons ensuite l'un des travaux fondateur de cette méthode. Les différentes démarches pouvant être adoptées lors d'une analyse stylistique sont ensuite détaillées. Nous terminons cette section par une synthèse dans laquelle nous étudions l'adaptation de cette méthode à notre cas.

3.2.1 Définition

Avant de définir la notion d'*analyse stylistique*, tâchons de préciser ce qu'est le *style*. Au sens littéraire, le style correspond à « *l'ensemble des moyens d'expression (vocabulaire, images, tournures de*

phrases, rythme) qui traduisent de façon originale les pensées, les sentiments, toute la personnalité d'un auteur » (TLFI). Le style rassemble ainsi l'ensemble des éléments du texte caractérisant son auteur. La stylistique vise à étudier l'ensemble des « modes de composition » et « procédés littéraires » utilisés dans un texte par son auteur (TLFI), c'est-à-dire analyser dans un texte l'ensemble des éléments traduisant la présence de l'auteur.

Bally (1952, p.59) introduit la stylistique en ces termes :

> « *La tâche de la stylistique consiste à rechercher quels sont les types expressifs qui, dans une période donnée, servent à rendre les mouvements de la pensée et du sentiment des sujets parlants, et à étudier les effets produits spontanément chez les sujets entendants par l'emploi de ces types.* »

La première partie de cette définition reprend celle exposée précédemment : la stylistique consiste à analyser les éléments caractérisant l'auteur, sa personnalité et ses sentiments dans un texte. Cependant, la deuxième partie ajoute une nouvelle dimension à cette discipline : l'étude de l'impact du style d'un auteur chez son lecteur. Le style est étudié dans différentes disciplines, telles que la linguistique et la littérature, en vue d'analyser un texte sous différents angles. Il se manifeste à différents niveaux dans un texte : syntaxique, lexical, sémantique, etc. À l'initiative de Biber (1989), de nombreuses études statistiques sur les textes se sont basées sur des représentations regroupant différents niveaux d'analyse. Les travaux de Karlgren (1999) et Argamon et al. (2007) ont établi de ces études statistiques un lien entre l'informatique et la stylistique.

Karlgren (1999) définit le style comme la variation entre les différentes façons d'exprimer une même idée. Un auteur se trouve toujours face à plusieurs choix lors de la rédaction d'un texte : le choix des termes, des structures syntaxiques, du public visé, etc. Ils sont faits selon des préférences personnelles, mais aussi en fonction des lecteurs visés, et des textes similaires, c'est-à-dire véhiculant la même idée. Dans un certain sens, le style est un moyen d'orienter le lecteur dans sa compréhension du sens du texte.

Karlgren (1999, p. 1) définit alors la stylistique comme :

> « *the variation in a text that is not primarily topical* [1]*, that has not to do with meaning* [2]*.* »

Afin de mieux comprendre cette définition, il paraît important de préciser ce que sont les variations de thème et de sens dans un texte. Selon Ducrot et Schaeffer (1999, p. 345), « *le thème (. . .) d'un acte d'énonciation, c'est ce dont parle le locuteur, c'est l'objet du discours, ou comme disaient les linguistes du début du siècle, le sujet psychologique, le propos* ».

Meaning peut se traduire en français par *signification* ou *sens*. La différence entre ces deux termes étant débattue, nous nous sommes basés sur les définitions de Rossignol (2005, p. 196, 197) :

Signification : contenu d'une unité linguistique, défini en faisant abstraction des contextes et des situations ;

Référence : être ou état de choses « du monde réel » que veut évoquer un mot ou un syntagme ;

Sens : le sens doit être distingué, d'une part, de la *référence*, car il est défini au sein du système de la langue, sans référence au « monde réel », et d'autre part de la *signification*, qui est définie pour un mot de manière statique et absolue tandis que son sens ne prend corps qu'en contexte, c'est-à-dire par l'usage.

Selon Rastier (1989), la signification ne serait qu'un artefact des linguistes. Si la langue se définit selon des textes, un mot peut avoir un sens différent dans chaque contexte d'usage. Pourtant, les dictionnaires

[1] thématique ou sujet
[2] sens ou signification

et lexiques recensent un sens « standard » pour chaque mot. C'est celui-ci que Rossignol (2005) appelle *signification*.

Le *sens d'un mot*, quant à lui, ne se définit que par rapport à un contexte d'énonciation. La traduction du mot *meaning*, utilisé par Biber pour un texte, semble donc correspondre plutôt au sens, c'est-à-dire à la signification « absolue » d'un texte, indépendante du contexte paraissant impossible à décrire.

Argamon et al. (2007) distinguent dans un texte le « quoi » (ou thème) du « comment » (style). Selon eux, le style de l'auteur influe sur :

- Les émotions exprimées dans le texte ;
- Le genre du texte ;
- Le type de discours du texte ;
- La personnalité de l'auteur.

L'ensemble de ces caractéristiques, très hétérogène, a toutefois un point commun : elles sont indépendantes du thème du texte. Elles peuvent se caractériser par des variations de surface dans les textes. Le style d'un auteur est marqué par des caractéristiques indiquant son choix vers un mode d'expression parmi un ensemble de modes pour un contenu donné. L'auteur choisit donc certains mots, une certaine syntaxe, des stratégies de discours, etc. (*ibid.*).

Karlgren (1999, p. 1) affirme que la stylistique consiste à analyser les variations entre deux textes, indépendamment de leur thème et de leur sens. Cependant, comme il le remarque, la démarcation entre les variations thématiques et les variations stylistiques n'est pas nette. Certains thèmes sont intrinsèquement liés à un style particulier, par exemple les textes de lois qui sont toujours écrits dans un « jargon » juridique. De plus, le sens et le style semblent aussi être liés. Comment savoir si le sens qu'un auteur veut donner à son texte n'influence pas le style présent dans ce texte ? À l'inverse, le style d'un texte a t-il un impact sur son sens ? Théoriquement, l'analyse des variations stylistiques devrait être indépendante du thème et du sens ; mais en pratique ce n'est pas le cas.

Dans le cadre de l'analyse stylistique automatique, ce problème n'est pas forcément gênant. En effet, ces liens entre le style, le thème et le sens peuvent être enrichissants. Des variations thématiques peuvent être un complément aux variations stylistiques afin de caractériser un phénomène linguistique. La thématique d'un texte peut être un indice lors de la détermination de son genre. Un texte portant sur l'architecture logicielle fera plus probablement partie des genres *article de recherche* ou *rapport technique* que *roman* ou *recette de cuisine*.

L'analyse stylistique est donc une discipline visant à caractériser les variations stylistiques entre plusieurs textes, afin d'en dégager des caractéristiques mesurables (« *stylistic items* »).

3.2.2 Objectif et démarche de cette analyse

L'analyse stylistique consiste à déterminer quels sont les facteurs et les caractéristiques des variations de style observables entre des textes. D'une grande quantité de données textuelles doit émerger un ensemble d'éléments caractéristiques d'un phénomène stylistique. Ces *éléments stylistiques* (*stylistic items*, Karlgren (1999)), appelés aussi *critères*, sont basés sur la structure des textes et leurs caractéristiques linguistiques. Ils sont rassemblés et organisés de façon à former une *typologie*, c'est-à-dire une classification et une description des textes s'appuyant sur leurs caractéristiques internes.

Si l'analyse stylistique est effectuée dans le but d'automatiser la reconnaissance de phénomènes linguistiques, ces critères doivent être des quantités mesurables. On les appelle alors *critères opératoires*. Leur sélection se fait au détriment de critères représentant des notions subjectives ou non-opératoires. Nous appelons non-opératoires les critères subjectifs, comme l'expression de l'ironie dans un texte, ou non-calculables, comme la présence ou l'absence d'une introduction dans un texte.

Une analyse stylistique se base sur un ensemble de textes ou sur un corpus, avec pour objectif la caractérisation d'une ou plusieurs classes (correspondant alors à des catégories pré-établies ou à des phénomènes linguistiques). Dans le cas où les textes ne sont pas classés, elle permet de générer des classes de similarités, c'est la *démarche inductive*. Dans le cas où les documents sont classés, on parle de *démarche déductive* (Habert, 2000). Biber fut l'un des pionniers de la discipline, nous présentons dans un premier temps ses travaux, puis abordons chacune des démarches inductives et déductives.

3.2.3 Les travaux de Biber

Biber présente en 1988 une analyse multidimensionnelle de la langue anglaise, se basant sur des textes oraux et écrits issus de 23 genres (cf. table 3.1). Afin de caractériser ces genres, et plus généralement les textes écrits et oraux de l'anglais, Biber cherche à établir une liste la plus complète possible[3] de critères linguistiques. La typologie ainsi formée comporte 67 critères linguistiques organisés en 16 catégories, telles que les adverbes de temps et de lieu, les formes passives, les modaux… (voir la partie 3.4.1 pour les détails).

Genres de l'écrit	**Genres de l'oral**
Reportage de presse	Conversation en face-à-face
Éditoriaux	Conversations téléphoniques
Revues de presse	Conversations publiques, débats et interviews
Religion	Émissions
Compétences et hobbies	Discours spontanés
Textes traditionnels populaires	Discours préparés
Biographies	
Documents officiels	
Prose académique	
Fiction « générale »	
Fiction policière	
Science fiction	
Fiction d'aventure	
Fiction romantique	
Humour	
Lettres personnelles	
Lettres professionnelles	

Table 3.1 – Genres présents dans l'étude de Biber

Cette typologie, volontairement généraliste, contient un panel de critères couvrant différents niveaux d'analyse linguistique dont les variations peuvent caractériser les différents aspects d'un texte ou d'une classe de texte. À partir de ces critères, Biber souhaite déterminer un ensemble de *dimensions*. Les textes sont évalués selon certains paramètres, pouvant souvent être considérés comme des continuums (ex. : formel / informel). Une dimension correspond à un groupement de critères qui co-occurrent à haute fréquence dans le corpus. Six dimensions sont identifiées dans le corpus de Biber :

- Informational vs involved production ;

[3] Liste la plus complète possible mais non exhaustive, il semble difficile de dresser une liste exhaustive de tous les critères linguistiques d'un texte.

- Narrative vs non-narrative concern ;
- Explicit vs situation dependant reference ;
- Overt expression of persuasion ;
- Abstract - non abstract information ;
- On-line information elaboration.

C'est ensuite selon les variations des textes dans ces différentes dimensions que Biber caractérise l'anglais écrit ou oral, mais aussi les différents genres présents dans le corpus. Cette typologie est ensuite utilisée dans le cadre d'une étude sur les types de textes (Biber, 1989). Biber distingue les genres, catégorisations s'effectuant à partir de critères externes, des types de textes, définis à partir des caractéristiques linguistiques des textes. En effet, alors que les genres ne sont pas des classes homogènes en termes linguistiques (un texte du genre science fiction peut être d'un type abstrait, ou d'un type technique), les types peuvent être pleinement définis par un ensemble de caractéristiques linguistiques. En adoptant une démarche inductive, il s'attelle à faire émerger au sein d'un corpus des traits discriminants permettant d'identifier différents *types de textes*. Son hypothèse est que les traits permettant d'identifier des groupements de textes doivent venir des textes eux-mêmes. Ainsi, « *les textes doivent être le matériau des travaux d'ingénierie linguistique, l'objet, la source d'observation et non le moyen de vérifier des hypothèses* » (Beauvisage, 2001, p. 5). Il utilise ensuite la statistique multidimensionnelle afin de faire émerger des oppositions/associations de traits permettant ainsi d'opposer/rapprocher certains textes sous la forme de classes. Les textes, représentés dans un espace à n dimensions (n étant le nombre de critères utilisés) se trouvent entourés de textes similaires. De multiples regroupements de ce type apparaissent dans l'espace.

De cette analyse émergent huit *types de textes* :

- l'interaction intime personnelle ;
- l'interaction informationnelle ;
- l'exposé scientifique ;
- l'exposé savant ;
- la fiction narrative ;
- le récit ;
- le reportage situé ;
- l'argumentation impliquée.

De nombreux travaux se sont inspirés de ceux de Biber, en adoptant une démarche inductive (Bretan et al., 1998; Folch et al., 2000) ou déductive (Malrieu et Rastier, 2002; Vinot et al., 2003). Ce sont notamment sa démarche et ses méthodes qui sont utilisées, ainsi que sa typologie et ses dimensions dans les travaux que nous présentons dans les sections suivantes.

3.2.4 La démarche inductive

Se basant sur un corpus composé de documents non classés, la démarche inductive consiste à les analyser selon certains traits. Ces traits dépendent des objectifs fixés, c'est-à-dire des objectifs de l'analyse, de la nature souhaitée des classes qui en sont déduites. L'ensemble des traits jugés significatifs dans cette analyse forment alors une *typologie inductive*. Celle-ci fait apparaître des corrélations entre documents permettant de déterminer des classes de similarité au sein du corpus. Cette démarche peut permettre de confirmer une typologie existante. On peut parler ici de cadre *non-supervisé* (voir schéma 3.1).

Folch et al. (2000) s'attellent au problème d'hétérogénéité dans les corpus. En effet, il est maintenant possible, grâce à internet par exemple, de collecter de grandes quantités de données textuelles afin de constituer des corpus. Cependant, ces données, pour certains traitements, se doivent d'être homogènes

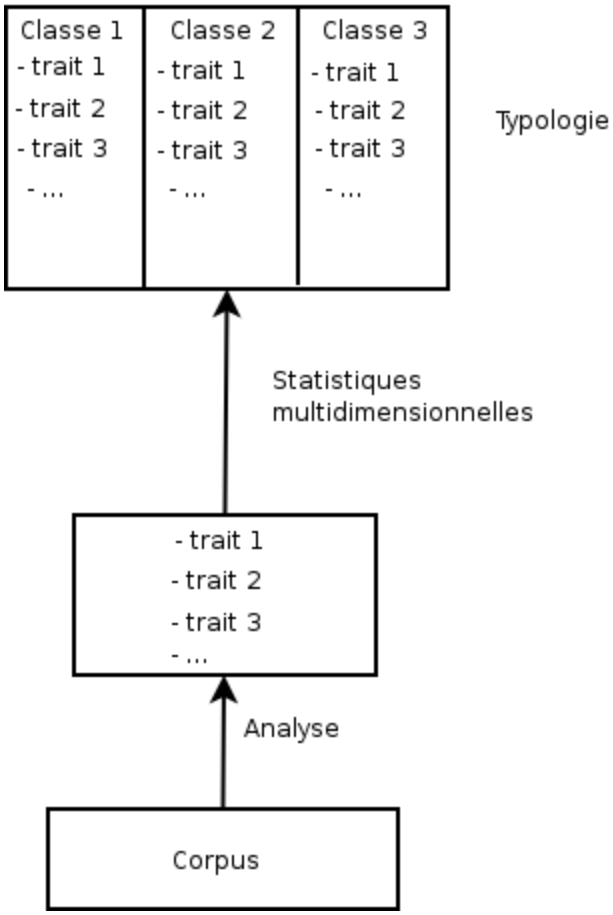


Figure 3.1 – Démarche inductive

(notamment du point de vue lexical, syntaxique, morphologique...). Ils proposent donc des méthodes et outils de profilage de textes, permettant de « calibrer » différentes parties d'un corpus, selon des caractéristiques linguistiques propres à chaque texte : vocabulaire, catégories morpho-syntaxiques, etc. Leur projet porte sur trois axes :

- l'analyse stylistique des textes ;
- le développement d'une architecture de travail sur corpus (comprenant étiqueteurs et outils d'extraction[4]) ;
- le développement d'un ensemble d'outils d'analyse statistique multidimensionnelle.

C'est au premier axe que nous nous intéressons ici. L'objectif de Folch et al. (2000) étant d'adapter la démarche de Biber au français afin d'obtenir des parties de corpus homogènes. Aux 67 critères dégagés par Biber et adaptés au français s'ajoutent des critères inspirés des travaux en analyse du discours de Sueur (1982), ainsi que ceux de Bronckart et al. (1985). Ils collectent ainsi environ 200 critères linguistiques appliqués à la langue française, appartenant aux catégories suivantes :

- Caractères : ponctuation, lettres capitales, chiffres, etc. ;
- Critères lexicaux : ensembles fermés de mots fonctionnels ;

[4] « extractor » dans le texte

- Catégories linguistiques à fine granularité : critères issus de (Sueur, 1982; Bronckart et al., 1985) ;
- Structure textuelle : titre, présence d'images, tableaux...

Ces 200 critères, auxquels ils appliquent la statistique multidimensionnelle, permettent d'observer quelles parties d'un corpus sont les plus homogènes, celles qui partagent le plus grand nombre de caractéristiques linguistiques et structurelles communes.

Ces *types de textes*, tout comme les groupements homogènes, relèvent d'une analyse linguistique, ils correspondent à des corrélations de critères linguistiques dans des textes. Les genres ou les types de discours sont au contraire des catégories existantes, socio-culturelles. Ces catégorisations sont basées sur les besoins humains, et sont souvent appréhendées de façon intuitive. On peut alors se demander si une démarche inductive, c'est-à-dire qui ne se base pas sur un classement des textes, peut permettre d'aboutir à des classifications pré-établies telles que la classification en genres.

3.2.5 La démarche déductive

Se basant sur un corpus dont les documents sont classés (manuellement), l'objectif de la démarche déductive est de créer une typologie des textes du corpus permettant de caractériser leur appartenance à une des classes du corpus. Cette typologie, appelée typologie déductive est le fruit d'une analyse des éléments des différentes classes, couplée à une analyse stylistique/linguistique de ses éléments. Cette démarche peut être assimilée au cadre *supervisé*, où les classes visées sont connues à l'avance (voir schéma 3.2).

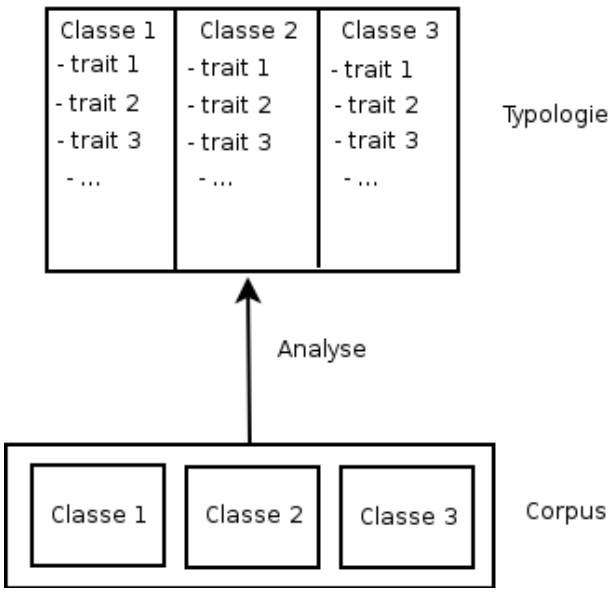


Figure 3.2 – Démarche déductive

Quelle que soit la démarche adoptée, deux méthodes permettent d'analyser les documents :

- la méthode contrastive, dans laquelle les documents de différentes classes sont comparés, afin de détecter quels éléments caractérisent leurs différences ;
- la méthode unitaire, dans laquelle chaque classe est analysée individuellement (par ses documents), afin d'en dégager des caractéristiques.

Il existe de nombreux travaux, particulièrement en classification textuelle, adoptant la démarche déductive, mais seuls quelques uns présentent de façon détaillée la démarche et les fruits de leurs analyses stylistiques. Les premiers travaux auxquels nous nous intéressons portent sur la classification de textes en genres. En effet, la notion de genre est assez difficile à définir, mais aussi à caractériser, c'est pourquoi elle est très étudiée.

Malrieu et Rastier (2002) se sont intéressés à la caractérisation automatique des genres à l'aide de critères morphosyntaxiques et à la spécification des « fonctionnements propres aux genres ». C'est en effectuant une analyse des variations morphosyntaxiques sur un corpus représentant une variété de genres que ceux-ci sont caractérisés. Leur corpus comprend plus de 2 500 ouvrages, soit 164 millions de mots, répartis sur 4 discours (scientifique, juridique, essayiste, littéraire), chacun d'entre eux étant ensuite divisé en champs génériques puis en genres. L'analyse de ce corpus les amène alors à distinguer 3 types de critères :

- Variables bibliographiques : titre de l'ouvrage, nom d'auteur, date de la première publication ;
- Variables quantitatives : la taille en Ko, des chiffres absolus concernant quelques grandes catégories morphosyntaxiques, le pourcentage de chaque catégorie par rapport à la catégorie superordonnée ;
- Pour chaque catégorie morphosyntaxique : moyennes par ouvrage ; moyennes par taille ; valeurs minimale et maximale de chaque variable par discours, champ générique et genre.

Une partie de ces critères est commune à ceux de Biber, comme le temps des verbes, les pronoms... Une fois ces valeurs attribuées à chaque texte, différentes expériences sont menées afin de caractériser les genres, mais aussi pour voir l'influence de ces critères sur les discours et les champs génériques : l'approche univariée (permettant de mesurer les distances relatives entre textes du même genre, du même discours ou du même champs générique), l'approche multivariée (recherche d'ensemble de critères apportant la plus grande variance entre les textes) et certaines techniques de classification.

Karlgren et Cutting (1994) se sont eux aussi intéressés aux genres, en essayant de créer un système de classification automatique. Leur objectif est de trouver un nombre minimal de critères permettant de classer correctement les documents en genres. Pour cela, ils se basent sur la liste de critères de Biber (1989), dont ils ne conservent que ceux qui sont fortement opératoires. À l'aide de ces derniers, ils effectuent une analyse discriminante : partant d'un ensemble de textes pré-classés et de leurs variations sur un certain nombre de critères, ils créent un ensemble de fonctions discriminantes permettant de distinguer les classes.

De nombreux autres travaux portent sur l'étude des genres, notamment ceux de Marina Santini (Santini et al., 2006; Santini, 2007) et Beauvisage (2001). Nous ne les détaillerons pas ici, la démarche étant sensiblement similaire aux travaux présentés ci-dessus.

Vinot et al. (2003) se sont penchés sur la détection de contenus racistes sur le Web. La principale différence avec les travaux exposés précédemment est que les documents ne doivent pas ici être classés selon leur genre, mais selon le point de vue énoncé. Ils se basent sur l'hypothèse suivante : « *la combinaison d'indices venant de plusieurs niveaux d'unités linguistiques (caractères, morphèmes, catégories syntaxiques, expressions complexes, isotopies sémantiques, code HTML, etc.) et basée sur une analyse plus globale des documents Web permet de mieux cerner et profiler le contenu de ces documents* ». Se basant sur un corpus de documents du Web classés manuellement selon l'opinion exprimée par l'auteur dans le texte, leur analyse stylistique vise à créer une typologie permettant de reproduire la classification de départ. C'est donc en effectuant une analyse déductive qu'ils dégagent un ensemble de critères portant sur différentes caractéristiques lexicales, syntaxiques... des documents, mais aussi en tenant compte de leur structure.

Un autre cadre de l'analyse stylistique est le domaine de la création et analyse de corpus. En effet, un certain nombre de travaux, sous l'impulsion de Biber se sont penchés sur la notion de *représentativité*

des corpus (voir chapitre 3). Selon Biber (1994), la représentativité réfère à la généralisation permettant de considérer qu'un échantillon représente toute la population. Ainsi, tout corpus ou sous-corpus (*sub-corpus*, Sinclair (1996a)) doit être pleinement représentatif des particularités linguistiques de l'ensemble considéré. Pour cela, une technique consiste à séparer le corpus en sous groupes ou strates correspondant à la population cible (*Stratified sampling*). Ces strates peuvent correspondre à :

- des *catégories de situation* (*situationnally defined text categories*) : basées sur des critères externes au corpus (intentions de l'auteur, position du texte dans la communauté...) ;
- des *catégories linguistiques* (*linguistically defined categories*) : basées sur des critères internes au corpus (donc émergeant après construction).

Ces catégories peuvent être filtrées manuellement ou automatiquement. Dans les deux cas, une analyse stylistique des documents peut être effectuée afin d'établir une liste de caractéristiques propres à la(es) catégorie(s), facilitant la classification. En effet, une fois ces strates ou catégories définies, les documents choisis pour être inclus dans le corpus doivent y être classés.

La démarche déductive permet de reproduire automatiquement des classification effectuées a priori, mais aussi de caractériser des catégories textuelles définies théoriquement et reconnues intuitivement (les genres par exemple).

3.2.6 Synthèse

L'analyse stylistique consiste à déterminer quels sont les facteurs et les caractéristiques des variations de style observables entre des textes. D'une grande quantité de données textuelles doit émerger un ensemble d'éléments caractéristiques d'un phénomène stylistique. Deux démarches peuvent être adoptées : la démarche inductive, se basant sur les caractéristiques des textes afin d'en dégager des classes de similarité ; et la démarche déductive, partant sur une classification des textes, et cherchant à dégager un ensemble de critères caractérisant ces différentes classes.

Au départ de toute analyse stylistique se trouve un ensemble de textes, ou corpus :

> « *les textes doivent être le matériau des travaux d'ingénierie linguistique, l'objet, la source d'observation et non le moyen de vérifier des hypothèses* » (Biber, 1989).

L'analyse inductive est généralement utilisée dans les cas où l'objectif est de rassembler un ensemble de textes selon certaines particularités linguistiques, comme dans l'étude de Biber (1989). Elle peut aussi être utilisée dans le cas où les textes doivent être rassemblés en classes de similarité (ou homogénéité) (Folch et al., 2000). Cette notion de groupe de textes homogènes est pertinente dans le cadre de la construction de corpus représentatifs. Cette démarche vise à répertorier des ensembles de textes similaires, sans pour autant accorder un sens à la classification générée.

Au contraire, les classes sont un élément central de la démarche déductive. Selon Malrieu et Rastier (2002), *« aucune typologie des textes fondée sur des critères définis indépendamment des genres (comme oral vs écrit, public vs privé, etc.) n'a permis d'isoler les genres. »* . On peut généraliser cette affirmation dans ce sens : dans le cadre d'une démarche déductive, la typologie créée doit se reposer sur les classes que l'on souhaite observer. En effet, que ce soit une classification en genres, en discours, en points de vue, la typologie dégagée de l'analyse est entièrement basée sur les classes.

Cependant, on trouve de nombreux liens entre ces deux démarches. Tout d'abord, Biber (1989) fut l'un des premiers à s'atteler à la création d'une typologie de la langue générale anglaise, en dégageant près de 70 traits linguistiques discriminants, aboutissant à 8 types de textes. Ce travail constitue une référence dans le domaine et sa démarche ainsi que ses critères sont cités et utilisés par un grand nombre de travaux, qu'ils soient inductifs ou déductifs. De plus, en s'inspirant dans une démarche déductive

d'éléments issus des connaissances théoriques, une part inductive est introduite. Selon Malrieu et Rastier (2002), l'opposition entre démarche inductive et démarche déductive fait référence à deux conceptions bien différentes du texte : l'une considère un texte comme un ensemble de chaînes de caractères, la deuxième comme une œuvre, « au sens général du terme ». Nous cherchons ici à caractériser les types de discours scientifiques et vulgarisés. Partant de classes existantes, nous allons donc adopter une démarche déductive. Nous allons voir dans la section suivante que l'analyse stylistique que nous réalisons fait appel à des connaissances théoriques linguistiques et contient donc une part d'induction.

3.3 Application de l'analyse stylistique

L'objectif de notre analyse stylistique est de caractériser les types de discours scientifiques et vulgarisés pour des documents issus du Web en français et japonais traitant de domaines de spécialité. Nous nous basons pour cela sur notre corpus d'étude, présenté dans la section 2.5 du chapitre 2. Les documents de ce corpus, issus du domaine médical, ont été classés manuellement selon leur type de discours. Nous adoptons donc une démarche déductive afin de caractériser la distinction entre ces deux classes. Elle sera de plus contrastive : en comparant conjointement des couples de documents scientifiques et vulgarisés, nous pourrons identifier les différences récurrentes et s'en inspirer pour créer la typologie. Cette démarche se distingue clairement de la vision en « sacs de mots » des corpus : le document représente l'unité fondamentale de cette analyse. Nous nous basons sur la globalité des éléments définissant le document afin de caractériser son type de discours.

La typologie rassemble l'ensemble des traits caractéristiques identifiés lors de l'analyse. Nous souhaitons qu'elle soit :

- *linguistiquement motivée* : de nombreux travaux se basant sur la typologie de Biber font l'hypothèse suivante : plus les critères sont nombreux, plus les chances de caractériser une classe seront élevées. L'un des problème de cette méthode est qu'il est parfois impossible de savoir quels sont les traits les plus discriminants. Nous cherchons ici, non seulement à créer une typologie pouvant être utilisée par des systèmes de classification automatique, mais surtout à caractériser les deux types de discours. Il nous paraît donc indispensable de bien définir la liste des traits insérés dans la typologie.
- *robuste* : l'analyse stylistique peut être considérée comme un processus nous permettant de passer d'un ensemble d'indices empiriques à une liste de traits motivés qui devront être opératoires, c'est-à-dire pourront être implémentés, afin de classer automatiquement les documents. Nous choisirons donc, dans la mesure du possible, des marqueurs de surface.
- *générique* : nous travaillons sur deux langues, il est donc nécessaire de veiller à ce que la typologie soit suffisamment générique pour couvrir tous les types de discours dans ces deux langues. De plus, elle devra être adaptée à d'autres domaines de spécialité, et pas seulement celui du corpus d'étude.

Nous présentons dans la section suivante les différents niveaux d'analyse des documents, puis les critères correspondants.

3.4 Structure de la typologie

Notre analyse stylistique, adoptant une démarche déductive et contrastive, s'est aussi appuyée sur des travaux fondateurs de la discipline, afin de dégager les différents axes d'analyse des documents. Nous parlons ici des travaux de Biber et Sinclair, évoqués précédemment.

3.4.1 Structure dimensionnelle

Nous avons présenté sa méthodologie dans la section 3.2.3, détaillons maintenant la typologie des textes anglais élaborée par Biber. Elle est composée de 67 critères linguistiques organisés en 16 catégories. Par souci de clarté, nous n'en donnons que quelques exemples pour chaque catégorie, la typologie complète se trouve en annexe A.

- Marqueurs de temps : passé, présent, *perfect aspect* ;
- Adverbes de lieu et de temps ;
- Pronoms et « pro-verbes » : pronoms personnels, démonstratifs, indéfinis, *pro-verb « do »* ;
- Questions : Questions en WH ;
- Formes nominales : *nominalizations (-tion, -ment, -ness, -ity)*, gérondifs ;
- Formes passives ;
- Formes statives (qui indiquent un état permanent) ;
- Éléments de subordination : compléments de verbes THAT, complément d'adjectif THAT, propositions WH, infinitifs, etc. ;
- Adverbes, adjectifs et locutions prépositionnelles : tous les adverbes, adjectifs attributs, adjectifs prédicatifs ;
- Spécificités lexicales : ration type/token, longueur moyenne des mots ;
- Classes lexicales : adverbes conjonctifs, dépréciatifs, amplificatifs, emphatiques, etc. ;
- Modaux : modaux de possibilité, nécessité, prédictifs ;
- Classes spéciales de verbes : verbes publics (assert, declare. . .), verbes privés (assume, believe. . .), etc. ;
- Formes réduites et structures peu usitées : contractions, omission de *that*, etc. ;
- Coordination : coordination phrastique, coordination de propositions indépendantes ;
- Négation.

Cette typologie contient un large panel de critères dont les variations peuvent caractériser les différents aspects d'un texte ou d'une classe de texte. C'est pour cette raison qu'elle est si souvent utilisée. À partir de ces critères, Biber souhaite trouver un ensemble de *dimensions*, groupements de critères qui co-occurrent à haute fréquence dans un corpus. Six dimensions sont identifiées dans le corpus de Biber :

- Informational vs involved production ;
- Narrative vs non-narrative concern ;
- Explicit vs situation dependant reference ;
- Overt expression of persuasion ;
- Abstract - non abstract information ;
- On-line information elaboration.

C'est ensuite selon les variations des textes dans ces différentes dimensions que Biber caractérise l'anglais écrit ou oral, mais aussi les différents genres présents dans le corpus.

3.4.2 Structure externe/interne

Dans son rapport sur les typologies de textes, Sinclair (1996b) distingue différentes catégories de critères, correspondant à différents niveaux d'analyse des documents :

- Niveau externe : les participants, le contexte social, les fonctions communicatives du langage, etc.
  - Origine : critères concernant l'origine du texte, susceptibles d'affecter sa structure ou son contenu ;
  - Structure : critères concernant l'apparence du texte, sa structure et sa relation aux éléments non textuels:

  - Buts : critères concernant les raisons de la création du texte et les effets désirés ;
- Niveau interne : la récurrence des phénomènes langagiers (*language patterns*) dans les parties du langage.
  - Thème : critères concernant le sujet, les domaines de connaissance du texte ;
  - Style : éléments du texte et de la langue qui peuvent être en corrélation avec les critères externes.

Ces différents niveaux d'analyse peuvent permettre de mieux organiser l'analyse stylistique en observant successivement les documents sous différents angles.

3.4.3 Structure de notre typologie des types de discours scientifiques et vulgarisés

Nous nous appuyons sur la typologie de Sinclair qui distingue deux dimensions : la dimension externe et la dimension interne. La dimension externe sera principalement composée de caractéristiques structurelles des documents. Quant aux caractéristiques linguistiques, correspondant à la dimension interne, elles seront séparées en deux catégories : les caractéristiques lexicales et les caractéristiques modales.

Nous détaillons et justifions l'usage de ces trois catégories dans la suite de cette section.

3.4.3.1 Les caractéristiques structurelles

Les textes de notre corpus, étant extraits du Web, comportent d'une part de multiples informations linguistiques, lexicales, syntaxiques ; mais aussi des informations externes, relatives à la structure même des documents, comme on peut le voir dans les documents `HTML`. Contrairement à de nombreux travaux se basant uniquement sur le texte et ses aspects linguistiques, nous exploitons la structure des documents afin de caractériser leur discours. Riboni (2002) montre par exemple que l'exploitation de certaines balises, couplée à l'analyse du texte en lui-même, permet d'améliorer les résultats d'une classification thématique. Notre typologie étant destinée à être automatisée, nous souhaitons que les critères la composant soient opérationnels, c'est-à-dire dont la reconnaissance peut-être effectuée automatiquement. Ainsi, nous n'avons conservé dans la partie externe que les critères structurels, l'origine et les buts des textes du Web étant très difficiles, voire impossibles à déterminer sans intervention humaine.

3.4.3.2 Les caractéristiques modales

La présence du locuteur et son rapport à l'interlocuteur sont différents d'un texte à un autre. Dans une discussion sur un forum, elle est fortement marquée, de même que celle de l'interlocuteur. Dans un article scientifique, elle est marquée tandis que l'interlocuteur est quasiment absent. Nous pensons que ces marques, caractéristiques de l'énonciation, peuvent être discriminantes pour distinguer les deux types de discours.

Il paraît évident que le ton du locuteur et sa façon de s'adresser à l'interlocuteur évolue selon son degré de spécialisation (sur le thème qu'il aborde) et celui de son interlocuteur. Ces évolutions sont caractéristiques de l'énonciation.

L'énonciation correspond à l'acte de production d'un énoncé par un locuteur à destination d'un allocutaire dans une situation de communication (contexte spacio-temporel particulier) (Riegel et al., 1994, p. 575). Les situations de communication dans lesquelles un acte d'énonciation est effectué sont caractérisées par (*ibid.*) :

- le locuteur et l'allocutaire, acteurs de la communication échangeant des informations ;
- un temps et un lieu ;

• l'environnement des protagonistes.

L'étude du phénomène d'énonciation vise à analyser les différents éléments d'une situation de communication et son contexte. Ces études se basent sur deux principaux types d'indices : les déitiques et les modalités. Les déitiques sont « *les unités dont le sens implique obligatoirement un renvoi à la situation d'énoncation pour trouver le référent visé* » (Riegel et al., 1994, p. 577). Les modalités sont des éléments exprimant l'attitude du locuteur par rapport à son énoncé.

Selon Bally (1952), une phrase peut s'analyser sous deux aspects : le *dictum*, ou contenu représenté et le *modus*, position du locuteur par rapport à la réalité du contenu exprimé. C'est ici au *modus*, ou à la réalité, que nous nous intéressons. Riegel et al. (1994, p. 580) distinguent les modalités d'énonciation et les modalités d'énoncés.

Les modalités d'énonciation « *renvoient au sujet de l'énonciation en marquant l'attitude énonciative de celui-ci dans sa relation à son allocutaire* », par exemple l'interrogation ou l'injonction ;

Les modalités d'énoncés « *renvoient au sujet de l'énonciation en marquant son attitude vis-à-vis du contenu de l'énoncé* », c'est-à-dire à la manière dont le locuteur apprécie son énoncé ;

Nous pensons que les modalités d'énoncés vont nous permettre de caractériser les types de discours scientifique et vulgarisé. En effet, l'attitude du locuteur au sein de son énoncé est bien différente dans un texte scientifique et dans un texte vulgarisé. Nous avons distingué dans cette partie deux théories de la modalité qui nous paraissent pertinentes : la théorie de la modalité Locutive de Charaudeau (1992), se basant sur la position du locuteur vis-à-vis de son interlocuteur, ainsi que la théorie de la modalité Irrealis de Givón (1994), se basant sur la réalité ou vérités des évènements énoncés par le locuteur. Nous les présentons dans la section 3.5.2.

