Characterisation of scientific and popular science discourse in French, Japanese and Russian

Lorraine Goeuriot(1), Natalia Grabar(2,3), Béatrice Daille(1)

(1) LINA/Nantes (2) INSERM, UMR_S 872, (3) Health on the Net Foundation
(1) Université de Nantes, France; (2) Université René Descartes, France; (3) SIM/HUG Genève, Suisse
natalia.grabar@biomath.jussieu.fr, {lorraine.goeuriot, beatrice.daille}@univ-nantes.fr

Abstract

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

1. Introduction

Comparable corpora are sets of texts in different languages that are not translations of each others but share some characteristics (Bowker and Pearson, 2002). These characteristics 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 corpus depends on its finality. This affects the *degree of comparability*, notion used to quantify how two corpora can be comparable. We aim to create a tool to assist comparable corpora compilation from the Web. This kind of corpus is increasingly used for multilingual and translingual information retrieval. It overcomes the lack of translated bilingual resources. Moreover, these corpora are richer than parallel corpora because they supply more information with respect to linguistic and cultural particularities of each language. We work on *language for special purposes* corpora (especially coming from scientific domains) in three languages with high linguistic dissimilarities: French, Japanese and Russian. We want to guarantee a high degree of comparability in our corpus. Thus, the first common characteristic that we choose is the document's topic, the second one being the distinction between scientific and popular science discourses, or levels of communication. We base our choice on Ducrot and Schaeffer (1999) discourse definition: *every enunciator's utterances set characterised by a global topic unity.*

In order to automate a part of comparable corpora compilation, it is necessary to recognise automatically those two characteristics. A document's topic can be recognised by the keywords used for its web search, while a document's discourse is harder to identify, and should be supported by a classification system. To characterise scientific and popular science discourses, we compile a training corpus in three languages and analyse it. Inspired from Karlgren (1998), we apply a stylistic and contrastive analysis on the corpus. We then manually create a multilingual typology of science and popular science discourse from this analysis. This typology has three levels: structural, modal and lexical; and is composed of features characterising the discourse of documents (Biber et al., 1998). We use the typology to adapt machine learning algorithms to the corpus. In this paper, our objective is to check if this three levels typology can actually characterise a Web document's discourse, and refine the notion of *corpus comparability*. We may observe the performance of the typology for each one of the three languages composing the corpus.

2. Corpus compilation

Our training corpus is a comparable one, involving French, Japanese and Russian languages. As we work on multilingual information retrieval, we want to garantee a high degree of comparability in our corpus, without weakening cultural and linguistic characteristics of each language. We consider two comparability levels:

- the first comparability level is assured by a common topic in the corpus' documents. The topic we choose, ”diabetes and diet”, belongs to medical domain. This topic refers to a wide public and guarantees a diversified compilation of documents from the Web. From an applicative point-of-view, a common topic implies common linguistic features and vocabulary for each languages.
- since we work on medical domain, two major discourses (or levels of communication) appears: scientific and popular science.

In this section, we present our methodology for corpus compilation and principal characteristics of our corpus.

2.1. Methodology for corpus compilation

We work on a trilingual comparable corpus involving French, Japanese and Russian languages. Documents are extracted from the Web. Corpus compilation consists of three steps:

	French		Japanese		Russian	
	SC	PS	SC	PS	SC	PS
Nb. documents	65	183	119	419	45	150
Nb. words	425 800	267 900			318 596	175126
Nb. characters	2 668 783	2 845 114	493 587	1 154 773	2298306	2 165 768

Table 1: Corpus characteristics

1. Web documents search, according to the chosen topic;
2. Selection of relevant web pages;
3. Classification of these web pages according to their discourse.

Web page search is made using classical Web search tools: (1) National Web search engines; (2) Specialised portals; (3) Link collections. Methods (1) and (2) require keywords. In order to collect a wide range of documents, queries are composed of various combinations of keywords as *diet*, *diabetes* and *obesity* extended with i) synonymous terms found in thesaurus; ii) semantically linked terms found in web documents. In the case of specialised web portals, keywords are adapted.

Among documents gathered by web search, we manually selected relevant documents to the topic. Selected pages were then classified according to their discourse. Manual classification is based on the following heuristics:

- a scientific document is written by specialists, for 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. Our manual classification method is still empirical, thus we did not consider ambiguous documents (unclassified or for which judgements were different) in our training corpus.

2.2. Corpus characteristics

Table 1 presents the main features of the corpus: the number of documents and the number of words for each language and discourse (SC: scientific; PS: popular science).

