Compilation of Specialized Comparable Corpora in French and Japanese

Lorraine Goeuriot, Emmanuel Morin and Béatrice Daille
LINA - Université de Nantes
France
firstname.lastname@univ-nantes.fr

Abstract

We present in this paper the development of a specialized comparable corpora compilation tool, for which quality would be close to a manually compiled corpus. The comparability is based on three levels: domain, topic and type of discourse. Domain and topic can be filtered with the keywords used through web search. But the detection of the type of discourse needs a wide linguistic analysis. The first step of our work is to automate the detection of the type of discourse that can be found in a scientific domain (science and popular science) in French and Japanese languages. First, a contrastive stylistic analysis of the two types of discourse is done on both languages. This analysis leads to the creation of a reusable, generic and robust typology. Machine learning algorithms are then applied to the typology, using shallow parsing. We obtain good results, with an average precision of 80% and an average recall of 70% that demonstrate the efficiency of this typology. This classification tool is then inserted in a corpus compilation tool which is a text collection treatment chain realized through IBM UIMA system. Starting from two specialized web documents collection in French and Japanese, this tool creates the corresponding corpus.

1 Introduction

Comparable corpora are sets of texts in different languages, that are not translations, but share some characteristics (Bowker and Pearson, 2002). They represent useful resources from which are extracted multilingual terminologies (Déjean et al., 2002) or multilingual lexicons (Fung and Yee, 1998). Comparable corpora are also used in contrastive multilingual studies framework (Peters and Picchi, 1997), they constitute a precious resource for translators (Laviosa, 1998) and teachers (Zanettin, 1998), as they provide a way to observe languages in use.

Their compilation is easier than parallel corpora compilation, because translated resources are rare and there is a lack of resources when the languages involved do not include English. Furthermore, the amount of multilingual documents available on the Web ensures the possibility of automatically compiling them. Nevertheless, this task can not be summarized to a simple collection of documents sharing vocabulary. It is necessary to respect the common characteristics of texts in corpora, established before the compilation, according to the corpus finality (McEnery and Xiao, 2007). Many works are about compilation of corpora from the Web (Baroni and Kilgarriff, 2006) but none, in our knowledge, focuses on compilation of comparable corpora, which has to satisfy many constraints. We fix three comparability levels: domain, topic and type of discourse. Our goal is to automate recognition of these comparability levels in documents, in order to include them into a corpus. We work on Web documents on specialized scientific domains in French and Japanese languages. As document topics can be filtered with keywords in the Web search (Chakrabarti et al., 1999), we focus in this paper on automatic recognition of types of discourse that can be found in scientific documents: science and popular science. This classification tool is then inserted in a specialized comparable corpora compilation tool, which is developped through the Unstructured Information Man-

Proceedings of the 2nd Workshop on Building and Using Comparable Corpora, ACL-IJCNLP 2009, pages 55–63,
Suntec, Singapore, 6 August 2009.

agement Architecture (UIMA) (Ferrucci and Lally, 2004).

This paper is structured as follows. After an introduction of related works in section 2, stylistic analysis of our corpus will be presented in section 3. This analysis will lead to the creation of a typology of scientific and popular science discourse type in specifialized domains. The application of learning algorithms to the typology will be described in section 4, and the results will be presented in section 5. We will show that our typology, based on linguistically motivated features, can characterize science and popular science discourses in French and Japanese documents, and that the use of our three comparablility levels can improve corpora comparability. Finally, we describe the development of the corpus compilation tool.

2 Background

"A comparable corpus can be defined as a corpus containing components that are collected using the same sampling frame and similar balance and representativeness" (McEnery and Xiao, 2007, p. 20). Comparability is ensured using characteristics which can refer to the text creation context (period, author...), or to the text itself (topic, genre...). The choice of the common characteristics, which define the content of corpora, affects the *degree of comparability*, notion used to quantify how two corpora can be comparable. The choice of these characteristics depends on the finality of the corpus. Among papers on comparable corpora, we distinguish two types of works, which induces different choices:

- General language works, where texts of corpora usually share a domain and a period. Fung and Yee (1998) used a corpus composed of newspaper in English and Chinese on a specific period to extract words translations, using IR and NLP methods. Rapp (1999) used a English / German corpus, composed of documents coming from newspapers as well as scientific papers to study alignment methods and bilingual lexicon extraction from non-parallel corpora (which can be considered as comparable);

- Specialized language works, where choice of criteria is various. Déjean et al. (2002) used a corpus composed of scientific abstracts from *Medline*, a medical portal, in English and German. Thus they used documents sharing a domain and a genre to extract bilingual terminology. Chiao (2002) used a corpus of documents of medical domain on a specific topic to work on the extraction of specialized terminologies.