3.4.3.3 Les caractéristiques lexicales

Charaudeau (1992, p. 65) note qu'une situation de spécialisation « *engendre un vocabulaire spécialisé, compris de manière précise par les seuls spécialistes [. . .] dans différents domaines de spécialité* »; alors que la situation de vulgarisation engendre « *un vocabulaire parallèle au vocabulaire spécialisé, [ou] une transformation du sens du mot spécialisé* ». Les caractéristiques lexicales, contrairement aux caractéristiques modales, permettent d'analyser le contenu de l'énoncé. L'analyse de la terminologie et du lexique employés dans des textes pourraient nous servir d'indices dans la détermination de leur degré de spécialisation.

3.5 Typologie des discours scientifiques et vulgarisés dans les langues française et japonaise

Nous présentons dans cette section la typologie issue de notre analyse stylistique. Celle-ci est organisée en trois parties, correspondant aux trois niveaux d'analyse présentés précédemment. Nous appelons critère un élément de la typologie, et marqueur(s) le ou les éléments nous permettant de détecter les critères dans nos documents. Les critères de cette typologie sont issus de l'observation des documents, mais aussi de travaux plus détaillés dont nous nous sommes inspirés. Nous n'avons conservé dans ce cas que les critères que nous jugions pertinents (c'est-à-dire susceptibles d'apparaître dans les documents et de caractériser l'un des types de discours). Certains d'eux ont dû être adaptés à notre étude, particulièrement aux domaines de spécialité, mais aussi à la langue. Pour cela, des linguistes japonais et russes ont travaillé sur l'adaptation des critères et de leurs marqueurs aux deux langues.

3.5.1 Caractéristiques structurelles

Les caractéristiques structurelles sont tous les éléments relatifs à la structure, l'origine et l'aspect graphique du document. Cette catégorie rassemble tous les éléments extérieurs au texte.

Patron d'URL : patrons génériques d'URL en France, par exemple `http://www.univ-***.fr` pour les universités ou `http://www.chu-***.fr` pour les hôpitaux.

Format de document : principaux formats des documents du Web (`html, pdf`...). Les articles de recherche sont souvent au format `pdf` ou `Microsoft doc`, tandis que les articles de journaux en ligne par exemple sont souvent dans des formats générant du `html` (`php, asp`...).

Méta-informations : présence de méta-informations dans le code des documents *HTML*. Ces méta-informations permettent de conserver au sein des fichiers certains renseignements sur leurs créations tels que le créateur du fichier, la date de création, mais aussi des informations sur le contenu du fichier comme les mots clés[5].

Titre de la page : présence d'un titre correspondant à la balise `TITLE` des documents *HTML*. Seuls quelques sites ne possèdent pas de titre.

Techniques de mise en page : utilisation de CSS ou de tableaux pour la mise en page des documents *HTML*. Les CSS sont encore réservés aux experts, bien que cet usage ait tendance à se démocratiser. Les créateurs de pages Web n'ayant pas ou peu de connaissances font d'avantage appel aux tableaux ou aux cadres[6].

Fonds : couleur ou image de fond des documents *HTML*. La balise `BACKGROUND` permet de mettre en fond de page une image ou une couleur. Les sites réalisés par des professionnels n'utilisent généralement qu'une couleur de fond, afin de ne pas surcharger les pages.

Images : présence d'images dans les documents *HTML*. Les images peuvent avoir différents rôles au sein d'une page. Certaines sont informatives (schémas, graphiques), certaines illustrent la page (photos) et d'autres servent uniquement à « décorer » la page ou à attirer l'attention du visiteur. Ces dernières, appelées *images de navigation* sont les plus fréquentes : elles servent à mettre en valeur un lien hypertexte, à attirer l'attention sur une phrase ou une partie de la page... Ces différents types d'images peuvent être distingués (notamment par leur taille), mais nous avons observé que les sites utilisant le plus d'images sont souvent les sites « grand public » ou les sites ayant vocation à l'être. Le nombre d'images nous paraît donc suffisamment discriminant.

Paragraphes : structuration du texte sous forme de paragraphes dans les documents *HTML*. Très couramment utilisés, ils sont toutefois beaucoup plus nombreux dans les pages créées avec des systèmes *wysiwyg* (*what you see is what you get*, système à interface graphique permettant de générer des pages `HTML`).

Listes : structuration du texte sous forme de listes d'items dans les documents *HTML*.

Liens : présence de liens hypertexte dans les documents *HTML*. Les liens peuvent être externes au texte (dans le menu, l'entête...) ou internes au texte. Les textes spécialisés contiennent généralement peu de liens, ou ceux-ci sont rassemblés à la fin du document.

[5] Ces mots-clés étaient utilisés il y a quelques années par les moteurs de recherche lors de l'indexation des documents. Ainsi, certains abus se produisaient : des pages Web contenaient des listes de mots clés très longue, recouvrant la plupart des recherches les plus populaires. Un gain en popularité au détriment de la pertinence des recherches. Ces méta-informations sont moins utilisées maintenant.

[6] Ce critère est toutefois dépendant des effets de mode sur la toile. L'usage des CSS a quelque peu détrôné les tableaux et cadres.

Typographie : présence de balises typographiques dans les documents *HTML* (italique, gras…). Utilisées généralement pour mettre en emphase certains passages ou mots d'un texte.

Nombre de phrases, nombre de caractères : estimation de la longueur des documents et de leurs phrases. Les textes scientifiques ont tendance à être plus longs.

Certains de ces critères se basent sur la structure `HTML` des documents. Le nombre des critères structurels se trouve donc réduit pour les documents `PDF`.

3.5.2 Caractéristiques modales

Deux théories nous ont parues pertinentes dans le cadre de notre étude, nous les présentons ici avec les critères qui leur sont associés.

3.5.2.1 Théorie de la modalité de Charaudeau

Selon Charaudeau (1992), tout acte de communication doit être considéré comme un « *dispositif au cœur duquel se trouve un sujet parlant, en relation avec un partenaire* ». Ce dispositif se compose de :

la situation de communication : « *cadre à la fois physique et mental dans lequel se trouvent les partenaires de l'échange langagier* » ;

les modes d'organisation du discours : « *principes d'organisation de la matière linguistique* », dépendants de « *la finalité communicative que se donne le sujet parlant* » ;

la langue : « matériau verbal » ;

le texte : « résultat matériel de l'acte de communication ».

Charaudeau décrit l'énonciation comme le phénomène qui « *témoigne de la façon dont le sujet parlant s'approprie la langue pour l'organiser en discours* ». Selon Benveniste (1970), « *avant l'énonciation, la langue n'est que la possibilité de la langue* ». Le discours se construit donc à travers l'énonciation et la langue n'est qu'un concept tant que ce processus n'est pas mis en œuvre. Comme l'illustre la figure 3.3, ce phénomène d'énonciation amène le sujet parlant à se situer vis-à-vis de son interlocuteur, vis-à-vis du monde qui l'entoure et de son propos (Charaudeau, 1992, p. 572).

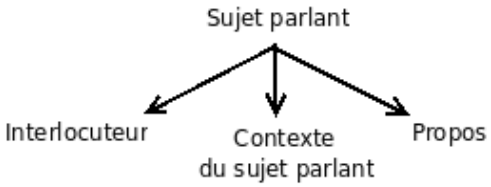


Figure 3.3 – Phénomène d'énonciation

Toute modalité implique un acte locutif qui est spécifié par un certain nombre de modalités. Nous distinguons la modalité d'énonciation de la modalité d'énoncé. La modalité d'énonciation renvoie au sujet de l'énonciation en marquant l'attitude énonciative de celui-ci dans sa relation à son allocutaire. La modalité d'énoncé renvoie au sujet de l'énonciation en marquant son attitude vis-à-vis du contenu de l'énoncé. Elle exprime la manière dont le locuteur apprécie le contenu de l'énoncé. Nous présentons dans les sections suivantes les différents actes locutifs caractéristiques des discours scientifiques et vulgarisés

en français et en japonais, ainsi que les différentes caractéristiques du mode d'organisation du discours. La modalisation est une partie du phénomène d'énonciation. Elle permet d'« *expliciter ce que sont les positions du sujet parlant par rapport à son interlocuteur, à lui-même et à son propos* » (Charaudeau, 1992, p. 572).

$$Modalisation = \{(Loc. \rightarrow Interloc.), (Loc. \rightarrow Loc.), (Loc. \rightarrow Propos)\}$$

La modalisation est composée d'*actes locutifs*, position particulière du locuteur dans son acte de locution, qui sont spécifiés par des *modalités énonciatives*.

On compte parmi les actes locutifs :

L'acte allocutif, dans lequel « *le locuteur implique l'interlocuteur dans son acte d'énonciation et lui impose le contenu de son propos.* »;

L'acte élocutif, dans lequel « *le locuteur situe son propos par rapport à lui même, dans son acte d'énonciation. Il révèle sa propre position quant à ce qu'il dit.* »;

L'acte délocutif, dans lequel « *le locuteur laisse s'imposer le propos en tant que tel, comme s'il n'en était nullement responsable. Locuteur et interlocuteur sont absents de cet acte d'énonciation.* ».

La figure 3.4 reprend ces trois actes locutifs.

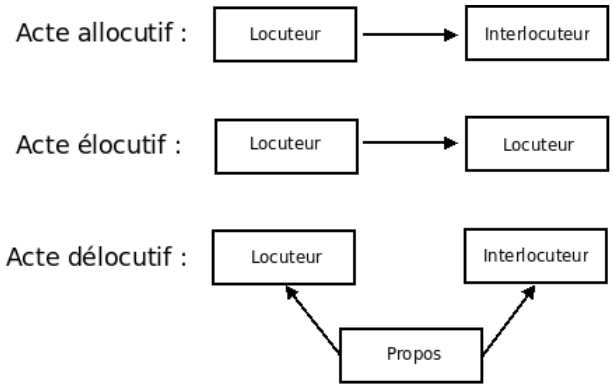


Figure 3.4 – Les trois principaux actes locutifs de Charaudeau

Toute modalité implique un acte locutif, et tout acte locutif est spécifié par un certain nombre de modalités.

Du point de vue linguistique, les modalités peuvent être caractérisées par :

1. des marques formelles explicites : verbes, adverbes, adjectifs, noms en construction personnelle ou impersonnelle, statuts de phrase (impératif, interrogatif, exclamatif) ;
2. une organisation particulière du discours : les marqueurs « implicites » du discours permettent de l'interpréter correctement. L'énoncé « Je reviendrai demain » peut par exemple être interprété comme une promesse ou une menace selon les marqueurs implicites (intonation, gestes, contexte…) (Charaudeau, 1992, p. 573).

Dans l'énoncé de l'acte allocutif, la présence de l'interlocuteur réside sous la forme de pronoms personnels (« tu », « vous »), du nom propre et commun qui l'identifient. Cet énoncé contient également de l'impératif et de l'interrogatif.

À l'inverse, dans l'acte élocutif le locuteur n'indique pas l'interlocuteur mais lui-même, c'est sa propre position qu'il révèle dans son énonciation. Sa présence est marquée sous la forme de pronoms personnels (« je », « nous », « on » en tant que nous), du nom propre et du nom commun qui identifient le locuteur. L'acte élocutif tient également compte de l'exclamatif et de l'optatif (expression d'un souhait).

Enfin, l'acte délocutif ne fait pas référence au locuteur ni à l'interlocuteur. Ainsi les énoncés de cet acte prennent les formes impersonnelles et référentielles.

Nous nous intéressons ici aux textes issus de domaines spécialisés, en particulier le domaine médical, dont nous pouvons estimer que la majorité utilise l'acte délocutif. En analysant dans de tels textes la modalité locutive, nous souhaitons observer si l'existence du locuteur ou de l'interlocuteur dans les textes différencient les types de discours scientifique et vulgarisé et si c'est le cas, détecter les marqueurs pertinents pour les distinguer. Ainsi, dans notre étude, nous ne traitons pas l'acte délocutif, puisqu'il ne concerne pas directement le locuteur et l'interlocuteur.

Charaudeau (1992) fait un inventaire très complet de toutes les modalités possibles. Nous avons dû les sélectionner selon plusieurs critères :

- Leur adaptabilité aux domaines de spécialité ;
- Leur adaptabilité aux langues française et japonaise ;
- La possibilité de trouver des marqueurs permettant de détecter automatiquement ces modalités dans les textes.

Pour cela, chaque modalité a été examinée, et pour chacune nous avons cherché des applications et des exemples dans les domaines de spécialité (par le biais de notre corpus d'étude), et cela pour chaque langue. Les exemples illustrant chaque modalité sont issus de notre corpus d'étude. Nous avons pour cela sélectionné aléatoirement dix documents par type de discours pour chaque langue, dans lesquels nous avons cherché des marqueurs de chaque modalité.

Les actes allocutifs

Modalité de l'interpellation :

Les caractéristiques de la modalité de l'interpellation sont :

- Le locuteur identifie une personne (son interlocuteur ou autre) parmi « un ensemble d'interlocuteurs possibles » ;
- Il attend que son interlocuteur réagisse à son interpellation.

Ex. : *Madame, Monsieur, vous êtes atteint d'un diabète de « type 2 ». Ce document est destiné à vous aider (…)*

みなさん、こんにちは! (Bonjour tout le monde !)

Modalité de l'injonction :

La caractéristique de la modalité de l'injonction est que le locuteur pose, dans son énoncé, une action à dire ou à faire. Le locuteur impose fortement à son interlocuteur de réaliser cette action.

Ex. : *Ne pas dépasser l'équivalent en alcool de deux verres de vin par jour.*

(…) 必ず医師の指示に従ってください。(Suivez le conseil de votre médecin)

Modalité de l'autorisation :

Le locuteur pose une action à réaliser. Il sait ou suppose que l'interlocuteur désire l'exécuter. Jugeant que les circonstances sont convenables, il lui donne le droit d'exécuter cette action.

Ex. : *Si le goût du sucre vous manque vraiment, vous pouvez avoir recours aux édulcorants.*

どの日から参加されても結構です ((Vous) pouvez (nous) assister quand vous voulez.)

Modalité de l'avertissement :

La modalité de l'avertissement se définit par les points suivants :

- Le locuteur énonce une action à réaliser par lui-même qui « peut être assortie d'une condition » (Charaudeau, 1992, p. 587) ;
- Il sait ou suppose que l'interlocuteur ignore son intention ;
- Par la déclaration de son intention, le locuteur protège le risque de dégradation de sa situation.

Ex. : *Attention, si vous êtes traité par sulfamides hypoglycémiants, évitez les boissons alcoolisées en dehors des repas*
塩分制限のある方は注意!! (Attention si (vous) êtes sous le contrôle de la salinité !!)

Modalité du jugement :

La modalité du jugement porte sur une action réalisée, dont le locuteur suppose que la responsabilité appartient à l'interlocuteur. Il juge si cet acte est bien/mauvais, en déclarant son approbation / réprobation. Sa qualification de l'interlocuteur y est indiquée.
Ex. : *Ce menu vous tente mais vous hésitez et vous faites bien [. . .]*
1日の総カロリー数の制限を守ることですね、それさえ守っていれば何を食べてもいいと思いますが、1日3食バランスよく食べることを忘れずに。 (Il faut respecter la limite de calories quotidienne. Si vous la respectez, (vous) pouvez manger tout ce que vous voulez, mais n'oubliez pas de manger trois repas équilibrés par jour.)

Modalité de la suggestion :

Le locuteur pose une action à réaliser ou à ne pas réaliser. La caractéristique de cette modalité, à savoir l'énonciation par rapport à l'action à réaliser, se partage avec celle de la théorie Irrealis mentionné ci-dessous. Tandis que la théorie locutive concentre cette modalité en tant qu'une catégorie dans l'acte allocutif. Irrealis s'intéresse à tous les énoncés qui parlent d'un événement qui n'est pas réalisé. Le locuteur sait, ou il suppose, que l'interlocuteur n'est pas content de sa situation, et lui propose d'exécuter l'action afin de l'améliorer. Les verbes devoir et falloir au conditionnel fonctionnent également comme la modalité de la suggestion.
Ex. : *Revenons donc aux malheureux engagés à leur insu sur la voie du diabète.*
適正なエネルギー量の食事をとりましょう (Prenons un repas qui contient un nombre de calories raisonnable)

La modalité de proposition :

Comme précédemment, en utilisant la modalité de la proposition, le locuteur pose une action à réaliser. La différence avec la modalité de suggestion est que la réalisation de cette action permet le profit de 1) l'interlocuteur ou de 2) l'interlocuteur + le locuteur. Le bénéficiaire est toujours l'interlocuteur, le résultat dépend de son acceptation.
Ex. : *Vous pouvez décaler l'horaire du petit déjeuner dans la matinée ou le fractionner.*
アンセイでは、豊富なメニューと管理栄養士による食事のプランや運動メニューなどをご提案いたしております。 (Chez Ansei, (nous) (vous) proposons les menus riches ainsi que des plannings de repas programmés par le diététicien ou de l'exercice.)

La modalité de l'interrogation :

Avec cette modalité, le locuteur pose une information à acquérir. Il demande à l'interlocuteur de dire ce qu'il sait ; il révèle ainsi son ignorance par rapport à ce qu'il demande. Il impose donc à l'interlocuteur un rôle de répondeur, et se donne le droit de questionner. La demande du locuteur peut deux significations : l'un est la demande des informations et l'autre est la vérification de la connaissance de l'interlocuteur par rapport au contenu concerné dans l'énoncé.
Ex. : *Comment adapter votre régime face à une situation particulière ?*
運動の主な効果は? (Quel est l'effet principal des exercices ?)

Modalité de la requête :

Avec cette modalité, le locuteur se trouve dans une situation défavorable et il demande à son interlocuteur de faire quelque chose pour lui.
Ex. : *Pouvez-vous nous parler des problèmes de vision associés au diabète ?*
ご都合でご参加できなくなった場合には必ず取消のご連絡をお願いします。(Si (vous) avez un empêchement, (nous) (vous) prions de (nous) contacter pour annuler votre présence.)

Les actes élocutifs

Modalité du constat :
Avec cette modalité, le locuteur décrit un fait sans donner son opinion ou jugement.
Ex. : *Or aujourd'hui, on observe dans de nombreux pays (…) une augmentation particulièrement importante du diabète de type 2 (…), lequel est reconnu pour accompagner la fréquence accrue de l'obésité infantile et de l'adulte.*
私どもはまた、分離大豆たん白質は、直接作用によるCCK分泌促進作用が強いことも明らかにした (Nous avons constaté également que les protéines dissociées du soja recevaient directement l'effet fort de la sécrétion de CCK.)

Modalité du savoir/l'ignorance :
Avec cette modalité, une information est présupposée et le locuteur affirme la connaître ou non.
Ex. : *On sait de longue date qu'une alimentation trop riche et l'obésité sont des facteurs de risque de diabète de type 2*
食事療法はテキメン。今まで理想的なカロリーなんて知らなかったから。 (Le régime alimentaire est efficace. Jusqu'à maintenant (je) ne connaissais pas la quantité de calories idéale.)

Modalité de l'opinion :
Le locuteur exprime son point de vue par rapport à un fait ou à une information qui réside dans sa croyance. Il évalue également la vérité de son propos.
Ex. : *Il ne nous semble pas justifié de proposer un dépistage systématique du diabète de type 2 dans la population générale*
(…) たいへん意義深いものであると思われる。(Il (nous) semble très significatif.)

Modalité de l'appréciation :
Dans les énoncés de la modalité de l'appréciation, le locuteur montre son évaluation d'un fait, en révélant ses propres sentiments. Cette évaluation ne vient pas du jugement d'après la raison mais elle vient d'un jugement d'après l'affect.
Ex. : *Depuis, mon diabète va bien mieux*
本特集が、この分野の研究の理解・発展の一助となれば私とすればこのうえない幸いである。 ((Je) serai heureux si ce numéro spécial aidait à la compréhension et au développement de la recherche du domaine.)

Modalité de l'obligation :
Les énoncés de la modalité de l'obligation concernent une action à faire par le locuteur lui-même. La réalisation de cette action dépend de lui mais le caractère de l'obligation est différent selon le lien entre l'action et le locuteur.
Ex. : *Nous devons mener une importante réflexion quant à la gestion de ces différents facteurs*
栄養教育を押しつけるのではなく,教育効果を上げるために患者の心を知り,心理学的なアプローチを行ってゆくための知識と技術を私達医療スタッフは修得し なくてはならない。 (Nous, l'équipe médicale, devons connaître les patients afin d'avoir une approche psychologique et éducative.)

Modalité de la possibilité :

Comme pour la modalité de l'obligation, la modalité de la possibilité concerne une action à faire dont la réalisation dépend de lui. Le locuteur mentionne son aptitude ou sa disposition d'esprit afin d'achever cette action.

Ex. : *On peut parfaitement avoir du diabète depuis de nombreuses années sans le savoir.*
アルコールを中止するなんて考えることもできません。 ((Je) ne peux pas arrêter l'alcool.)

Modalité du vouloir :

La modalité du vouloir concerne une action à faire dont la réalisation ne dépend pas toujours de lui.

Ex. : *Cette aide informatique gagnerait à être incluse dans les fiches médicales informatisées utilisées par les médecins*
(...) わが国における本剤導入の一助としたい ((Nous) voulons encourager l'importation de ce médicament dans notre pays)

Modalité de la déclaration :

Avec cette modalité, le locuteur montre l'existence et la vérité d'un savoir à son interlocuteur.

Ex. : *on reconnaît de plus en plus le tissu adipeux comme un organe endocrinien*
「実は飲んでいません」 (En fait, (je) n'ai pas pris (le médicament).)

Les modes d'organisation du discours Les modes d'organisation du discours constituent « *les principes d'organisation de la matière linguistique* ». Ces principes dépendent de la finalité que se donne le sujet parlant dans la communication. Ils consistent à « *utiliser certaines catégories de langue pour les ordonner en fonction des finalités discursives de l'acte de communication* ». On compte quatre modes d'organisation : l'énonciatif, le descriptif, le narratif et l'argumentatif. Les modes d'organisation du discours sont caractérisés par un ensemble de procédés discursifs. Nous nous intéressons en particulier aux suivants :

- Citations : présence d'extraits cités (placés entre guillemets). Forme de discours rapporté, permettant de faire référence à un texte, mise en emphase d'un mot ou d'une expression, etc. ;
- Type de phrase : phrases affirmatives, interrogatives, exclamatives ;
- Connecteurs pragmatiques : « *mots qui ne sont pas destinés à apporter des informations, mais à marquer le rapport du locuteur à la situation* » Ducrot (1980) (mais, donc, alors que, néanmoins...). Organisation de la logique argumentative dans le discours ;
- Fins de phrases : marqueurs de politesse en fin de phrase.

3.5.2.2 La modalité Irrealis

Nous avons vu que la modalité de Charaudeau, telle que nous l'utilisons dans notre typologie, se base en grande partie sur la présence du locuteur ou de l'interlocuteur. En langue française, cela s'exprime principalement grâce aux pronoms.

Présentation La notion de modalité Irrealis (*irrealis modality*) se base sur la dimension binaire : Realis et Irrealis. D'une manière générale, ce concept remonte, notamment pour les langues européennes, à la distinction entre l'indicatif et le subjonctif. L'indicatif touche à la notion de Realis, le subjonctif, à celle d'Irrealis. Comme on le verra dans cette section, ce schéma prototypique, qui ne reflète pas exactement la variété énonciative, encadre tout de même la stratégie basique de cette modalité.

La modalité Irrealis examine la nature de la réalité ou de la vérité d'un état ou d'un événement (Givón, 1994, p. 321). La distinction entre Realis/Irrealis est établie en recherchant si la situation exprimée par le locuteur appartient à l'intérieur ou à l'extérieur de la réalité. Selon Palmer (2001, p. 4), la majorité des propositions en sens phrastique se groupe en Realis ou en Irrealis. La théorie de la modalité Irrealis considère les indices marquant l'irréalité d'un événement dans l'énoncé.

De ce point de vue, elle conduit à s'interroger sur les attitudes épistémiques et évaluatives. Par l'attitude épistémique, on examine le degré de certitude indiqué par la croyance ou la probabilité par rapport à un événement. Par l'attitude évaluative, on examine la force d'opération indiquée par le désir ou la préférence du locuteur par rapport à un événement (Givón, 1994, p. 266, 328).

Les marqueurs d'Irrealis se trouvent dans plusieurs catégories grammaticales. Par exemple, Givón (1994) présente les catégories suivantes pour l'anglais :

- Temps et aspect ;
- Adverbes modaux ;
- Verbes complétifs ;
- Actes de parole non-déclaratifs (suggestion, demande, avertissement, impératif) ;
- Phrases adverbiales ;
- Auxiliaires modaux.

Cette typologie permet d'examiner les différents degrés de la certitude du locuteur par rapport à la réalité de son propos. Comme on le verra dans le reste de cette section, nous avons appliqué cette classification à notre étude franco-japonaise.

Avec la modalité Irrealis, l'objectif du locuteur envers son énoncé est moins important pour interpréter la modalité, mais son attitude envers la réalité de la proposition se situe au centre de l'interprétation. Elle nous paraît être une théorie intéressante pour notre étude, et ce pour deux raisons : premièrement au niveau de la comparaison avec la modalité locutive et deuxièmement au niveau de l'application au français et au japonais.

Critères Irrealis Givón a créé sa théorie de la modalité pour la langue anglaise. Une adaptation de ses différents critères a dû être effectuée, de façon à ne conserver, comme pour les critères de la modalité de Charaudeau, que ceux étant adaptables aux deux langues, aux domaines de spécialité, et caractérisables par des marqueurs simples. Pour chaque critère, nous présentons un exemple tiré de (Givón, 1994), puis un exemple tiré du corpus.

Temps futur :

Un événement qui aura lieu après le moment de l'énonciation est irréel et potentiel (Chung et Timberlake, 1985, p. 206). En français, le temps futur se forme par le futur simple, le futur proche et le futur antérieur.

Ex. : *le conseil nutritionnel pourra aller vers des aliments plutôt riches en potentiel anti-oxydants*

L'approche de la réalisation de l'événement se marque par les locutions telles que être sur le point de, être en passe de (Riegel et al., 1994, p. 253). Les verbes au présent peuvent également exprimer une activité à réaliser dans l'avenir immédiat ou plus lointain (Ex. :Je viens tout de suite.)

En japonais, le passé est également employé pour parler du futur. Le locuteur souhaite vérifier si quelque chose se réalisera dans l'avenir et il sait que cette réalisation a été décidée dans le passé ou qu'on le sait depuis longtemps. Dans l'exemple suivant, le contenu de la question est la réunion du lendemain. On utilise ici le verbe être au passé afin d'accentuer le fait que le locuteur veut vérifier la connaissance qu'il possède mais dont il n'est pas certain :

Ex. : 会議は明日も2時からでしたか。(La réunion était (=sera) à partir de 2 h demain aussi ?)

Adverbes modaux :

L'adverbe modal est un adverbe qui modifie le degré de certitude ou de potentialité de la réalisation d'un événement. Les adverbes tels que *probablement, peut-être, sans doute* ou *certainement* expriment le jugement de certitude du locuteur par rapport à la proposition de l'énoncé.
Ex. : *La destruction auto-immune des cellules produisant l'insuline est probablement un processus lent[. . .]*
おそらく多大な苦痛や心理的ストレスが生じると思われます。 (Il est probable qu'une grande douleur ou un stress psychique soient provoqués.)

Irrealis dans les compléments de verbe :

Quand un verbe s'accompagne d'un complément, ce dernier a une nature Irrealis dans les cas suivants (la partie complétive que l'on traite ici est un verbe ou la subordonnée).

Le complément du verbe de demande :

Il s'agit des verbes qui comportent la nature de demande du locuteur à l'interlocuteur de faire quelque chose afin de réaliser une action indiquée dans le verbe complétif.
Ex. : *Demandez conseil à votre « guide » (diététicien ou médecin) pour le choix des graisses.*
糖尿病とともに歩む人生について教えてください。 (Renseignez-(moi) sur la vie diabétique.)

Le complément de verbes de perception :

Ce groupe de verbes exprime notamment le jugement du locuteur par rapport au contenu exprimé dans le complément.
Ex. : *Les enfants prédiabétiques présentent en effet une croissance plus rapide, que l'on croyait liée à des facteurs génétiques prédisposants[. . .]*
例えば60歳の人では、2.64 %が平均、2.15 %以下は糖尿病を疑う。 (Par exemple, pour les gens de plus de soixante ans, la moyenne (du coefficient fluctuant de l'intervalle du pouls) est 2.64 %, et (si quelqu'un a) moins de 2.15 %, (on) doute qu'il est diabétique.)

Ordre :

L'ordre est aussi classé dans la catégorie Irrealis. Quand le locuteur ordonne ou demande quelque chose à son interlocuteur, l'activité n'est pas encore réalisée.

Impératif de la première personne plurielle :

Givón distingue l'impératif en trois groupes selon le destinataire de l'énoncé. L'impératif à la première personne du pluriel est différent de l'impératif à la première personne du singulier, parce que le locuteur montre sa volonté de participer à l'événement qu'il énonce :
Ex. : *Allons manger cette galette.*
塩分を制限されている方は注意して! (Faites attention si vous devez contrôler le sel !)

Ordre pour la troisième personne :

Quand le locuteur ordonne quelque chose à son interlocuteur et que sa réalisation dépend d'une troisième personne, le contenu de l'ordre est destiné à cette troisième personne.
Ex. : *Revenons donc aux malheureux engagés à leur insu sur la voie du diabète.*
食物せんいを多く含んだ食事を食べさせてあげてください。 (Faites manger (aux patients) un repas qui contient beaucoup de cellulose végétale.)

Question 'oui/non' :

Dans la question dont la réponse se forme par oui ou non, quand le locuteur montre sa connaissance par rapport au contenu de son propre énoncé, cet énoncé contient la nature irréelle. Ce genre

d'énoncé implique une certitude moindre de la connaissance du locuteur.
Ex. : *Obésité et diabète, même combat ?*
でも、カロリー制限をした食事療法は必要なんですよね? (Un régime alimentaire contrôlant les calories est nécessaire, n'est-ce pas ?)

Proposition temporelle d'Irrealis :

Comme l'énoncé anglais when you get a loan, I'll sell you my car, cette catégorie propose une nature Irrealis dans l'énoncé où l'événement sera réalisé quand la condition indiquée dans la subordonnée est satisfaite.

Conditionnel simple :

Le conditionnel simple se forme, en français, par si + présent / présent.
Ex. : *Si cet indice est supérieur ou égal à 30, l'individu est considéré comme obèse.* En japonais, le conditionnel simple se forme par le verbe passé + « ra » dans la subordonnée / présent.
進行したら薬物療法が必要になるよ。 (Si (la maladie) fait des progrès, il sera nécessaire de suivre un traitement médicamenteux.)

Conditionnel (hypothèse probable) :

La probabilité de l'hypothèse se forme par la structure, en français, si + présent / futur.
Ex. : *Cela ne vous prendra pas plus de temps si vous marchez d'un pas un peu plus rapide.*
En japonais, l'adverbe « moshi » (possible à omettre) + verbe + « nara » dans la subordonnée précède la proposition principale qui se forme par le verbe présent ou futur :
Ex. : 自覚症状がなくても、慢性合併症を防ぐために、糖尿病といわれたら、すぐに治療を開始しなければなりません。 (Même si (vous) n'avez pas de symptômes subjectifs, si (vous) êtes diagnostiqué comme diabétique, il faut immédiatement commencer le traitement, afin d'éviter toute complication chronologique.)

Conditionnel d'irréel :

Cette catégorie exprime un état contraire à la réalité. En français, elle se forme par les structures différentes selon le temps :

- si + imparfait / conditionnel présent. Exemple : *Si j'avais connu plus tôt l'existence de cette association, j'aurais certainement bien mieux vécu.*
- si + passé composé / passé composé ou présent. Exemple : *Contrôlez votre glycémie au doigt avant et après votre activité physique de façon à juger si l'adaptation de votre alimentation a été suffisante, insuffisante ou exagérée.*
- si + plus-que-parfait / conditionnel passé ou présent. Exemple : *Si j'avais connu l'association à cette époque et si j'avais été immédiatement pris en charge comme je l'ai été en 1996, j'aurais appris que cette maladie traîtresse vous conduit inexorablement à de nombreuses complications.*

En japonais, l'hypothèse ou l'état contraire à la réalité se forment par la combinaison de certaines terminaisons du verbe et de certaines conjonctions dans la subordonnée. Les conjonctions utilisées pour exprimer une situation irréelle sont « tara, dara, reba, toshitemo, mo, to, nara » qui correspondent généralement aux terminaisons convenables :
Ex. : 生・老・病・死は人間を取り囲む環境がどのように変化しても,避けることのできない永遠のテーマであろう. (La vie, la vieillesse, la maladie et la mort sont les thèmes éternels et inévitables, quelque soit l'environnement qui entoure l'être humain.)

Auxiliaires modaux :

Les auxiliaires modaux en français tels que vouloir, pouvoir, devoir introduisent évidemment la

nature irréelle. En les rajoutant, Givón présente la dimension sémantique, c'est-à-dire les expressions qui fonctionnent comme les auxiliaires modaux tels que les verbes souhaiter, aimer.
Ex. : *Le patient diabétique doit là encore veiller à respecter les recommandations et modérer ses apports en protéine.*
En japonais, les expressions modales telles que « られます » ou « られる » (pouvoir), « ことがある » ou « ことができます » (pouvoir), « べきです » ou « べきだ » (être obligé de), « たいです » ou « たい » (vouloir), « たいものです » ou « たいものである » ou « たいものだ » (souhaiter) etc. Elles sont attachées à l'infinitif du verbe qui les précédent :
Ex. : 糖尿病とはいえ、栄養の約半分は炭水化物から摂取すべきである。 (Même si (on) est diabétique, il faut que les glucides représentent au moins la moitié de l'alimentation.)

3.5.3 Caractéristiques lexicales

Nous avons vu dans la section 3.4.3.3 que le degré de spécialisation d'un document se caractérisait aussi par le lexique utilisé dans les documents. Nous présentons dans cette section les critères nous paraissant discriminants dans la distinction des types de discours scientifique et vulgarisé.

Vocabulaire spécialisé : le vocabulaire scientifique et technique est l'une des caractéristiques des langues de spécialité (Mortureux, 1995) (ex: triglycérides, lipotoxicité…). Dans les documents scientifiques il est extrêmement fréquent. Les documents vulgarisés emploient aussi des termes spécialisés, mais ceux-ci sont généralement uniquement introduits puis substitués par des synonymes ou des termes de la langue générale.

Noms propres : généralement des noms de personnes ou de lieux sont utilisés. Les auteurs des documents et de nombreuses citations bibliographiques sont indiqués par le nom dans les articles scientifiques.

Caractères numériques : très souvent des quantités numériques dans un corpus médical, des dates dans le cas de citations bibliographiques par exemple.

Unités de mesure : principales unités de mesures, fréquemment utilisées dans les domaines scientifiques (grammes, mètres, calories…).

Bibliographie : présence d'une bibliographie à la fin du document, souvent sous la forme d'une liste de noms suivie d'un titre, d'une date, d'un nom de journal ou de conférence… Les bibliographies sont le plus souvent présentes dans les documents scientifiques.

Citations bibliographiques : références à des éléments de la bibliographie dans le texte. Celles-ci sont sous différentes formes : (*nom*, *date*), [*identifiant*]… Ces citations sont généralement présentes dans les documents scientifiques.

Ponctuation : utilisation des principaux symboles de ponctuation.

Marqueurs de glose : la glose correspond à l'explication d'une idée par une autre dans un discours Steuckardt et Niklas-Salminen (2005). Ceci est souvent indiqué grâce à un ensemble de marqueurs, parmi lesquels on trouve *à savoir, autrement dit, c'est-à-dire, par exemple,* etc.

Parenthèses : les parenthèses peuvent être utilisées dans différents buts. Elles peuvent servir à détailler une idée, donner un exemple, expliciter un acronyme…

Autres alphabets (latin, hiragana, katakana) : différents alphabets utilisés dans des textes en langue japonaise. L'alphabet katakana permet d'écrire en japonais les mots d'origine étrangère, les onomatopées, ils permettent aussi de mettre des mots en évidence dans un texte. L'alphabet hiragana est

utilisé afin d'écrire les mots japonais auxquels aucun kanji ne correspond, les mots pour lesquels l'auteur ne connaît pas l'écriture en kanji ou dans le cas où l'écriture en kanji est trop formelle. Les kanjis sont hérités des caractères chinois et permettent avec les deux autres alphabets d'écrire l'ensemble des mots en japonais. Les caractères latins (romajis) sont utilisés assez rarement, ils permettent d'écrire certains noms étrangers et sont utilisés pour écrire les formules mathématiques.

Symboles : caractères ne faisant pas partie des alphabets ni des principaux symboles de ponctuation (émoticones, •, ○). Ces derniers sont rarement utilisés sur des articles scientifiques, des rapports de recherche ou des documents institutionnels. Ils sont parfois introduits par des logiciels facilitant la création de pages Web. Quelques symboles se trouvent dans les formules mathématiques des articles scientifiques.

3.6 Conclusion

Nous avons présenté dans ce chapitre une analyse stylistique de notre corpus d'étude. Notre démarche est déductive et contrastive : en partant de documents répartis dans deux classes (scientifique et vulgarisée), ils sont analysés par couples (un de chaque classe) afin de déterminer quels éléments caractérisant chaque classe peuvent être discriminants. Ces caractéristiques relèvent de trois niveaux : les caractéristiques structurelles, correspondant aux éléments graphiques et structurels des documents ; les caractéristiques modales, correspondant aux éléments relatifs à la modalité dans les documents et les caractéristiques lexicales. Nous avons choisi d'utiliser deux théories de la modalité : la théorie Locutive de Charaudeau et la théorie Irrealis de Givon. Elles sont comparées dans le chapitre 5. Cette typologie est donc composée de caractéristiques du type de discours scientifique ou vulgarisé. Elle est de plus multilingue et motivée linguistiquement : les niveaux d'analyse et la sélection rigoureuse des critères permettent réellement de caractériser un phénomène (plutôt que de lister des critères potentiellement discriminants).