This corpus is composed of more than 1.5 million words in three languages. Since the number of words is hard to evaluate in a japanese document, we indicated the number of characters. All the documents collected represent more than three alphabets (Cyrillic, Latin, Hiragana, Katakana...) and several charset encodings. Thus we chose to use Unicode, the only encoding allowing Latin, Cyrillic and Japanese alphabets. Our documents are from different formats: classical web formats and other formats (`pdf`, `ps`, `doc`, etc.). Every document is stored in the corpus in his original format and converted into text. All the Web genres (Bretan et al., 1998) are not present in the French corpus, which holds mostly reports and articles (press and scientific). On the contrary, Japanese and Russian corpus holds a large panel of genres (scientific report, articles, cooking recipes or job offerings, etc.).

3. Stylistic analysis of the corpus

Stylistic analysis is a linguistic field with a scientific application in the textual classification domain. In this domain, works are based on automatic classification methods. Most known algorithms are *suppost vector machines* (`SVM`), neural networks, Bayesian classifiers, decision trees (Sebastiani, 2002). We distinguish two types of stylistic analysis. First, the inductive approach consists in the analysis of a corpus, and the detection of correlations specifying similarity classes (which can vary according to the chosen correlations). This method leads to create inductive typologies. In the automatic classification field, this method is called *unsupervised classification*, or *clustering*. The second approach is the deductive one: an analysis is made on a classified corpus to find elements characterising the classes. The characteristics are collected to constitute a typology. In this approach, two methods lead to create a typology: an analysis of the documents one at a time, or a contrastive analysis of documents from two different classes. We choose the deductive and contrastive analysis.

Textual classification algorithms are applied to several kinds of typologies. Most frequent kinds are topic or genre typologies (Bretan et al., 1998). Since works on discourse typologies are less frequent, we adapt deductive and contrastive approaches on genres and themes to build our discourse typology. As documents from our corpus are collected from the web, they present a proper structure that we cannot overlook. Thus, we analysed both structure and content of the documents, our discourse typology is based on these two types of information.

Sinclair worked on text and corpus typologies and introduced the concept of levels in a textual typology (Sinclair, 1996b; Sinclair, 1996a). According to him, it is appropriate to consider two levels of criteria:

- *external criteria*: ”features of the non-linguistic environment or society in which the text occurs”,
- *internal criteria*: ”differentiating features of the language of the texts”.

Since our corpus is made of web documents, we consider external criteria as all the features related to the creation of documents and their structure (non-linguistic features). Internal criteria are the linguistic features of the documents.

Feature	Fr	Jap	Ru
URL pattern	×		
Document's format	×	×	×
Meta tags	×	×	×
Page's title	×	×	×
Pages layout	×	×	×
Pages background	×	×	×
Images	×	×	×
Paragraphs	×	×	×
Item lists	×	×	×
Number of sentences	×	×	×
Typography	×	×	×
Document's length	×	×	×

Table 2: Structural characteristics

Feature	Fr	Jap	Ru
Allocutive modality			
Allocutive personal pronouns	×	×	
Injunction modality	×	×	×
Authorization modality	×		×
Judgement modality	×		
Suggestion modality	×	×	×
Interrogation modality	×	×	×
Interjection modality	×		×
Request modality	×	×	×
Elocutive modality			
Elocutive personal pronouns	×	×	
Noticing modality	×	×	×
Knowledge modality	×	×	×
Opinion modality	×	×	×
Will modality	×	×	×
Promise modality	×	×	×
Declaration modality		×	×
Appreciation modality	×		×
Commitment modality	×		×
Possibility modality	×		×
Interdiction modality			×

Table 3: Modal characteristics

However, stylistic analysis enlights several granularity levels among internal criteria. First, in order to distinguish between scientific and popular science documents, we need to consider the speaker in his speech: the modality. Then scientific discourse can be characterised by the vocabulary used, words length and other lexical features. Therefore our typology is based on three analysis levels:

Structural criteria : textual and graphical structure of documents;

Modal criteria : elements characterising the modality in texts;

Lexical criteria : lexical features of texts.

Feature	Fr	Jap	Ru
Specialized vocabulary	×	×	×
Numerals	×	×	×
unit of measurement	×	×	×
Words length	×		×
Bibliography	×	×	×
Bibliographic quotes	×	×	×
Punctuation	×	×	×
Sentences end		×	
Brackets	×	×	×
Other alphabets (latin, hiragana, katakana)		×	×
Symbols		×	

Table 4: Lexical characteristics

Structural features (table 2) gathers mainly graphical elements of documents (format, images, tables...), and structural elements, by dint of HTML tags (paragraphs, item lists, titles...). All of these features fits to the three languages involved in our study.