In general language works, documents of comparable corpora often share characteristics like domain or topic. As they are usually extracted from newspapers, it is important to limit them to a certain period to guarantee their comparability.

In specialized corpora, first levels of comparability can be achieved with the domain and the topic. Moreover, several communicative settings appear in specialized language (Bowker and Pearson, 2002): expert-expert, expert-initiate, relative expert to the uninitiated, teacher-pupil. Malrieu and Rastier (2002) specify several levels of textual classification, each of which corresponding to a certain granularity. The first level is *discourse*, defined as a set of utterances from a enunciator characterized by a global topical unit (Ducrot and Todorov, 1972). The second level is *genre*, defined as text categories distinguished by matured speakers. For example, to literary discourse correspond several genres: drama, poetry, prose... Inspired by these communicative settings and textual categories, we choose to distinguish two communicative settings or *type of discourse* in specialized domains: science (texts written by experts to experts) and popular science (texts written to non-experts, by experts, semi-experts or non-experts). This comparability level, the type of discourse, reflects the context of production or usage of the documents, and guarantees a lexical homogeneity in corpora (Bowker and Pearson, 2002, p. 27). Furthermore, Morin et al. (2007) proved that comparable corpora sharing a topic and a type of discourse are well adapted for multilingual terminologies extraction.

Our goal is to create a tool to compile comparable corpora in French and Japanese which documents are extracted from the Web. We investigate automatic categorization of documents according to their type of discourse. This categorization is based on a typology of elements characterizing these types of discourse. To this end, we carry out a stylistic and contrastive analysis (Karlgren, 1998). This analysis aims to highlight linguistically motivated features through several dimen-

sions (structural, modal and lexical), whose combination characterizes scientific or popular science discourse. A specialized comparable corpus can be compiled from a single type of discourse document collection through several steps. Last part of this paper focuses on the automation of these steps using the IBM Unstructured Information Management Architecture (UIMA).

3 Analysis of Types of Discourse

The recognition of types of discourse is based on a stylistic analysis adapted from a deductive and contrastive method, which purpose is to raise discriminant and linguistically motivated features characterizing these two types of discourse. Main difficulty here is to find relevant features which fit every language involved. These features, gathered in a typology, will be used to adapt machine learning algorithms to compilation of corpora. This typology thus needs to be robust, generic and reusable in other languages and domains. Genericity is ensured by a broad typology composed of features covering a wide range of documents characteristics, while robustness is guaranteed with operational (computable) features and treatment adaptable to Web documents as well as texts.

Sinclair (1996) distinguishes two levels of analysis in his report on text typologies: external level, characterizing the context of creation of the document; and internal level, corresponding to linguistic characteristics of document. Because our corpora are composed of documents extracted from the Web, we consider external level features as all the features related to the creation of documents and their structure (non-linguistic features) and call them *structural features*. Stylistic analysis raises several granularity levels among linguistic characteristics of the texts. We thus distinguish two levels in the internal dimension. Firstly, in order to distinguish between scientific and popular science documents, we need to consider the speaker in his speech: the modality. Secondly, scientific discourse can be characterized by vocabulary, word length and other lexical features. Therefore our typology is based on three analysis levels: structural, modal and lexical.

3.1 Structural Dimension

When documents are extracted from the Web, the structure and the context of creation of the documents should be considered. In the framework of Web documents classification, several elements bring useful information: pictures, videos and other multimedia contents (Asirvatham and Ravi, 2001); meta-information, title and *HTML* structure (Riboni, 2002). While those information are not often used in comparable corpora, they can be used to classify them. Table 1 shows structural features.