Nous présentons dans le chapitre suivant la mise en œuvre de cette typologie, c'est-à-dire l'implémentation des différents critères. Cette mise en œuvre nous permet d'apprendre un modèle de classification, mais aussi de tester la robustesse et la généricité de notre typologie.

CHAPITRE 4

Classification automatique des documents français et japonais selon leur type de discours

4.1 Introduction

« *La catégorisation des textes fait partie des activités cognitives spontanées des sujets. [...] Sans l'existence de catégories, notre appréhension des énoncés produits serait probablement impossible* » (Adam, 1992, p. 6). Toute collection de données textuelles doit, pour être exploitée par l'homme, être ordonnée (bibliothèques, bases de données...). La quantité sans cesse croissante de données textuelles électroniques, notamment sur le Web, a engendré un gros besoin de classification et soulevé le problème de la classification automatique. Les données étant très nombreuses (et à chaque instant plus nombreuses), la classification manuelle est devenue impossible. Il a alors été nécessaire de créer des systèmes capables de produire une classification au sein de ces données permettant à l'homme d'y accéder. Le domaine de la classification automatique est né de ces difficultés. Nous cherchons dans ce chapitre à adapter les méthodes de ce domaine à la classification de documents du Web sur des domaines spécialisés en type de discours scientifique ou vulgarisé.

La figure 4.1 présente les différentes étapes nécessaires à la mise en œuvre de la typologie et son application aux méthodes de classification automatique. Les deux premières étapes sont manuelles. La première étape consiste à analyser le corpus d'apprentissage afin de créer une typologie adaptée à la classification souhaitée. Il s'agit, dans notre cas, d'analyser les documents du corpus répartis en deux classes : les documents scientifiques et les documents vulgarisés afin d'en dégager des caractéristiques propres à ces deux classes. Ces caractéristiques forment une typologie des discours scientifiques et vulgarisés dans les domaines de spécialité en français et japonais (voir chapitre 4).

Les systèmes de classification automatique se basent généralement sur une représentation vectorielle des documents, chaque élément des vecteurs correspondant à un critère de la typologie. La reconnaissance de chacun des critères doit donc être implémentée, afin de pouvoir construire ces vecteurs. Sebastiani (2005) nomme cette étape l'indexation des documents, elle est en partie manuelle (implémentation de chacun des vecteurs) et automatique (application d'un programme de création du vecteur correspondant à chaque document). Une fois les représentations vectorielles des documents réalisées, les méthodes

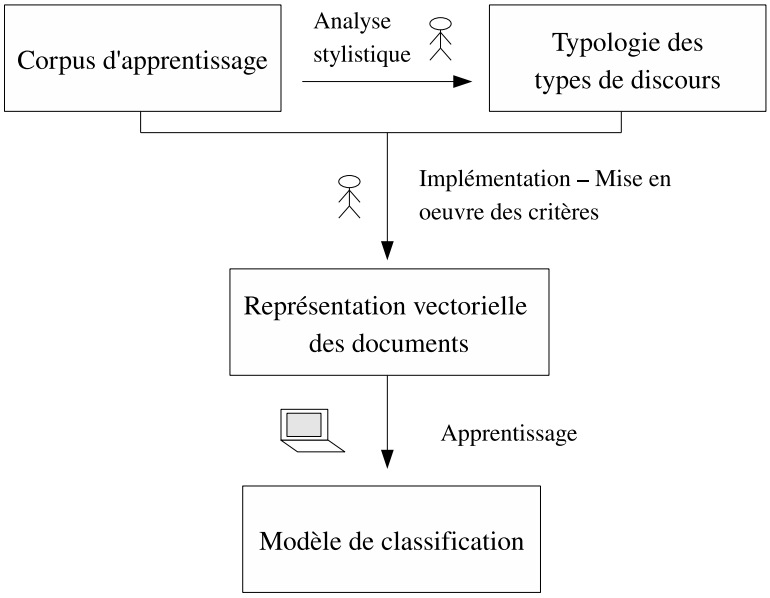


Figure 4.1 – Différentes étapes de la mise en œuvre de la typologie

d'apprentissage automatique peuvent être appliquées. Nous présentons dans la section 4.2 les différentes étapes de l'élaboration d'un modèle de classification : indexation, apprentissage, classification, évaluation. Nous présentons dans la section 4.3 l'application de chacune de ces étapes à notre problème et les algorithmes de classification choisis.

4.2 Méthodes de classification automatique de textes

Nous présentons dans cette section un bref état de l'art des méthodes de classification automatique textuelle. Nous verrons dans un premier temps les principales étapes de l'élaboration d'un classifieur, puis nous détaillons chacune de ces étapes.

4.2.1 Présentation

Sebastiani (2002, 2005) propose des descriptions et états de l'art complets sur le classification automatique textuelle. Nous nous appuyons ici sur son travail et ses notations.

Définissons dans un premier temps une fonction de classification :

Definition 1. *Une fonction de classification est une fonction* $\Phi : \mathcal{D} \times \mathcal{C} \rightarrow \{V, F\}$, *avec* $\mathcal{C} = \{c_1, \ldots c_{|C|}\}$ *un ensemble de classes et* $\mathcal{D} = \{d_1, \ldots d_{|D|}\}$ *un ensemble de documents, telle que :*

$$\Phi(d_i, c_j) = \begin{cases} V \text{ si } d_i \in c_j \\ F \text{sinon} \end{cases}$$

Dans le cas où $\Phi(d_i, c_j) = V$, d_i est appelé un *exemple positif* de la classe c_j, sinon il est appelé *exemple négatif*.

Nous nous plaçons ici dans le cadre supervisé, c'est-à-dire que les classes sont prédéfinies. La classification automatique de textes peut donc être définie comme suit (Sebastiani, 2005, p. 111) :

Definition 2 (Classification automatique)**.** *La classification automatique de textes est une tâche dont le but est d'approximer une fonction de classification cible* $\Phi : \mathcal{D} \times \mathcal{C} \rightarrow \{V, F\}$ *déterminée par un expert, en utilisant la fonction* $\hat{\Phi} : \mathcal{D} \times \mathcal{C} \rightarrow \{V, F\}$ *appelée* classifieur.

Sebastiani distingue les classifieurs mono-catégorie (*single-label*) des classifieurs multi-catégories (*multi-label*) :

un classifieur mono-catégorie est un classifieur tel que $\forall d_i \in \mathcal{D}, \exists! c_j \in \mathcal{C} : \Phi(d_i, c_j) = V$;

un classifieur binaire (bi-catégories) est un cas particulier des classifieurs à label unique : $\mathcal{C} = \{c\}$: $\forall d_i \in \mathcal{D}, \Phi(d_i, c) = V$ ou $\Phi(d_i, \bar{c}) = V$;

un classifieur multi-catégories est un classifieur tel que $\forall d_i \in D, \exists \mathcal{C}_{d_i} = \{c_j : \Phi(d_i, c_j) = V\}$.

Généralement, les classes sont symboliques, aucune information sur leur signification n'est disponible. La classification est alors basée sur les informations extraites des documents. De plus, la tâche de classification est subjective : deux experts, qu'ils soient humains ou artificiels ayant à statuer sur l'appartenance d'un document à une classe peuvent ne pas être du même avis. Les outils et recherches en apprentissage automatique ne prétendent pas fournir de compromis à ces ambiguïtés, mais plutôt de reproduire la classification d'un expert parmi d'autres (Sebastiani, 2005).

L'élaboration d'un système de classification automatique comporte trois grandes étapes (*ibid.*) :

1. **L'indexation des documents** (*document indexing*) : représentation vectorielle des documents utilisés pour apprendre le modèle de classification ;
2. **L'apprentissage du classifieur** (*classifier learning*) : apprentissage du modèle sur un corpus d'apprentissage composé de documents pré-classés ;
3. **L'évaluation du classifieur** (*classifier evaluation*) : évaluation du modèle sur un corpus de test.

L'ensemble de ces étapes est repris sur la figure 4.2. Nous présentons dans la suite de cette section chacune de ces étapes en détails.

4.2.2 Indexation des documents

Les textes ne peuvent pas être interprétés directement par un classifieur, il est donc nécessaire de lui fournir une représentation adaptée des données. L'indexation des documents consiste à convertir l'ensemble des documents d'un corpus en une représentation compacte de leur contenu pouvant être directement interprétée par un algorithme d'apprentissage automatique et par un classifieur une fois construit. Les méthodes utilisées pour extraire ces traits des documents sont héritées de la Recherche d'Information. Les documents des corpus sont généralement représentés sous la forme de vecteurs :

$$\vec{d_i} = \langle w_{1i}, \ldots, w_{|\mathcal{T}|i} \rangle$$

où $\mathcal{T}$ est l'ensemble des critères caractérisant les documents, et w_{ji} quantifie la valeur du trait t_j dans le document d_i. Une méthode d'indexation est caractérisée par :

- la définition des critères ;
- la méthode pour implémenter ces critères.

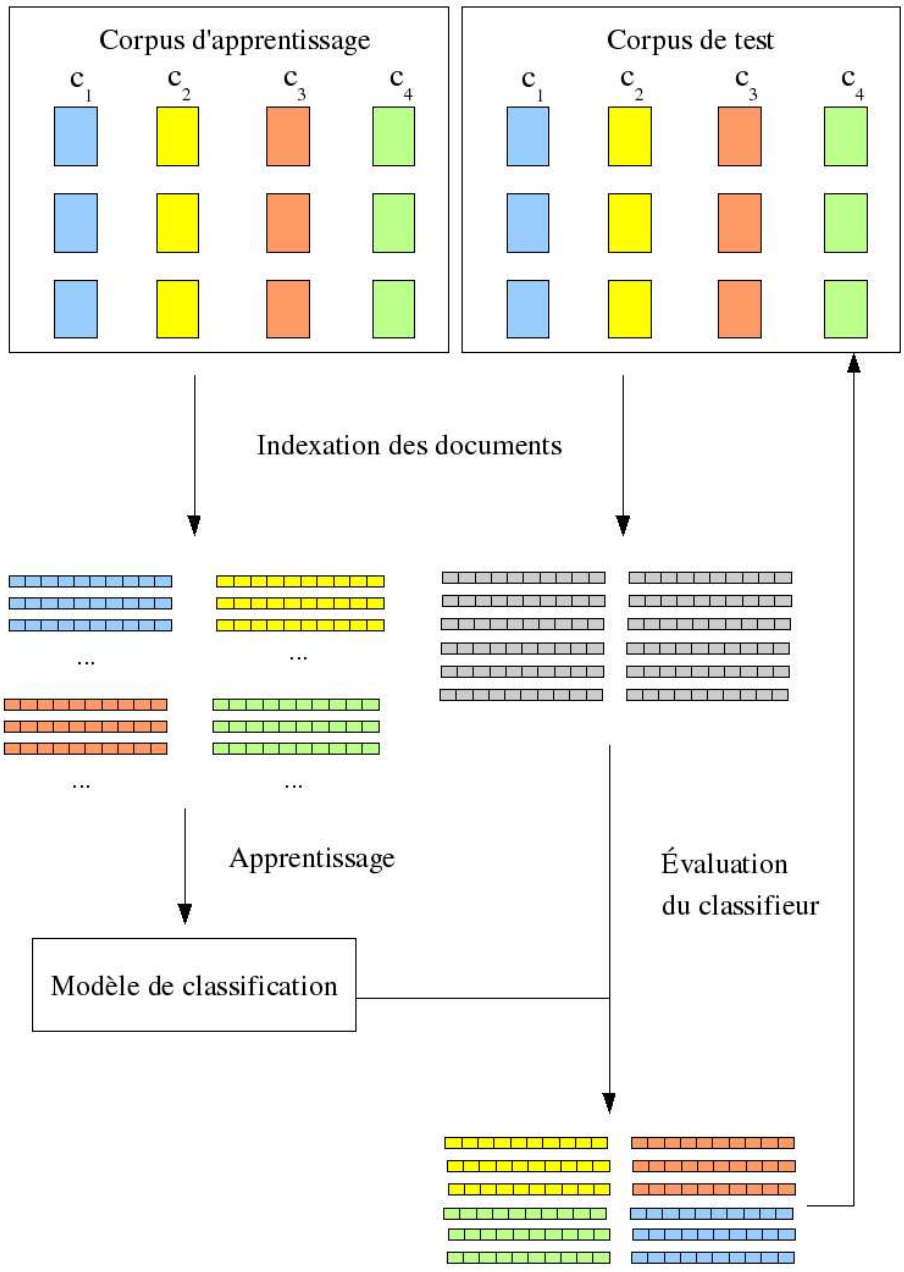


Figure 4.2 – Étapes de l'élaboration d'un classifieur

Le choix de la représentation des données, et donc des critères correspondant, dépend des unités porteuses de sens des textes (*lexical semantics problem*) et des règles propres à la langue de combinaison de ces unités (*compositional semantics problem*) (Sebastiani, 2002, p. 2). De nombreux travaux en apprentissage automatique négligent ces aspects et se basent sur une représentation en « sac de mots » des textes.

C'est pourquoi la représentation vectorielle la plus fréquente se base sur des vecteurs d'occurrence des termes dans les textes (Sebastiani, 2005). Les critères utilisés peuvent être les mots simples (hors mots outils) des textes, les lemmes, mais aussi des unités plus longues telles que des n-grammes, des phrases ou des unités plus complexes extraites grâce à des méthodes statistiques ou au *shallow parsing*.

Quelques études comparatives ont été menées sur ces différentes approches (le corpus comme un sac de mots ou le corpus comme un ensemble de textes) avec des critères plus complexes. Lewis (1992) montre que dans certains cas, l'utilisation de représentations plus complexes que les vecteurs de termes n'était pas forcément plus efficace. Selon elle, bien que les représentations basées sur les unités complexes telles que les phrases soient plus riches sémantiquement, elles sont de moindre qualité du point de vue statistique. Les phrases peuvent contenir plusieurs flexions du même mot ou des synonymes, qui ne seront pas liés et qui font baisser la probabilité pour une phrase d'être fréquente. Selon Sebastiani (2002, p. 11), une combinaison des deux approches pourrait être la meilleure solution, à l'instar de Tzeras et Hartmann (1993) qui améliorent leurs résultats en utilisant les groupes nominaux, obtenus grâce à des critères statistiques et syntaxiques.

Le poids calculé de ces critères peut être binaire ($w_{ij} \in \{0, 1\}$) – il signale alors la présence ou l'absence du critère – ou réel ($w_{ij} \in \mathbb{R}$) – il correspond alors à la fréquence d'apparition du critère par exemple. Les poids réels sont implémentés à l'aide de méthodes statistiques ou probabilistes. Les calculs basés sur la fréquence d'apparition d'un critère dans un document dépendent fortement de la longueur des documents. Il est donc souvent nécessaire de normaliser les poids. Deux hypothèses sont très utiles lors du calcul du poids des critères et sont issues de la méthode $tf \times idf$, très utilisée en classification automatique textuelle (Sebastiani, 2005) :

- plus un critère est fréquent dans un document, plus il est important pour décrire ce document (*term frequence*) ;
- plus un critère apparaît dans les documents d'un corpus, moins il est discriminant : sa contribution à la caractérisation d'un document ou d'une classe sera moindre (*inverse document frequency*).

Un nombre trop important de critères peut poser des problèmes, notamment pour certains classifieurs devenant plus lents ou moins efficaces avec trop de critères. Il est parfois nécessaire de réduire le nombre de ces critères, en sélectionnant les plus pertinents.

4.2.3 Apprentissage du modèle de classification

Un classifieur est généré automatiquement par un système inductif (le système d'apprentissage). L'induction peut se définir comme un « *type de raisonnement consistant à remonter, par une suite d'opérations cognitives, de données particulières (faits, expériences, énoncés) à des propositions plus générales, de cas particuliers à la loi qui les régit, des effets à la cause, des conséquences au principe, de l'expérience à la théorie* » (TLFi). Ainsi, à partir d'un ensemble de documents pré-classés, le système inductif apprend des règles basées sur les critères des documents permettant de les classer. Le but étant de pouvoir reproduire le classement initial, et ainsi pouvoir classer tout nouveau document.

Pour construire un classifieur pour un ensemble de classes $\mathcal{C}$, il est nécessaire de disposer d'un corpus $\mathcal{D}$ composé de documents tel que $\Phi(d_i, c_j)$ est connue $\forall (d_i, c_j) \in \mathcal{C} \times \mathcal{D}$.

Généralement, les corpus $\mathcal{D}$ sont divisés en deux parties : le corpus d'apprentissage, utilisé pour générer le classifieur, et le corpus de test, utilisé pour évaluer le classifieur. Les travaux en classification textuelle expérimentale ajoutent parfois une troisième partie : le corpus de validation, utilisé pour « régler » le classifieur une fois celui-ci généré (Sebastiani, 2005).

Deux types de systèmes d'apprentissage peuvent être distingués :

- les méthodes générant des classifieurs à valeurs binaires : $\hat{\Phi} : \mathcal{D} \times \mathcal{C} \rightarrow \{V, F\}$;

- les méthodes générant des classifieurs à valeurs réelles, CVS (*Categorization Statue Value*) : $CVS : \mathcal{D} \times \mathcal{C} \rightarrow [0, 1]$.

Les méthodes générant des classifieurs à valeurs réelles peuvent être utilisées lorsque la qualité de la classification est importante, les classifieurs fournissent alors un degré d'appartenance à une classe. Sinon, il est nécessaire de déterminer un seuil τ_i permettant de passer d'une valeur réelle représentant le degré d'appartenance d'un document à une classe c_i à une valeur binaire.

4.2.4 Évaluation du système de classification

Sebastiani (2005) cite trois mesures de qualité d'un classifieur :

- les performances d'apprentissage (*training efficiency*) : temps moyen de création d'un classifieur ;
- les performances de classification (*classification efficiency*) : temps moyen de classification d'un document ;
- l'efficacité (*effectiveness*) : pourcentage de résultats corrects.

En classification textuelle, l'efficacité est la mesure la plus utilisée. Pour les classifieur à label unique, l'efficacité est mesurée grâce à l'exactitude (*accuracy*), pourcentage de classifications correctes. Dans le cas de la classification binaire, la mesure de l'exactitude n'est pas adaptée. En effet, il arrive souvent qu'une classe soit plus importante qu'une autre. Un classifieur mettant tous les documents dans cette classe obtiendrait alors un bon résultat. Un couple de mesures est utilisé dans ce cas : la précision (π_i) et le rappel (ρ_i) :

$$\pi_i = P(\hat{\Phi}(d_i, c_j) = V | \Phi(d_i, c_j) = V)$$

$$\rho_i = P(\Phi(d_i, c_j) = V | \hat{\Phi}(d_i, c_j) = V)$$

La précision représente donc la probabilité que la classification dans c_j d'un document pris au hasard soit correcte, tandis que le rappel représente la probabilité qu'un document appartenant à la classe c_j y soit réellement classé. La précision π et le rappel ρ d'un classifieur peuvent donc être calculés en faisant la moyenne des π_i et ρ_i sur l'ensemble des catégories.

Ces probabilités peuvent être estimées grâce à une *table de contingence* pour c_i sur un corpus de test donné (Sebastiani, 2002, p. 33). Cette table de contingence est présentée dans le tableau 4.1. FP_i correspond au nombre de documents classés par erreur dans c_i, FN_i correspond au nombre de documents appartenant à c_i mais n'y étant pas classés, etc. Les estimations de la précision et du rappel pour c_i sont alors :

$$\hat{\pi}_i = \frac{TP_i}{TP_i + FP_i}, \hat{\rho}_i = \frac{TP_i}{TP_i + FN_i}$$

Catégorie c_i		Jugement de l'expert	
		OUI	**NON**
Jugement du classifieur	**OUI**	TP_i	FP_i
	NON	FN_i	TN_i

Table 4.1 – Table de contingence pour une classe c_i (Sebastiani, 2002)

Pour calculer la précision et le rappel globaux, Sebastiani (2002) propose deux méthodes. La première, appelée micro-moyenne (*microaveraging*), consiste à faire la somme des résultats obtenus dans le tableau 4.1, ce qui donne le tableau 4.2.

Ensemble de catégories $\mathcal{C} = \{c_1, \ldots, c_{\lvert\mathcal{C}\rvert}\}$		Jugement de l'expert	
		OUI	**NON**
Jugement du classifieur	**OUI**	$TP = \sum_{i=1}^{\lvert\mathcal{C}\rvert} TP_i$	$FP = \sum_{i=1}^{\lvert\mathcal{C}\rvert} FP_i$
	NON	$FN = \sum_{i=1}^{\lvert\mathcal{C}\rvert} FN_i$	$TN = \sum_{i=1}^{\lvert\mathcal{C}\rvert} TN_i$

Table 4.2 – Table de contingence globale (Sebastiani, 2002)

à partir de cette table, la précision et le rappel s'obtiennent de la même façon que précédemment, c'est-à-dire :

$$\hat{\pi}^{\mu} = \frac{TP}{TP + FP}, \hat{\rho}^{\mu} = \frac{TP}{TP + FN}$$

μ est utilisé pour indiquer qu'il s'agit de microaveraging.

La seconde méthode, dite de macro-moyenne (*macroaveraging*), consiste à évaluer la précision et le rappel pour chaque classe, et d'en faire une moyenne :

$$\hat{\pi}^{M} = \frac{\sum_{i=1}^{\lvert\mathcal{C}\rvert} \hat{\pi}_i}{\lvert\ \mathcal{C}\ \rvert}, \hat{\rho}^{M} = \frac{\sum_{i=1}^{\lvert\mathcal{C}\rvert} \hat{\rho}_i}{\lvert\ \mathcal{C}\ \rvert}$$

M est utilisé pour indiquer qu'il s'agit de macroaveraging.

Il existe d'autres mesures, que nous ne détaillerons pas ici puisqu'elles sont assez peu utilisées, plus de détails sont disponibles dans Sebastiani (2002, p. 33-34).

4.2.5 Synthèse

Le domaine de la classification automatique est très vaste : tout type de document peut être classé, selon n'importe quelle catégorie. Nous nous sommes ici restreints à la classification textuelle. Bien que travaillant sur des documents issus du Web, ils sont de par leur contenu similaires à de simples textes. Nous avons vu dans cette partie le fonctionnement global de l'élaboration d'un système de classification automatique de textes. Celle-ci se compose de trois principales étapes : l'indexation des documents, l'apprentissage du classifieur et l'évaluation du classifieur.

Nous présentons dans la section suivante l'adaptation de chacune de ces étapes à notre problème : la classification automatique de documents issus du Web en langues française et japonaise, dans des domaines de spécialités, selon leur type de discours, scientifique ou vulgarisé.

4.3 Élaboration d'un système de classification des types de discours scientifique et vulgarisé sur des documents français et japonais

Nous présentons dans cette section l'élaboration d'un système de classification automatique de documents spécialisés issus du Web en français et japonais selon leur type de discours : scientifique ou vulgarisé. Nous détaillons la mise en place de chacune des étapes présentées précédemment : l'indexation des documents (section 4.3.1), le choix des méthodes d'apprentissage (section 4.3.2), la création des classifieurs (section 4.3.3).

4.3.1 Indexation des documents

Nous avons vu dans la section 4.2.2 qu'un système d'apprentissage se base sur une représentation compacte et vectorielle des documents. Les documents sont alors décrits de la façon suivante :

$$\vec{d_i} = (w_{1i}, \ldots, w_{|\mathcal{T}|i})$$

où $\mathcal{T}$ correspond à un ensemble de critères caractérisant les documents, et chaque w_{ij} au poids associé à chacun de ces critères. Il est donc nécessaire de savoir comment décrire les documents, en fonction de la classification souhaitée. Avant toute chose, fixons la terminologie que nous employons dans cette partie. Nous distinguons trois notions dans la description des documents :

les critères, définition théorique d'un des éléments caractérisant les documents ;

les marqueurs, définition opérationnelle d'un critère, c'est-à-dire moyen de mise en œuvre de ce critère ;

les poids des critères, valeur numérique pour chaque critère, calculée sur la base des marqueurs.

Comme nous l'avons vu dans la section 4.2.2, nombreux sont les travaux en classification textuelle se basant uniquement sur une représentation en « sac de mots » des documents, où chaque élément des vecteurs correspond à la fréquence pondérée d'un terme du texte. Dans cette représentation, un critère correspond à l'un des termes du texte, le(s) marqueur(s) associé(s) peu(ven)t alors être sa forme lemmatisée, ses différentes formes fléchies ou un ensemble de synonymes. Le poids de ce critère correspond par exemple au nombre d'occurrences de ses marqueurs dans le texte.

Nous souhaitons cependant éviter la représentation en « sac de mots », partant du principe que le texte est l'unité fondamentale de notre corpus, et que le discours, contrairement à la thématique qui peut être caractérisée lexicalement, s'ancre à différents niveaux dans un texte. Nous nous basons donc ici sur les critères définis dans notre typologie (voir chapitre 3). Cette typologie est composée de trois niveaux d'analyse :

- structurel : concernant les caractéristiques propres au contexte de création du document, ainsi que ses caractéristiques graphiques ;
- modal : concernant les caractéristiques relatives à la modalisation dans les documents ;
- lexical : concernant le vocabulaire et les caractéristiques lexicales des documents.

Lewis (1992) affirme que les représentations des vecteurs de termes se basant sur l'unité mot des textes est la plus efficace. Cependant, Tzeras et Hartmann (1993) améliorent leurs résultats en accompagnant les termes de marqueurs plus complexes, basés sur les groupes nominaux. Afin de trouver un compromis entre une représentation structurée des documents et une représentation à l'aide de marqueurs simples, nous avons donné la priorité à une analyse en surface des documents, à l'aide de patrons lexicaux et lexico-syntaxiques. Ces marqueurs se basent sur les documents dans leur forme originale, ainsi qu'une version texte brute et une version enrichie, étiquetée et lemmatisée.

L'étape la plus importante pour l'indexation des documents consiste à déterminer, pour chacun des critères de la typologie, le marqueur correspondant. La première contrainte de ce travail porte sur la nature des critères et a donc été travaillée en amont de la phase d'indexation. Il s'agit de déterminer quels sont les critères *opératoires*. La notion d'opérabilité est assez complexe. Définissons une condition minimale : un critère est opératoire s'il existe des marqueurs lui correspondant. Ainsi, l'opérabilité dépend fortement de la complexité de l'implémentation des marqueurs. Puisque nous privilégions ici une analyse en surface des documents, nous souhaitons avoir des marqueurs dont la reconnaissance a une complexité linéaire.

Le choix de cette complexité linéaire pour nos marqueurs peut toutefois introduire deux phénomènes :

Le bruit, apparaissant lorsque des occurrences détectées grâce aux marqueurs ne correspondent pas au critère recherché. Dans ce cas, les marqueurs sont souvent trop généraux ce qui permet de détecter tous les critères en introduisant dans les résultats des faux exemples, ce qui correspond à une forte précision mais un faible rappel ;

Le silence, apparaissant lorsque certaines occurrences recherchées ne sont pas détectées. Des marqueurs trop précis peuvent causer le silence, cela correspond à une faible précision et un fort rappel.

Afin de minimiser chacun de ces deux phénomènes, il est nécessaire de trouver un compromis entre des critères très précis permettant de ne pas introduire de bruit et des critères très généraux afin d'éviter le silence. Pour cela, nous cherchons à créer des listes exhaustives de marqueurs relativement simples. Nous étudions cependant le bruit et le silence introduits par certains critères dans le chapitre 5.

Nous présentons dans les parties suivantes les marqueurs pour chacune des catégories de critères.

4.3.1.1 Critères structurels

Les critères structurels décrivent principalement les caractéristiques non linguistiques des documents. Celles-ci portent principalement sur la structure des documents (mise en page, typographie, images, etc.), ainsi que certains aspects relatifs à leur mise en ligne sur le Web (patron d'URL, format, etc.). Selon le format des documents traités, toutes les informations ne sont pas disponibles. Elles le sont pour les documents au format `HTML`, majoritaires sur le Web. Par contre, pour les fichiers PDF, aucune information sur la structure n'est disponible (pas de structuration en paragraphes, pas d'hyperliens, etc. dans leur version au format texte). Nous recherchons néanmoins la présence de méta-informations dans les fichiers PDF (contenant généralement l'auteur, le logiciel créateur, la date de création...).

Critères	**Marqueurs**	
	Français	**Japonais**
Patron d'URL	*http://www.univ-***.fr* *http://www.chu-***.fr* ...	
Longueur des documents	Nb. de mots dans le texte	Nb. de caractères dans le texte
Format des documents	Extension du fichier	
Méta-informations	Utilisation de balises `<META>`	
Titre de la page	Utilisation de balises `<TITLE>`	
Mise en page	Utilisation de tableaux sans bordures, feuilles de style CSS	
Fond de page	Images ou couleurs de fond	
Images	Nombre de balises `<IMG>`	
Liens	Nombre de liens, balises `<A HREF>`	
Paragraphes	Nombre de balises `<P>`	
Listes d'items	Nombre de listes d'items, balises `<UL>` et `<LI>`	
Nombre de phrases	Nombre de symboles de ponctuation de fin de phrase	
Typographie	Nombre de balises `<B>, <I>`	

Table 4.3 – Marqueurs des caractéristiques structurelles

Les marqueurs relatifs à la structure des documents HTML correspondent tous à des balises. Quant aux informations sur le format et l'URL, nous nous basons sur des patrons lexicaux. La longueur des documents est évaluée grâce au nombre de mots et caractères dans les textes pour le français et le japonais (respectivement). Ce critère nous apporte des informations pour la classification, et nous permet aussi de normaliser les documents (cf. section 4.3.1.4).

L'ensemble des marqueurs pour chacun des critères structurels est présenté dans le tableau 4.3. Tous ces marqueurs se basent sur les documents dans leur format original, ainsi que sur les méta-informations de chaque fichier.

4.3.1.2 Critères modaux

Nous avons présenté dans le chapitre précédent les caractéristiques modales de notre typologie. Nous y proposions deux modalités : la modalité locutive de Charaudeau (1992), basée sur les liens entre le locuteur, son interlocuteur, son propos, et le contexte dans lequel il se trouve et la modalité Irrealis de Givón (1994), basée sur la réalisation des actions énoncées.

Dans les deux ouvrages de référence, la théorie est présentée et illustrée sur la langue générale. Nous avons donc, dans un premier temps, sélectionné les critères pouvant s'adapter aux langues de spécialité. Dans un second temps, nous avons, pour chaque critère et pour chaque langue, trouvé un ensemble de marqueurs adaptés[1]. Nous nous sommes pour cela aidés du corpus d'apprentissage, ayant déjà servi à l'analyse stylistique.

Pour déterminer l'ensemble des marqueurs, nous nous sommes basés sur les documents du corpus dans leur format texte, ainsi que dans leur format enrichi, c'est-à-dire étiqueté et lemmatisé. Notre objectif étant de trouver des marqueurs simples, basés sur des patrons lexicaux et lexico-syntaxiques, nous avons privilégié pour chaque critère une analyse simple, en superficie.

La modalité locutive Nous comptons trois catégories de critères dans cette théorie de la modalité (cf. chapitre 3). Les modes d'organisation du discours sont présentés dans le tableau 4.5. Dans cette catégorie se trouvent les critères permettant d'organiser, d'articuler son discours. Pour certains critères, les marqueurs sont très simples, comme pour la ponctuation ou les citations. D'autres par contre demandent quelques recherches, notamment les connecteurs pragmatiques. Nous nous sommes alors basés sur Ducrot (1980) pour trouver les principaux connecteurs. Les marqueurs de la modalité allocutive et élocutive sont présentés dans le tableau 4.4. Certains de ces marqueurs se basent sur des mots clés, comme l'interpellation, tandis que d'autres se basent sur des patrons lexico-syntaxiques, faisant appel aux étiquettes grammaticales. Par exemple, lorsque les marqueurs sont des verbes, il faut s'assurer du sujet de la phrase (*je* et *nous* pour la modalité élocutive, *tu* et *vous* pour la modalité allocutive). La langue japonaise n'utilisant pas ou peu de pronoms, il a fallu adapter ce travail. C'est grâce à l'utilisation de particules en fin de verbes, au mode de conjugaison ainsi qu'aux marqueurs de politesse que nous détectons la présence du locuteur et de l'interlocuteur.

La modalité Irrealis La modalité Irrealis, bien que ne se servant pas des pronoms, se base beaucoup sur les catégories grammaticales et les temps des verbes. Ainsi, nous utilisons les documents en texte ainsi qu'en format enrichi pour implémenter les marqueurs. L'absence de pronoms dans cette théorie simplifie la tâche pour la langue japonaise. Givón (1994) ayant créé sa théorie sur la langue anglaise, nous avons dû effectuer une double adaptation : au français et au japonais. Les marqueurs français sont

[1] Ce travail a été effectué en collaboration et a donné lieu à deux publications (Gœuriot et al., 2007, 2008).

Critères	**Marqueurs français**	**Marqueurs japonais**
Modalité allocutive		
Pronoms personnels	*Tu, vous*	あなた, 貴方, 貴女, アナタ
Injonction	Verbes 2ème pers impératif *ordonner* (sujet 1ère pers + complément) 2ème pers)	なさる, なさい, ように, ること
Autorisation	Verbe *pouvoir* (2ème pers) *permettre de, autoriser à* (sujet 1ère pers + complément 2ème pers)	
Avertissement	*Attention*, verbes *avertir, prévenir, informer* (sujet 1ère pers + complément 2ème pers)	
Jugement	*Bravo, heureusement, malheureusement, malencontreusement* *Félicitations* *féliciter, reprocher, applaudir approuver, accuser, condamner* (2ème pers.)	
Suggestion	*Conseiller, recommander, inviter, proposer, suggérer, solliciter*(sujet 1ère pers + complément 2ème pers)	ほうがよ, ほうがい, ほうがよい, お勧めし
Interpellation	*Madame, monsieur, docteur, mademoiselle*	皆さん, こんにちは, S先生, 皆様, みなさん, 皆様, 皆さま
Requête	*Vouloir, demander, prier* (sujet 1ère pers + complément 2ème pers)	お願いします
Modalité élocutive		
Pronoms personnels	*Je, nous, on*	私, わたし, わたしたち, 私達, わたし達, 私たち, 我々, われわれ, 私ども, わたしども, 私共, わたし共
Constat	*Remarquer, apercevoir, constater, découvrir, noter, observer, voir* (1ère pers)	痛感する
Savoir	*Savoir, connaître, ignorer* (1ère pers)	知らなかった, 知っていたが, 知らなかった, 知らないが, 私は知りませんでした, 知らなかったから, 知っていました
Opinion	*Penser, paraître, sembler, estimer* (1ère pers)	と思う, 考える, と考えられる, 思う, 考える
Volonté	*Vouloir, souhaiter, avoir envie* (1ère pers)	たい
Déclaration	*Avouer, confesser, reconnaître, révéler, dévoiler, affirmer, prétendre, confirmer, attester, certifier* (1ère pers)	宣言致しま, 告げ
Obligation	Verbe *devoir* (1ère pers), *obligé, obligatoirement*	
Interdiction	*Interdire, défendre* (1ère pers)	...な, ...してはいけません, 禁止です, 禁止だ, ならない, なりません

Table 4.4 – Marqueurs des caractéristiques modales (théorie de Charaudeau)

Critères	**Marqueurs français**	**Marqueurs japonais**
Citations	« … », « … », ’…’	<…>,「…」, 『…』, « … », 《…》
Types de phrases		
Déclarative Interrogative Exclamative	., … ? !	.。. ?? !!
Connecteurs pragmatiques	alors, car, comme, d’ailleurs, donc, enfin, mais, puisque, sinon…	また, 及び, および, しかし, 又, 或は, あるいは, 又は, または…
Fins de phrases		でした, ください, ましょう, です, しょう, ました, ません ます, 下さい

Table 4.5 – Marqueurs des modes d’organisation du discours

essentiellement des verbes ou leur conjugaison (ex. : temps futur, ordre…), et les marqueurs japonais se basent eux-aussi sur les verbes et les particules. L’ensemble de ces marqueurs est présenté dans les tableaux 4.6 et 4.7.

