Mining Expert Comments on the Application of ILO Conventions on Freedom of Association and Collective Bargaining

Gilbert Ritschard, Djamel A. Zighed, Lucio Baccaro,
Irini Georgiou, Vincent Pisetta and Matthias Studer

No 2007.02

Cahiers du département d'économétrie
Faculté des sciences économiques et sociales
Université de Genève

Novembre 2007

Département d'économétrie
Université de Genève, 40 Boulevard du Pont-d'Arve, CH -1211 Genève 4
http://www.unige.ch/ses/metri/

Mining Expert Comments on the Application of ILO Conventions on Freedom of Association and Collective Bargaining

November 6, 2007

Gilbert Ritschard[1], Djamel A. Zighed[2], Lucio Baccaro[3,4], Irini Georgiou[1,3], Vincent Pisetta[2], and Matthias Studer[1]

[1] University of Geneva, Switzerland
gilbert.ritschard@unige.ch, matthias.studer@metri.unige.ch
[2] University of Lyon 2, France
abdelkader.zighed@univ-lyon2.fr, v-pisett@mail.univ-lyon2.fr
[3] International Institute of Labour Studies, IILS (ILO), Geneva
[4] MIT, Cambridge MA
lucio.baccaro@gmail.com

Abstract. This paper explains how text mining was used within the context of a research project on social dialogue regimes, jointly undertaken by the University of Geneva, the University of Lyon 2 and the International Institute of Labour Studies of the International Labour Organisation (ILO). The research project, which was made possible through the generous support of the Geneva International Academic Network Foundation (GIAN), is seeking to provide a better understanding of the structural determinants (e.g., economic, social, cultural and institutional), as well as the socio-economic outcomes of "social dialogue regimes." An important part of the research is based on the analysis of the reports of the Committee of Experts on the Application of Conventions and Recommendations (CEACR). The CEACR is one of the main bodies within the ILO supervisory system. It is responsible for supervising the application by member States of ILO Conventions and recommendations. Its observations concerning the progress made by members States in the implementation of ratified ILO Conventions are published in its reports. Text mining was used for the purpose of extracting useful information from these reports in semi-automatic way. This paper discusses the text mining approach that was followed, the different steps of the mining process and presents a synthetic analysis of the results obtained.

Keywords: Text mining, legal text, text coding, social dialogue, violation of international standards.

Table of Contents

1 Introduction

The ILO Committee of Experts on the Application of Conventions and Recommendations (CEACR) publishes each year a volume of about 800 pages containing multiple observations on the application of Conventions by ratifying countries. For the purpose of our research project on Social Dialogue Regimes supported by the GIAN Foundation (Baccaro et al., 2003), our research team focused on comments regarding Convention 87 (Freedom of Association and Protection of Right to Organise) and 98 (Right to Organise and Collective Bargaining) formulated in the period 1990 to 2002. The aim was to develop a procedure which enables the automatic identification of the main problems singled out by the Committee in the application of the standards set forth in these two Conventions.

The large number of texts that had to be examined made it necessary to chose to take advantage of text mining methods to retrieve automatically the relevant information from the CEACR comments. In the following lines we give a rough overview of how we proceeded. We started by listing the main types of violations (which we referred to as 'key concepts') of the standards contained in the two Conventions. At that stage, we availed ourselves of the legal expertise of some members of the research team, as well as of members of the ILO International Labour Standards Department. Then, we read and labelled manually a sample of CEACR reports on Conventions 87 and 98. Manual labelling consisted in assigning the relevant key concepts to portions of text, in which a violation was reported by the Committee. Using this labelled learning sample, we then resorted to machine learning algorithms to generate prediction rules that would allow us to predict automatically labels (types of violations) from the texts. By applying these rules to the entire corpus of about 1200 available texts — including those that were not in the learning sample — we obtained predictions of the types of violations referred to in each report. These predictions take the form of probabilities. In other words, for each text, the text mining methodology produces a probability as to whether the text in question reports about a given type of violation.

Like any statistical and data mining method, the text mining process we followed is not error free. It's main value, as shown below, lies in the fact that it is rather infrequent for it to predict that the Committee reports a given type of violation for a country in a given year when that is not the case, or reversely, to fail to identify an existing type of violation. Nonetheless, possible errors, though infrequent, may have serious consequences if somebody is interested in the situation prevailing in a specific country. For this reason, it is believed that at least at the present stage the developed methodology should be used to assist in and speed up the process of expert analysis, rather than substitute for it. It is also worth mentioning that the main objective of our research on Social Dialogue Regimes is not to provide an individual analysis of the regime prevailing in each country, but rather to produce a synthetic analysis of possible relationships between different constructs. Hence, the predictions provided by the text mining will merely serve as material for quantitative analyses. As long as the overall error remains reasonable, synthetic results should only be marginally affected. For instance, we will establish that when there is a problem with excessive restrictions regarding the way trade unions can be organised there is very often also a problem with unduly limited rights to conduct industrial actions. Such synthetic results should hold despite possible errors in the assessment of the situation in one or more specific countries.