Feature	French	Japanese
URL pattern	×	
Document's format	×	×
Meta tags	×	×
Title tag	×	×
Pages layout	×	×
Pages background	×	×
Images	×	×
Links	×	×
Paragraphs	×	×
Item lists	×	×
Number of sentences	×	×
Typography	×	×
Document's length	×	×

Table 1: Structural dimension features

3.2 Modal Dimension

The degree of specialization required by the recipient or reader is characterized by the relation built in the utterance between the speaker or author and the recipient or reader[1]. The tone and linguistic elements in texts define this relation. The modalisation is an interpretation of the author's attitude toward the content of his/her assertion. Modalisation is characterized by many textual markers: verbs, adverbs, politeness forms, etc. Presence of the speaker and his position towards his speech are quite different in scientific and popular science discourse. Thus we think modalisation markers can be relevant. For example, the speaker directly speaks to the reader in some popular science documents: "*By eating well, you'll also help to prevent diabetes problems that can occur later in life, like heart disease*". Whereas a scientific document would have a neutral tone: "*Obesity plays a central role in the insulin resistance syndrome, which includes hyperinsulinemia, [. . .] and an increased risk of atherosclerotic cardiovascular disease*".

Most of the modal theories are language dependent, and use description phenomena that are specific to each language. Conversely, the theory exposed in (Charaudeau, 1992) is rather indepen-

[1] Since we work on a scientific domain, we will consider the speaker as the author of texts, and the recipient as the reader.

dent of the language and operational for French and Japanese (Ishimaru, 2006). According to Charaudeau (1992, p.572), modalisation clarifies the position of the speaker with respect to his reader, to himself and to his speech. Modalisation is composed of locutive acts, particular positions of the author in his speech, and each locutive act is characterized by modalities. We kept in his theory two locutive acts involving the author:

Allocutive act: the author gets the reader involved in the speech (ex.: "*You have to do this.*");

Elocutive act: the author is involved in his own speech, he reveals his position regarding his speech (ex.: "*I would like to do this.*").

Each of these acts are then divided into several modalities. These modalities are presented in table 2 with English examples. Some of the modalities are not used in a language or another, because they are not frequent or too ambiguous.

3.3 Lexical Dimension

Biber (1988) uses lexical information to observe variations between texts, especially between genres and types of texts. Karlgren (1998) also use lexical information to characterize text genres, and use them to observe stylistic variations among texts. Thus, we assume that lexical information is relevant in the distinction between science and popular science discourse. Firstly, because a specialized vocabulary is a principal characteristic of specialized domain texts (Bowker and Pearson, 2002, p. 26). Secondly, because scientific documents contain more complex lexical units, nominal compounds or nominal sentences than popular science documents (Sager, 1990).

Table 3 presents the lexical dimension features. Note that these features show a higher language dependency than other dimension features.

Feature	French	Japanese
Specialized vocabulary	×	×
Numerals	×	×
Units of measurement	×	×
Words length	×	
Bibliography	×	×
Bibliographic quotes	×	×
Punctuation	×	×
Sentences end		×
Brackets	×	×
Other alphabets (latin, hiragana, katakana)		×
Symbols		×

Table 3: Lexical dimension features

Dimension	Method
Structural	Pattern matching
Modal	Lexical and lexico-syntactic patterns
Lexical	Lexical patterns

Table 4: Markers detection methods

4 Automatic Classification by Type of Discourse

The process of documents classification can be divided into three steps: document indexing, classifier learning and classifier evaluation (Sebastiani, 2002). Document indexing consists in building a compact representation of documents that can be interpreted by a classifier. In our case, each document d_i is represented as a vector of features weight: $\vec{d_i} = \{w_{1i}, \ldots, w_{ni}\}$ where n is the number of features of the typology and w_{ij} is the weight of the j^{th} feature in the i^{th} document. Each feature weight is normalized, dividing the weight by the total. Documents indexing is characterized by our typology (section 3) and features implementation.

4.1 Features Implementation

In order to get a fast classification system, we privileged for the implementation of our typology features shallow parsing such as lexical markers and lexico-syntactic patterns (method for each dimension is detailed in table 4).