Critères	**Marqueurs français**
Temps futur	Verbes au futur et formes « *être en passe de*, « *sur le point de* »
Adverbes modaux	*Probablement, peut-être, certainement, à peu près, possiblement, assurément, selon toute apparence, éventuellement, sûrement, vraisemblablement, nullement, aucunement, pas du tout, point, rien, apparemment, sans doute, presque*
Compléments de verbes	
de manipulation	*Préférer, suggérer, espérer, demander, permettre, risquer, prier, souhaiter, commander*
de perception	*Croire, trouver, penser, soupçonner, suggérer, espérer, souhaiter*
Discours non-déclaratif	
ordre	Impératif, forme faire + INF
requête	*Voulez-vous, pouvez-vous, etc.*
questions oui-non	*Est-ce … ? Faut-il … ?*
propositions temporelles	Conditionnel
adverbes modaux	*Pouvoir, vouloir, devoir, aller, paraître, sembler*

Table 4.6 – Marqueurs des caractéristiques modales du français (théorie Irrealis)

4.3.1.3 Critères lexicaux

La catégorie lexicale de la typologie contient un ensemble de critères assez hétérogènes. En effet, souhaitant que notre typologie soit adaptable à tout domaine de spécialité et éventuellement extensible à d’autres langues, ces critères rassemblent des éléments pouvant paraître très différents les uns des autres. L’objectif est finalement de couvrir le plus de manifestations possibles de spécialisation dans un document. Par exemple, la présence d’une bibliographie et de citations bibliographiques dans un texte est un indice assez fort sur son genre (article de recherche par exemple), qui donne ainsi une idée du

Critères	**Marqueurs japonais**
Temps futur	だろう, でしょう
Adverbes modaux	間違いなく, おそらく, 多分, どうやら, 必ず, 絶対, 確かに, まず, まさか, さぞ, どうも, いまに, いずれ, もうすぐ, やがて, 例え, のちほど, いまにも, ぜひ, もし, 仮に, かりに, たとえ, よもや, これから, 万一, きっと, いくら, いかに
Compléments de verbes	
modaux	Particules en fin de mot : たい, たかった, ようとした, ようとしました, 彼女は別の仕事を見つけたかった。 たい, たかった, 私は、その箱を開けようとした。 ようとした, ようとしました
de manipulation	Impératif du verbe + と, 命じる, 禁じる, 頼む, 教える, 説明する Infinitif du verbe + よう, 命じる, 禁じる, 頼む, 教える, Infinitif du verbe + ことを+命じる, 説明する
de perception	と思う, と想像する, と信じる, ことを想像する, ことを疑う, かも知れない, かもしれない, と考えられます, と信じ, ことを信じ, と考えられる, を疑う, と思い
Discours non-déclaratif	
ordre	Forme étiquetées : 命令e, 命令yo, 命令i, 命令ro, な な 助詞-終助詞, な ない ナ 助動詞
requête	te-forme du verbe +してくれますか te-forme du verbe +できますか te-forme du verbe +もらえますか, forme négative pour demander plus poliment, ou montrer la colère (te-forme du verbe +してくれませんか te-forme du verbe +できませんか te-forme du verbe +もらえませんか)
exhortation	未然ウ接続
jussive	せなさい, せてください, せろ
questions oui-non	よね en fin de phrase
propositions temporelles	仮定形

Table 4.7 – Marqueurs des caractéristiques modales du japonais (théorie Irrealis)

type de discours auquel cela peut référer. À l'instar de Namer et Baud (2007), nous nous basons sur les racines gréco-latines des termes afin de déterminer leur spécificité. Nous partons donc de l'hypothèse qu'un document spécialisé fait plus appel aux termes formés de racines gréco-latines. Nous cherchons donc dans les textes les mots ayant des affixes et suffixes couramment utilisés, que nous avons collecté dans (Béchade, 1992). Les marqueurs correspondant aux critères tels que les caractères numériques, la ponctuation, les parenthèses, etc. ne sont que de simples collectes de toutes leurs expressions en français et en japonais. Les marqueurs de glose ont été collectés dans (Steuckardt et Niklas-Salminen, 2005). L'emploi d'autres alphabets est très peu utilisé en français, mais beaucoup en japonais : utilisation de deux alphabets pour écrire en japonais, un alphabet pour les noms étrangers et parfois l'alphabet latin pour certaines expressions étrangères.

Les marqueurs des critères lexicaux sont présentés dans le tableau 4.8.

Critères	**Marqueurs français**	**Marqueurs japonais**
Vocabulaire spécialisé	nano-, semi-, post-, inter-… -gramme, -ique, -isme, -algie…	性, 腺様
Caractères numériques	0, 1, 2, 3, 4, 5, 6, 7, 8, 9	0, 1, 2, 3, 4, 5, 6, 7, 8, 9 零, 一, 二, 三, 四, 五, 六, 七, 八, 九, 十, 〇, 百, 千
Unités de mesure	mètre, gramme, degré, calorie, leurs composés et abréviations	キロ グラム, ミリ グラム, キロ メートル, センチ メートル leurs composés et abréviations
Bibliographie	« bibliographie », « références »	参考文献, 引用文献　文献
Citations bibliographiques	(nom, AAAA), (nom, AA), [AA]	
Ponctuation	? , . ; : ! …	, . ・ 。 、 … ?? !!
Marqueurs de glose	par exemple, exemple, ainsi, cependant, comme, en revanche, entre autres, notamment, par contre, …	つまり, 特に, 中でも, 例えば、また
Parenthèses	(…), -…-	(…), (…), 〔…〕, […], {…}
Autres alphabets		hiragana, katakana, latin
Symboles		caractères non-alphanumériques : émoticones, ●, ○

Table 4.8 – Marqueurs des caractéristiques lexicales

4.3.1.4 Finalisation de l'indexation

Une fois l'ensemble des marqueurs déterminé, la phase d'indexation se déroule de la façon suivante : pour chaque document un vecteur est calculé. La forme de ces vecteurs peut varier à quelques détails près selon le classifieur utilisé, nous évoquons ce problème dans la section suivante. La création des vecteurs pour chacun des documents est présentée dans l'algorithme 1.

Cet algorithme est appliqué à l'ensemble des documents du corpus. S'ils font partie du corpus d'apprentissage, nous indiquons dans le vecteur son type de discours pour que l'algorithme d'apprentissage utilise cette information. Les poids de chaque critère sont normalisés. Pour cela, nous utilisons la lon-

Algorithme 1 : Construction de vecteur $\vec{d_i}$

Entrée : Indice i du document d_i
Sortie : Vecteur $\vec{d_i}$
début
  // Initialisation
  $d_i^o \leftarrow OuvrirOriginal(i)$; $d_i^m \leftarrow OuvrirMeta(i)$; $d_i^t \leftarrow OuvrirTexte(i)$;
  $d_i^e \leftarrow OuvrirEtiquette(i)$;
  $\vec{d_i} \leftarrow \vec{0}$;
  $l_i \leftarrow LongueurDocument(d_i^t)$;
  // Construction du vecteur
  si $d_i \in CorpusApprentissage$ **alors**
    $\vec{d_i}[0] \leftarrow TypeDiscours(d_i)$;
  sinon
    $\vec{d_i}[0] \leftarrow 0$;
  pour tous les $j \in [1 \ldots \mid \mathcal{T} \mid]$ **faire**
    suivant *Critère* j **faire**
      cas où $j \in$ *Critères Structurels*
        si $j \in$ *Méta-informations* **alors**
          $\vec{d_i}[j] \leftarrow CalculePoids(j, d_i^m)$;
        sinon
          $\vec{d_i}[j] \leftarrow CalculePoids(j, d_i^o)$;
      cas où $j \in$ *Critères Modaux*
        si $j \in$ *Étiquettes-lemmes* **alors**
          $\vec{d_i}[j] \leftarrow CalculePoids(j, d_i^e)$;
        sinon
          $\vec{d_i}[j] \leftarrow CalculePoids(j, d_i^t)$;
      cas où $j \in$ *Critères Lexicaux*
        si $j \in$ *Étiquettes-lemmes* **alors**
          $\vec{d_i}[j] \leftarrow CalculePoids(j, d_i^e)$;
        sinon
          $\vec{d_i}[j] \leftarrow CalculePoids(j, d_i^t)$;
  // Normalisation du vecteur
  pour k *allant de 1 à NbCritères* **faire**
    $\vec{d_i}[k] \leftarrow \frac{\vec{d_i}[k]}{l_i}$;
  retourner $\vec{d_i}$;
fin

gueur du document dans sa version texte. Nos critères étant peu nombreux et rigoureusement sélectionnés, nous n'avons pas eu besoin d'appliquer de méthode de *dimensionality reduction*.

4.3.2 Choix des méthodes d'apprentissage

À partir d'une indexation de documents, de nombreux algorithmes peuvent apprendre un modèle. Parmi les plus connus, citons les réseaux de neurones, les classifications de Bayes, les machines à vecteurs de support, etc. Sebastiani (2002) a mené une étude comparative sur ces systèmes. Il teste dans cette étude différents systèmes de classification automatique se basant sur des corpus composés de dépêches Reuters, en faisant varier la quantité de documents, de documents d'apprentissage et de catégories. Il semble dans cette étude que, pour un nombre de classes limitées, les meilleures techniques soient les machines à vecteur de support ainsi que les arbres de décision (Sebastiani, 2002, p. 38). Nous cherchons ici à produire un classifieur binaire à partir de représentations vectorielles de documents basées sur une quarantaine de critères. Nous présentons dans les sections suivantes les deux algorithmes choisis, les systèmes utilisés et les choix effectués pour l'évaluation de ces classifieurs.

4.3.2.1 Machines à vecteurs de support

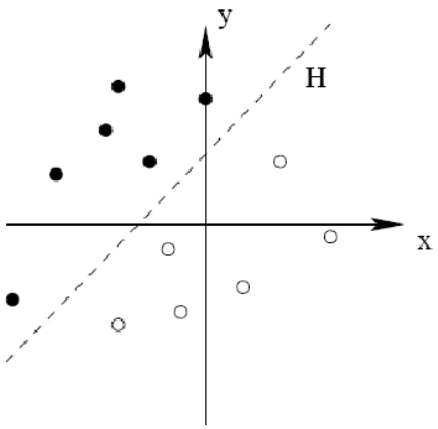


Figure 4.3 – Le cas le plus simple : dans un espace bidimensionnel, une droite sépare les deux ensembles d'exemples

Le modèle des machines à vecteurs de support[2] a été introduit par Vapnik (1998) et a été appliqué à la classification textuelle par Joachims (2002). Cette méthode se base sur une représentation de chaque élément à classer sous la forme d'un point dans un espace multidimensionnel. Chaque *critère* correspond à une dimension ; chaque élément est donc représenté par les poids de chacun de ses critères. Le principe général de cette méthode est de chercher, dans un espace multidimensionnel, un hyperplan H, combinaison des $\sigma_i \in \mathcal{T} = \{\sigma_1, \ldots, \sigma_{|\mathcal{T}|}\}$, tel que σ_i sépare les exemples positifs des exemples négatifs. Tous nos exemples représenteront des cas à deux dimensions, par simplification. La figure 4.3 présente le cas le plus simple, où une droite sépare les exemples positifs (en blanc), des exemples négatifs (en noir).

[2] « *Support vector machines* », traduit par Cornuéjols et Miclet (2002).

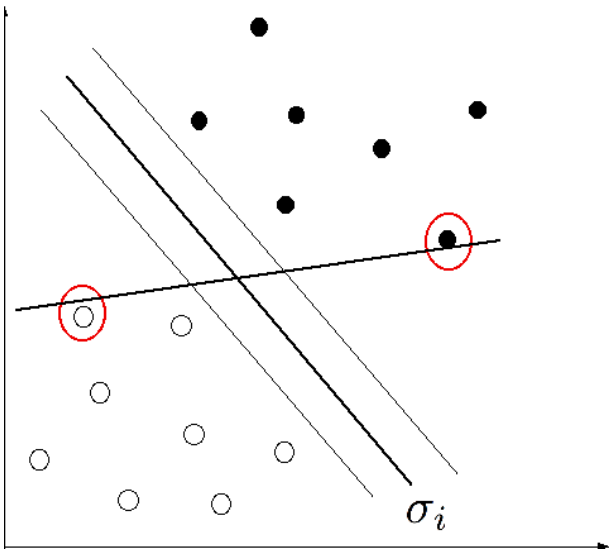


Figure 4.4 – Cette méthode cherche à trouver l'hyperplan séparant l'espace des données en deux en ayant une marge maximale

Cet espace σ_i doit de plus présenter une marge la plus grande possible, c'est-à-dire permettant les translations les plus grandes de σ_i. La figure 4.4 présente un cas simple d'espace bidimensionnel. Les deux droites tracées séparent l'espace en deux, mais seule celle notée σ_i a des marges maximales.

Les vecteurs les plus proches de l'hyperplan, déterminant ainsi la taille de la marge, sont appelés *vecteurs de support*(Joachims, 2002, p. 37). Dans un cas non-linéaire, l'algorithme ajoute une dimension à l'espace vectoriel, générant ainsi un *espace de re-description* de plus grande dimension dans lequel il va chercher à retomber sur un cadre linéaire. Cette technique est abordée plus en détail dans Joachims (2002). Nous utilisons le système `SVMlight`[3], développé par Thorsten Joachims en 2002.

4.3.2.2 Arbres de décision

La théorie des arbres de décision est basée sur la théorie de Breiman élaborée en 1984. Cette théorie se démarque des classifieurs probabilistes et fait partie des algorithmes qualifiés de *symboliques*.

Dans le cadre de la classification textuelle, un arbre de décision est composé de nœuds, représentant des critères de classification, de branches représentant des conditions sur les poids des critères et de feuilles représentant les classes. Le parcours d'une branche correspond ainsi à une combinaison de poids de critères que la représentation vectorielle d'un document doit contenir pour appartenir à la classe correspondant à la feuille. Le tableau 4.9 décrit un exemple présenté par Quinlan (1993), permettant de décider, en fonction des conditions météorologiques si un jeu est possible ou pas. Un arbre correspondant à cet exemple est présenté dans la figure 4.5.

Cette méthode d'apprentissage se base sur une décomposition d'un problème de classification en une suite de tests (imbriqués) portant sur un critère ou une combinaison linéaire de plusieurs critères, afin de créer des règles de classification sous forme d'arbres. L'objectif final étant de créer une séquence hiérarchique de tests, aussi courte que possible, divisant successivement l'ensemble des données d'apprentis-

[3] `http://svmlight.joachims.org/`

Numéro	Ensoleillement	Température (°F)	Humidité (%)	Vent	Jouer
1	soleil	75	70	oui	oui
2	soleil	80	90	oui	non
3	soleil	85	85	non	non
4	soleil	72	95	non	non
5	soleil	69	70	non	oui
6	couvert	72	90	oui	oui
7	couvert	83	78	non	oui
8	couvert	64	65	oui	oui
9	couvert	81	75	non	oui
10	pluie	71	80	oui	non
11	pluie	65	70	oui	non
12	pluie	75	80	non	oui
13	pluie	68	80	non	oui
14	pluie	70	96	non	oui

Table 4.9 – Exemple de Quinlan (1993)

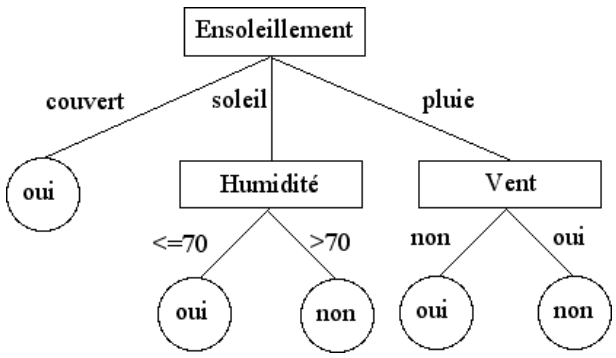


Figure 4.5 – Exemple d'arbre de décision

sage en sous-ensembles disjoints. Sebastiani (2002) décrit l'une des méthodes permettant de générer un arbre de décision pour une classe c_i, basée sur une stratégie *divide and conquer* :

> (i) Si tous les vecteurs d'apprentissage appartiennent à la même classe c_i ou $\bar{c_i}$ alors l'arbre ne contient qu'une feuille ;
> (ii) Sinon sélectionner un critère c_k partitionnant l'espace des vecteurs d'apprentissage en espaces ayant la même valeur pour c_k (*segmentation*). c_k devient alors un nœud et les deux espaces correspondent aux sous-arbres.
> (iii) Réitérer l'étape (ii) sur les sous-arbres jusqu'à avoir classé tous les exemples du corpus d'apprentissage ;
> (iv) Élaguer l'arbre si nécessaire.

Les points clés de cette méthode sont : le choix du critère c_k et l'élagage de l'arbre (suppression des branches qui ne sont pas nécessaires, quitte à ajouter quelques erreurs). Le choix du critère c_k s'effectue grâce à des mesures de gain ou d'entropie (ex. : χ^2, entropie de Shannon, etc.), afin de sélectionner à

chaque étape le critère le plus discriminant possible. L'élagage de l'arbre permet de pallier le problème de *sur-apprentissage*. En effet, les arbres de décision tendent à être très complexes en collant le plus possible aux données. La détermination de la taille de l'arbre est alors essentielle pour diminuer la complexité des arbres générés (Breiman et al., 1984). Deux méthodes permettent de réduire la taille des arbres : le pré- et le post-élagage. Le pré-élagage consiste à fixer un critère permettant de stopper la construction de l'arbre. Ce critère d'arrêt peut par exemple évaluer l'apport informationnel de la segmentation initiée. Le post-élagage consiste à réduire la taille de l'arbre une fois celui-ci construit, en estimant par exemple le taux d'erreur introduit par l'élagage d'une branche. L'ensemble de ces techniques est abordé plus en détail dans Breiman et al. (1984); Quinlan (1993). Nous utilisons le système `C4.5`[4], développé par Quinlan (1993).

4.3.3 Création des classifieurs et protocole d'évaluation

4.3.3.1 Indexation des documents

Nous avons présenté dans la section 4.3.1.4 la phase d'indexation, créant pour chaque document sa représentation vectorielle. Les systèmes `C4.5` et `SVMlight` nécessitent un format particulier, présenté dans le tableau 4.10.

Le format requis pour le système `SVMlight` nécessite : d'indiquer en début de vecteur la classe à laquelle appartient le document, précéder chaque poids de l'indice correspondant au critère. Le dernier élément du vecteur est facultatif, nous avons choisi d'y indiquer le numéro de fichier. Si certains critères ont un poids nul ou que leur calcul n'est pas pertinent, il suffit de les omettre.

Le format requis pour le système `C4.5` nécessite : d'indiquer en fin de vecteur la classe à laquelle appartient le document, donner les poids de chaque critère (0 si non pertinent), séparés par des virgules.

`SVMlight`	Document d'apprentissage scientifique d_i	$+1\ 1 : w_{1i}\ 2 : w_{2i} \ldots n : w_{ni}$ #nomfichier
	Document d'apprentissage vulgarisé d_j	$-1\ 1 : w_{1j}\ 2 : w_{2j} \ldots, n : w_{nj}$ #nomfichier
	Document d'évaluation d_k	$0\ 1 : w_{1k}\ 2 : w_{2k} \ldots n : w_{nk}$ #nomfichier
`C4.5`	Document d'apprentissage scientifique d_i	$w_{1i}, w_{2i}, \ldots w_{ni}$, S
	Document d'apprentissage vulgarisé d_j	$w_{1j}, w_{2j}, \ldots, w_{nj}$, V
	Document d'évaluation d_k	$w_{1k}, w_{2k}, \ldots, w_{nk}$, S ou V

Table 4.10 – Formats d'indexation pour `SVMlight` et `C4.5`

Exemple :
Le document correspondant au vecteur suivant : sera représenté par la chaîne

Fichier	Classe	Critère 1	Critère 2	Critère 3	Critère 4
filename	SC	0.556	0.7	21	2.8

$$+1\ 1 : 0.556\ 2 : 0.7\ 3 : 21\ 4 : 2.8\#filename$$

pour `SVMlight` et

$$0.556, 0.7, 21, 2.8, S$$

[4] `http://www.rulequest.com/Personal/c4.5r8.tar.gz`

pour `C4.5`.

Pour ces deux systèmes, les vecteurs doivent être stockés dans des fichiers : un fichier pour le corpus d'apprentissage, un pour le corpus de test. Le système `C4.5` a de plus besoin d'un fichier de description, dans lequel les critères sont succinctement décrits. Le fichier doit débuter par les différentes classes à apprendre :

```
Scientifique, Vulgarisé
```

Puis doit suivre la description des critères, où seule la nature (continu ou discret) de chaque critère est nécessaire, par exemple :

```
format-fichier : html, pdf, ps
compte-images : continuous
```

Une fois les fichiers de description des vecteurs définis, ils peuvent être utilisés afin d'apprendre les classifieurs et les tester.

4.3.3.2 Application des systèmes d'apprentissage

Deux systèmes d'apprentissage sont utilisés : `SVMlight` et `C4.5`. Ne disposant au départ que d'un corpus, le corpus [DIAB_CP], nous avons décidé d'utiliser la méthode dite « par validation croisée » (*N-fold cross validation*) (Cornuéjols et Miclet, 2002, p. 113). Cette méthode consiste à :

1. Partitionner le corpus en N sous-corpus de tailles égales ;
2. Retenir le i^{eme} sous-corpus, apprendre le classifieur sur les $N-1$ autres sous-corpus ;
3. Évaluer le classifieur avec le i^{eme} sous corpus ;
4. Réitérer les étapes 2 et 3 avec i allant de 1 à N.

Nous choisissons $N = 5$. Les résultats sont alors présentés sous la forme de moyenne sur les cinq classifieurs et le meilleur classifieur (meilleurs rappel et précision) est sélectionné pour la phase d'évaluation.

Pour apprendre un classifieur avec `SVMlight`, il faut utiliser la commande :

`./svm_learn [options]` *fichier-apprentissage svm-classifieur*

Les options de la commande `svm_learn` sont nombreuses, elles permettent de paramétrer le classifieur. Joachims (2002) a mis en place un réglage de ces paramètres par défaut permettant à de nombreux cas de fonctionner sans paramétrage. Cependant, il arrive que les classifieurs, si les classes sont de taille inégale, classent tous les documents dans la classe la plus importante en taille. Ce problème peut être réglé en paramétrant l'option *j*. Cette option permet de faire varier le coût des erreurs lors de l'apprentissage. Plus de détails sur ces options sont disponibles dans (Joachims, 2002, p. 197).

Le fichier *svm-classifieur* correspond au classifieur créé, qui peut être par la suite testé en utilisant la commande :

`./svm_classify [options]` *fichier-test svm-classifieur fichier-resultat*

Le système `C4.5` permet d'apprendre un classifieur (et générer un arbre de décision) grâce à la commande :

`c4.5 -f` *file-id*

L'option `-f` permet de spécifier un identifiant de fichier `file-id`, auquel doivent correspondre : le fichier de description des critères, `file-id.names`, le fichier d'apprentissage `file-id.data`. À partir de ces fichiers sont générés : `file-id.unpruned` l'arbre de décision non-élagué et `file-id.tree` l'arbre de décision élagué. L'option `-u` permet de fournir un fichier de test `file-id.test`.

4.3.3.3 Évaluation des classifieurs

Nous choisissons d'évaluer nos classifieurs à l'aide des métriques de rappel et de précision (voir section 4.2.4). Nous souhaitons évaluer nos classifieurs sous différents angles. Dans un premier temps, afin de mettre en œuvre la méthode par validation croisée, chaque sous-corpus doit être testé avec le classifieur correspondant. Dans un second temps, nous souhaitons « mettre à l'épreuve » notre système de classification en le soumettant à un corpus portant sur une autre thématique. Enfin, nous souhaitons utiliser ces systèmes de classification afin de mesurer la pertinence de chacune de nos catégories de critères.

Le chapitre 5 porte sur la phase d'évaluation, les résultats de la classification et leur analyse.

4.4 Conclusion

Nous avons présenté dans la première partie de ce chapitre la méthode d'élaboration d'un système de classification automatique. Celle-ci se déroule en trois étapes : l'indexation des documents, l'apprentissage du classifieur et l'évaluation du classifieur. L'indexation des documents consiste à générer pour chaque document une représentation vectorielle, chaque élément des vecteurs correspondant à la valeur d'un critère. Ces critères peuvent être des fréquences de termes, de patrons lexico-syntaxiques… Ces représentations vectorielles sont ensuite utilisées afin qu'un système d'apprentissage automatique apprenne à reproduire la classification souhaitée à partir des valeurs des critères. Enfin, le classifieur généré est évalué à l'aide de documents n'ayant pas servi à l'apprentissage. La seconde partie de ce chapitre est consacrée à la mise en œuvre de cette méthode aux classes scientifiques et vulgarisées de domaines spécialisés. Dans un premier temps nous avons présenté la création des représentations vectorielles des documents de notre corpus d'apprentissage ([DIAB_CP]) à l'aide des critères de la typologie présentée dans le chapitre 3. Ces vecteurs sont ensuite utilisés afin d'apprendre les modèles de classification à l'aide des systèmes `SVMlight` (machines à vecteurs de support) et `C4.5` (arbres de décision). Les résultats obtenus par ces classifieurs sont présentés dans le chapitre 5.

CHAPITRE 5

Résultats et évaluation de la classification

5.1 Introduction

Nous présentons dans ce chapitre l'évaluation des classifieurs dont l'élaboration a été décrite dans le chapitre 4. Nous souhaitons les évaluer d'un point de vue technique : résultats obtenus par chaque classifieurs, mais aussi évaluer quelle influence ont les critères de la typologie sur la classification. Leur apprentissage a été effectué sur le corpus d'apprentissage [DIAB_CP], portant sur le diabète et l'alimentation. Nous les évaluons sur un second corpus, appelé [BC_CP] et portant sur le cancer du sein (décrit dans la section 5.2). Les résultats de cette évaluation figurent dans la section 5.3. Dans un second temps nous testons la pertinence de chacune des catégories de critères de la typologie dans la section 5.4, ce qui nous permet d'améliorer nos classifieurs en ne conservant que les catégories de critères les plus discriminantes. Nous analysons ensuite plus en détails les classifieurs obtenus et l'influence des critères sur la classification (section 5.5). Afin de prouver l'efficacité de notre classification, nous la comparons à une méthode classifique de catégorisation textuelle : la méthode par vecteurs de termes (section 5.6. Nous terminons ce chapitre par une discussion sur l'aspect binaire de la distinction des types de discours scientifique et vulgarisé, dans laquelle nous montrons qu'il existe un continuum entre ces deux classes (section 5.7).

5.2 Corpus d'évaluation

Dans le chapitre 2, nous présentions le corpus d'étude ayant tout d'abord servi à mettre en application les études menées sur la construction des corpus comparables. Dans un second temps, ce corpus a été utilisé afin de mener à bien une analyse stylistique, permettant de créer une typologie des discours scientifiques et vulgarisés dans des documents issus du domaine médical, en français et japonais. À partir de ce corpus et de cette typologie, des classifieurs ont pu être générés à l'aide des systèmes `SVMlight` et `C4.5`.

Ce corpus d'étude, utilisé comme corpus d'apprentissage, va nous permettre de tester une première fois les classifieurs générés. Il est composé de documents issus du Web, traitant de la thématique « diabète et alimentation ». Notre objectif étant d'intégrer un classifieur à un système d'assistance à la construction de corpus comparables, ce classifieur doit être suffisamment robuste pour classer correctement des documents spécialisés issus de différentes thématiques. Nous avons alors construit un corpus comparable d'évaluation en français et japonais portant sur une nouvelle thématique.

Nous présentons dans cette section les différentes étapes de la construction de ce corpus : délimitation du cadre, sélection et collecte des données, normalisation, annotations. Nous présentons pour finir les caractéristiques du corpus ainsi constitué.

5.2.1 Délimitation du cadre

Nous cherchons, comme pour le corpus [DIAB_CP], à collecter des documents scientifiques et vulgarisés. Le choix de la thématique de ce corpus d'évaluation a été fait en fonction des thématiques pour lesquelles de nombreux textes sont publiés. Nous restons dans le domaine médical et cherchons donc une thématique sur laquelle il existe un grand nombre de publications scientifiques et vulgarisées. Nous avons choisi le thème du « cancer du sein », les travaux dans cette thématique nous semblant très actifs, que ce soit en France ou au Japon. Des publications scientifiques telles que des articles de recherche, rapports, note aux praticiens… sont accessibles directement sur le Web ou via certains portails médicaux et scientifiques, ainsi que des brochures à l'usage des patientes.

Comme pour le corpus [DIAB_CP], nous situons la comparabilité à deux niveaux : le thème « cancer du sein » et le type de discours, scientifique ou vulgarisé. Nous nous fixons une taille limite d'une centaine de documents par langue.

5.2.2 Sélection et collecte des données

Nous souhaitions nous limiter aux portails médicaux pour construire notre corpus. En effet, ces sites garantissent une certaine qualité des documents (qualité éditoriale et garantie du contenu pertinent des documents) et en rassemblent de grandes quantités. Cette solution est donc plus rapide. Pour étendre les mots clés, nous nous sommes principalement basés sur la méthode consistant à utiliser des mots sémantiquement liés collectés lors de la recherche de documents. Nous utilisons les mêmes critères de classification que dans le chapitre 2 :

- 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 ».

Cependant, nous disposons, en utilisant les portails, d'informations complémentaires permettant de faciliter la classification. La figure 5.1 montre quel type d'informations le portail CISMeF fournit sur chaque document lors de recherches. Le champ *résumé* nous permet de vérifier que le document correspond à notre thématique. À la fin de ce résumé est souvent indiqué le pays d'origine de l'auteur du document. Cette information nous permet de filtrer les documents écrits par des auteurs français (et donc non traduits). Le champ *site éditeur* permet de plus de vérifier que le document n'est pas issu d'une organisation internationale (auquel cas il est très probable qu'il soit une traduction). Le champs *Type* renseigne souvent le genre du document, nous utilisons ce champ ainsi que le contenu du document afin de déterminer son type de discours. Par exemple, un document dont le type est *brochure information patient* sera classé comme vulgarisé. Un document dont le type est *information scientifique et technique* sera classé come scientifique. Pour les types tels que *article de périodique*, nous consultons le contenu afin de déterminer le type de discours.

Nous n'avons collecté que les documents aux formats `PDF` et `HTML`, en utilisant l'outil `wget`.

5.2.3 Normalisation, annotation et documentation

Les documents sélectionnés sont convertis en format texte, transormés en encodage UTF-8 et nettoyés si nécessaire (afin de ne conserver que le texte). Nous annotons ensuite le corpus avec des étiquettes morpho-syntaxiques. Cela est réalisé en français avec les logiciels *Brill* (Brill, 1994) et *Flemm* (Namer,

Figure 5.1 – Exemples de description de documents sur CISMeF

2000), en japonais avec le logiciel *Chasen* (Matsumoto et al., 1999). Pour chacun des textes sélectionnés, nous conservons :

- la source du texte : son URL (champ `fileDesc` dans la TEI) ;
- la méthode de recherche utilisée : moteur de recherche et mots clés par exemple (champ `fileDesc` dans la TEI) ;
- la date de téléchargement du document (champ `profileDesc` dans la TEI) ;
- la langue du document (champ `profileDesc` dans la TEI).

5.2.4 Caractéristiques du corpus

Le tableau 5.1 présente les principales caractéristiques du corpus ainsi constitué, que nous appellons [BC_CP] : le nombre de documents et le nombre de mots dans chacune des langues et pour chaque type de discours (SC = scientifique, VU = vulgarisé).

	Français		Japonais	
	SC	VU	SC	VU
Nb. documents	50	42	48	51
Nb. mots	443 741	71 980	211 122	123 277

Table 5.1 – Caractéristiques du corpus [BC_CP]

Ce corpus rassemble plus de 800 000 mots dans deux langues. Il contient moins de documents que le corpus [DIAB_CP], ce qui ne pose pas de problème puisqu'il est seulement utilisé pour l'évaluation des classifieurs. Les corpus comparables tels que nous les construisons n'ont pas vocation à être très

volumineux.

Le système d'indexation présenté dans le chapitre précédent, permettant d'obtenir la représentation vectorielle d'un corpus pour qu'un classifieur puisse interpréter les données, a été appliqué à ce corpus.

5.3 Résultats de la classification

Nous présentons dans cette section les résultats des classifieurs créés. Ces classifieurs sont créés à partir des représentations vectorielles des documents utilisant la totalité des critères de la typologie présentée dans le chapitre 3. Ces classifieurs seront testés dans un premier temps sur le corpus d'apprentissage [DIAB_CP] puis sur le corpus d'évaluation [BC_CP]. Le premier corpus nous permet de valider les classifieurs, tandis que le second nous permet d'évaluer la robustesse de nos classifieurs (et donc de la typologie) sur une autre thématique. Les métriques utilisées sont le rappel et la précision.

Corpus	Système	Classe	Français		Japonais	
			Préc.	Rapp.	Préc.	Rapp.
[DIAB_CP]	SVMlight	SC	0.92	0.65	0.77	0.58
		VU	0.73	0.95	0.87	0.95
		MOY	**0.83**	**0.80**	**0.82**	**0.77**
	C4.5	SC	0.88	0.97	0.50	0.39
		VU	0.96	0.78	0.30	0.42
		MOY	**0.92**	**0.86**	**0.40**	**0.41**
[BC_CP]	SVMlight	SC	0.90	0.57	0.74	0.47
		VU	0.66	0.93	0.65	0.86
		MOY	**0.78**	**0.75**	**0.70**	**0.67**
	C4.5	SC	0.67	0.91	0.35	0.58
		VU	0.83	0.49	0.05	0.02
		MOY	**0.75**	**0.70**	**0.20**	**0.30**

Table 5.2 – Précision et rappel pour chaque langage et chaque classifieur sur les deux corpus

Le tableau 5.2 présente les résultats de la classification pour chaque corpus, chaque classe, chaque langue et chaque système de classification. Nous avons utilisé la méthode dite par validation croisée (cf. chapitre 4 section 4.3.3.2). Les résultats sur le corpus [DIAB_CP] présentés dans ce tableau sont la moyenne des 5 tests. Les résultats sur le corpus d'évaluation ont été obtenus en sélectionnant le meilleur des 5 classifieurs obtenus. En premier lieu, nous remarquons que les résultats obtenus avec le corpus d'évaluation sont plus faibles que ceux obtenus avec le corpus d'apprentissage, d'environ 5 à 10 points. Cela semble assez logique : les documents ayant servi à l'apprentissage (appartenant au corpus [DIAB_CP]) sont mieux classés que les documents inconnus du classifieur (appartenant au corpus [BC_CP]). Nous constatons ensuite que SVMlight donne de meilleurs résultats en moyenne quelle que soit la langue et le corpus : les résultats sont doublés pour le japonais et ils sont légèrement supérieurs sur le corpus [BC_CP]. De plus, les résultats pour le français sont généralement supérieurs aux résultats obtenus sur la langue japonaise. Avec le classifieur SVMlight, ceux-ci sont légèrement plus faibles, tandis qu'avec C4.5 une baisse considérable est observée (de 92 % de précision à 40 % sur le corpus [DIAB_CP]). Cette baisse peut être due aux critères de la typologie. En effet, les deux classifieurs

n'ont pas la même tolérance face aux critères peu discriminants : si ceux-ci n'altèrent pas l'apprentissage avec `SVMlight`, ils apportent beaucoup de bruit avec `C4.5` et peuvent gêner l'apprentissage. Avec `SVMlight`, le rappel de la classe scientifique et la précision de la classe vulgarisée ont tendance à être plus élevés. Sur le corpus [BC_CP], la précision pour la classe scientifique est de 90 % et le rappel pour la classe vulgarisée est de 93 % par exemple. Ce phénomène apparaît lorsque beaucoup de documents sont classés dans la classe vulgarisée. Cette tendance est inversée avec `C4.5`. Ces tendances peuvent s'expliquer par la proportion déséquilibrée de documents scientifiques et vulgarisés dans le corpus d'apprentissage (cf table 2.1 dans le chapitre 2). Pour terminer, nous jugeons certains de ces résultats satisfaisants : nous obtenons plus de 70 % de documents correctement classés pour le français quel que soit le classifieur. Avec `SVMlight`, les résultats sont relativement bons pour la langue japonaise avec en moyenne 70 % de précision et 67 % de rappel sur le corpus d'évaluation. Néanmoins, les résultats obtenus avec `C4.5` sur la langue japonaise sont très insuffisants : 40 % de précision et 41 % de rappel sur le corpus d'apprentissage, et 20 % de précision et 30 % de rappel sur le corpus d'évaluation. Ces résultats particulièrement faibles nous amènent à nous poser des questions sur l'efficacité de certains critères de la typologie. Nous avons vu que `C4.5` était sensible aux critères peu discriminants, qui posent problème lors de l'apprentissage du modèle. Nous examinons dans la section suivante chacune des catégories de critères afin de déterminer leur pertinence et ainsi créer une nouvelle typologie plus efficace.

5.4 Étude des catégories de critères de la typologie

Nous avons vu dans le chapitre 3 les différentes catégories de critères déterminées suite à l'analyse stylistique sur le corpus [DIAB_CP]. Notre typologie est composée de quatre catégories, correspondant à différents niveaux d'analyse des documents. Ces catégories sont les suivantes : critères structurels, critères modaux (théorie de Charaudeau et théorie Irrealis) et critères lexicaux. Nous avons présenté dans le chapitre 4 les marqueurs associés à chacun de ces critères. Nous allons dans cette partie analyser ces différentes catégories, en créant différents classifieurs grâce au corpus [DIAB_CP] que nous testons ensuite sur le corpus [BC_CP].

5.4.1 Pertinence de chaque catégorie de critères

Quatre expériences ont été réalisées dans un premier temps, nous présentons celles-ci dans le tableau 5.3.

Expérience	Critères structurels	Critères modaux		Critères lexicaux
		(Charaudeau)	(Givon)	
1	×			
2		×		
3			×	
4				×

Table 5.3 – Quatre expériences menées afin de tester la pertinence de chaque catégorie de critères

Le tableau 5.4 présente les résultats de la classification pour chaque catégorie de critères, c'est-à-dire les expériences 1 à 4. Pour cela, nous avons généré pour chaque document quatre représentations vectorielles : une par catégorie de critères. De nouveaux classifieurs ont donc été appris pour chacune

de ces catégories en suivant la méthode de validation croisée sur le corpus [DIAB_CP] à l'aide des deux systèmes de classification `SVMlight` et `C4.5`.