The remainder of the paper is organised as follows: Section 2 describes the setting of our research on Social Dialogue Regimes (Baccaro et al., 2003). Section 3 provides a short introduction to text mining, while Section 4 specifies the nature of the CEACR reports to be mined and the kind of information we seek to extract by means of text mining. In Section 5 we detail the process we followed for obtaining quantitative representations of unstructured CEACR comments. These quantitative representations are used in Section 6 for deriving prediction rules from a learning sample of labelled comments. Some details on the used classification tree approach are also given in that Section. The output of the text mining process is discussed in Section 7 using synthetic analyses of the obtained predictions. Essentially, we investigate the relationship between different types of violations by means of multiple correspondence analysis, clustering analysis and statistical implicative analysis Finally, the concluding Section 8 summarises the present study and stresses the scope and limits of this experiment with text mining of legal texts.

2 The Social Dialogue Regimes Research Framework

The text mining methodology articulated in this paper is part of a broader research project aimed at understanding the structural determinants (e.g., economic, social, cultural and institutional), as well as the socio-economic outcomes of "social dialogue regimes," i.e., socio-political regimes in which workers have freedom to establish organisations of their own choosing, negotiate collectively over working conditions, and participate through their associations in the design and implementation of policies that affect their lives.

At present, not much is known about the conditions in which social dialogue regimes emerge and reproduce themselves over time, and about the socio-economic outcomes (including macro-economic performance) associated with them, in spite of the topic's importance. Two opposing views dominate this research field. On one hand, it is often argued that questions about worker rights, collective bargaining and negotiated policy-making can only be meaningfully addressed after countries reach a crucial stage of economic development. In the meantime, these considerations should not be allowed to interfere with the developing countries' main source of comparative advantage, namely low labour costs. On the other hand, it is also argued that unionisation, collective bargaining, and tripartism (discussions among government, workers and employers) should be encouraged and perhaps even imposed by national and international policy-making authorities. According to this second view, negotiated regulation is not just desirable for its ethical and political properties. It also has beneficial economic effects, because it contributes to rule out the possibility of self-defeating approaches to economic development, based on low labour costs/low productivity strategies. Far from being the consequence of economic development, as in the previous view, labour rights, organisations and institutions are in this case regarded as its precondition.

This debate is of crucial importance, but has so far remained at a very general, almost ideological level. This state of affairs appears to be linked with a basic methodological problem. Good measurements of key constructs like worker rights, industrial relations processes, or negotiated policy-making are not available, particularly for developing countries. Consequently, the researchers' capacity to apply the tools of social science to the questions laid out above is necessarily limited.

The research project on Social Dialogue Regimes seeks to contribute to filling the aforementioned methodological gap. In fact, the first step in the project is the production of cross-country indicators of social dialogue in about 50 countries — both developed and developing — at two points in time (i.e., 1990 and 2000). Social dialogue indicators fall into three categories: indicators of associational and collective bargaining rights; indicators of industrial relations structures and processes (unionisation and collective bargaining); and indicators of negotiated policy-making or tripartism. The aim of the latter is to capture the extent to which economic and social policies are co-determined by governments and the social partners (that is, trade unions and employer organisations), rather than being implemented by governments alone. Multiple methods will be used in the production of the indicators, and the data mining procedure illustrated in this paper is one of them. Another source of information the project will avail itself of is questionnaires addressed to country experts.

The second step in the project combines quantitative analysis with qualitative analysis. First, statistical models — incorporating the newly-developed indicators as both dependent and independent variables — will seek to uncover the broad structural determinants of labour rights, industrial relations processes, and tripartism, as well as some of the socio-economic outcomes associated with them. Second, four in-depth case studies will explore the particular causal mechanisms through which particular factors combine with each other to produce outcomes, and to investigate country cases that appear to defy general trends. The selection of case studies will be linked to the results of the quantitative analysis.

This research project promises to generate important information on the desirability of particular labour market institutions and on the conditions in which these institutional configurations are sustainable. This kind of information is crucial for both national and international policy-makers, and particularly for the ILO's mission of "Decent work for all," as it seeks to identify policy solutions that reconcile economic well-being and development with workers' dignity, rights and protections.

3 Text Mining

Before presenting in detail the text mining process followed for the purpose of our research, it is useful to provide a brief review of the general principles text mining.

3.1 What Is Text Mining?

Text mining (Feldman and Dagan, 1995; Kodratoff, 1999; Fan et al., 2006) refers to the process of analysing text to extract information that is useful for particular purposes (Witten and Frank, 2005, pp 351-356). It is supposed to be more than just finding documents or pages containing a given keyword — which is what a simple indexing or search engine would do. For instance, if we are looking for text commenting on violations of the principle of trade union pluralism, we will not be satisfied with just texts containing the keyword "trade union pluralism", but we may want to consider also all terms or expressions more or less related to this notion.