Structural Features We used 12 structural features introduced in section 3.1. Most of these features are achieved through pattern matching. For example, URL patterns can determine is the document belongs to websites such as hospital (`http://www.chu-***.fr`) or universities websites (`http://www.univ-***.fr`), etc. As for paragraphs, images, links, etc., one simple search of `HTML` tags was made.

Modal Features Locutor presence markers in a text can be implicit or ambiguous. We focused here on simple markers of his presence in order to avoid *noise* in our results (high precision but weak recall). Thus we don't recognize all modal markers in a text but those recognized are correct. There are pronouns which are specific to the speech act: for instance, for the elocutive act, the French pronouns *je* (I) and *nous* (we), and the Japanese pronouns 私 (I), 私達 (we)

Feature	Example	French	Japanese
Allocutive modality			
Allocutive personal pronouns	*You*	×	
Injunction modality	*Don't do this*	×	×
Authorization modality	*You can do this*	×	
Judgement modality	*Congratulations for doing it!*	×	
Suggestion modality	*You should do this*	×	×
Interrogation modality	*When do you arrive?*	×	×
Interjection modality	*How are you, Sir?*	×	
Request modality	*Please, do this*	×	×
Elocutive modality			
Elocutive personal	*I, we*	×	×
Noticing modality	*We notice that he left*	×	×
Knowledge modality	*I know that he left*	×	×
Opinion modality	*I think he left*	×	×
Will modality	*I would like him to leave*	×	×
Promise modality	*I promise to be here*	×	×
Declaration modality	*I affirm he left*		×
Appreciation modality	*I like this*	×	
Commitment modality	*We have to do this*	×	
Possibility modality	*I can inform them*	×	

Table 2: Modal dimension features

and 我々 (we). The modalities are also computed with lexical markers. For example, the modality of knowledge can be detected in French with verbs like *savoir, connaître* (know), and in Japanese with the verb 知る (know), with polite form 知っています and with neutral form 知っている.

Lexical Features Some of our lexical criteria are specific to the scientific documents, like bibliographies and bibliographic quotations, specialized vocabulary or the measurement units. To measure the terminological density (proportion of specialized vocabulary in the text) in French, we evaluate terms with stems of Greek-Latin (Namer and Baud, 2007) and suffix characters of relational adjectives that are particularly frequent in scientific domains (Daille, 2000). We listed about 50 stems such as *inter-*, *auto-* or *nano-*, and the 10 relational suffixes such such as *-ique* or *-al*. For Japanese, we listed prefix characteristics of names of disease or symptoms (先天性 (congenital), 遺伝性(hereditary), etc.). These stems can be found in both type of discourse, but not in the same proportions. Specialized terms are used in both type of discourse in different ways. For example, the term "ovarectomie" (*ovarectomy*) can be frequent in a scientific document and used once in a popular science documents to explain it and then replaced by "ablation des ovaires" (*ovary ablation*). Sentences end are specific ending particles used in japanese, for example the particle か is often used at the end of an interrogative sentence.

4.2 Learning Algorithms

Classifier learning is a process which observes features weight of documents classified in a class c or $\bar{c}$ and determine characteristics that a new document should have to be classified in one of these two classes [2]. Given a document indexing, there are some well-known algorithms that can achieve this process (neural network, Bayes classifiers, SVM, etc.) of which Sebastiani (2002) carried out a research about the assemblage and comparison. Applied to a Reuters newswires corpus, these techniques showed variable performances in the usage level of supervised or unsupervised approaches, of the size of the corpus, of the number of categories, etc. We decided to use *SVMlight* (Joachims, 2002) and *C4.5* (Quinlan, 1993), since both of them seem to be the most appropriate to our data (small corpora, binary classification, less than 100 features).

5 Experiments

In this section, we describe the two comparable corpora used and present the two experiments carried out with each of them. The first comparable corpus is used to train the classifier in order to learn a classification model based on our typology (*i.e.* training task). The second comparable corpus is used to evaluate the impact of the classification model when applied on new documents (*i.e.* evaluation task).

[2] This is the binary case. See (Sebastiani, 2002) for other cases.