Système	Expérience	Catégorie de critères	Français		Japonais	
			Préc.	Rapp.	Préc.	Rapp.
`SVMlight`	1	Critères structurels	0.90	0.67	0.59	0.71
		Critères modaux :				
	2	- Charaudeau	0.60	0.50	0.50	0.49
	3	- Givon	0.78	0.76	0.23	0.23
	4	Critères lexicaux	0.91	0.75	0.58	0.53
`C4.5`	1	Critères structurels	0.85	0.85	0.41	0.44
		Critères modaux :				
	2	- Charaudeau	0.89	0.91	0.39	0.44
	3	- Givon	0.66	0.65	0.15	0.21
	4	Critères lexicaux	0.85	0.85	0.47	0.45

Table 5.4 – Résultats pour chaque catégorie de critères sur le corpus [BC_CP]

Quels que soient les classifieurs et les langues, les critères permettant d'obtenir les meilleurs résultats sont les critères structurels et lexicaux. Par exemple, ils permettent d'atteindre une précision de plus de 90 % en français avec le système `SVMlight`. Par contre, les critères modaux donnent en moyenne de moins bons résultats : entre 60 % et 78 % en français pour `SVMlight`.

Les critères structurels et lexicaux sont très efficaces en français, avec les deux classifieurs mais relativement moins bons pour le japonais bien qu'ils soient les meilleurs pour cette langue. Avec `C4.5`, 85 % des documents scientifiques français sont correctement classés en utilisant les critères structurels ou lexicaux seuls, tandis que pour le japonais, seulement 41 % ou 47 % le sont.

Les critères modaux de Charaudeau permettent d'obtenir des résultats similaires à ceux des critères structurels et lexicaux en utilisant `C4.5`. Par contre, les résultats obtenus en utilisant les critères modaux de Givon sont moins bons avec ce classifieur. Avec `SVMlight`, les résultats sont plus contrastés : avec les critères de Givon, les résultats sont meilleurs pour le français (78 % et 76 % de précision et rappel contre 60 % et 50 %) ; tandis que pour le japonais, ce sont les critères de Charaudeau qui permettent d'atteindre les meilleurs résultats (50 % et 49 % de précision et rappel contre 23 % et 23 %).

Ces résultats nous permettent de faire l'hypothèse que les critères structurels et lexicaux sont les plus pertinents, pour les deux langues et les deux classifieurs. Nous souhaitons tester la combinaison de ces deux catégories, ainsi que mesurer l'apport des caractéristiques modales à cette combinaison.

5.4.2 Évaluation de combinaisons de critères

Dans cette section, nous présentons quatre nouvelles expériences. Dans la première, nous testons l'efficacité de la combinaison critères structurels - critères lexicaux. Cette expérience servira de base aux deux expériences suivantes, visant à tester l'impact des catégories modales sur la combinaison des critères lexicaux et structurels.

Expérience	Critères structurels	Critères modaux		Critères lexicaux
		(Charaudeau)	(Givon)	
5	×			×
6	×	×		×
7	×		×	×
8	×	×	×	×

Table 5.5 – Huit expérience menées afin de tester la pertinence de chaque catégories de critères

5.4.2.1 Résultats de l'expérience 5

Dans cette expérience nous testons la typologie sans aucun critère modal (uniquement les critères structurels et lexicaux). Celle-ci nous sert de base pour l'étude de l'influence de chacune des modalités dans la performance des classifieurs générés. Les résultats sont présentés dans le tableau 5.6. Les résultats obtenus sont assez proches de ceux obtenus pour les critères seuls, mises à part quelques baisses sur le rappel pour les documents japonais, qui peuvent s'expliquer par la différence du nombre d'occurrences entre la partie scientifique et la partie vulgarisée dans le sous-corpus japonais.

Système	Classe	Français		Japonais	
		Préc.	Rapp.	Préc.	Rapp
`SVMlight`	SC	0.90	*0.57*	**0.71**	*0.47*
	VU	*0.66*	**0.93**	**0.63**	**0.84**
`C4.5`	SC	*0.61*	**0.91**	**0.95**	**0.49**
	VU	**0.78**	*0.34*	**0.69**	**0.98**

Table 5.6 – Résultats obtenus pour l'expérience 5 : critères modaux + critères lexicaux

Nous pouvons remarquer que dans l'ensemble les résultats de la combinaison des critères structurels et lexicaux sont meilleurs.

5.4.2.2 Expériences 6, 7 et 8

L'expérience 8 correspond à la typologie complète, celle dont nous présentions les résultats en section 5.3. Dans les expériences 6 et 7, nous ajoutons à la « base » présentée dans l'expérience 5 (critères structurels et lexicaux) alternativement les deux ensembles de critères modaux correspondant aux théories Locutive de Charaudeau et Irrealis de Givon. Le but de cette expérience est de confirmer les observations sur les expériences 2 et 3 sur les catégories de critères et leur pertinence en les insérant dans une typologie composée de deux ensembles de critères qui s'avèrent efficaces. Nous cherchons à connaître quelle théorie est la plus pertinente dans notre contexte (deux langues, documents issus du Web sur des thématiques spécialisées, deux classes).

Pour les documents en français, le système `SVMlight` donne des résultats identiques à ceux de l'expérience 5. Les critères modaux sont donc neutres dans ce contexte. Le système `C4.5` apporte quant à lui des améliorations notables en utilisant les critères modaux de Charaudeau, qu'ils soient seuls (expérience 6) ou combinés avec ceux de Givon (expérience 8). C'est dans l'expérience 6 qu'ils sont les meilleurs.

Système	Expérience	Classe	Français		Japonais	
			Préc.	Rapp.	Préc	Rapp
SVMlight	6	SC	0.90	0.57	0.71	0.47
		VU	0.66	0.93	0.63	0.84
	7	SC	0.90	0.57	0.71	0.47
		VU	0.66	0.93	0.63	0.84
	8	SC	0.90	0.57	**0.74**	0.47
		VU	0.66	0.93	**0.65**	0.86
C4.5	6	SC	**0.68**	0.91	0.95	0.49
		VU	**0.84**	**0.51**	0.69	0.98
	7	SC	0.60	0.91	*0.33*	**0.56**
		VU	0.76	0.32	*0.05*	*0.02*
	8	SC	**0.67**	0.91	*0.35*	**0.58**
		VU	**0.83**	**0.49**	*0.05*	*0.02*

Table 5.7 – Résultats obtenus pour les expériences 6 et 7 : critères structurels et lexicaux + critères modaux de Charaudeau ou Irrealis

Pour les documents en japonais, SVMlight donne des résultats identiques, sauf pour l'expérience 8 dans laquelle la précision passe de 71 % et 63 % à 74 % et 65 % (ce qui représente un document de plus correctement classé). Cette très faible amélioration nous amène à penser que les critères modaux sont neutres avec ce système. L'utilisation des critères modaux de Givon dégrade très fortement les résultats, qu'ils soient seuls (expérience 7) ou combinés avec les critères de Charaudeau (expérience 8). Ces derniers permettent en revanche d'obtenir des résultats identiques à l'expérience 5.

Les critères de la théorie de Charaudeau permettent donc d'améliorer les résultats pour le français et d'obtenir des résultats stables pour les documents en japonais. Par contre, les critères de Givon font baisser considérablement les résultats pour le japonais, qu'ils soient seuls ou combinés avec les critères de Charaudeau. La combinaison de l'expérience 6 : critères structurels, lexicaux et modaux de Charaudeau semble être la meilleure pour les deux langues et les deux classifieurs (aucune baisse des résultats et amélioration de certains).

5.5 Étude des classifieurs, pertinence des critères

Dans cette section nous étudions la pertinence des critères de la typologie. Nous cherchons à savoir quels sont les critères les plus discriminants, pour quel type de discours et quelle langue.

5.5.1 Arbres de décision

La figure 5.2 représente un arbre de décision obtenu en apprenant le classifieur sur la partie française du corpus [DIAB_CP] et la figure 5.3 un arbre obtenu sur la partie japonaise du corpus. Les valeurs notées sur les arbres correspondent à des quantités normalisées, ce qui explique qu'elles paraissent faibles.

Quelques branches de l'arbre de décision pour la partie française se distinguent. Tout d'abord, la racine de l'arbre correspond au nombre de phrases du document. Si le nombre de phrases normalisé est supérieur à 48, l'arbre classe correctement 40 % des documents scientifiques. Pour un nombre de

phrases normalisé inférieur à 48, si le document contient une bibliographie et des connecteurs logiques, l'arbre classe environ 15 % des documents scientifiques correctement. Sans bibliographie, les documents n'ayant aucun patron d'URL connu, ayant des balises paragraphe et aucune marque de la modalité de la déclaration représentent 79 % des documents vulgarisés. Enfin, si les documents décrits ci-dessus ne font pas appel aux balises paragraphe et ne comptent aucun marqueur de la modalité d'obligation, l'arbre détecte un peu moins de 20 % des documents scientifiques.

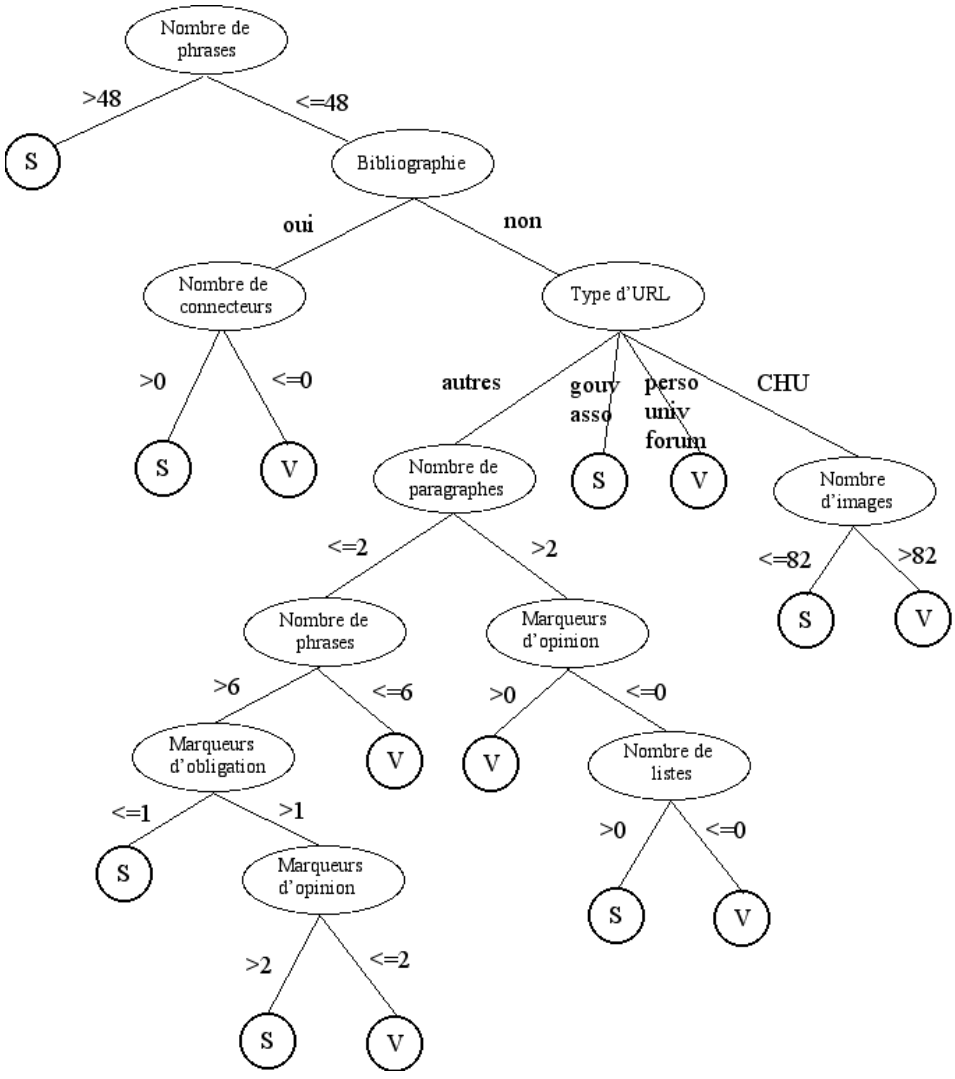


Figure 5.2 – Arbre obtenu avec la dernière typologie choisie pour la langue française

Dans l'arbre japonais, nous notons de la même façon quelques branches nous paraissant les plus efficaces. Un document contenant un nombre normalisé de parenthèses inférieur à 201 et un nombre de balises paragraphe <P> inférieur à 0.044, le classifieur classe correctement 60 % des documents vulgarisés. Si le nombre normalisé de parenthèses est supérieur à 201 et le nombre de caractères inférieur à 7206, l'arbre classe correctement 66 % des documents vulgarisés. Pour 13 % des documents vulgarisés, le nombre de parenthèses normalisé est compris entre 201 et 212, le nombre de caractères supérieur à

7206 et les marqueurs d'interrogation et de volonté sont inférieurs à 0.089 et 0.004. 34 % des documents vulgarisés ont les caractéristiques suivantes : plus de 201 parenthèses, un nombre de caractères supérieur à 7206, plus de 0.089 marqueurs de l'interrogation et pas de bibliographie.

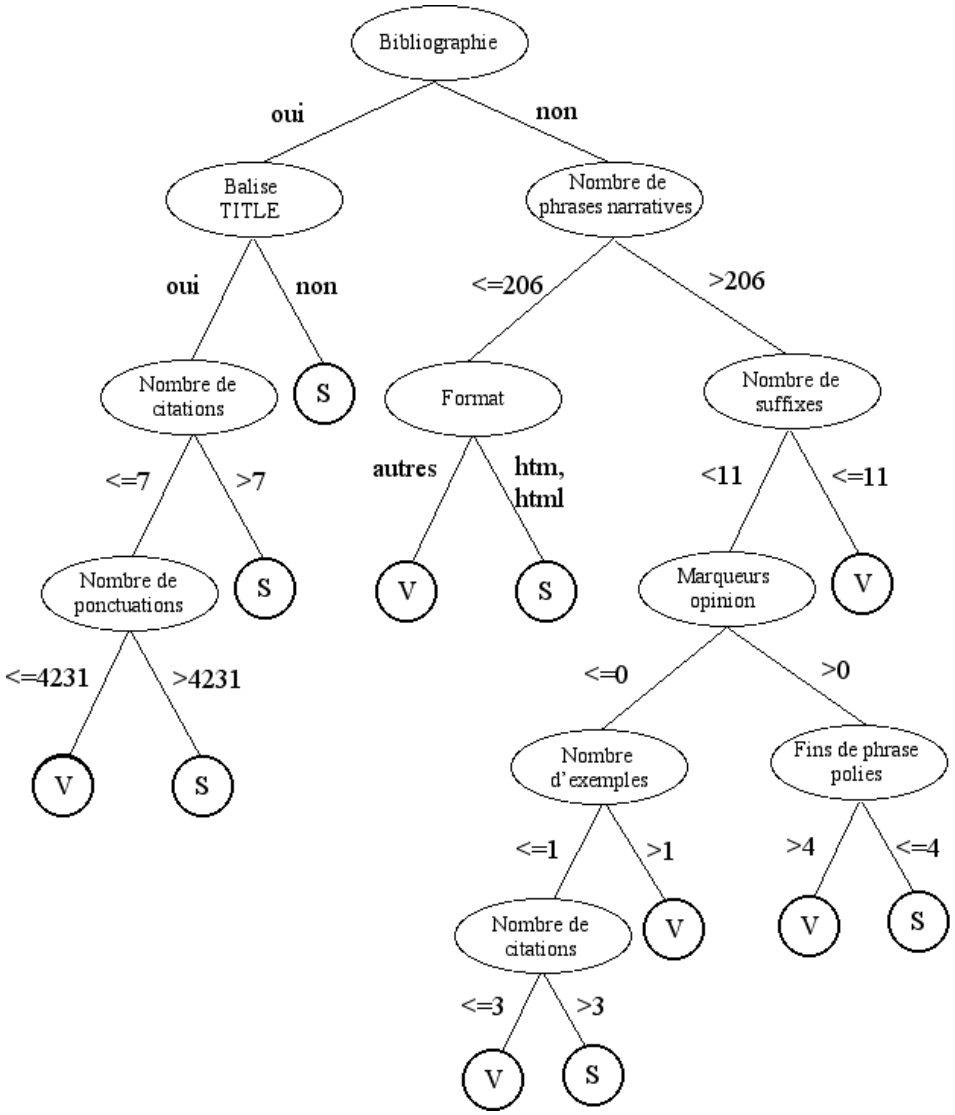
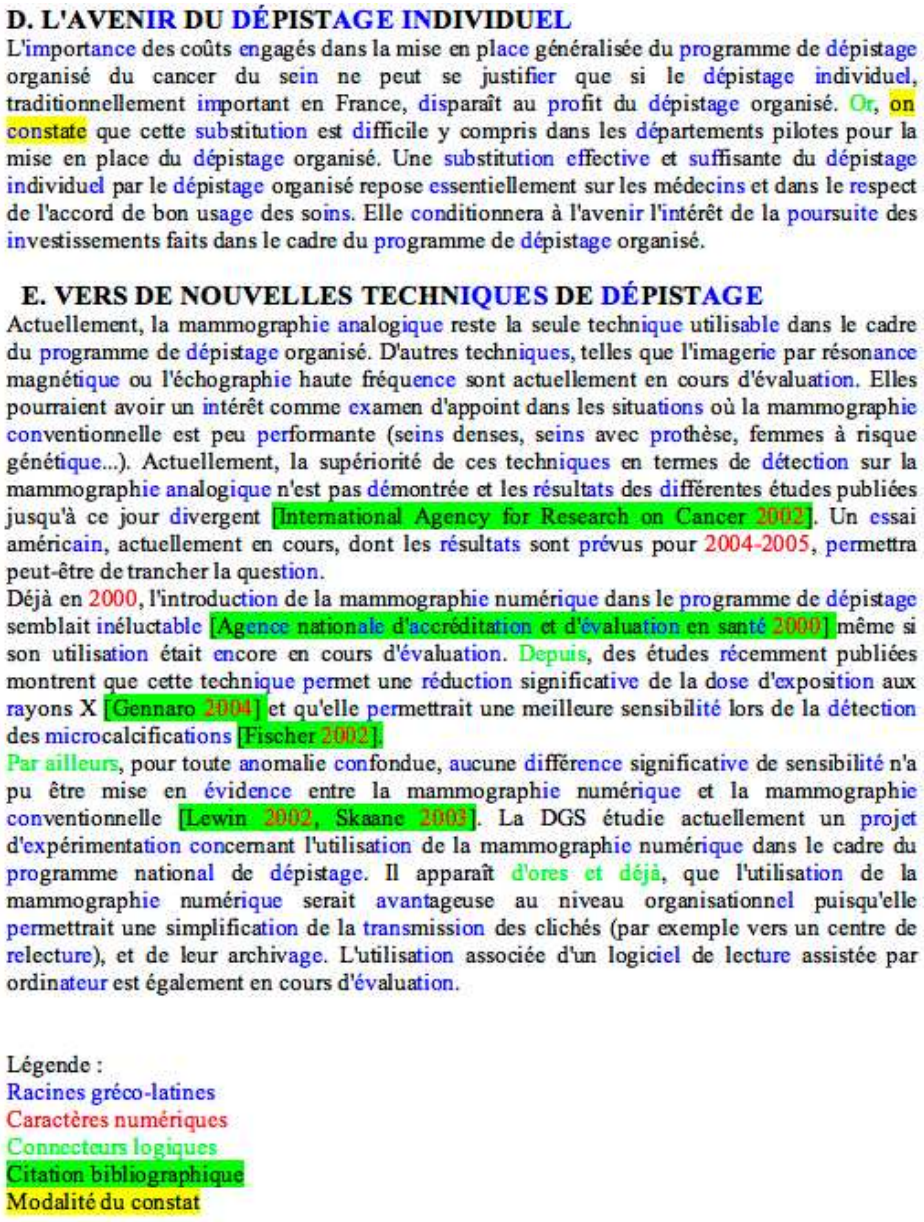


Figure 5.3 – Arbre obtenu avec la dernière typologie choisie pour la langue japonaise

5.5.2 Illustration sur quelques documents du corpus

Nous souhaitons dans cette section illustrer la présence de critères dans deux documents extraits de notre corpus [BC_CP]. Nous avons pour cela choisi d'extraire quelques paragraphes des deux documents obtenant avec SVM les scores minimaux et maximaux. Sur la figure 5.4 se trouve un extrait de texte scientifique sur lequel nous avons surligné et coloré certains critères lexicaux et modaux. Ce document est un rapport au gouvernement sur l'état des campagnes et des technologies de dépistage du cancer du sein en

France. Ce document contient donc, comme nous pouvons le constater sur la figure, de nombreuses références bibliographiques. Les seuls caractères numériques présents dans cet extrait sont des dates, mais d'autres passages dans ce document contiennent différentes quantités numériques. Le discours s'articule autour de quelques connecteurs logiques. Nous notons de plus l'emploi de nombreux termes utilisant des racines gréco-latines. Enfin, nous trouvons un marqueur de la modalité élocutive du constat, ce qui paraît logique pour un document scientifique : l'auteur ne s'adresse pas directement au lecteur mais s'exprime régulièrement à la première personne du singulier ou du pluriel.



D. L'AVENIR DU DÉPISTAGE INDIVIDUEL

L'importance des coûts engagés dans la mise en place généralisée du programme de dépistage organisé du cancer du sein ne peut se justifier que si le dépistage individuel, traditionnellement important en France, disparaît au profit du dépistage organisé. Or, on constate que cette substitution est difficile y compris dans les départements pilotes pour la mise en place du dépistage organisé. Une substitution effective et suffisante du dépistage individuel par le dépistage organisé repose essentiellement sur les médecins et dans le respect de l'accord de bon usage des soins. Elle conditionnera à l'avenir l'intérêt de la poursuite des investissements faits dans le cadre du programme de dépistage organisé.

E. VERS DE NOUVELLES TECHNIQUES DE DÉPISTAGE

Actuellement, la mammographie analogique reste la seule technique utilisable dans le cadre du programme de dépistage organisé. D'autres techniques, telles que l'imagerie par résonance magnétique ou l'échographie haute fréquence sont actuellement en cours d'évaluation. Elles pourraient avoir un intérêt comme examen d'appoint dans les situations où la mammographie conventionnelle est peu performante (seins denses, seins avec prothèse, femmes à risque génétique...). Actuellement, la supériorité de ces techniques en termes de détection sur la mammographie analogique n'est pas démontrée et les résultats des différentes études publiées jusqu'à ce jour divergent [International Agency for Research on Cancer 2002]. Un essai américain, actuellement en cours, dont les résultats sont prévus pour 2004-2005, permettra peut-être de trancher la question.

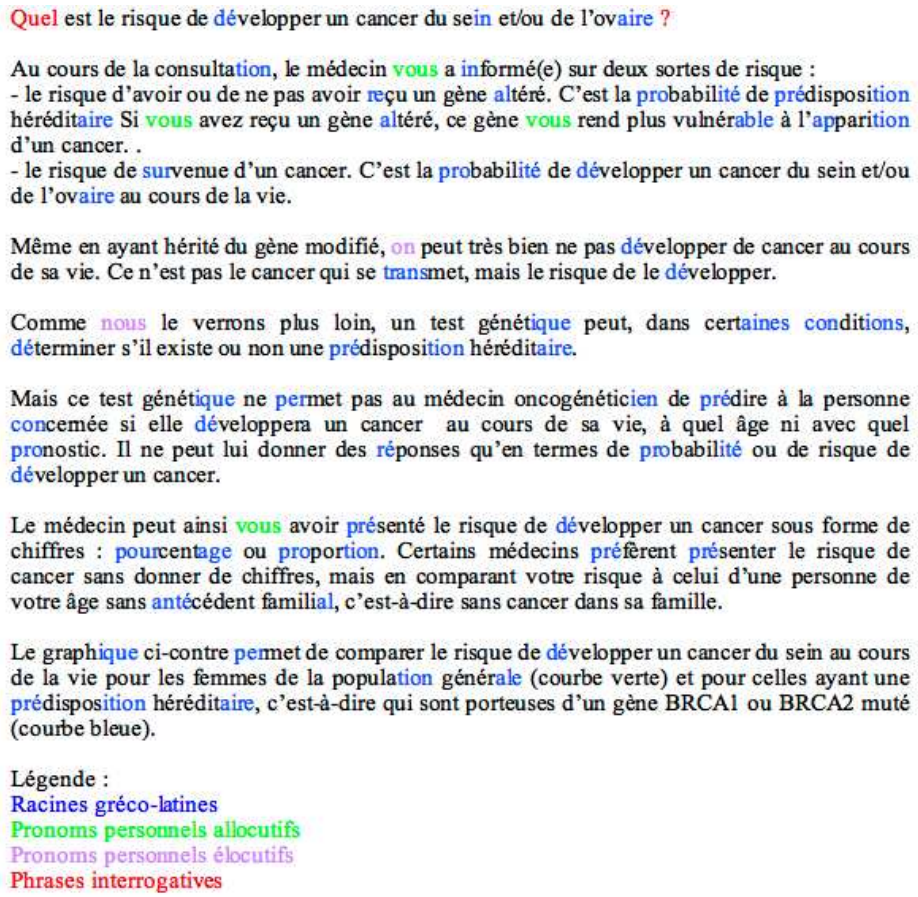
Déjà en 2000, l'introduction de la mammographie numérique dans le programme de dépistage semblait inéluctable [Agence nationale d'accréditation et d'évaluation en santé 2000] même si son utilisation était encore en cours d'évaluation. Depuis, des études récemment publiées montrent que cette technique permet une réduction significative de la dose d'exposition aux rayons X [Gennaro 2004] et qu'elle permettrait une meilleure sensibilité lors de la détection des microcalcifications [Fischer 2002].

Par ailleurs, pour toute anomalie confondue, aucune différence significative de sensibilité n'a pu être mise en évidence entre la mammographie numérique et la mammographie conventionnelle [Lewin 2002, Skaane 2003]. La DGS étudie actuellement un projet d'expérimentation concernant l'utilisation de la mammographie numérique dans le cadre du programme national de dépistage. Il apparaît d'ores et déjà, que l'utilisation de la mammographie numérique serait avantageuse au niveau organisationnel puisqu'elle permettrait une simplification de la transmission des clichés (par exemple vers un centre de relecture), et de leur archivage. L'utilisation associée d'un logiciel de lecture assistée par ordinateur est également en cours d'évaluation.

Légende :
Racines gréco-latines
Caractères numériques
Connecteurs logiques
Citation bibliographique
Modalité du constat

Figure 5.4 – Exemples de critères pertinents étiquetés sur un extrait de fichier scientifique du corpus

Le deuxième document présenté sur la figure 5.5 est extrait d'une brochure d'information destinée aux patientes faisant un dépistage du cancer du sein. Nous observons tout d'abord plus de pronoms personnels dans ce documents dans le premier : des pronoms personnels allocutifs (« vous ») et des élocutifs (« nous », « on »). Ici, le locuteur s'adresse directement au lecteur. Nous notons un peu moins de racines gréco-latines dans cet extrait. Cette brochure se présente sous la forme de questions-réponses, nous notons donc des phrases interrogatives, comportant des pronoms interrogatifs ainsi que des symboles de ponctuation interrogative.



Quel est le risque de développer un cancer du sein et/ou de l'ovaire ?

Au cours de la consultation, le médecin vous a informé(e) sur deux sortes de risque :
- le risque d'avoir ou de ne pas avoir reçu un gène altéré. C'est la probabilité de prédisposition héréditaire Si vous avez reçu un gène altéré, ce gène vous rend plus vulnérable à l'apparition d'un cancer. .
- le risque de survenue d'un cancer. C'est la probabilité de développer un cancer du sein et/ou de l'ovaire au cours de la vie.

Même en ayant hérité du gène modifié, on peut très bien ne pas développer de cancer au cours de sa vie. Ce n'est pas le cancer qui se transmet, mais le risque de le développer.

Comme nous le verrons plus loin, un test génétique peut, dans certaines conditions, déterminer s'il existe ou non une prédisposition héréditaire.

Mais ce test génétique ne permet pas au médecin oncogénéticien de prédire à la personne concernée si elle développera un cancer au cours de sa vie, à quel âge ni avec quel pronostic. Il ne peut lui donner des réponses qu'en termes de probabilité ou de risque de développer un cancer.

Le médecin peut ainsi vous avoir présenté le risque de développer un cancer sous forme de chiffres : pourcentage ou proportion. Certains médecins préfèrent présenter le risque de cancer sans donner de chiffres, mais en comparant votre risque à celui d'une personne de votre âge sans antécédent familial, c'est-à-dire sans cancer dans sa famille.

Le graphique ci-contre permet de comparer le risque de développer un cancer du sein au cours de la vie pour les femmes de la population générale (courbe verte) et pour celles ayant une prédisposition héréditaire, c'est-à-dire qui sont porteuses d'un gène BRCA1 ou BRCA2 muté (courbe bleue).

Légende :
Racines gréco-latines
Pronoms personnels allocutifs
Pronoms personnels élocutifs
Phrases interrogatives

Figure 5.5 – Exemples de critères pertinents étiquetés sur un extrait de fichier vulgarisé du corpus

5.5.3 Analyse de l'évolution de quelques critères

Nous étudions dans cette section l'évolution de quelques critères discriminants selon le score attribué à chaque document par SVM. Pour cela, nous avons observé les valeurs des différents critères sur l'ensemble du corpus et n'avons conservé que quelques uns des critères présentant une évolution que nous jugions intéressante. Les valeurs de chacun des critères présentés dans les sections suivantes sont les valeurs normalisées. Chaque graphique figurant dans les sections suivantes représente l'évolution des valeurs d'un critère selon le score SVM attribué aux documents correspondants. Ces schémas permettent de se rendre compte de la difficulté de la tâche de classification. Rares sont les critères permettant de

trancher. Généralement, leur évolution est assez floue, ou n'est discriminante que sur un sous-ensemble d'une classe.

5.5.3.1 Critères français

Le graphique 5.6 représente la proportion de balises `IMG` et `P` dans le corpus, permettant d'insérer dans un document `HTML` des images et des paragraphes (respectivement). Notons que les deux courbes évoluent de façon sensiblement équivalente selon le score des documents. Les documents classés comme vulgarisés (au score négatif) semblent contenir plus de ces balises que les documents classés comme scientifiques. Cela peut s'expliquer par la proportion de fichiers `PDF`, plus importante dans la partie scientifique du corpus que dans la partie vulgarisée.

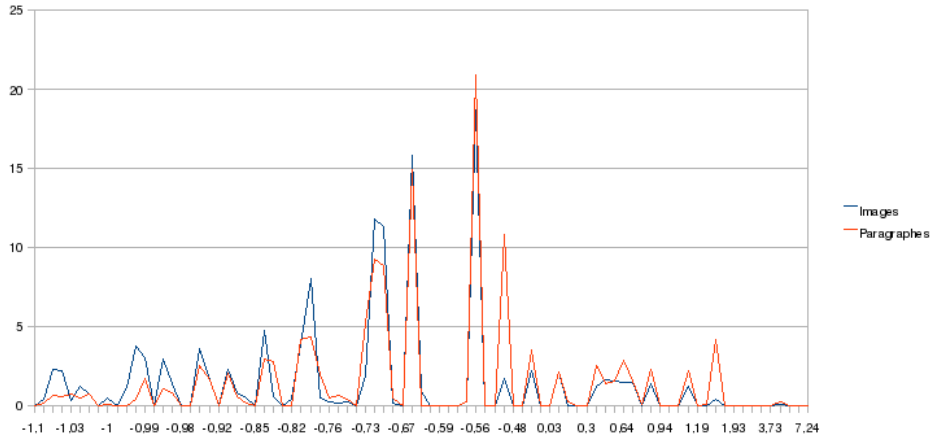


Figure 5.6 – Proportion de balises `IMG` et `P` en fonction du score des documents

Le graphique 5.7 représente la proportion de pronoms personnels allocutifs et élocutifs dans le corpus. Les documents utilisant des pronoms obtiennent en majorité un score négatif. Les pronoms élocutifs semblent être moins utilisés que les pronoms allocutifs (sauf pour un document). Quelle que soit le type de pronom personnel, si un document classé comme scientifique en utilise, leur quantité sera en moyenne inférieure à celle des documents classés comme vulgarisés.

Le graphique 5.8 présente la proportion de racines gréco-latines dans les documents en fonction de leur score `SVM`. Cette courbe peut se décomposer en trois parties. Les documents dont le score est inférieur à 0.7 environ contiennent globalement moins de racines gréco-latines. Entre 0.7 et 0 se trouvent les documents contenant le plus de racines, avec des pics à 200 et 300 (valeurs normalisées). Au dessus de 0, les valeurs sont constantes et légèrement supérieures à la première partie de la courbe. Il paraît logique que les documents les plus scientifiques contiennent plus de racines gréco-latines que les autres. Cette courbe ne contredit pas forcément cette hypothèse, elle signifie juste que les document les plus scientifiques *selon SVM* ne correspondent pas aux valeurs maximales.

Le graphique 5.9 représente l'évolution de la proportion de caractères numériques dans les documents du corpus. Nous constatons tout d'abord que la quantité de caractères numériques est très faible pour les documents obtenant les scores minimaux. Une forte hausse du nombre de caractères numériques apparaît entre -0.5 et 0. Cette brusque augmentation correspond en réalité à deux documents

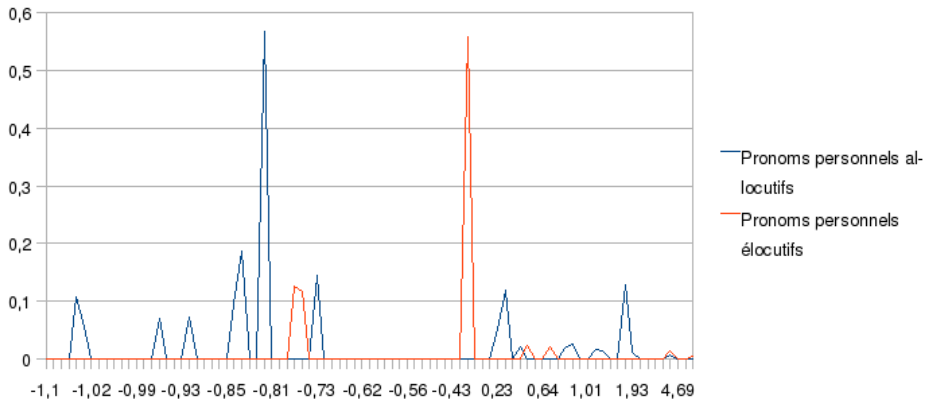


Figure 5.7 – Proportion de pronoms en fonction du score des documents

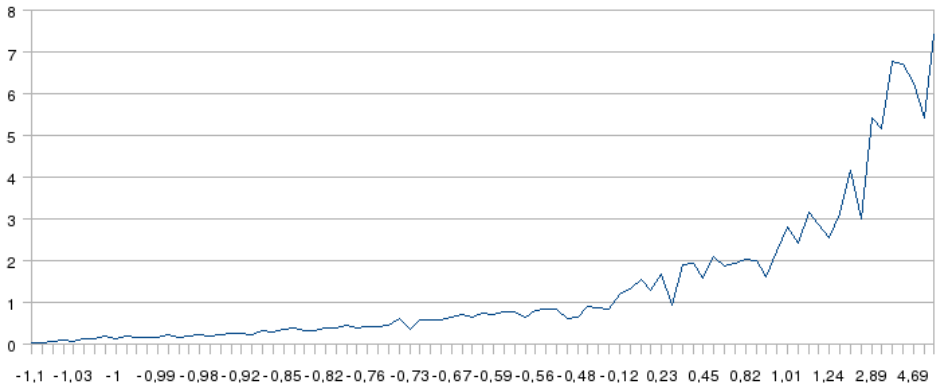


Figure 5.8 – Proportion de racines gréco-latines en fonction du score des documents

classés comme vulgarisés pour lesquels la proportion est largement supérieure. Ces documents peuvent simplement contenir des tableaux avec beaucoup de quantités numériques par exemple.

Le graphique 5.10 montre l'évolution de la présence de marqueurs des modalités d'opinion, de déclaration et d'obligation dans les documents du corpus. Nous notons tout d'abord que la quantité de marqueurs de ces modalités est assez faible en général (sauf pour quelques documents dont les scores se situent entre -0.6 et 0). Globalement, la quantité de marqueurs de l'obligation semble être légèrement supérieure aux autres marqueurs, nous remarquons toutefois que l'évolution des trois modalités semble corrélée. De plus, les marqueurs de ces modalités sont liés à l'utilisation des pronoms personnels allocutifs ou élocutifs, leurs évolutions sont corrélées entre elles, ainsi qu'avec la courbe des pronoms (graphique 5.7).