Modal characteristics (table 3) correspond to linguistic elements characterizing modality in documents, *ie* speaker attitude in his own speech. These features are directly inspired from Charaudeau (1992) and adapted to our corpus (Krivine et al., 2006; Nakao, 2008). Among the acts of locution from Charaudeau (1992), we kept those which were operational (easy to identify automatically). For example, opinion modality can be detected in French using verbs like *penser* (to think), *paraître* (to appear), *sembler* (to seem). Most of these features appear in the three languages, but some (judgement, interdiction) are specific to one of them, because the typology instantiation is based on isolated studies of each languages.

Table 4 presents lexical features. Contrary to modal and structural features, lexical features present a high reliance on the language, like the use of hiragana and katakana characters in Japanese texts or Latin alphabets in Russian texts. Some of the features are specific to scientific articles style, like bibliographies, bibliographic quotes, specialized vocabulary or units of measurement.

4. Automatic discourse classification

Our objective is to apply machine learning algorithms to our three levels typology, in order to classify automatically documents according to their discourse: science or popular science. In this section, we present machine learning algorithms and our evalutation methods.

4.1. Learning machine algorithms

A machine learning system uses documents as vectors, where elements of vectors are values for each criteria of the typology. Vector's length corresponds to the frequency (rough or balanced) of each criteria in a document. Starting from a learning corpus, where documents are divided into two classes (science and popular science), and a list of criteria, machine learning algorithms create a classification model. This model is then check proofed with other data. There are several textual classification algorithms

		French		Japanese		Russian	
		Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
SVMlight	Science	1,00	0,36	0,20	0,41	1,00	0,52
	Popular science	0,80	1,00	0,72	0,80	0,75	1,00
C4.5	Science	0,89	0,80	0,13	0,12	0,50	0,38
	Popular science	0,91	0,94	0,84	0,86	0,74	0,82

Table 5: Precision and recall of classification in each category with classifiers SVM light and C4.5 for every language

		French		Japanese		Russian	
		Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
SVMlight	Structural features	0,90	0,67	0,59	0,71	0,85	0,74
	Modal features	0,60	0,50	0,50	0,49	0,28	0,50
	Lexical features	0,91	0,75	0,58	0,53	0,98	0,97
C4.5	Structural features	0,85	0,85	0,41	0,44	0.62	0,68
	Modal features	0,89	0,91	0,39	0,44	0,34	0,68
	Lexical features	0,85	0,85	0,47	0,45	0,45	0,52

Table 6: Results for each feature category

(neural networks, Bayasian classifier, support vector machines, etc.). Sebastiani (2002) gathered and compared them: working on a Reuters news corpus, these algorithms have different performances depending on the supervised or unsupervised method, corpus size, number of classes, number of features... We chose to use support vector machines with *SVMlight* (Joachims, 2002) and decision trees with *C4.5* (Quinlan, 1993). As we want a fast classification system, we analyse superficially our documents and their content. Hence, we privileged simple techniques based on lexical or lexico-syntactic patterns to implement our features. Thus our categorization system is based on the detection of patterns: based on `HTML` tags for graphical features or on linguistic elements characterizing modality or lexical patterns.

4.2. Evaluation

As we only have a learning corpus, we need to split it into two parts: a training one and a testing one. To do this, we use *N-fold cross validation* (Cornuéjols and Miclet, 2002). This method consists in splitting our corpus into N equal parts. Then the i_th fold is chosen for the testing part, the other $n-1$ folds are used to learn a model. Once the model is tested, this operation is reiterated, i varying from 1 to N. We set $n = 5$. For each iteration, we use 80% of our documents to learn a classification model, the other 20% are used to test the models created. Results presented in section 5. are averages of the 5-fold cross validation.

We use precision and recall metrics to evaluate our results:

- Recall is the number of well classified documents in a class C divided by the number of documents pertaining to that class;
- Precision is the number of well classified documents in C divided by the number of documents classified in C.