5.1 Comparable Corpora

The corpora used in our experiments are both composed of French and Japanese documents harvested from the Web. The documents were taken from the medical domain, within the topic of *diabetes and nutrition* for training task, and *breast cancer* for the evaluation task. Document harvesting was carried out with a domain-based search and a manual selection. Documents topic is filtered using keywords reflecting the specialized domain: for example *alimentation*, *diabète* and *obésité* [3] for French part and 糖尿病 and 肥満 [4] for the Japanese part of the training task corpus. Those keywords are directly related to the topic or they can be synonyms (found on thesaurus) or semantically linked terms (found in Web documents collected). Then the documents were manually selected by native speakers of each language who are not domain specialists, and classified with respect to their type of discourse: science (SC) or popular science (PS). Manual classification is based on the following heuristics, to decide their type of discourse:

- A scientific document is written by specialists to specialists.
- We distinguish two levels of popular science: texts written by specialists for the general public and texts written by the general public for the general public. Without distinction of these last two levels, we privileged documents written by specialists, assuming that they may be richer in content and vocabulary (for example advices from a doctor would be richer and longer than forum discussions).

Our manual classification is based on the two previous heuristics, and endorsed by several empirical elements: website's origin, vocabulary used, etc. The classification of ambiguous documents has been validated by linguists. A few documents for which it was difficult to decide on the type of discourse, such as those written by people whose specialist status was not clear, were not retained.

We thus created two comparable corpora:

- [DIAB_CP] related to the topic of *diabetes and nutrition* and used to train the classifier.
- [BC_CP] related to the topic of *breast cancer* and used to evaluate the effectiveness of the classifier.

Table 5 shows the main features of each comparable corpora: the number of documents, and the number of words[5] for each language and each type of discourse.

			# docs	# words
[DIAB_CP]	FR	SC	65	425,781
		PS	183	267,885
	JP	SC	119	234,857
		PS	419	572,430
[BC_CP]	FR	SC	50	443,741
		PS	42	71,980
	JP	SC	48	211,122
		PS	51	123,277

Table 5: Basic data on each comparable corpora

5.2 Results

We present in this section two classification tasks:

- the first one consists in training and testing classifiers with [DIAB_CP], using N-fold cross validation method that consists in dividing the corpus into n sub-samples of the same size (we fix $N = 5$). Results are for 5 partitioning on average;
- the second one consists in testing on [BC_CP] the best classifier learned on [DIAB_CP], in order to evaluate its impact on new documents.

Tables 6 and 7 show results of these two tasks. On both table we present precision and recall metrics with the two learning systems used. On table 6, we can see that the results concerning the French documents are quite satisfactory altogether, with a recall on average of 87%, and a precision on average of 90% as for the classifier *C4.5* (more than 215 documents are well classified from 248 French documents of [DIAB_CP]). The results of the classification in Japanese are also good with the classifier C.4.5. More than 90% of documents are correctly classified, and the precision reaches on average 80%. Some of the lower results can be explained, especially in Japanese by the high range of document genres in the corpus (research papers, newspapers, scientific magazines, recipes, job offers, forum discussions...).

[3] *nutrition*, *diabetes*, and *obesity*

[4] *diabetes* and *overweight*

[5] For Japanese, the number of words is the number of occurrences recognized by ChaSen (Matsumoto et al., 1999)

		French		Japanese	
		Prec.	Rec.	Prec.	Rec.
svml	SC	1.00	0.36	0.70	0.65
	PS	0.80	1,00	0.72	0.80
c4.5	SC	0.89	0.80	0.76	0.96
	PS	0.91	0.94	0.95	0.99

Table 6: Precision and recall for each language, each classifier, on [DIAB_CP]

Table 7 shows results on [BC_CP]. In general, we note a decrease of the results with [BC_CP], although results are still satisfactory. French documents are well classified whatever the classifier is, with a precision higher than 75% and a recall higher than 75%, which represent more than 70 well classified documents on 92. Japanese documents are well classified too, with 76% precision and 77% recall on average, with 23 documents wrong classified on 99. This classification model is effective when it is applied to a different medical topic. This classification model seems efficient to recognize scientific discourse from popular science one in French and Japanese documents on a particular topic.

		French		Japanese	
		Prec.	Rec.	Prec.	Rec.
svml	SC	0.92	0.53	0.90	0.61
	PS	0.64	0.95	0.66	0.98
c4.5	SC	0.70	0.92	0.76	0.70
	PS	0.87	0.56	0.75	0.80

Table 7: Precision and recall for each language, each classifier, on [BC_CP]

6 Comparable Corpora Compilation Tool

Compilation of a corpus, whatever type it is, is composed of several steps.