5.5.3.2 Critères japonais

Le graphique 5.11 présente la proportion de citations bibliographiques dans les documents en fonction de leur score SVM. Bien que la proportion soit globalement très faible, nous remarquons que les seuls

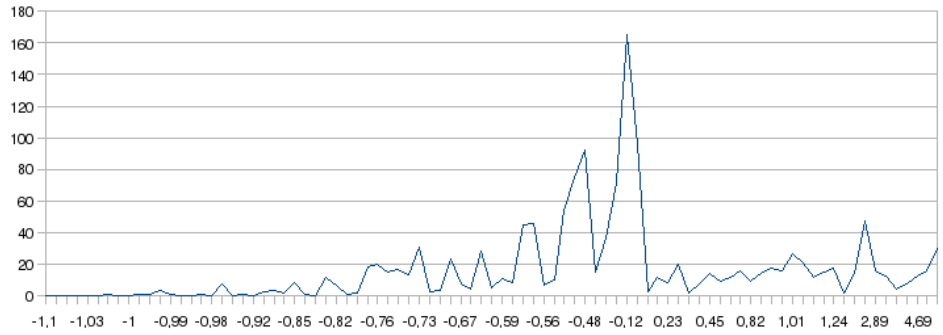


Figure 5.9 – Proportion de caractères numériques en fonction du score des documents

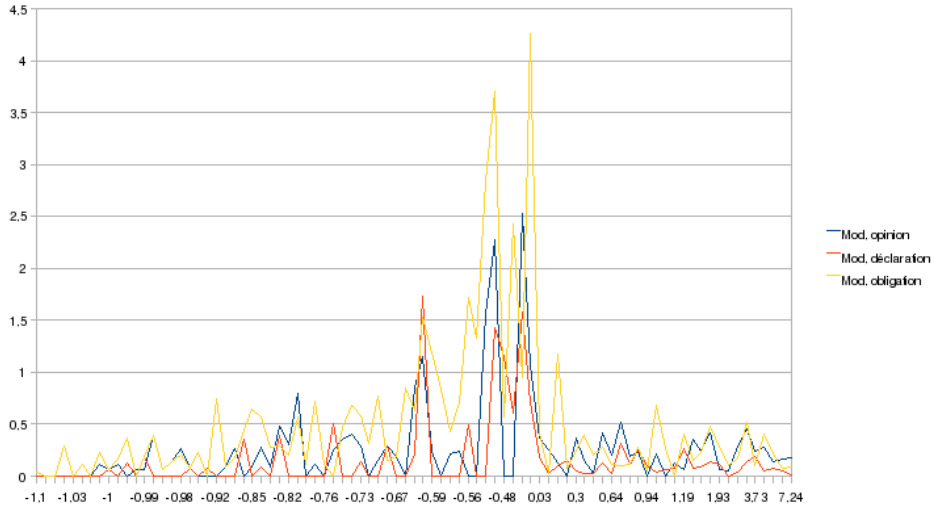


Figure 5.10 – Proportion de marqueurs des modalités d'opinion, de déclaration et d'obligation en fonction du score des documents

documents dans lesquels des citations bibliographiques ont été détectées sont classés comme scientifiques, ce qui paraît assez logique. Ce critère paraît donc fortement discriminant pour déterminer un sous-ensemble des documents scientifiques.

Le graphique 5.12 illustre l'évolution du nombre normalisé de parenthèses en fonction du score des documents. Cette courbe est très irrégulière, mais nous observons globalement une baisse de l'intervalle (`valeur minimale`, `valeur maximale`). De plus, les documents ayant un score supérieur à 0.6 ont la proportion de parenthèses la plus basse.

À l'inverse, nous remarquons sur le graphique 5.13 que l'utilisation de particules neutres ou polies en fin de phrase semble être caractéristique des documents vulgarisés. En effet, seuls quelques documents scientifiques utilisent des fins de phrases neutres. Les fins de phrases polies semblent être légèrement plus fréquentes en moyenne pour les documents classés comme vulgarisés et elles ne sont jamais utilisés dans les documents scientifiques.

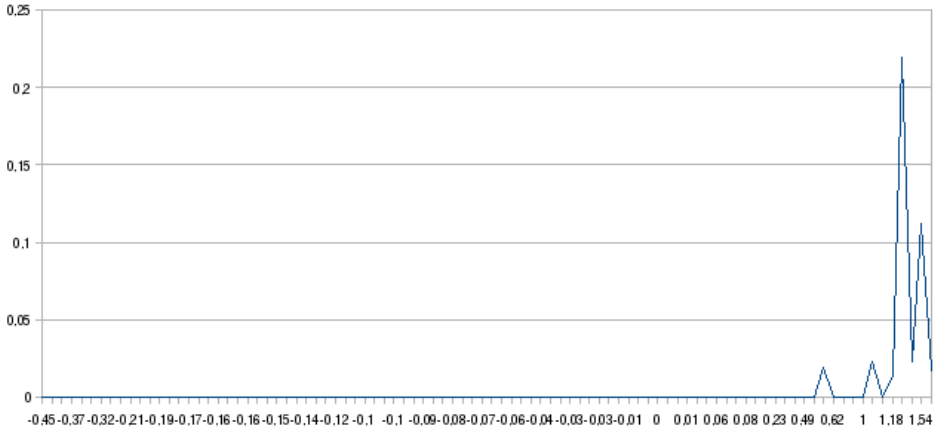


Figure 5.11 – Proportion de citations bibliographiques en fonction du score des documents

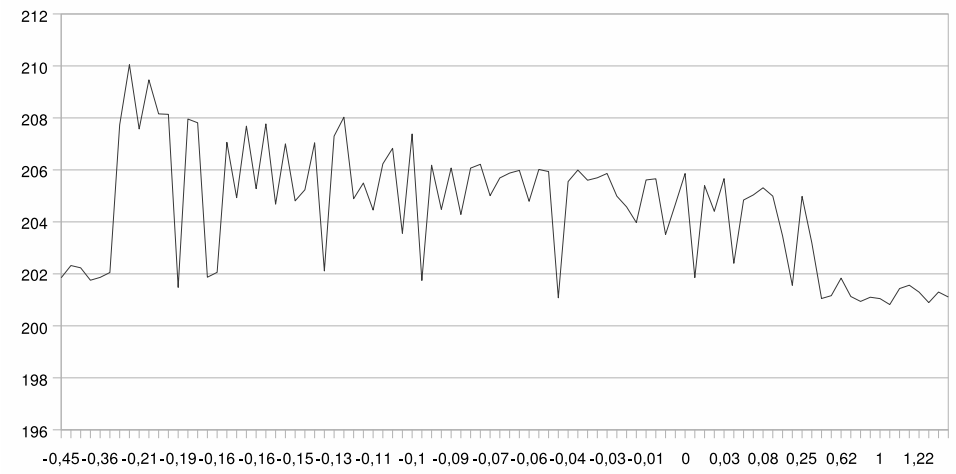


Figure 5.12 – Proportion de parenthèses en fonction du score des documents

La courbe 5.14 présente la proportion de phrases narratives, interrogatives et exclamatives dans le corpus. Nous constatons tout d'abord que les phrases narratives sont beaucoup plus fréquentes que les phrases interrogatives. Les phrases exclamatives quant à elles sont quasiment inexistantes. Néanmoins les quelques occurrences de phrases exclamatives (peu visibles sur le graphique) sont, à quelques exceptions près, dans des documents classés comme vulgarisés. Les courbes des deux autres types de phrases sont assez peu discriminantes. Bien qu'ayant chacune une légère tendance à croître ou décroître, les écarts entre deux documents ayant quasiment le même score sont trop élevés pour pouvoir conclure.

La figure 5.15 présente la proportion de marqueurs identifiant les pronoms dans les documents du corpus. Ces marqueurs sont peu fréquents dans les deux cas bien que quelques documents utilisent plus de pronoms allocutifs. Les documents classés comme scientifiques utilisent peu de pronoms élocutifs, seuls ceux obtenant un score proche de zéro en utilisent.

Sur la courbe 5.16 est présentée l'utilisation de balises `IMG` et `P` dans les documents japonais. Ces

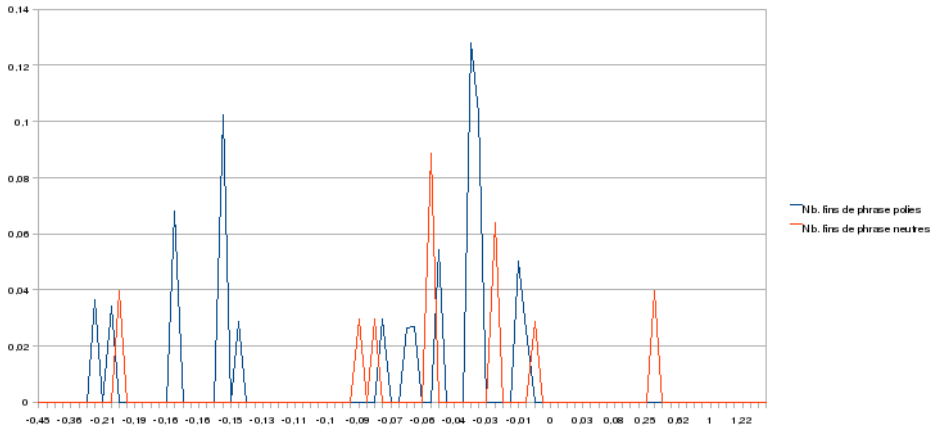


Figure 5.13 – Proportion de fins de phrase polies et neutres en fonction du score des documents

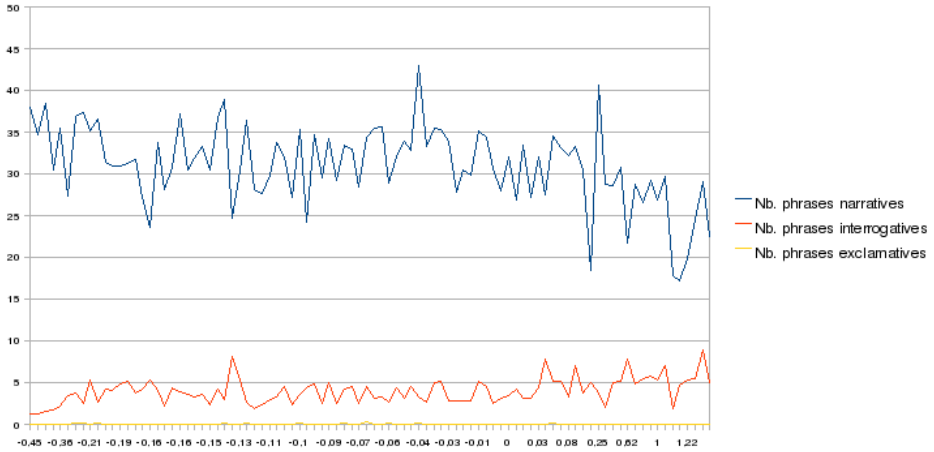


Figure 5.14 – Proportion de phrases narratives, interrogatives et exclamatives en fonction du score des documents

critères n'étant pas présents dans les documents PDF, les courbes ont des formes assez crantées. La proportion de ces balises semble plus faible dans les documents scientifiques, ce qui peut s'expliquer par la proportion de documents PDF dans la partie scientifique du corpus.

5.5.4 Étude du bruit et du silence générés par les critères sur quelques exemples

Dans cette section, nous étudions sur quelques exemples les notions de bruit et de silence présentées dans la section 4.3.1 du chapitre 4. Le bruit apparaît lorsque de mauvais exemples se trouvent dans les occurrences des critères. Le silence apparaît lorsque certains exemples ne figurent pas dans les occurrences. Le bruit peut être introduit pour diverses raisons :

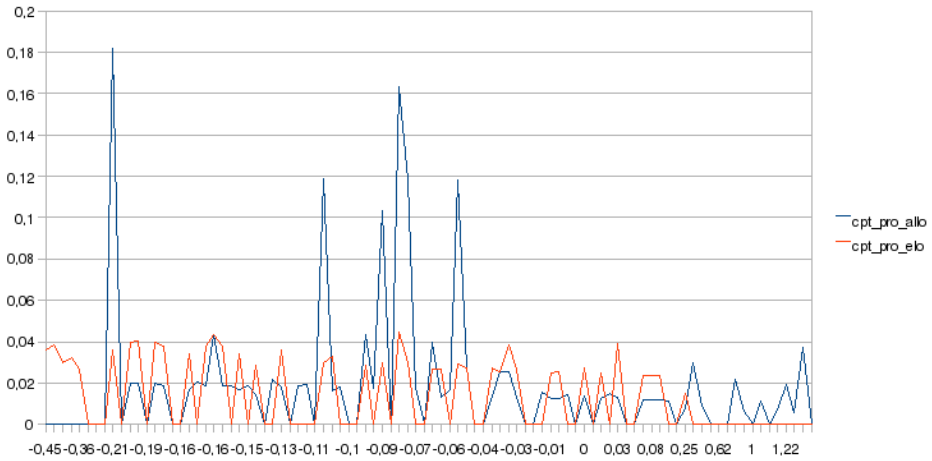


Figure 5.15 – Proportion de pronoms en fonction du score des documents

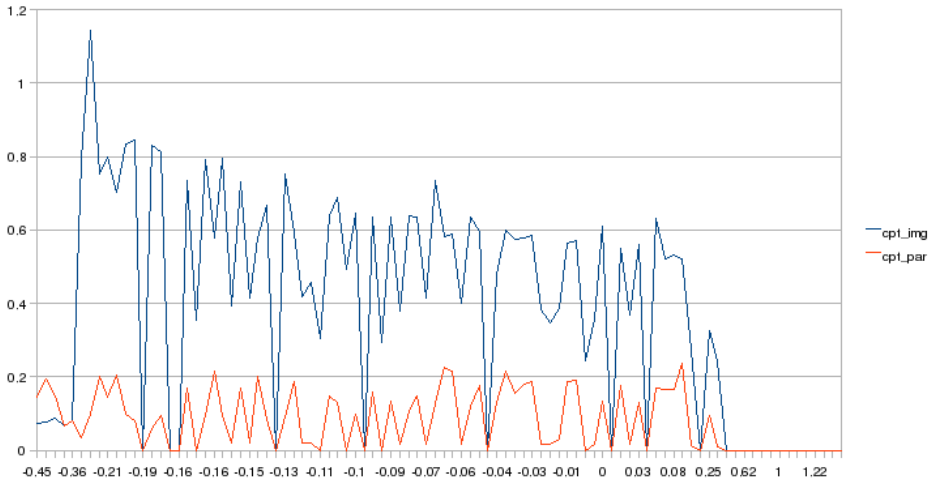


Figure 5.16 – Proportion de balises `IMG` et `P` en fonction du score des documents

- polysémie des marqueurs. Exemple : le verbe *pouvoir* à la 2ème personne doit indiquer la modalité de l'autorisation, cela n'est pas le cas dans la phrase « Pouvez-vous nous parler des problèmes de vision associés au diabète ? » ;
- ambiguïté des étiquettes du logiciel Flemm. Par exemple, quand la conjugaison est la même pour le présent de l'indicatif et pour l'impératif, les deux étiquettes sont présentes (« il est conseillé de bien choisir les graisses que vous consommez » / « Consommez-le en quantité contrôlée et régulière »).

Le silence peut être constaté lorsque la liste des marqueurs pour un critère ne couvre pas toutes les occurrences possibles. C'est le cas lorsque certaines occurrences sont implicites par exemple ou correspondent à des phénomènes très rares que nous n'avons pas recensé. Par exemple, tous les termes spécialisés ne peuvent pas être détectés grâce à notre liste non-exhaustive de racines gréco-latines (œstradiol,

tamoxifène, adénocarcinome...).

Nous avons choisi de minimiser le bruit au détriment du silence. Les occurrences des critères dans les documents sont moins nombreuses et parfois nulles mais nous avons une certaine assurance sur ce que les représentations vectorielles signifient réellement. De plus, cet objectif correspond à notre volonté de nous distinguer des méthodes utilisant beaucoup de critères sans attacher beaucoup d'importance à leur sens en utilisant une typologie linguistiquement motivée.

Notre priorité était de minimiser le bruit dans notre système et nous avons alors tenté d'écarter les marqueurs pouvant caractériser d'autres phénomènes que celui que nous étudions. Nous avons de plus tenté de filtrer les critères en utilisant les informations apportées par des étiquettes morpho-syntaxiques : conjugaisons, pronoms...

5.5.5 Synthèse

Nous avons dans cette section présenté plus en détail les classifieurs générés et la pertinence de nos critères.

La section 5.5.1 nous permet d'avoir un aperçu des arbres générés par `C4.5` avec notre typologie sur notre corpus [DIAB_CP], ainsi qu'une idée des critères les plus discriminants de la typologie. Nous nous rendons tout d'abord compte que les critères discriminants relèvent des trois catégories de la typologie. De plus, nous notons que les critères discriminants ne sont pas les mêmes selon la langue. En français, ce sont le nombre de phrases, la bibliographie, le nombre de paragraphes, les marqueurs de la modalité de l'opinion, tandis qu'en japonais ce sont la balise `<TITLE>`, la bibliographie, le nombre de suffixes, le nombre de phrases.

La section 5.5.2 nous permet de constater sur quelques documents obtenant les plus forts scores positifs ou négatifs avec `SVM` la similitude entre les critères empiriques ayant permis la classification manuelle et les critères discriminants de la typologie. En effet, en surlignant quelques critères modaux et lexicaux il est possible de voir que la classification manuelle, sans forcément formaliser les critères de classification, se base sur certains des critères discriminants de notre typologie.

La section 5.5.3 montre toutefois qu'une poignée de critères est insuffisante pour classer une telle quantité de documents. Même si certains d'entre eux nous paraissent, *a priori*, très discriminants, les courbes de leur évolution sont généralement irrégulières. Bien que nos corpus soient représentatifs et donc homogènes du point de vue thématique, ils sont hétérogènes du point de vue lexical ou modal et le sont encore plus du point de vue d'un seul critère.

5.6 Comparaison de notre méthode avec la méthode classique des vecteurs de termes

Nous présentions dans le chapitre précédent une approche très courante en classification automatique textuelle se basant uniquement sur les termes apparaissant dans les documents. Nous souhaitons tester cette méthode afin de comparer ses résultats aux notres. Cette méthode consiste à sélectionner un ensemble de termes représentatifs du corpus d'apprentissage. Une liste de termes les plus fréquents est sélectionnée, en omettant les mots outils, trop fréquents et peu représentatifs. Il existe plusieurs approches pour sélectionner les « termes », ils peuvent être des mots, des lemmes, des bigrammes... Nous avons choisi ici, par simplicité, de ne collecter que les mots les plus fréquents, hors mots outils, de notre corpus d'apprentissage [DIAB_CP]. Nous ne testons cette méthode que sur les corpus français. La liste

de ces mots est disponible en annexe B. La représentation vectorielle des documents est donc composée de la fréquence de chaque terme, pondérée par le nombre de caractères des documents. Deux classifieurs ont été créés à partir du corpus [DIAB_CP] à l'aide des systèmes `SVMlight` et `C4.5`. Nous avons ensuite évalué ces deux classifieurs sur les corpus [DIAB_CP] et [BC_CP] à l'aide des métriques du rappel et de la précision. Les résultats de ces évaluations sont présentés dans le tableau 5.8.

Corpus	Système	Méthode vecteurs de termes		Notre méthode	
		Préc.	Rapp.	Préc.	Rapp.
[DIAB_CP]	`SVMlight`	0.76	0.79	0.83	0.80
	`C4.5`	0.89	0.93	0.95	0.97
[BC_CP]	`SVMlight`	0.61	0.60	0.78	0.75
	`C4.5`	0.68	0.65	0.76	0.71

Table 5.8 – Résultats obtenus grâce à notre méthode et à celle par vecteurs de terme

Cette méthode permet d'obtenir des résultats satisfaisants sur le corpus [DIAB_CP], avec une précision et un rappel moyens de 89 % et 93 % avec `C4.5` et 76 % et 79 % avec `SVMlight`. Les termes représentant les documents étant extraits de ce corpus, ce résultat semble normal. Ces résultats sont très proches de ceux que nous obtenons avec notre méthode. Nous observons une baisse sensible de ces résultats lorsque les classifieurs sont appliqués au corpus [BC_CP] : un rappel de 65 % et 60 % et une précision de 68 % et 61 % respectivement pour les systèmes `C4.5` et `SVMlight`. Bien que ces résultats soient corrects, ils nous amènent à nous interroger sur la pertinence de cette méthode dans notre cas. Nous souhaitons que nos classifieurs soient capables de reconnaître les documents scientifiques des documents vulgarisés quel que soit le thème de ces documents. Si les résultats baissent de 15 points en passant d'une thématique à une autre dans le domaine médical, quelle baisse observerait-on sur un document portant sur la physique nucléaire ? Le vocabulaire peut être commun à un domaine, mais il ne l'est pas pour toute la communauté scientifique. Cette méthode paraît donc trop limitée pour notre cas. De plus, elle se base sur une représentation du texte se limitant aux mots qui le composent. Nous souhaitons dans ce travail ancrer les marqueurs des types de discours à différents niveaux d'analyse, se basant sur le texte lui-même. Cette approche semble plus robuste et permet de caractériser le type de discours indépendamment du domaine sur lequel portent les textes.

5.7 Discussion sur la catégorisation en type de discours scientifique et vulgarisé

Nous nous interrogeons dans cette section sur la légitimité de notre classification binaire en type de discours scientifique et vulgarisé. En effet, ce problème se pose dès la construction des corpus. Pour beaucoup de documents, la classification est claire. Les différents indices glanés au cours de la collecte et l'observation d'un document permettent souvent d'en déterminer le type de discours. Mais pour une autre partie des documents, le type de discours est difficile voire impossible à déterminer et les avis diffèrent selon les personnes. Existe-t-il alors 3 classes ? Une scientifique, une vulgarisée et une troisième composée de documents mi-scientifiques, mi-vulgarisés ? La distinction entre ces deux types de discours paraît plus complexe.

Tout d'abord, en observant ces deux classes, nous remarquons qu'elles sont très hétérogènes. Des rapports techniques côtoient des articles scientifiques et des cours universitaires dans la classe scientifique, tandis que des articles de revues grand public, des recommandations aux patients côtoient des conversations sur forums dans la classe vulgarisée. Ces différents genres de documents ne partagent pas les mêmes objectifs, ni le même public. Comme le remarquent à juste titre Bowker et Pearson (2002, p.28), « *because LSP users have different levels of expertise, there are also different levels of LSP communication* ». Elles distinguent ainsi les experts (chercheurs, praticiens, ...), les semi-experts (experts d'un domaine lié, étudiants...) et les non-experts (les autres). Plutôt que de compter 2 catégories comme nous le faisons, nous pouvons distinguer avec ces 3 utilisateurs : la communication d'expert à expert, d'expert à semi-expert, d'expert à non-expert, mais aussi la communication de semi-expert à expert, de semi-expert à semi-expert, de semi-expert à non-expert et enfin de non-expert à non-expert. Certaines de ces catégories sont obsolètes, ou ne sont pas forcément considérées comme spécialisées (notamment la communication de non-expert à non-expert). Cette distinction ne permet cependant pas de résoudre les ambiguïtés : les semi-experts peuvent être des étudiants de Licence 1 à Master 2, ce qui n'implique pas le même niveau de spécialisation et de langue. Nous pourrions alors distinguer au sein de ces utilisateurs plusieurs catégories selon le niveau de connaissance par exemple. Ainsi, nous pensons qu'il existe un continuum entre le type de discours scientifique et le type de discours vulgarisé. Ce continuum peut être vu comme un degré de spécialisation ou à l'inverse, de vulgarisation. Il est souvent nécessaire de catégoriser ces deux types de discours, d'en définir ses limites et cette tâche est complexe parce que ces limites sont floues.

Le système de classification `SVMlight` attribue un « score » à chaque document classé. Ce « score » peut être vu comme un degré de spécialisation. En effet, en observant les résultats de classification et les scores attribués à chaque document, nous remarquons, à quelques exceptions près, que la complexité des documents augmente avec le score des documents. Ainsi, les scores les plus élevés sont attribués à des rapports d'évaluation et des articles scientifiques. Ces textes ont pour propriété des structures propres, un vocabulaire très spécialisé, de nombreuses données numériques... À l'inverse, les documents obtenant les plus petits scores sont des brochures d'information pour les patients, des articles de presse grand public... Ces documents sont généralement courts, contiennent des couleurs et des images... Les scores du milieu sont attribués à des cours, des articles de presse spécialisée... Ces documents partagent des caractéristiques des deux catégories : un vocabulaire spécialisé mais beaucoup de marqueurs de glose, des images et des couleurs par exemple.

Afin de vérifier nos hypothèses, nous avons collecté quelques documents que nous jugions « ambigus »: un article tiré de la revue « La recherche »[1], l'article de Wikipédia consacré au cancer du sein[2], un cours d'université sur le cancer du sein[3] et un cours sur le TALN[4].

Les scores que nous attribuons aux documents pour l'apprentissage sont $+1$ pour les documents scientifiques et -1 pour les documents vulgarisés. Les scores attribués par le classifieur varient entre ces deux bornes, mais peuvent les dépasser. L'article de la revue « La recherche » porte sur le thème de l'hydrologie, plus particulièrement sur l'exploitation des ressources du Rhône. Cet article est technique, illustré et assez long. Il obtient un score de -0.64. L'article Wikipedia sur le cancer du sein est long, comporte un grand nombre de termes spécialisés mais aussi de nombreux marqueurs de glose, aucun pronom, quelques images et des références bibliographiques. Il est au format `HTML`. Il obtient un score de 0.23. Le cours sur le cancer du sein est au format `HTML`. Il n'est pas très long, contient beaucoup de

[1] `www.larecherche.fr/html/2008/pdf/399_Rhone.pdf`
[2] `http://fr.wikipedia.org/wiki/Cancer_du_sein`
[3] `http://www.uvp5.univ-paris5.fr/CAMPUS-GYNECO-OBST/cycle3/poly/20000faq1.asp`
[4] `http://sites.univ-provence.fr/veronis/cours/INFZ18/veronis-INFZ18.pdf`

listes d'items, de tableaux, pas d'images. Il contient des questions et peu de connecteurs logiques. Le classifieur lui attribue un score de -0.15. Enfin, le cours sur le TALN est au format PDF, il est destiné à des débutants donc contient peu de vocabulaire spécialisé, beaucoup de reformulations, d'illustrations. En revanche il contient de nombreuses références bibliographiques et est assez long. Il obtient un score de 0.2.

Aucun de ces documents ne dépasse les bornes $[-1, 1]$ et un seul d'entre eux dépasse les bornes $[-0.5, 0.5]$. Ces documents obtiennent un score qui correspond à leur degré de spécialisation, c'est-à-dire moyen pour la majorité. À l'inverse, certains documents de notre corpus, très spécialisés, peuvent obtenir des scores allant jusqu'à 4.

Le travail de classification dans ces deux types de discours ne peut être effectué sans avoir conscience de ce continuum. Il est important de bien définir au préalable à quel degré de spécialisation commencent les catégories scientifique et vulgarisée.

5.8 Conclusion

La première partie de ce chapitre était consacrée à l'évaluation des classifieurs présentés dans le chapitre 4. Dans cette perspective, un corpus comparable d'évaluation [BC_CP] a été créé. Il est composé de documents français et japonais extraits du Web, portant sur le thème du cancer du sein et contient deux classes : documents scientifiques et documents vulgarisés. Les classifieurs appris sur le corpus [DIAB_CP] avec les systèmes C4.5 (arbres de décision) et SVMlight (machines à vecteurs de support) ont alors été testés sur ce même corpus puis sur le corpus d'évaluation. Les résultats obtenus sont globalement satisfaisants, sauf pour les documents japonais avec le classifieur C4.5. Les résultats faibles obtenus dans ce cas nous ont amené à nous poser des questions sur la pertinence des critères de notre typologie. Différents tests ont alors été effectués afin de déterminer quelles catégories de critères étaient les plus pertinentes. Cette étude nous a mené à ôter de notre typologie les critères de la modalité Irrealis, donnant de très faibles résultats pour la langue japonaise. La typologie que nous conservons afin d'apprendre les modèles de classification se compose donc des critères structurels, de la modalité locutive et lexicaux. Nous avons souhaité par la suite étudier plus en détails les classifieurs ainsi générés et la pertinence des critères conservés. Une première étude des arbres de décision obtenus nous permet d'avoir un premier aperçu des critères les plus discriminants dans les deux langues : le nombre de phrases, la bibliographie, certaines balises HTML... pour le français et le nombre de parenthèses, de balises paragraphe, de marqueurs d'interrogation... en japonais. Nous avons ensuite évoqué les notions de bruit et de silence dans nos résultats. En effet, lors de la sélection des marqueurs nous avons choisi de minimiser le bruit dans notre analyse. Nous constatons sur quelques exemples que certains critères apportent toutefois du bruit. Cela est principalement dû à la polysémie de certains marqueurs et à l'ambiguïté de certaines étiquettes lexico-syntaxiques. De plus, notre volonté de minimiser le bruit introduit du silence dans le système : pour que la typologie ne contienne pas de marqueurs trop généraux ou ambigus (représentant d'autres critères de la typologie ou d'autres caractéristiques linguistiques), certaines occurrences ne sont pas détectées dans les textes. De plus, certaines occurrences implicites ne peuvent être détectées automatiquement. Nous sommes partis du principe qu'un nombre d'occurrences restreint était cependant plus pertinent qu'un grand nombre d'occurrences contenant des faux-positifs. Enfin, nous avons observé l'évolution de certains critères selon le score attribué aux documents par SVM. L'évolution des critères pris individuellement ne paraît pas suffisamment discriminante pour distinguer les deux classes et cette étude nous pousse à croire que notre typologie à trois niveaux propose une combinaison de critères efficace. Afin de confirmer cette idée, nous comparons dans la section suivante notre typologie à

la méthode classique de vecteurs de termes. Si cette méthode semble efficace sur le corpus d'apprentissage, elle prouve ses limites sur un corpus à la thématique différente. Une typologie linguistiquement motivée semble donc bien être mieux adaptée à la caractérisation du type de discours dans les langues de spécialité.

Pour terminer, nous nous sommes interrogés dans la dernière partie sur la légitimité de notre classification binaire. Nous proposons alors une alternative, sous la forme d'un degré de spécialisation des documents. En nous appuyant sur les scores SVM attribués aux documents lors de la classification, nous montrons qu'un continuum peut effectivement être observé entre ces deux classes.

CHAPITRE 6

Création d'un système d'aide à la construction de corpus comparables

6.1 Introduction

Dans les chapitres précédents, nous avons prouvé l'intérêt des corpus comparables dans le cadre d'études multilingues des langues de spécialité. La constitution de corpus comparables spécialisés de qualité, c'est-à-dire représentatifs d'un domaine et au degré de comparabilité élevé, est une tâche complexe. Nous avons proposé une méthode de constitution et décrivons les différentes étapes dans le chapitre 2. Dans ce chapitre 6 nous présentons la mise en œuvre d'un outil d'aide à la construction de corpus comparables. Cet outil doit permettre d'assister l'utilisateur dans la création d'un corpus comparable spécialisé en français et en japonais. Il doit atteindre plusieurs objectifs. Tout d'abord, il doit permettre de construire des corpus portant sur un domaine et un thème fixés par l'utilisateur. De plus, il doit garantir la construction d'un corpus de bonne qualité. D'un point de vue pratique, l'utilisateur doit garder une part de liberté : il doit pouvoir contrôler certaines parties cruciales de la constitution du corpus, en particulier la sélection des documents. Enfin, cet outil doit être stable, portable et facilement utilisable et modifiable. Nous présentons dans cette section les différentes étapes de la construction de corpus comparables spécialisés puis leur mise en place ainsi que les choix techniques sous-jacents.

6.1.1 Construction de corpus comparables

Nous souhaitons créer un outil d'aide à la construction de corpus comparables spécialisés. Cet outil doit faciliter la tâche de l'utilisateur lors de la construction d'un corpus. Les étapes de la constitution d'un corpus sont les suivantes :

Élaboration du « cahier des charges » : un corpus est généralement construit afin de mener une étude. Cette étape consiste à définir clairement les besoins et les critères que devra suivre le corpus pour qu'il soit représentatif de la population à analyser. Pour cela, la projection de la population dans le corpus doit être spécifiée en premier lieu : comment la population, dans notre cas une langue de spécialité, peut-être représentée dans un corpus ? Pour les corpus comparables spécialisés, cela revient à définir les langues, le domaine, le thème et le type de discours du corpus. Du point de vue pratique, il est aussi nécessaire de fixer une taille *a priori* au corpus (dépendant de l'exploitation du corpus), ainsi que la (les) ressource(s) dont seront extraits les documents (Web, collections de documents, portails spécialisés…). Enfin une méthode d'échantillonnage des données doit être définie, permettant de déterminer la distribution au sein des sous-catégories dans le corpus (afin

qu'elles soient toutes représentées). Pour les corpus comparables, cela peut revenir à définir la proportion de documents par type de discours, mais aussi la répartition en genre au sein des types de discours...

Sélection et collecte des données : une fois le cahier des charges élaboré, cette étape consiste à chercher et collecter des documents correspondants aux critères établis. Selon la ressource et la méthode d'échantillonnage, des documents correspondant au domaine et au thème doivent être sélectionnés. La plupart des ressources proposent des systèmes de recherche basés sur des mots-clés. Ceux-ci permettent de filtrer les résultats. Ensuite, les documents sélectionnés doivent être classés selon leur type de discours.

Normalisation et annotation des documents : cette étape consiste à effectuer plusieurs tâches de traitement sur les textes afin de pouvoir les exploiter. Dans le cas de documents extraits du Web, une première phase de nettoyage doit être réalisée afin d'extraire des documents le texte. La deuxième phase consiste à annoter les documents, c'est-à-dire ajouter aux textes extraits des informations (étiquettes morpho-syntaxiques par exemple) ;

Documentation du corpus : le corpus, une fois constitué et annoté, doit être documenté. Une trace des informations sur sa constitution doit être conservée afin qu'il puisse être réutilisé, par exemple les ressources utilisées, la date de constitution, les outils utilisés lors de sa constitution...

Nous abordons dans la partie suivante l'automatisation de ce processus d'un point de vue technique.

6.1.2 Mise en place et choix techniques

Parmi les tâches que nous avons présentées, certaines sont automatisables, tandis que d'autres ne le sont pas. En effet, l'élaboration du cahier des charges et la délimitation du cadre dépendent des objectifs applicatifs fixés par le créateur du corpus, seuls son expertise et son jugement peuvent permettre de déterminer quels critères permettent d'obtenir le corpus le plus représentatif. Dans le cadre de notre étude, l'utilisateur doit fixer lui même le domaine, le thème, la distribution au sein des types de discours, la taille du corpus et les ressources. La sélection des données peut être automatique ou semi-automatique (sélection faite par le système ou sélection de l'utilisateur parmi une pré-sélection établie par le système). Nous avons choisi de laisser dans un premier temps à l'utilisateur les tâches de définition du cahier des charges, de sélection et de collecte des documents. Cela permet d'éviter d'insérer dans le corpus des documents non-pertinents (pouvant figurer parmi les résultats d'une requête) et donc de garantir la qualité du corpus construit.

Nous travaillons donc sur l'automatisation de la normalisation, de l'annotation et de la documentation des corpus comparables spécialisés en français et japonais. Notre système d'aide à la création de corpus comparables est une chaîne de traitement composée de plusieurs composants ou unités de traitements. Nous souhaitons que cette chaîne soit stable, portable, facilement utilisable et surtout facilement modifiable. Pour cela, nous avons choisi de développer notre système en utilisant la plateforme `UIMA` (Unstructured Information Management Architecture). Celle-ci, développée par *IBM Research Division*, permet de développer en Java ou C++ des chaînes de traitements de corpus en faisant abstraction de certaines modalités techniques. De plus, `UIMA` permet une grande modularité : création de chaînes de modules, ajout et suppression facilitée, outils d'exécution des chaînes de traitement... De plus, cette plateforme est actuellement utilisée par l'équipe TALN du LINA afin de disposer d'un outil adapté permettant la mise en commun des composants réalisés par chacun. Nous avons donc souhaité nous inscrire dans ce projet afin de rendre accessible notre outil. Nous présentons dans une première partie le fonctionnement de la plateforme `UIMA` puis exposons les différentes étapes de la réalisation de notre chaîne.

6.2 Présentation de UIMA

Dans cette section nous décrivons tout d'abord le principe des application de gestion de données non structurées (UIM applications), leurs objectifs et définissons ce que sont des données non structurées. Dans un second temps, nous présentons l'architecture du système UIMA et ses principales composantes.

6.2.1 Principe et objectifs de UIMA

Les applications de gestion de données non-structurées (UIM, Unstructured Information Management) sont des systèmes logiciels qui analysent des grandes collections de données non-structurées afin d'en extraire, d'organiser et de représenter des connaissances pertinentes (Ferrucci et Lally, 2004b). Les données structurées (*structured information*) sont des informations dont le sens est non ambigu et représenté explicitement dans leur structure, par exemple des bases données relationnelles. Par opposition, les données non-structurées (*Unstructured data*) sont des données dont le sens et la forme sont libres, une interprétation est nécessaire pour déduire et représenter le sens, par exemple du texte, du son, des images, etc. (Ferrucci et Lally, 2004a, p. 455). Il est nécessaire, pour traiter des quantités sans cesse croissantes de données non-structurées, de développer des outils permettant de les gérer et d'en extraire des connaissances. Un corpus de notices pharmaceutiques peut par exemple être utilisé afin d'extraire les interactions médicamenteuses les plus dangereuses. Les différents outils de gestion de données non structurées visent à les représenter sous forme de données structurées. Pour cela, ils utilisent des technologies telles que le traitement statistique de la langue, la recherche d'informations, l'apprentissage automatique, les ontologies,etc. ainsi que des données structurées telles que des bases de données relationnelles.