5. Results and discussion

We applied *SVMlight* and *C4.5* to our corpus. Classification results are presented in tables 5 and 6. In table 5, we can see that our classification model is better for popular science discourse than scientific one, whatever the language and the classifier are. Results for French documents are satisfying with an average recall of 87%, and an average precision of 90% whatever the classifier is (which represents more than 200 well classified documents on the 250 of the corpus). Classification results for Japanese documents are good for popular science documents, but quite bad for scientific ones. At last, classification of Russian documents gives good results with *SVMlight*, with a recall higher than 75% and a precision of 87%. Low results for classification of Japanese and Russian scientific documents can be explained by the high proportion of popular science documents in the corpus (see table 1). Vector normalisation is performed to overcome this ratio but a larger learning corpus would provide more information, represent larger range of scientific documents and generate a fuller language model. Classification results of French scientific documents are surprising: occurrences are almost twice numerous in scientific corpus, despite documents are less numerous. Results for Japanese documents are in average lower than the others, it can be explained by the high genre diversity in the corpus. Thus it is harder to characterise a discourse with a stable set of features. We may not have enough manual classification criteria and they may have to be refined.

Table 6 presents results of each features category for each language and classifier. Each category seems to be relevant and important for the classification. In fact, whatever the classifier is, results in each language show that it is possible to classify more than a half of the documents using only features of one category. However, none of the category can be distinguished from the others within this test, and best categories varies according to the classifier used. With *SVMlight*, structural and lexical categories seem to be better in the three languages. With *C4.5*, the best category varies according to the language: modality for French, lexicon for Japanese and structure for Russian. Most relevant features are: URL pattern, delocutive pronouns and narrative sentences for French; documents length, elocutive pronouns and politeness maker in sentences end for Japanese; documents title, images and specialized vocabulary for Russian.

6. Conclusion and perspectives

Starting with a trilingual comparable corpora composed of Web documents in French, Japanese and Russian, we made a contrastive stylistic analysis and created a typology for science and popular science web documents. Our typology is based on three aspects of web documents: structural, modal and lexical. This typology, implemented through machine learning algorithms support vector machines (*SVMlight*) and decision trees (*C4.5*) gives good results. Thus we can estimate the discourse of a web document. Furthermore, these results show that each level of the typology is relevant to characterise the discourse. In fact, each of them provide satisfying results, and their combination improves the results. Web documents discourse can therefore be characterised through these three aspects. We note that some of the features appear in three languages, so we assume that our typology may be universal. There seems to be recurrent elements appearing in documents from three different languages characterising science or popular science discourse, and our typology seems to be carriable from a language to another.

One of the major limit of our work seems to come from the limited size of the corpus, especially for Japanese and Russian scientific documents. Moreover, low results for Japanese scientific documents can be explained by a high genre diversity in the corpus. Through this observation, we wonder about our corpus composition and our manual classification. Looking to hierarchical classification of Malrieu and Rastier (2002), a distinction into discourse and genre would help us refining our results. Addind a genre distinction in the corpus, our classes would be homogeneous and would guarantee a higher degree of comparability in our comparable corpora. Finally, our binary classification may not be legitimate, we think it would be more interesting to consider science and popular science classes as a continuum. This would lead us to evaluate a scientific degree instead of a class belonging.

7. Acknowledgement

This research program has been funded by CNRS as part of the TCAN program for 2004-2006 with NII and INaLCO research laboratories. We thank Estelle Dubreil, Sonia Krivine and Masaru Tomimitsu for their participationin comparable corpus compilation.

8. References

Douglas Biber, Susan Conrad, and Randi Reppen. 1998. *Corpus Linguistics - Investigating Language Structure and Use*. Cambridge University Press.

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

Ivan Bretan, Johan Dewe, Anders Hallberg, Niklas Wolkert, and Jussi Karlgren. 1998. Web-specific genre visualisation. In *Proceedings of the 3rd World Conference on the WWW and Internet*.

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

Antoine Cornuéjols and Laurent Miclet. 2002. *Apprentissage artificiel : Concepts et algorithmes*. Eyrolles.

Oswald Ducrot and Jean-Marie Schaeffer. 1999. *Nouveau dictionnaire encyclopédique des sciences du langage*. Seuil.

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.

Sonia Krivine, Masaru Tomimitsu, Natalia Grabar, and Monique Slodzian. 2006. Relever des critères pour la distinction automatique entre les documents médicaux scientifiques et vulgarisés en russe et en japonais. In Piet Mertens, Cédrick Fairon, Anne Dister, and Patrick Watrin, editors, *Actes de la 13ème conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, volume vol.1, pages 522–531, Leuven, apr.

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

Yukie Nakao. 2008. Multilingual modalities for specialised languages. In *Proceedings of the LREC workshop "Multilingual and Comparative Perspectives in Specialized Language Resources"*. T*o be published*.

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

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

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

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