1. Corpus Specifications: they must be defined by the creator or user of the corpus. It includes decisions on its type, languages involved, resources from which are extracted documents, its size, etc. In the case of specialized comparable corpora, specifications concern languages involved, size, resources and documents domain, theme and type of discourse. This step depends on the applicative goals of the corpus and has to be done carefully.

2. Documents Selection and Collection: according to the resource, size and other corpus criteria chosen during the first step, documents are collected.

3. Documents Normalization and Annotation: cleaning and linguistic treatments are applied to documents in order to convert them into raw texts and annotated texts.

4. Corpus Documentation: compilation of a corpus that can be used in a durable way must include this step. Documentation of the corpus includes information about the compilation (creator, date, method, resources, etc.) and information about the corpus documents. Text Encoding Initiative (TEI) standard has been created in order to conserve in an uniformed way this kind of information in a corpus [6].

A corpus quality highly depends on the first two steps. Moreover, these steps are directly linked to the creator use of the corpus. The first step must be realized by the user to create an relevant corpus. Although second step can be computerizable (Rogelio Nazar and Cabré, 2008), we choose to keep it manual in order to guarantee corpus quality. We decided to work on a system which realizes the last steps, *i.e.* normalization, annotation and documentation, starting from a collection of documents selected by a user.

Our tool has been developed on Unstructured Information Management Architecture (UIMA) that has been created by *IBM Research Division* (Ferrucci and Lally, 2004). Unstructured data (texts, images, etc.) collections can be easily treated on this platform and many libraries are available. Our tool starts with a web documents or texts collection and is composed of several components realizing each part of the creation of the corpus:

1. the collection is loaded and documents are converted to texts (with conversion tools from pdf or html to text mainly);

2. all texts are cleaned and normalized (noise from the conversion is cleaned, all texts are converted into the same encoding, etc.);

[6] http://www.tei-c.org/index.xml

3. a pre-syntactic treatment is applied on texts (segmentation mainly) to prepare them for the following step;

4. morphologic and morpho-syntactic tagging tools are applied on the texts (Brill tagger (Brill, 1994) and Flemm lemmer (Namer, 2000) for French texts, Chasen (Matsumoto et al., 1999) for Japanese);

5. texts are classified according to their type of discourse: we use here the most efficient `SVMlight` classifier. In fact, two corpus are created, on for each type of discourse, then the user can choose one of them. A vectorial representation of each document is computed, then these vectors are classified with the classifier selected.

6. documentation is produced for the corpus, a certain amount of information are included and they can be easily completed by the user.

In reality, this tool is more a compilation assistant than a compilator. It facilitates the compilation task: the user is in charge of the most important part of the compilation, but the technical part (treatment of each document) is realized by the system. This guarantee a high quality in the corpus.

7 Conclusion

This article has described a first attempt of compiling smart comparable corpora. The quality is close to a manually collected corpus, and the high degree of comparability is guaranteed by a common domain and topic, but also by a same type of discourse. In order to detect automatically some of the comparability levels, we carried out a stylistic and contrastive analysis and elaborated a typology for the characterization of scientific and popular science types of discourse on the Web. This typology is based on three aspects of Web documents: the structural aspect, the modal aspect and lexical aspect. From the modality part, this distinction is operational even on linguistically distant languages, as we proved by the validation on French and Japanese. Our typology, implemented using *SVMlight* and *C4.5* learning algorithms brought satisfactory results of classification, not only on the training corpus but also on an evaluation corpus, since we obtained a precision on average of 80% and a recall of 70%. This classifier has then been included into a tool to assist specialized comparable corpora compilation. Starting from a Web documents collection selected by the user, this tool realizes cleaning, normalization and linguistic treatment of each document and "physically" creates the corpus.

This tool is a first attempt and can be improved. In a first time, we would like to assist the selection and collection of documents, which could be realized through the tool. Moreover, we would like to investigate needs of comparable corpora users in order to adapt our tool. Finally, others languages could be added to the system, which represents a quite time-consuming task: a classifier would have to be created so all the linguistic analysis and classification tasks would have to be done again for other languages.