Depuis quelques années, différents outils de traitement linguistique ont été développés, par exemple des segmenteurs, étiqueteurs syntaxiques, lemmatiseurs... Bien souvent, ces outils sont développés par une personne ou une équipe de personnes qui utilisent un langage, des structures, une syntaxe particuliers. Il est alors difficile de mettre en commun ces différents outils et plusieurs méthodes ont été adoptées pour mutualiser les efforts. Premièrement, des sites de dépôt ont permis de mettre à disposition de tous ces outils (Linguistic Data Consortium, European Language Resources Association...). La deuxième méthode consiste à fournir des systèmes logiciels au sein desquels peuvent être développés ces outils. Les premiers systèmes créés sont GATE (Bontcheva et al., 2002) et ATLAS (Laprun et al., 2002). Ces logiciels présentent l'avantage d'isoler les algorithmes centraux de traitement du langage des services système tels que le stockage des données, la communication entre composants... Les outils créés sur ces systèmes peuvent par la suite être facilement importés et utilisés, en faisant abstraction de nombreux aspects techniques (Hahn et al., 2008a). GATE, General Architecture of Text Engineering, est une plateforme de traitement de données textuelles. Elle se compose d'un environnement et d'une base de développement de composants et propose de nombreuses librairies et méthodes d'interaction. Elle est utilisée dans de nombreux projets[1], portant notamment sur le domaine de la gestion de connaissance et du Web sémantique, la création de librairies électroniques, la recherche d'informations dans le domaine de la bioinformatique... La plateforme ATLAS, Architecture and Tools for Linguistic Analysis Systems, correspond au même type d'outil. Cette plateforme propose un langage d'annotation ainsi qu'un système permettant de combiner et d'intégrer des informations produites par différents outils d'annotation. Une ontologie et un système de gestion de graphes d'annotations sont de plus disponibles. La plateforme UIMA a été développée par *IBM Research Division* et propose une architecture très flexible et extensible. Notre choix a porté sur cette plateforme pour plusieurs raisons. Tout d'abord, il nous permet de nous

[1] La liste des projets utilisant GATE se trouve à cette adresse : http://www.gate.ac.uk/projects.html.

inscrire dans le projet global de notre équipe de recherche. De plus, la communauté UIMA est l'une des plus active : de nombreuses contributions (librairies JAVA, composants...), publications et des ateliers (par exemple Hahn et al. (2008b)).

6.2.2 Architecture de UIMA

L'exécution d'une chaîne de traitement sur l'architecture UIMA est composée de deux phases principales : l'analyse des données et le retour d'informations. Ces deux phases sont résumées dans la figure 6.1. La première phase, l'analyse des données, consiste d'abord à charger une collection de données non-structurées. Ces données sont analysées, afin d'en extraire des informations qui sont ensuite stockées sous forme de données structurées. La seconde phase consiste à accéder à ces données structurées et à les représenter en fonction des besoins de l'utilisateur (ou du client).

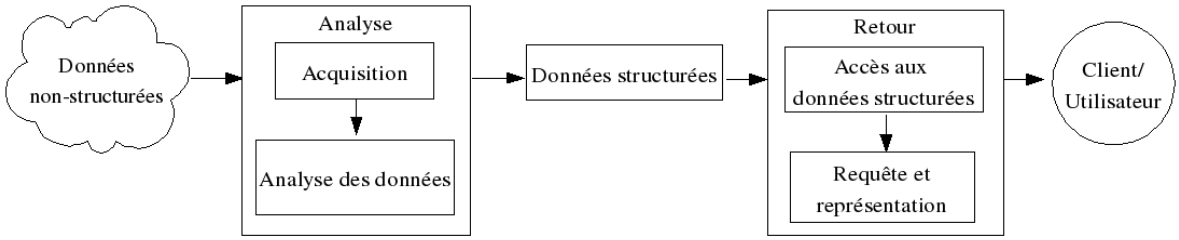


Figure 6.1 – Principales phases de l'architecture UIMA

Pour simplifier la présentation de cette architecture, nous décomposons cette section en deux parties : l'analyse au niveau du document (*document-level analysis*) et l'analyse au niveau de la collection de documents (*collection-level analysis*)[2].

6.2.2.1 Analyse au niveau du document

L'analyse d'un document correspond à la phase *analyse des documents* de la figure 6.1. Elle consiste à prendre un document et en extraire des données structurées ou méta-données (*meta data*) décrivant le document ou des parties du document. Une unité d'analyse de document est composée d'un ou plusieurs composants (*component*). Un composant analyse un document et ses méta-données et produit d'autres méta-données, des exemples de composants en TALN peuvent être des segmenteurs, des étiqueteurs syntaxiques, des détecteurs d'entités nommées... Le système UIMA traite deux types de composants : les composants primitifs et les composants agrégés. Les composants primitifs correspondent à une unité de traitement, tandis que les composants agrégés sont composés d'un ou plusieurs composants primitifs. La figure 6.2 présente un exemple de composant agrégé, qui fait appel à trois composants primitifs (segmentation, étiquetage et lemmatisation). Dans le cas où les données traitées sont textuelles, les composants sont alors appelés Text Analysis Engine (TAE).

[2]Nous nous inspirons de la présentation de UIMA faite par Ferrucci et Lally (2004a).

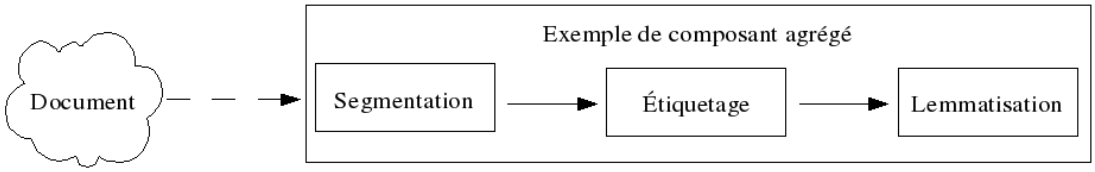


Figure 6.2 – Exemple de composant agrégé

Un *Type System* est associé à chaque TAE : il permet de définir et de rassembler l'ensemble des annotations qui seront extraites du texte pendant l'analyse. Un *type system* se définit par un ensemble de types d'annotations, appelés *feature structure*. Les documents et leurs méta-données transitent entre les différents composants dans une structure appelée *Common Analysis Structure* (CAS). Comme l'illustre la figure 6.3, un composant prend en entrée un CAS, analyse le texte, en extrait des informations (correspondant aux éléments du *type system* associé) et le met à jour avec ces informations. Le CAS résultant peut alors être traité par une autre application. L'avantage de cette méthode est que le développement d'un composant est indépendant du composant précédent, au sens où la sortie du composant précédent est connue et standard.

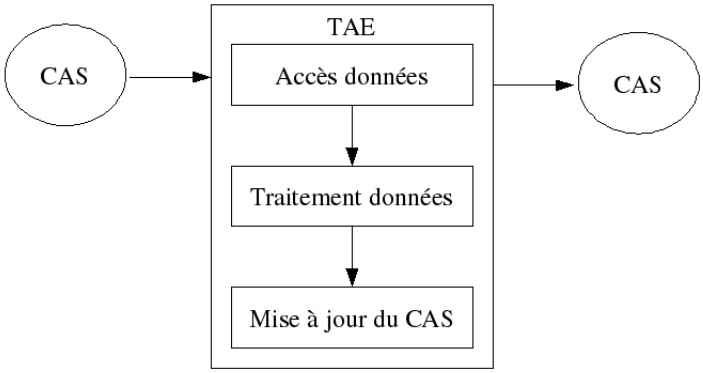


Figure 6.3 – Les CAS (Common Analysis Structure) permettent de transmettre les données d'un composant à l'autre

`UIMA` considère les textes traités comme des artefacts : les textes ne sont pas traités mais seulement analysés. Les informations extraites des textes sont ajoutées aux CAS sous la forme de méta-données ou annotations, qui peuvent porter sur le texte entier ou sur une partie, auquel cas leur localisation dans le texte est conservée avec l'annotation. `UIMA` propose une représentation `XML` des CAS, permettant de spécifier leur structure et une interface permettant l'accès, l'indexation et la mise à jour du contenu des CAS. Une interface orientée objet a été développée en JAVA : `JCas`. Elle permet la génération automatique de classes JAVA et des méthodes d'interaction entre les *Text Analysis Engine* et les CAS.

6.2.2.2 Analyse au niveau de la collection de documents

L'analyse au niveau du document n'est qu'une partie de l'architecture UIMA. Nous abordons dans cette section son fonctionnement au niveau d'une collection de documents. Une chaîne UIMA prend en entrée une collection de documents, les traite un à un à l'aide d'outils d'analyse niveau document puis rassemble les données extraites des différents documents afin de fournir à l'utilisateur les résultats, qui peuvent être par exemple un dictionnaire, une ontologie... Sur une collection de textes, l'unité fondamentale du système est appelée *Collection Processing Engine* (CPE). Ces CPE permettent d'appliquer des TAE (Text Analysis Engine) à toute une collection de textes. Le composant permettant d'accéder à la collection de texte et la transmettre aux TAE sous une forme appropriée est un *Collection Reader*. Le composant permettant de restituer à l'utilisateur les données extraites de la collection est un *CAS Consumer*. Le système UIMA propose de plus des API (Application Programming Interface) pour les CPE permettant de spécifier chacun des composants à utiliser : en début de chaîne le Collection Reader, puis une série d'outils d'analyse des textes (TAE) et enfin un CAS Consumer. Le système propose enfin des méthodes permettant de gérer les performances, le filtrage des données, les erreurs, etc.

6.2.2.3 Composants UIMA

Comme nous l'avons vu dans les sections précédentes, UIMA propose deux composants principaux de traitement : les TAE, au niveau des documents et les CPE, au niveau des collections. Le développement de chaînes sur UIMA consiste à implémenter et contrôler des combinaisons de ces composants. Deux propriétés doivent caractériser les composants UIMA, ils doivent être (Ferrucci et Lally, 2004a, p. 451) :

data-driven : le fonctionnement de chaque composant doit se baser uniquement sur les données qui lui sont fournies, ce qui est important pour que l'agrégation de composants ne nécessite aucune autre ressource ;

self-descriptive : chaque composant doit être accompagné de sa description (données en entrée, en sorties...).

La description des composants est contenue dans un fichier XML contenant :

- leur nom et un lien vers le code source du composant ;
- les paramètres de configuration ;
- le *type system* associé ;
- les spécifications des entrées et sorties du TAE.

Il existe différents types de composants, correspondant chacun à une tâche précise afin de faciliter leur combinaison : *Collection Reader*, *CAS Initialiser*, *TAE*, *CAS Consumer*. Un TAE primitif est composé d'un descripteur et d'un annotateur (programme d'analyse de documents). Un TAE agrégé est composé de couples (descripteur ; annotateur). Un CPE se définit par :

1. un *Collection Reader* ;
2. un *CAS Initializer* ;
3. un ou plusieurs TAE ;
4. un *CAS Consumer*.

C'est ensuite le système *UIMA* qui « orchestre » les instances de ces composants pour pouvoir exécuter le Collection Processing Engine.

6.2.3 Synthèse

Le schéma 6.4 reprend les notions présentées précédemment et décrit l'architecture globale de UIMA. Partant d'une collection de données, le CPE débute par un composant de type *Collection Reader*, permettant de lire, de charger et d'effectuer un pré-traitement sur les documents de la collection. Le *CAS Initializer* génère ensuite pour chaque document le CAS associé, qu'il transmet ensuite au TAE principal. Celui-ci peut être composé d'un ou plusieurs TAE, agrégés ou primitifs. Chaque TAE met à jour les CAS en y ajoutant des annotations. Les CAS extraits du TAE principal sont ensuite transmis à un ou plusieurs *CAS Consumers*, transformant les données du CAS dans un format exploitable par l'utilisateur (base de données, fichiers XML…).

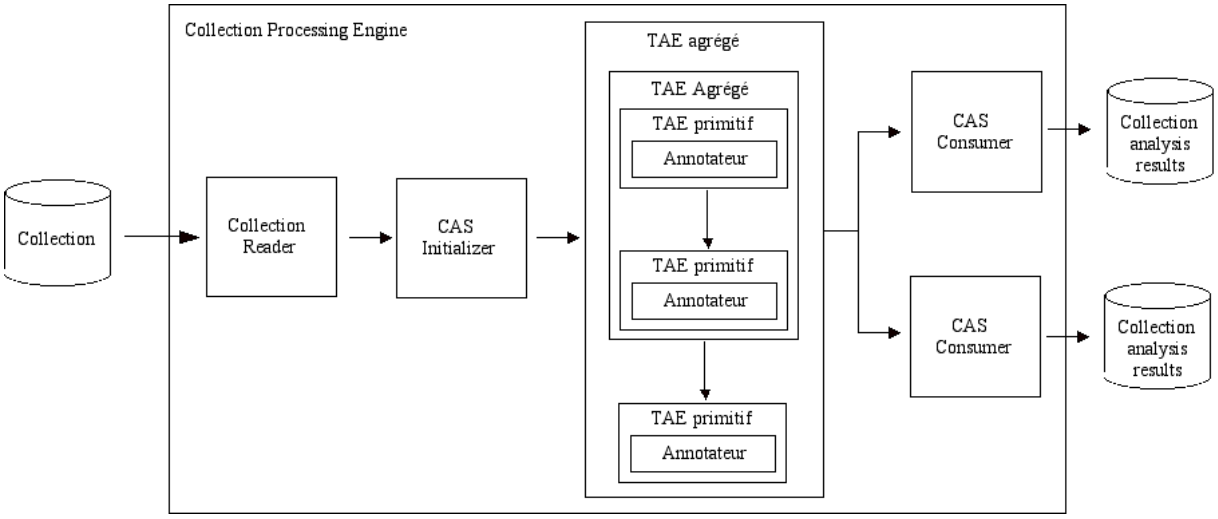


Figure 6.4 – Architecture globale d'un CPE

Notre but est de créer un système d'aide à la construction de corpus comparables dont le fonctionnement sera le suivant : partant d'une collection de documents extraits du Web et sélectionnés par un utilisateur, le système effectuera les tâches de nettoyage, d'annotation, de classification puis de documentation des documents afin de constituer le corpus. Cet outil peut donc se décomposer en un ensemble de traitements entre lesquels navigueront les documents. Pour cela, la plateforme UIMA est parfaitement adaptée. Elle propose une architecture permettant de charger une collection de documents puis de définir un ensemble d'outils d'annotations textuels effectuant différents traitements. La gestion des documents entre chaque étape de la construction du corpus est directement prise en charge par la plateforme. De plus, UIMA permet de traiter tout type de données non structurées et propose des outils permettant de stocker un document sous différents formats, ce qui peut s'avérer très utile lorsque l'on travaille sur des documents extraits du Web. Enfin, le système UIMA est capable de charger facilement et en très peu de temps des collections de plusieurs centaines de documents. Nous présentons dans la section suivante la mise en œuvre de notre outil sur la plateforme UIMA.

6.3 Mise en œuvre de l'outil d'aide à la création de corpus comparables

Notre objectif est de construire des corpus comparables spécialisés de qualité en français et japonais, c'est-à-dire représentatifs et au degré de comparabilité élevé. Pour cela, nous avons choisi de développer sur la plateforme `UIMA` un outil d'aide à la construction de corpus. Les documents de ces corpus partagent un domaine, un thème et un type de discours. Les étapes de la construction de ces corpus sont présentées dans la section 6.1.1. Afin de garantir la représentativité et la comparabilité, nous laissons l'utilisateur libre de sélectionner les documents. Cela permet de garantir l'adéquation entre les critères de construction des corpus et les documents choisis (par exemple éviter d'inclure des documents ne portant pas sur le domaine et la thématique choisis). Nous laissons l'utilisateur choisir des documents de n'importe quel type de discours, l'outil se charge de trier ensuite les documents grâce aux classifieurs que nous avons présentés dans les chapitres 4 et 5. Les tâches réalisées par l'outil sont donc : la classification des documents selon leur type de discours, la normalisation et l'annotation des documents, puis la création du corpus et de sa documentation. Les documents français et japonais sont traités séparément pour certaines tâches telles que la normalisation, l'annotation, et la classification. Deux collections de documents sont donc considérées, une par langue.

6.3.1 Architecture globale

Le schéma global de cet outil est présenté dans la figure 6.5.

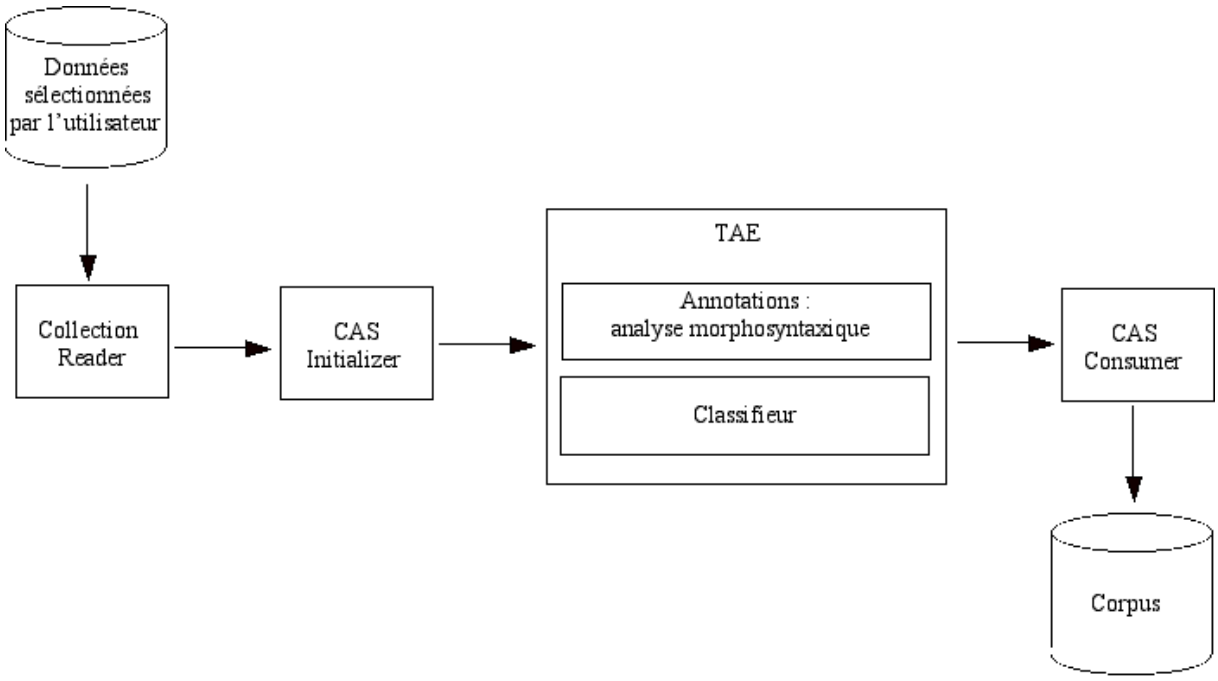


Figure 6.5 – Schéma global de l'outil d'aide à la construction de corpus comparables

La gestion de la collection de documents se fait grâce aux composants suivants :

Collection Reader : ce composant charge les documents de la collection, convertit chacun d'entre eux au format texte puis génère grâce au CAS Initializer les structures de données CAS pour chaque document ;

CAS Consumer : ce composant intervient à la fin du traitement. Son rôle est de repasser d'une collection de structure CAS à une réelle collection de documents (ici un corpus). Cela consiste ici à enregistrer sur le disque dur de l'utilisateur chaque version des documents (originale, texte, étiquetée), de répartir ces fichiers selon leur type de discours et de réaliser une documentation sur le corpus.

L'analyse au niveau des documents est réalisée lorsque ceux-ci sont stockés dans des CAS. Les différents composants sont les suivants :

La phase d'annotation et d'analyse morpho-syntaxique : cette partie est en réalité composée de deux étapes, l'une consacrée au nettoyage et au traitement présyntaxique (segmentation en phrase et en mots), l'autre consacrée à l'étiquetage morpho-syntaxique des textes. Le nettoyage consiste à éliminer une partie du bruit introduit dans les documents lors de la conversion afin de n'en conserver que le texte principal. Le traitement pré-syntaxique consiste à segmenter en phrases puis en unités lexicales (mots) les textes. Ces traitements sont nécessaires pour effectuer l'étiquetage morpho-syntaxique. Ce dernier est réalisé grâce aux outils Brill (Brill, 1994) et Flemm (Namer, 2000) pour le français, et ChaSen (Matsumoto et al., 1999) pour le japonais.

La phase de classification : composée de deux étapes, la création des représentations vectorielles des documents puis la classification de ces vecteurs grâce aux classifieurs SVM générés et présentés dans le chapitre 4. Cette classification a pour résultat un score qui sera utilisé par le CAS Consumer afin de répartir physiquement les documents selon leur type de discours.

Nous présentons dans la section suivante le *type system* défini pour nos documents puis nous présentons les deux niveaux d'analyse et leurs composantes dans les sections suivantes.

6.3.2 *Type System* définis

Le *Type System* que nous avons défini nous permet de stocker les informations générales relatives aux documents et à leur stockage dans le corpus. Il est représenté dans la figure 6.6 : un document est caractérisé par sa langue, son chemin absolu, son format, la date de son téléchargement, son URL (ou les informations sur la ressource dont il provient) et son score. Le score correspond au résultat de la classification `SVM`. Nous n'avons besoin en sortie de la chaîne que d'informations globales sur les documents, nous ne créons pas vraiment un outil d'annotation de textes. Ainsi, les annotations portent sur la globalité du document et pas sur des segments. Seules les annotations score SVM, langue et format sont utilisées pendant le traitement, les autres informations seront utilisées pour la documentation du corpus.

De plus, nous utilisons un *Type System* défini par l'équipe TALN du `UIMA` permettant de stocker un grand nombre d'informations, parmi lesquelles les catégories morpho-syntaxiques des mots ainsi que leurs lemmes. Une partie de ce *type system* est présentée dans la figure 6.7. Nous utilisons ce *type system* lors de la phase d'annotation des documents. Ce *type system* permet de spécifier pour chaque mot d'un texte son lemme, son stemme ainsi que sa catégorie grammaticale. Pour certaines catégories grammaticales, des attributs tels que le temps, le genre, le mode... peuvent être renseignés également. Ce *type system* a été conçu de façon à être exhaustif et à couvrir les besoins de l'équipe.

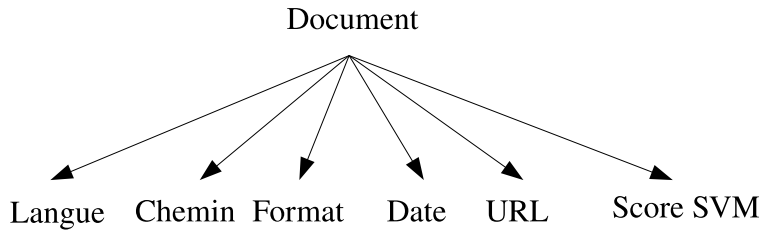


Figure 6.6 – *Type System* `Document`

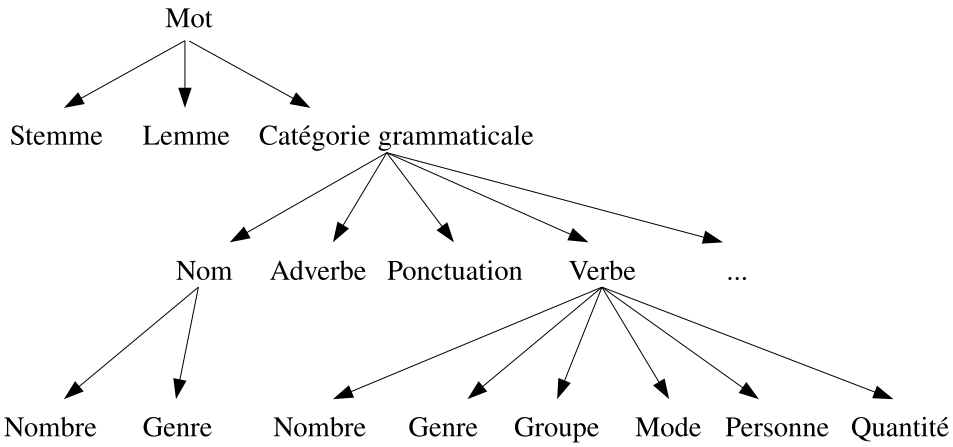


Figure 6.7 – *Type System* créé par l'équipe TALN

6.3.3 Analyse au niveau de la collection

`UIMA` permet de traiter des collections de documents : en chargeant chacun des documents, des composants les analysent ensuite un à un. Deux composants sont importants dans le traitement de la collection en tant que telle : le *collection reader* et le *CAS consumer*. Le *collection reader* intervient au début de la chaîne, il permet de charger chacun des documents de la collection et de les stocker dans une structure de données appelée CAS. Le *CAS consumer* permet de passer de la structure de données avec annotations à des fichiers. Nous présentons en détail dans les sections suivantes chacun de ces deux composants et leur fonctionnement dans notre outil de construction de corpus.

6.3.4 *Collection Reader*

Puisque nous avons laissé à l'utilisateur la charge de sélectionner et collecter les documents, la collection de départ est composée de documents extraits du Web ou d'autres bases de données textuelles. Ces documents peuvent être en HTML, au format PDF ou au format texte. Pour chacun de ces documents un CAS doit être créé dans lequel nous pouvons initialiser chacune des annotations définies dans notre

type system. De plus, les documents pouvant être dans différents formats, il est nécessaire de les convertir dans un format unique facilitant le traitement : le format texte.

Nous avons créé un Collection Reader basé sur celui de base proposé par `UIMA` et nous avons redéfini certaines méthodes (`initialize` et `getNext`) afin de les adapter à nos besoins. La gestion des différents formats de fichiers est gérée grâce aux vues, à un même document peuvent correspondre plusieurs vues. Afin de traiter des documents dans les langues française et japonaise, les documents doivent être convertis dans l'encodage UTF-8. La conversion et le traitement des encodages sont facilités par le langage JAVA qui permet de choisir l'encodage lors de l'ouverture des fichiers.

Le traitement de chacun des fichiers est le suivant :

1. Création du CAS, structure de données permettant de stocker un document, ses différentes vues et ses annotations ;
2. Lecture du document ;
3. Transformation en texte du document, réalisée à l'aide des outils `ExtractText` et `HTMLUtils` ;
4. Nettoyage des textes (nettoyage des dernières balises, caractères spéciaux...) : bien que le nettoyage des pages Web constitue un véritable challenge tant la tâche est complexe (Baroni et al., 2008), nous nous sommes basés sur des principes simples nous permettant d'éliminer la majorité du bruit laissé par les outils de conversion. Ce bruit se caractérise par des balises ou du code JavaScript, des symboles, sauts de lignes superflus, etc. ;
5. Stockage de la vue `Texte` : les textes ainsi nettoyés sont alors stockés dans la vue principale du CAS, celle qui sera la plus consultée ;
6. Pour les fichiers `HTML`, stockage de la vue `Original` : puisque nous ne traitons pas les versions originales des documents PDF mais seulement leur version texte, nous n'avons pas jugé nécessaire de stocker la version originale ;
7. Mise à jour des annotations `Langue`, `Chemin` et `Format` : ces informations sont fournies par l'utilisateur au moment du lancement de la chaîne.

Nous proposons de plus à l'utilisateur de fournir un fichier appelé `resources.info` dans lequel il peut donner des informations sur les documents qu'il a sélectionné. Ces informations, utilisées pour documenter le corpus, sont la date de téléchargement du document et la ressource dont il est extrait (et son URL s'il est extrait du Web). Le fichier est constitué d'une ligne par document, dans laquelle figurent le nom du fichier dans lequel le document est enregistré, la date de téléchargement et la ressource. Nous chargeons ce fichier au début de l'exécution du *collection reader* et l'utilisons afin de mettre à jour les CAS.

Une fois ces tâches effectuées, le *collection reader* passe le relais au *CAS initializer* qui est chargé de passer du niveau collection au niveau analyse en gérant les différents CAS créés. N'ayant pas modifié le composant de base, nous ne le présentons pas dans cette partie.

6.3.5 *CAS Consumer*

Nous avons créé un `CAS Consumer` propre à notre travail. Celui-ci est une extension du *CAS Consumer* de base, dans lequel nous avons redéfini quelques méthodes. Ce composant permet de passer d'une représentation structurée sous forme de CAS à une représentation adaptée à l'utilisateur, en l'occurrence un corpus. Trois répertoires sont créés : un répertoire de documents scientifiques, un répertoire de documents vulgarisés et un répertoire de documents ambigus. Pour chaque fichier, le traitement suivant est effectué :

1. Selon le score du document, REP = (scientifique ; vulgarisé ; ambigu) ;
2. Copie du fichier texte, original et étiqueté dans REP ;
3. Création du fichier XML correspondant adoptant le format TEI [3].

Pour la première étape, nous classons les documents obtenant un score compris entre $[-0.1, 0.1]$ dans le répertoire ambigu, les documents ayant un score supérieur à 0.1 dans le répertoire scientifique et les documents ayant un score inférieur à -0.1 dans le répertoire vulgarisé. Les fichiers XML au format TEI sont générés sous une forme minimale : seul un entête est utilisé, dans lequel nous stockons la langue, la date et le chemin du fichier original. Disposant de peu d'information sur les documents, nous ne voulions toutefois pas bloquer l'utilisateur et choisissons dans un premier temps de le laisser libre de compléter cette documentation.

Une fois cette étape terminée, les deux corpus scientifique et vulgarisé se trouvent dans chacun des deux répertoires. L'utilisateur reste libre d'y ajouter / supprimer des fichiers mais aussi d'y ajouter du contenu.

6.3.6 Analyse au niveau des documents

Dans cette section nous présentons l'analyse effectuée au niveau du document. Les documents sont transmis de composant en composant grâce aux CAS. Ainsi, seuls les CAS sont traités dans les composants. La figure 6.8 présente les deux phases du traitement des documents : l'analyse morpho-syntaxique puis la classification.

L'analyse morpho-syntaxique est elle-même réalisée en deux étapes : une première étape appelée présyntaxe puis une étape d'analyse morpho-syntaxique. Pour classer les documents, il faut d'abord créer les représentations vectorielles des documents et ensuite classer le vecteur à l'aide des classifieurs SVM. Nous présentons dans la suite de cette partie ces deux phases.

6.3.6.1 Analyse morpho-syntaxique

Aux textes de la collection sont appliqués un premier traitement présyntaxique puis une phase d'étiquetage morpho-syntaxique. Le traitement présyntaxique consiste à segmenter le texte en phrases puis en mots et à décoller la ponctuation. Ce traitement est nécessaire pour appliquer ensuite les logiciels d'étiquetage aux textes. La segmentation ne génère aucune annotation, elle ne fait que modifier le texte que nous stockons dans une variable. L'analyse morpho-syntaxique consiste à attribuer à chaque mot sa catégorie grammaticale et son lemme. Nous avons ici utilisé les logiciels Brill et Flemm pour le français, et Chasen pour le japonais. Nous avons utilisé les composants Brill et Flemm développés au LINA, dont le fonctionnement est le suivant : le logiciel est exécuté sur le texte, la sortie correspond à un texte étiqueté. Cette sortie est ensuite parcourue afin que chaque étiquette soit renseignée dans le *type system* LINA présenté dans la section 6.3.2. Chaque CAS est alors actualisé avec des annotations pour chaque mot des textes. Nous conservons toutefois la sortie des logiciels Brill et Flemm afin de l'inclure au corpus final. Nous n'avons trouvé aucun composant utilisant ChaSen. Nous avons donc créé notre propre composant, se contentant d'appeler le programme sur un texte et d'en récupérer la sortie étiquetée. Une fois que chaque document est stocké sous format original, texte et étiqueté, nous passons au TAE suivant réalisant la classification des documents.

[3] `http://www.tei-c.org.uk`

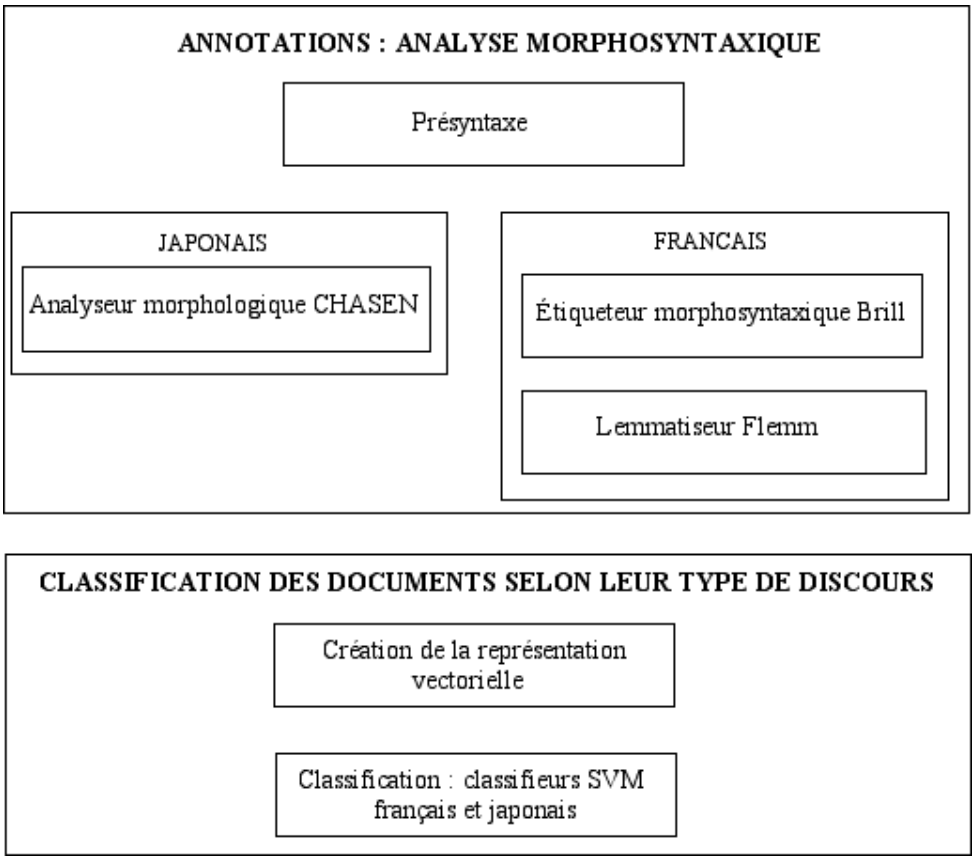


Figure 6.8 – Différentes étapes de l'analyse au niveau des documents

6.3.7 Composant *Classifier*

Ce composant est le dernier annotateur de notre chaîne. Il est composé de deux étapes : la première consiste à générer les représentations vectorielles de chaque document et la seconde consiste à classer le document. Pour générer la représentation vectorielle, nous utilisons notre algorithme d'analyse présenté dans le chapitre 4. Cet algorithme utilise les versions originales, textuelles et annotées des documents afin de les représenter sous forme de vecteurs. Chaque élément du vecteur correspond à la valeur d'un critère sur le document. Le vecteur ainsi constitué est utilisé sous forme de chaîne de caractères aux classifieurs SVM afin qu'ils leur attribue un score. Nous avons choisi d'utiliser le classifieur SVM utilisant la typologie composée des critères structurels, modaux (théorie de Charaudeau) et lexicaux pour deux raisons : ses résultats sont parmi les meilleurs en moyenne sur la précision et le rappel et il obtient les meilleurs résultats en précision. Nous privilégions la précision sur le rappel ici puisque notre objectif est de ne conserver qu'un seul type de discours dans le corpus, donc d'avoir une classification la plus exacte possible. Le résultats de la classification est ensuite stocké dans le CAS. Nous utilisons pour l'instant de façon directe les classifieurs SVM générés, mais nous envisageons d'utiliser des librairies d'apprentissage et de classification Weka[4]. Ces librairies permettent d'apprendre des modèles de classification en utilisant plusieurs algorithmes et de classer des documents. Cela permettrait d'apprendre de nouveaux

[4] `http://informatics.mayo.edu/text/index.php?page=weka`

classifieurs sur d'autres langues par exemple.

6.4 Conclusion

Nous avons présenté dans ce chapitre l'architecture UIMA et le développement d'un système d'aide à la construction de corpus comparables spécialisés sur UIMA. Nous avons proposé une première trame d'un système que nous envisageons bien plus complet prochainement. Notre outil permet d'effectuer automatiquement les parties de nettoyage, d'annotation et de documentation d'un corpus à partir de textes sélectionnés et collectés par l'utilisateur. Nous avons souhaité que cette chaîne de traitement soit stable, portable, facilement modifiable et utilisable. En ce sens, bien qu'elle ajoute certaines contraintes dans la programmation, l'utilisation de la plateforme UIMA présente de nombreux avantages : la facilité d'exécution grâce aux applications fournies avec UIMA, l'utilisation des nombreuses librairies JAVA... Le temps demandé par la prise en main de UIMA et de l'environnement de programmation est compensé par un véritable gain de temps, grâce à la structure *CAS*, permettant de ne pas gérer le transfert des données entre composants, mais aussi grâce aux composants disponibles sur le site de dépôt d'UIMA... Enfin, ce choix nous permet de nous inscrire dans le projet ambitieux de l'équipe visant à fournir une plateforme et des outils de mise en commun de nos travaux.

Il est difficile de chiffrer le temps que peut prendre la constitution manuelle d'un corpus comparable spécialisé, mais les tâches de nettoyage, d'annotation et de documentation peuvent être relativement longues, surtout sur de grosses collections de documents. Si l'utilisateur ne dispose d'aucun outil informatique, ces tâches peuvent représenter plusieurs jours de travail. Ici, le traitement est réalisé en quelques minutes. Cette chaîne n'a pas vocation à être totalement automatique : le matériau sur lequel nous travaillons, les textes et plus généralement la langue, sont trop complexes pour se passer de l'intervention humaine. Ainsi, nous considérons qu'un corpus de grande qualité ne peut être construit de façon totalement automatique. Notre outil permet toutefois de proposer aux constructeurs de corpus une aide et une plateforme permettant d'appliquer simplement des outils de traitement de texte et d'analyse linguistique sur un corpus.

Enfin, il sera aisé d'ajouter de nouveaux composants à cette chaîne : il suffit de partir de la composition des CAS pour redéfinir et intégrer un composant. Nous prévoyons par exemple d'ajouter à cette chaîne :

- un composant permettant de faciliter les tâches de sélection et de collecte des documents. Ce composant proposerait, pour un thème et un type de discours particuliers, une pré-sélection de documents. L'utilisateur pourrait alors choisir dans cette liste les documents à inclure dans le corpus et ceux-ci seraient alors collectés ;
- le choix dans les outils d'annotation des documents, par exemple Brill ou Tree Tagger pour le français ;
- un *type system* correspondant à notre typologie. Les valeurs de chacun des critères étant alors considérés comme des annotations, nous pourrions proposer un outil de visualisation de ces critères dans les textes (permettant une analyse des types de discours par exemple).

L'ajout de nouvelles langues à la chaîne paraît plus compliqué. En effet, s'il est simple de trouver des étiqueteurs morpho-syntaxiques, toute la phase d'analyse des types de discours présentée dans le chapitre 3 reste à faire afin de créer un classifieur.

Conclusion

Les corpus comparables sont des ensembles de textes dans différentes langues qui ne sont pas des traductions mais partagent un certain nombre de caractéristiques (Bowker et Pearson, 2002). Ces corpus sont très utilisés dans le cadre d'études sur les langues de spécialité afin d'extraire et d'actualiser terminologies et lexiques multilingues. Nous nous sommes intéressés dans cette thèse à la construction des corpus comparables spécialisés en français et en japonais. Ce travail de thèse s'est déroulé en trois grandes étapes. Tout d'abord, nous nous sommes penchés sur la définition des corpus comparables. Nous avons analysé la notion de comparabilité, les méthodes de calcul et l'impact du choix des critères de comparabilité sur celle-ci. Pour créer un système de construction automatique de corpus comparables, il a été nécessaire d'automatiser la reconnaissance des critères de comparabilité. Nous avons donc cherché à caractériser le type de discours de documents spécialisés afin de créer un système de reconnaissance automatique. Enfin, nous mis en commun les résultats des deux premières étapes afin de créer un système d'aide à la construction de corpus comparables. Nous présentons dans les parties suivantes un bilan de chacune des trois parties de cette thèse.

Comparabilité des corpus

Dans un premier temps, nous avons étudié la définition des corpus comparables. Le choix des caractéristiques communes et le degré de comparabilité d'un corpus dépendent de ses objectifs applicatifs. Ainsi, un corpus jugé très comparable pour une tâche particulière ne le sera pas forcément pour une autre. Nous nous sommes alors demandé s'il était possible de donner une définition générale de la comparabilité. Nous avons de plus cherché quel pouvait être le lien entre le choix des caractéristiques communes et la comparabilité.

Nous avons collecté l'ensemble des caractéristiques communes aux textes et le contexte de leur utilisation. Ceci nous a permis de dégager deux séries de caractéristiques récurrentes : les corpus comparables portant sur des domaines de la langue générale font plus souvent appel à des textes ayant en commun un thème, une période et un genre ou un médium, tandis que les textes des corpus issus de domaines spécialisés ont en commun un thème ou un domaine, un genre et/ou un type de discours. Nous avons de plus proposé une analyse de la notion de comparabilité et des moyens de la quantifier. La comparabilité mesure les similarités entre documents d'un corpus. En pratique, la comparabilité s'instancie différemment selon le contexte d'utilisation du corpus : dans le cadre de l'extraction de traductions par exemple, celle-ci correspond alors au vocabulaire commun entre les textes. Cette analyse nous permet de conclure qu'en dehors de tout contexte d'utilisation, seules les caractéristiques communes des textes peuvent nous permettre de statuer sur la comparabilité d'un corpus. Dans un cadre pratique, à une utilisation précise peut correspondre une méthode de calcul de la comparabilité. Cependant, nous jugeons important que l'unité dans ce calcul soit située au niveau du texte et pas au niveau du mot, afin d'éviter la vision « sac de mots » des corpus. Dans un second temps nous nous sommes concentrés sur les corpus comparables spécialisés. Nous utilisons pour définir nos corpus la théorie des langues de spécialité de Bowker et Pearson (2002). L'étude de la comparabilité nous permet de déterminer que les textes constitutifs de ces corpus doivent partager un thème (scientifique) et un type de discours (scientifique ou vulgarisé). Le thème permet de circonscrire un champ scientifique et le type de discours permet de filtrer le niveau de

communication des documents du corpus, garantissant un niveau de langue, un lexique et une syntaxe communs. Ces corpus ont alors un fort degré de comparabilité pour les études des langues de spécialité.

Caractérisation du type de discours

Notre objectif étant de créer un système d'aide à la construction de corpus comparables, la reconnaissance de ces caractéristiques communes doit être automatisée. Que les documents soient collectés sur une base de données textuelles ou sur le Web, leur thème peut être filtré grâce à des mots clés. Pour le type de discours, la création d'un système de classification automatique a été nécessaire. Pour cela, nous avons construit un corpus comparable d'apprentissage sur le thème du diabète et de l'alimentation dont les documents sont classés selon leur type de discours. Une analyse stylistique contrastive sur ce corpus nous a permis de dégager un ensemble de critères caractérisant le type de discours des documents. Afin que cette typologie soit motivée linguistiquement, nous avons choisi des critères correspondant à trois niveaux d'analyse : le niveau structurel, correspondant aux caractéristiques non linguistiques des documents (format, structure…), le niveau modal, correspondant aux marqueurs de la modalité dans le document (présence du locuteur et de l'interlocuteur), et le niveau lexical (vocabulaire, alphabets…). À partir de cette typologie, nous avons créé des modèles de classification avec les systèmes *SVMlight* (séparateurs à vastes marges) et *C4.5* (arbres de décision). Afin d'évaluer ces classifieurs, nous avons constitué un corpus d'évaluation portant sur le thème du cancer du sein. Les résultats obtenus sont satisfaisants, plus de 70 % des documents sont correctement classés, quelle que soit la langue. Nous obtenons une précision proche de 80 % pour le français et de 70 % pour la japonais, nous permettant par la suite de minimiser les erreurs de classification dans les corpus générés avec notre système. Une analyse des résultats et de l'influence des critères sur la classification nous a permis de constater que certains critères étaient très discriminants (des critères structuraux tels que les balises paragraphe et lexicaux tels que le vocabulaire spécialisé) tandis que d'autres, en théorie pertinents, se révélaient être très peu efficaces en pratique (les critères de la modalité Irrealis par exemple). Bien que certains d'entre eux soient très discriminants, c'est bien leur combinaison relevant de trois niveaux d'analyse qui rend nos classifieurs efficaces. De l'efficacité de ces classifieurs nous pouvons conclure que notre typologie bilingue caractérise bien les types de discours scientifiques et vulgarisés. Celle-ci est de plus linguistiquement motivée et adaptable à d'autres langues.

Construction automatique de corpus comparables spécialisés

La troisième partie de cette thèse était consacrée à la réalisation d'un outil d'aide à la construction de corpus comparables. La tâche de construction d'un corpus est lourde et coûteuse en temps. L'objectif de ce système est d'optimiser certaines étapes de la construction. Nous avons proposé une méthode de construction de ces corpus, en adaptant les méthodes classiques à notre cas : des corpus multilingues au degré de comparabilité élevé et représentatifs d'un domaine de spécialité. Cette méthode est composée de 4 étapes :

1. Définition du cahier des charges : définition de la population (domaine et thème) et des caractéristiques du corpus (ressources, taille, échantillonnage) ;
2. Sélection et collecte des documents ;
3. Normalisation et annotation des documents ;
4. Documentation du corpus.

L'utilisateur réalise les deux premières tâches : celles-ci sont très subjectives et dépendent des objectifs applicatifs que s'est fixé l'utilisateur. La qualité du corpus est de plus directement liée à ces deux tâches. Ce système permet donc, à partir d'un ensemble de documents portant sur un thème particulier préalablement collectés de construire un corpus comparable composé de documents scientifiques ou vulgarisés. Il effectue donc les deux dernières tâches, qui se déroulent concrètement de la façon suivante : conversion et nettoyage, annotation et classification des documents, puis documentation du corpus. Ce système a été développé sur la plateforme UIMA (Unstructured Information Management Architecture).

Bilan et perspectives

Nous avons présenté dans cette thèse un premier travail de synthèse sur la définition des corpus comparables et de la comparabilité. De cette synthèse découle une méthode de constitution de corpus comparables spécialisés adaptée des méthodes classiques. Plusieurs communautés utilisent et exploitent ces corpus avec des objectifs différents : des linguistiques, des traducteurs et des chercheurs en TALN. Nous avons cherché dans cette thèse à rassembler ces communautés et à mettre en commun leurs besoins. Nous proposons ici un compromis entre la vision très floue des corpus comparables (les considérant comme des *sacs de mots*) de certains informaticiens et une vision très précise et contrainte des linguistes. Ainsi, l'outil de construction que nous proposons permet à toutes ces communautés de constituer de façon plus rapide des corpus comparables de qualité, au degré de comparabilité élevé et pleinement représentatifs d'un domaine.

Nous avons dans cette conclusion présenté les principaux apports et enseignements de cette thèse. Nous souhaitons aborder, dans ce qui suit, les pistes de recherche nous paraissant les plus intéressantes et prometteuses.

D'un point de vue technique, notre système d'aide à la construction de corpus comparable pourrait être étoffé et amélioré. Nous l'avons voulu évolutif, c'est pourquoi nous avons utilisé la plateforme *UIMA*. Cette plateforme permet de mettre de côté les considérations techniques telles que la manipulation des données et son transit entre composants afin de ne s'intéresser qu'au traitement des documents. Des apports quant à la sélection des documents sont envisageables, par exemple une collecte automatique de documents depuis le Web en fonction de mots-clés. Des outils de visualisation des documents et de leur représentation vectorielle peuvent également être mis à disposition de l'utilisateur. Un système de mise en commun des corpus ainsi créés pourrait permettre de mutualiser les ressources et favoriser les liens entre équipes.

D'un point de vue linguistique, l'adaptation des différentes théories de la modalité mériterait d'être approfondie. En théorie, chaque langue et chaque culture semblent correspondre à une caractérisation particulière de la modalité. Nous pensions ici que la théorie Locutive était adaptée au français et la théorie Irrealis au japonais. En pratique, cela ne semble pas être le cas. La mise en œuvre de la modalité Irrealis par les marqueurs donne de très mauvais résultats sur le japonais. Il serait intéressant d'analyser ce phénomène afin de déterminer si ce sont les marqueurs choisis qui posent problème ou si cette modalité n'est pas adaptée à un traitement automatique, mais aussi pourquoi une théorie est plus efficace qu'une autre dans le cadre de la détermination du type de discours.

Nous avons émis des doutes dans le chapitre 5 sur la légitimité de notre classification binaire scientifique/vulgarisé. Proposant de considérer un degré de spécialisation plutôt que cette distinction binaire, nous nous sommes appuyés sur les scores attribués par notre classifieur aux documents afin de montrer que celui-ci évolue graduellement. Des documents que nous avons jugé entre les deux classes obtiennent effectivement des scores moyens. Cependant, une étude approfondie mériterait d'être menée afin d'obser-

ver ce phénomène sur d'autres documents et d'analyser cette notion de score ou degré de spécialisation.

Nous avons ici étudié la construction de corpus comparables spécialisés en français et japonais. Une extension de ce travail à d'autres langues serait souhaitable pour notre outil. Pour cela, une adaptation de la typologie aux nouvelles langues est nécessaire, impliquant l'investissement d'un expert linguiste. À chaque critère de la typologie doit correspondre un ensemble de marqueurs, résultant souvent d'une analyse sur corpus. Un corpus comparable (ou sous-corpus monolingue si le corpus comparable existe) doit donc être créé, et sur celui-ci sera appris le modèle de classification de la langue.

Nous pensons enfin que le travail de définition et formalisation des corpus comparables doit être poursuivi et étendu. En effet, même si ces corpus sont souvent utilisés, peu de travaux détaillent leur vision des corpus comparables, ce qui entraîne des ambiguités. La synthèse proposés dans cette thèse pourrait servir de base à une définition commune de ces corpus rassemblant la communauté (naissante). Cette mise en commun permettrait de la même façon d'uniformiser la construction de ces corpus et d'encourager la création de *corpus comparables de référence*. Le workshop « *Building and Using Comparable Corpora* » à LREC 2008 a pour la première fois rassemblé cette communauté dans une conférence. La qualité des discussions et présentations, ainsi que l'intérêt porté à cette uniformisation vont dans cette direction et son plus qu'encourageants.

Bibliographie

Adam, J.-M. (1992). *Les textes : types et prototypes.* Paris: Nathan.

Argamon, S., Whitelaw, C., Chase, P., Dhawle, S., Hota, S. R., Garg, N., et Levitan, S. (2007). Stylistic text classification using functional lexical features. *Information Science and Technology*, **58**(6), 802–822.

Aston, G. et Burnard, L. (1998). *The BNC Handbook: exploring the british national corpus with Sara*. Edinburgh University Press.

Bally, C. (1952). *Le langage et la vie*. Droz et Giard, Zurich, M. Niehans, Genève, 3e ed. augm. edition. Romanica Helvetica. Séries linguistica Vol. I.

Baroni, M., Chantree, F., Kilgarriff, A., et Sharoff, S. (2008). Cleaneval: a competition for cleaning webpages. In N. Calzolari, editor, *Actes de la 6ème édition de Language ressources and Evaluation Conference (LREC 2008)*.

Beauvisage, T. (2001). Morphosyntaxe et genres textuels. *Traitement Automatique des Langues (TAL)*, **42**(2), 579–608.

Benveniste, E. (1966). *Problèmes de linguistique générale . I*, volume vol 1. Gallimard – 1976.

Benveniste, E. (1970). L'appareil formel de l'énonciation. *Langages*, **17**.

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

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing. Journal of the Association for Litterary and Linguistic Computing*, **8/4**, 243–257.

Biber, D. (1994). Representativeness in corpus design. In A. Zampolli, N. Calzolari, et M. Palmer, editors, *Current Issues in Computational Linguistics: Essays in Honour of Don Walker*, pages 377–407. Giardini Editori e Stampatori and Kluwer Academic Publishers, Pisa and Dordrecht.

Bontcheva, K., Cunningham, H., Maynard, D., Tablan, V., et Saggion, H. (2002). Developing reusable and robust language processing components for information systems using gate. In *Proceedings of the 13th International Workshop on Database and Expert Systems Applications (DEXA 2002)*, pages 223–227. IEEE Computer Society.

Bowker, L. (1998). Using specialised native-language corpora as a translation resource: a pilot study. *Meta*, **43**(4), 631–651.

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

Breiman, L., Friedman, J., Stone, C., et Olshen, R. (1984). *Classification and Regression Tree*. California: Wadsworth International.

Bretan, I., Dewe, J., Hallberg, A., Wolkert, N., et Karlgren, J. (1998). Web-specific genre visualization. In H. A. Maurer et R. G. Olson, editors, *Proceedings of WebNet 98 - World Conference on the WWW and Internet & Intranet*, Orlando, Florida, USA. AACE.

Brill, E. (1994). Some advances in transformation-based part of speech tagging. In *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI'94)*, pages 722–727, Seattle, WA, USA.

Bronckart, J. P. (1996). Genres de textes, types de discours et opérations discursives. *Enjeux*, **37-38**, 31–47. Namur.

Bronckart, J. P., Bain, D., Schneuwly, B., Davaud, C., et Pasquier, A. (1985). *Le fonctionnement des discours : un modèle psychologique et une méthode d'analyse*. Lausanne: Delachaux and Niestlé.

Brown, P., Pietra, S. D., et Mercer, R. (1991). Word sense disambigua- tion using statistical methods. In D. Appelt, editor, *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 264–270, Berkeley, California. Association for Computational Linguistics.

Béchade, H. (1992). *Phonétique et morphologie du français moderne et contemporain*. Presses Universitaires de France.

Calzolari, N. (1993). European efforts towards standardizing language resources. In *Proceedings of the European Association for Machine Translation (EAMT) Workshop*, pages 121–130.

Catizone, R., Russell, G., et Warwick, S. (1989). Deriving translation data from bilingual texts. In *Proceedings of the First International Lexical Acquisition Workshop*, pages 1–7, Detroit.

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

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

Chung, S. et Timberlake, A. (1985). Tense, aspect and mood. In T. Shopen, editor, *Language typology and syntactic description: Grammatical categories and the lexicon*, volume 3, pages 202–258. Cambridge University Press.

Cornuéjols, A. et Miclet, L. (2002). *Apprentissage artificiel - Concepts et algorithmes*. Eyrolles.

Culo, O., Schirra, S. H., Neumann, S., et Vela, M. (2008). Empirica l studies on language contrast using the english-german comparable and parallel croco corpus. In *Proceedings of the LREC workshop on Comparable Corpora*, pages 47–51.

Dagan, I. et Church, K. (1994). Termight : identifying and translating technical terminology. In *Proceedings of the 4th Conference on Applied Natural Language Processing (ANLP?94)*, pages 34–40, University of Stuttgart, Germany.

Daille, B., Gaussier, E., et Langé, J.-M. (1994). Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of the 15th International Conference on Computational Linguistics(COLING'94)*, pages 712–716, Kyoto, Japan.

Déjean, H. et Gaussier, E. (2002). Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica - Alignement lexical dans les corpus multilingues*.

Dubreil, E. (2006). *La dimension argumentative des collocations textuelles en corpus électronique spécialisé au domaine du TAL(N)*. Ph.D. thesis, Université de Nantes.

Ducrot, O. (1980). *Les Mots du discours*. Les Editions de Minuit.

Ducrot, O. et Schaeffer, J.-M. (3 septembre 1999). *Nouveau dictionnaire encyclopédique des sciences du langage*. Seuil.

Ferrucci, D. et Lally, A. (2004a). Building an example application with the unstructured information management architecture. *IBM Systems Journal*, **43**(3), 455–475.

Ferrucci, D. et Lally, A. (2004b). Uima: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, **10**, 327–348.

Finn, A. et Kushmerick, N. (2005). Learning to classify documents according to genre. *J. American Society for Information Science and Technology*. A paraître.

Folch, H., Heiden, S., Habert, B., Fleury, S., Illouz, G., Lafon, P., Nioche, J., et Prévost, S. (2000). TyP-Tex: Inductive typological text classification analysis for NLP systems tuning/evaluation. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, et G. Stainhaouer, editors, *Second International Conference on Language Resources and Evaluation (LREC)*, volume 1, pages 141–148, Athens, Greece. ELRA – European Language Resources Association.

Fung, P. (2000). *Parallel Text Processing: Alignment and Use of Translation Corpora*, chapter A Statistical View on Bilingual Lexicon Extraction-From Parallel Corpora to Non-parallel Corpora, pages 1–17. Kluwer.

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

Fung, P. et Yee, L. Y. (1998). An IR approach for translating new words from nonparallel, comparable texts. In C. Boitet et P. Whitelock, editors, *Proceedings of the 17th international conference on Computational linguistics*, volume 1, pages 414–420, Montreal, Quebec, Canada.

Gale, W. et Church, K. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, **19**, 75–102. SPECIAL ISSUE: Special issue on using large corpora.

Givón, T. (1994). Irrealis and the subjunctive. *Studies in Language*, **18**(2).

Godfrey, J., Holliman, E., et McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520, San Francisco.

Gœuriot, L., Grabar, N., et Daille, B. (2007). Caractérisation des discours scientifiques et vulgarisés en français, japonais et russe. In N. Hathout et P. Muller, editors, *Actes de la 14ème conférence sur le Traitement Automatique des Langues Naturelles (TALN 2007) (communications affichées et démonstrations)*, pages 93–102.

Gœuriot, L., Grabar, N., et Daille, B. (2008). Characterization of scientific and popular science discourse in french, japanese and russian. In N. Calzolari, editor, *Actes de la 6ème édition de Language ressources and Evaluation Conference (LREC 2008)*.

Habert, B. (2000). Des corpus représentatifs : de quoi, pour quoi, comment ? In M. Bilger, editor, *Linguistique sur corpus. Études et réflexions*, number 31 in Cahiers de l'université de Perpignan, pages 11–58. Presses Universitaires de Perpignan, Perpignan.

Habert, B., Nazarenko, A., et Salem, A. (1997). *Les linguistiques de corpus*. Armand Colin.

Habert, B., Grabar, N., Jacquemart, P., et Zweigenbaum, P. (2001). Building a text corpus for representing the variety of medical language. In P. Rayson, A. Wilson, T. McEnery, A. Hardie, et S. Khoja, editors, *Corpus Linguistics 2001*, pages 245–254, Lancaster. UCREL (University Centre for Computer Corpus Research on Language).

Hahn, U., Buyko, E., Landefeld, R., Mühlhausen, M., Poprat, M., Tomanek, K., et Wermter, J. (2008a). An overview of jcore, the julie lab uima component repository. In *Proceedings of the LREC workshop : Towards Enhanced Interoperability for large HLT systems: UIMA for NLP*, pages 1–7.

Hahn, U., Götz, T., Brown, E., Cunningham, H., et Nyberg, E., editors (2008b). *Proceedings of the LREC workshop : Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP*, Marrakech, Maroc.

Harris, Z. S. (1988). *Language and information*. Columbia University Press, New York.

Harris, Z. S. (1991). *A theory of language and information. A mathematical approach*. Oxford University Press.

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

Johansson, S., Leech, G., et Goodluck, H. (1978). *Manual of Information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers*. Department of English, University of Oslo.

Karlgren, J. (1999). Stylistic Experiments in Information Retrieval. In D. Kluwer, editor, *Natural Language Information Retrieval*. the Netherlands.

Karlgren, J. et Cutting, D. (1994). Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING)*, volume 2, pages 1071–1075, Kyoto, Japan.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, **6**(1), 97–133.

Kilgarriff, A. et Grefenstette, G. (2003). Introduction to the special issue on web as corpus. *Computational Linguistics*, **29**(3).

Knowles, F. et Roe, P. (1994). Facilitating the corpus-building process and maximising the 'analytical yield': A lsp-oriented case study. In *Papers in Computational Lexicography. Complex'94*, pages 137–146, Budapest - Hongrie.

Koehn, P. (2004). Europarl: A parallel corpus for statistical machine translation. In *MT Summit 2005*.

Kübler, N. (2008). A comparable learner translator corpus: creation and use. In *Proceedings of the LREC workshop on Comparable Corpora*, pages 73–78.

Kyto, M., Rissanan, M., et Wright, S., editors (1994). *Corpora across the centuries. Proceedings of the First International Colloquium on English Diachronic Corpora*, St Catharine's College Cambridge. Amsterdam and Atlanta: Rodopi.

Laprun, C., Fiscus, J., Garofolo, J., et Pajot, S. (2002). A practical introduction to atlas. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*. European Language Resources Association (ELRA).

Laviosa, S. (1998). The corpus-based approach: a new paradigm in translation studies. *Meta*, **43**(3), 474–479.

Lerat, P. (1995). Les langues spécialisées. *Linguistique nouvelle*.

Lewis, D. (1992). An evaluation of phrasal and clustered representations on a text categorization task. In N. Belkin, P. Ingwersen, et A. M. Pejtersen, editors, *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 37–50.

Lewis, D. M. (2005). *La linguistique de corpus*, chapter Corpus comparables et analyse contrastive : l'apport d'un corpus français/anglais de discours politiques à l'analyse des connecteurs adversatifs, pages 179–190. Presses Universitaires de Rennes.

L'Homme, M.-C. (2004). *La terminologie : principes et techniques*. Presses universitaires de Montréal.

Maingueneau, D. (1996). Les termes clés de l'analyse du discours. *Memo*, **20**.

Malrieu, D. et Rastier, F. (2002). Genres et variations morphosyntaxiques. *Traitement Automatique des Langues (TAL)*, **42**(2), 548–577.

Matsumoto, Y., Kitauchi, A., Yamashita, T., et Hirano, Y. (1999). Japanese morphological analysis system chasen version 2.0 manual. Rapport technique, Nara Institute of Science and Technology (NAIST).

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

Morin, E. (2007). *Synergie des approches et des ressources déployées pur le traitement de l'écrit*. Ph.D. thesis, Habilitation à Diriger les Recherches, Université de Nantes.

Morin, E. et Daille, B. (2004). Extraction de terminologies bilingues à partir de corpus comparables d'un domaine spécialisé. *Traitement Automatique des Langues (TAL)*, **45**(3), 103–122.

Mortureux, M.-F. (1995). Les vocabulaires scientifiques et techniques. In J.-C. Beacco et S. Moirand, editors, *Les enjeux des discours spécialisés*, pages 13–25. Presses universitaires de la Sorbonne.

Namer, F. (2000). Flemm : Un analyseur flexionnel du français à base de règles. *Traitement Automatique des Langues (TAL)*, **41**(2), 523–548.

Namer, F. et Baud, R. (2007). Defining and relatinf biomedical terms: Towards a cross-language morphosemantics-based system. *International Journal of Medical Informatics*, **76**(2-3), 226–233.

Oard, D. et Diekema, A. (1998). *Cross-Language Information Retrieval*, volume 33, pages 223–256. Information Today Inc. for the American Society for Information Science.

Palmer, F. R. (2001). *Mood and Modality*. Cambridge University Press.

Péry-Woodley, M.-P. (1995). Quels corpus pour quels traitements automatiques? *Traitement Automatique des Langues*, **36**(1-2), 213–232.

Péry-Woodley, M.-P. (2000). Une pragmatique à fleur de texte : approche en corpus de l'organisation textuelle. *Mémoire d'HDR, Carnets de grammaire N° 8*.

Peters, C., Picchi, E., et Biagini, L. (1996). Parallel and comparable bilingual corpora in language teaching and learning. In S. P. Botley, A. M. McEnery, et A. Wilson, editors, *Proceedings of Teaching and Language Corpora 1996*, pages 68–82. UCREL Technical Papers 9 (Special Issue), Lancaster University 1996.

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

Rapp, R. (1995). Identifying word translations in non-parallel texts. In *Proceedings of the 35th annual Meeting of the Association for Computational Linguistics (ACL 95)*, pages 320–322, Boston, Massachusetts, USA.

Rapp, R. (1999). Automatic identification of word translations from unrelated english and german corpora. In R. Dale et K. Church, editors, *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 519 – 526. Association for Computational Linguistics.

Rastier, F. (1989). *Sens et Textualité*. Hachette.

Rastier, F. (2002). Enjeux épistémologiques de la linguistique de corpus. In G. Williams, editor, *Actes des deuxièmes journées de Linguistique de Corpus*. Presses Universitaires de Rennes.

Riboni, D. (2002). Feature selection for web page classification. In H. Shafazand et A. M. Tjoa, editors, *Proceedings of the 1st EurAsian Conference on Advances in Information and Communication Technology (EURASIA-ICT)*, pages 473–478, Shiraz, Iran. Springer.

Riegel, M., Pellat, J.-C., et Rioul, R. (1994). *Grammaire méthodique du français*. Presses Universitaires de France.

Rossignol, M. (2005). *Acquisition sur corpus d'informations lexicales fondées sur la sémantique différentielle*. Ph.D. thesis, Université de Rennes 1.

Sager, N. (1986). Analyzing language in restricted domains. sublanguage description and processing. In R. Grishman et R. Kittredge, editors, *Sublanguage: Linguistic phenomenon, computational tool*. Lawrence Erlbaum.

Santini, M. (2007). *Automatic identification of genre in web pages*. Ph.D. thesis, University of Brighton.

Santini, M., Power, R., et Evans, R. (2006). Implementing a characterization of genre for automatic genre identification of web pages. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (ACL'06*, pages 699—-706, Sydney, Australia. The Association for Computer Linguistics.

Saralegi, X. et Alegría, I. (2007). Similitud entre documentos multilingües de carácter cientìfico-técnico en un entorno web. *Procesamiento del lenguaje natural*, **39**, 71–78.

Saralegi, X., Vicente, I. S., et Gurrutxaga, I. (2008). Automatic extraction of bilingual terms from comparable corpora in a popular science domain. In P. Fung et P. Zweigenbaum, editors, *Proceedings of the LREC workshop on Comparable Corpora*, pages 33–38.

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

Sebastiani, F. (2005). Text categorization. In A. Zanasi, editor, *Text Mining and its Applications to Intelligence, CRM and Knowledge Management*, pages 109–129. WIT Press, Southampton, UK.

Sharoff, S., Babych, B., et Hartley, A. (2006). Using comparable corpora to solve problems difficult for human translators. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 739–746, Sydney, Australia.

Sinclair, J. (1996a). Preliminary recommendations on corpus typology. Rapport technique, EAGLES (Expert Advisory Group on Language Engineering Standards).

Sinclair, J. (1996b). Preliminary recommendations on text typology. Rapport technique, EAGLES (Expert Advisory Group on Language Engineering Standards).

Steuckardt, A. et Niklas-Salminen, A. (2005). *Les marqueurs de glose*. Publications de l'Université de Provence.

Sueur, J. P. (1982). Pour une grammaire du discours. Élaboration d'une méthode; exemples d'applications. *Mots*, **5**, 145–185.

TEI Consortium (2007). Tei p5: Guidelines for electronic text encoding and interchange. Rapport technique, TEI Consortium.

Teubert, W. (1996). Comparable or parallel corpora? *International Journal of Lexicography*, **3**(9), 238–264.

TLFi (1960). Trésor de la langue française informatisé. `http://atilf.atilf.fr`.

Tzeras, K. et Hartmann, S. (1993). Automatic indexing based on bayesian inference networks. In R. Korfhage, E. Rasmussen, et P. Willett, editors, *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 22–34.

Vapnik, V. (1998). *Statistical Leaning Theory*. Wiley-Interscience.

Véronis, J. (2000). Alignement de corpus multilingues. In J.-M. Pierrel, editor, *Ingénierie des langues*, pages 151–171. Éditions Hermès.

Vinot, R., Grabar, N., et Valette, M. (2003). Application d'algorithmes de classification automatique pour la détection de contenus racistes sur l'Internet. In ATALA, editor, *Actes de la 10ème conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, pages 275–284, Batz sur mer.

Williams, G. (1999). *Les réseaux collocationnels dans la construction et l'exploitation d'un corpus dans le cadre d'une communauté de discours scientifique*. Ph.D. thesis, Université de Nantes.

Xiao, Z. et McEnery, A. (2002). A corpus-based approach to tense and aspect in english-chinese translation. Plenary talk given at the International Symposium on Contrastive and Translation Studies between Chinese and English.

Zanettin, F. (1998). Bilingual corpora and the training of translators. *Meta*, **43**(4), 616–630.

Zanettin, F. (2002). Corpora for translation practice. In E. Yuste-Rodrigo, editor, *Language Resources for Translation Work and Research, LREC 2002 Workshop Proceedings*, pages 10–14.

Liste des tableaux

Table des figures

Table des matières

Annexes

ANNEXE A

La typologie de Biber

A. Tense and aspect markers
    1. Past tense
    2. Perfect aspect
    3. Present tense
B. Place and time adverbials
    1. Place adverbials
    2. Time adverbials
C. Pronouns and pro-verbs
    C1. Personal pronouns
        1. First person pronouns
        2. Second person pronouns
        3. Third person pronouns
    C2. Impersonal pronouns
        1. Pronoun *it*
        2. Demonstrative pronouns
        3. Indefinite pronouns
    C3. Pro-verbs
        1. Pro-verb *do*
D. Questions
    1. Direct WH-questions
E. Nominal forms
    1. Nominalizations
    2. Gerunds
    3. Total other nouns
F. Passives
    1. Agentless passives
    2. *By*-passives
G. Stative forms
    1. *Be* as main verb

2. Existential *there*

H. Subordination

H1. Complementation

1. *That* verb complements
2. *That* adjective complements
3. WH-clauses
4. Infinitives

H2. Participial forms

1. Present participial clauses
2. Past participial clauses
3. Past participial WHIZ deletion relatives
4. Present participial WHIZ deletion relatives

H3. Relatives

1. *That* relative clauses on subject position
2. *That* relative clauses on object position
3. WH relative clauses on subject position
4. WH relative clauses on object position
5. Pied-piping relative clauses
6. Sentence relatives

H4. Adverbial clauses

1. Causative adverbial subordinators: *because*
2. Concessive aderbial subordinators: *although, though*
3. Concessive aderbial subordinators: *if, unless*
4. Other adverbial subordinators (having multiple functions)

I. Adjectives, adverbs and prepositional phrases

I1. Prepositional phrases

1. Total prepositional phrases

I2. Adjectives and adverbs

1. Attribute adjectives
2. Predicative adjectives
3. Total adverbs

J. Lexical specificity

1. Type/token ration
2. Word length

K. Lexical classes

1. Conjuncts

2. Downtoners
3. Hedges
4. Amplifiers
5. Emphatics
6. Discourse particles
7. Demonstratives

L. Modals
1. Possibility modals
2. Necessity modals
3. Predictive modals

M. Specialized verb classes
1. Public verbs
2. Private verbs
3. Suasive verbs
4. *Seem / appear*

N. Reduced forms and dispreferred structures
1. Contractions
2. Subordinator-*that* deletion
3. Stranted prepositions
4. Split infinitives
4. Split auxiliaries

O. Coordination
1. Phrasal coordination
2. Independant clause coordination

P. Negation
1. Synthetic negation
2. Analytic negation: *not*

ANNEXE B

Liste des mots utilisés pour la méthode par vecteurs de termes

diabète	type
insuline	diabétique
aliments	glycémie
poids	repas
glycémique	cas
risque	glucides
alimentation	sucre
boissons	traitement
diabétiques	insulinodépendant
adaptation	doses
index	graisses
produits	régime
physique	présentation
glucose	injection
patients	maladie
sang	pain
cholestérol	prise
obésité	ans
effet	complications
pratique	utilisation
hyperglycémie	fast
alimentaires	food
activité	vie
édulcorants	apport
alcoolisées	jour
santé	insulino

Table B.1 – Cinquante premiers mots les plus fréquents dans le corpus [DIAB_CP], utilisés dans le chapitre 5

Découverte et caractérisation des corpus comparables spécialisés

Lorraine GOEURIOT

Résumé

Les corpus comparables rassemblent des textes dans plusieurs langues qui ne sont pas des traductions mais partagent certaines caractéristiques. Ces corpus présentent l'avantage d'être représentatifs des particularités culturelles et linguistiques de chaque langue. Le Web peut théoriquement être considéré comme un réservoir à corpus comparables mais la qualité des corpus et des ressources qui en sont extraites réside dans la définition préalable des objectifs du corpus et du soin mis à sa composition (les caractéristiques communes aux textes dans le cas des corpus comparables). Notre travail porte sur la constitution de corpus comparables spécialisés en français et japonais dont les documents sont extraits du Web. Nous en proposons une définition et des caractéristiques communes : un domaine de spécialité, un thème et un type de discours (scientifique ou vulgarisé). Notre objectif est de créer un système d'aide à la construction de corpus comparables. Nous présentons d'abord la reconnaissance automatique des caractéristiques communes du corpus. Le thème peut être détecté grâce aux mots-clés utilisés lors de la recherche. Pour le type de discours nous utilisons les méthodes d'apprentissage automatique. Une analyse stylistique sur un corpus d'apprentissage nous permet de créer une typologie bilingue composée de trois niveaux d'analyse : structurel, modal et lexical. Nous l'utilisons ensuite afin d'apprendre un modèle de classification avec les systèmes *SVMlight* et *C4.5*. Ces modèles sont ensuite évalués sur un corpus d'évaluation et permettent de classer correctement plus de 70 % des documents dans les deux langues. Nous intégrons ensuite le classifieur au sein d'une chaîne logicielle d'aide à la construction de corpus comparables implémentée sur la plateforme UIMA.

Mots-clés : corpus comparables, langues de spécialité, analyse stylistique, typologie multilingue, types de discours, apprentissage automatique

Abstract

Comparable corpora are sets of texts written in different languages that are not translations of each other but that share common characteristics. Their main advantage is to be fully representative of linguistics and cultural specificities of their respective language. The Web could theoretically be considered as a comparable corpora source. However, the quality of corpora and of their extracted resources depends on the preliminary definition of corpora and on the carefulness of their compilation (i.e. the definition of common features in comparable corpora). In this thesis, we focus on the compilation of specialized comparable corpora in French and Japanese which documents are extracted from the Web. We propose a definition of these corpora and a set of common features: a specialized domain, a topic and a type of discourse (science or popular science). Our goal is to create a tool to assist comparable corpora compilation. First, we present automatic recognition of common features. Topics can be easily identified with keywords used in Web searches. On the contrary, the detection of the type of discourse needs a wide stylistic analysis. This task is performed over a learning corpus, which leads to the creation of a bilingual typology based on three levels of analysis: structural, modal and lexical. Second, we use this typology to learn a classification model with *SVMlight* and *C4.5*. This classification model is tested over an evaluation corpus. Our test results indicate that more than 70 % of the documents are well classified. Finally, the classifier is integrated into a comparable corpora compilation assistant tool developed on UIMA system.

Keywords: comparable corpora, specialized languages, stylistic analysis, multilingual typology, type of discourse, machine learning