Acknowledgement

This research program has been funded by the French National Research Agency (ANR) through the C-mantic project (ANR-07-MDCO-002-01) 2008-2010. We thank Yukie Nakao for the japanese corpus and linguistic resources.

References

Arul Prakash Asirvatham and Kranthi Kumar Ravi. 2001. Web page classification based on document structure. *IEEE National Convention*.

Marco Baroni and Adam Kilgarriff. 2006. Large linguistically-processed web corpora for multiple languages. In *EACL'06*, pages 87–90. The Association for Computer Linguistics.

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

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

Eric Brill. 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.

Soumen Chakrabarti, Martin van den Berg, and Byron Dom. 1999. Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks (Amsterdam, Netherlands: 1999)*, 31(11–16):1623–1640.

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

Yun-Chuang Chiao and Pierre Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. In *COLING'02*, pages 1208–1212, Tapei, Taiwan.

Béatrice Daille. 2000. Morphological rule induction for terminology acquisition. In *COLING'00*, pages 215–221, Sarrbrucken, Germany.

Hervé Déjean, Éric Gaussier, and Fatia Sadat. 2002. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *COLING'02*.

Oswald Ducrot and Tzvetan Todorov. 1972. *Dictionnaire encyclopédique des sciences du langage*. Éditions du Seuil.

David Ferrucci and Adam Lally. 2004. Uima: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10:327–348.

Pascale Fung and Lo Yuen Yee. 1998. An ir approach for translating new words from nonparallel, comparable texts. In Christian Boitet and Pete Whitelock, editors, *COLING'98*, volume 1, pages 414–420, Montreal, Quebec, Canada.

Kumiko Ishimaru. 2006. *Comparative study on the discourse of advertisement in France and Japan: beauty products*. Ph.D. thesis, Osaka University, Japan.

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

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

Sarah Laviosa. 1998. Corpus-based approaches to contrastive linguistics and translation studies. *Meta*, 43(4):474–479.

Denise Malrieu and Francois Rastier. 2002. Genres et variations morphosyntaxiques. *Traitement Automatique des Langues (TAL)*, 42(2):548–577.

Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, and Yoshitaka Hirano. 1999. Japanese Morphological Analysis System ChaSen 2.0 Users Manual. Technical report, Nara Institute of Science and Technology (NAIST).

Anthony McEnery and Zhonghua Xiao. 2007. Parallel and comparable corpora: What is happening? In Gunilla Anderman and Margaret Rogers, editors, *Incorporating Corpora: The Linguist and the Translator*. Clevedon: Multilingual Matters.

Fiammetta Namer and Robert Baud. 2007. Defining and relating biomedical terms: Towards a cross-language morphosemantics-based system. *International Journal of Medical Informatics*, 76(2-3):226–233.

Fiametta Namer. 2000. Flemm : Un analyseur flexionnel du français à base de règles. *Traitement Automatique des Langues (TAL)*, 41(2):523–548.

Carol Peters and Eugenio Picchi. 1997. Using linguistic tools and resources in cross-language retrieval. In David Hull and Doug Oard, editors, *Cross-Language Text and Speech Retrieval. Papers from the 1997 AAAI Spring Symposium, Technical Report SS-97-05*, pages 179–188.

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

Reinhard Rapp. 1999. Automatic Identification of Word Translations from Unrelated English and German Corpora. In *ACL'99*, pages 519–526, College Park, Maryland, USA.

Daniele Riboni. 2002. Feature selection for web page classification. In Hassan Shafazand and A Min Tjoa, editors, *Proceedings of the 1st EurAsian Conference on Advances in Information and Communication Technology (EURASIA-ICT)*, pages 473–478, Shiraz, Iran. Springer.

Jorge Vivaldi Rogelio Nazar and Teresa Cabré. 2008. A suite to compile and analyze an lsp corpus. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA).

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

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

John Sinclair. 1996. Preliminary recommendations on text typology. Technical report, EAGLES (Expert Advisory Group on Language Engineering Standards).

Federico Zanettin. 1998. Bilingual comparable corpora and the training of translators. *Meta*, 43(4):616–630.