As opposed to numerical data, text data are essentially unstructured. Synonymy (different expressions with same meaning) and polysemy (different meanings for a same expression), among others, make them hard to analyse in an automatic way and necessitate heavy pre-processing. Also the focus may be on quite different aspects of texts ranging from their content to their structure, but also their representation, as well as the semantic and ontology of the concepts used. As we will see in sub-section 3.2, pre-processing deals mainly with the latter, i.e. text and concept representation.

The scope of text mining is very large and covers a great variety of tasks and application fields. For instance, it may consist in the statistical analysis of word frequencies, or of sentence structures and lengths, in studying relationships between texts, in automatically summarising articles or documents (Saravanan et al., 2003), in text categorisation (Sebastiani, 2002) by resorting either to clustering (e.g. for storing similar texts together) or supervised classification techniques (e.g. for detecting spam), in technological watch (Liu et al., 2001), in building ontologies, i.e. discovering typical terminologies of a domain and organising them into conceptual hierarchies (Gruber, 1993), in retrieving concepts from texts (Riloff and Hollaar, 1996), etc. Each of these tasks relies on specific methods that often share, however, the pre-processing phase.

3.2 Text Pre-Processing

Whatever the task, all text mining methods require important pre-processing for transforming the essentially unstructured text data into a suitable structured representation for further automatic processing. By structured representation we mean a representation where each useful notion is uniquely and unambiguously defined so that we can surely rely on the counts of its occurrences.

There are basically two main ways of representing a text: through n-grams and as a bag of words. The former ignores the meaning of the words and considers each subsequence of say 3 letters — 3-gram — that can be found in the words as a countable characteristic (Damashek, 1995; Mayfield and McNamee, 1998). The second (Salton et al., 1992, 1996) retains each different observed word as a characteristic and focuses essentially on its frequency in the text and among the texts. The latter approach is best suited for our supervised classification purpose where the semantic content of the text is of primary importance.

Now, texts contain a huge number of different words. Some of them may have a same or similar meaning (synonyms), may have a context dependent meaning (polysemy), or, as in the case of function or stop words (the, to, from, or, and, ...), will clearly be useless for discrimination purposes. The general practice is then to reduce the number of descriptors by dropping useless stop words and by merging synonyms into equivalence classes, using for instance the electronic lexicon WordNet (Fellbaum, 1998).

A first step for solving ambiguities is tagging words grammatically, which can be done automatically using for instance freely available tools such as Brill (Brill, 1995) or TreeTagger (Schmid, 1994). The grammatical tag permits indeed to distinguish for example between the noun, verb or adjective usage of the word “trade”, or the conjunction, verb or adjective usage of the word “like”. This grammatical tagging will also pinpoint stop words that could be dropped from the list of descriptors.

To avoid bothering with the various inflected forms of nouns, verbs and adjectives, other often applied pre-processing operations are lemmatisation and stemming (Plisson et al., 2004). The former consists in retaining just the base form — e.g. the infinite of a conjugated verb — of each encountered word, and the latter in extracting the lemma — the root — of each word. This can again be done almost automatically with freely available tools such as TreeTagger (Schmid, 1994).

In our case, since we wanted to facilitate the processing of new additional texts by legal experts with no experience in these pre-processing steps, we opted for an approach that will not require any pre-processing in its application phase. Therefore, we choose to not lemmatise the texts, and resorted to grammatical tagging only in the learning phase in order to facilitate the extraction of the useful terminology.

4 CEACR Comments and the Searched Information

The aim of this section is to clarify the nature of the CEACR reports to be mined. We first explain the role of the CEACR in the ILO supervisory system and then specify what kind of information we intend to extract from these reports.

4.1 ILO Supervisory Process

Both in terms of legitimacy and expertise the CEACR is, along with the Committee on Freedom of Association (CFA), one of the most competent body to report on the implementation of labour standards in the various ILO member

States. The CEACR, an independent body of prominent experts in the fields of labour law and industrial relations, is authorised to monitor the extent to which State Parties to ILO Conventions — namely, states that have ratified the relevant ILO Conventions — apply domestically the standards envisaged therein. It carries out its supervision at regular intervals on the basis of the reports submitted by State Parties themselves, and the comments made by workers' and employers' organisations (Articles 22 and 23, ILO Constitution). Although not empowered to interpret ILO Conventions, the CEACR has often assumed an interpretative role, shedding light into complex questions concerning the interpretation of the provisions of ILO Conventions. This has resulted to a comprehensive and coherent body of case law, to which also the CFA adheres, whenever it is confronted with questions of a technical nature.

The CEACR is not a judicial body empowered to give authoritative interpretation of ILO Conventions, the only body bestowed with such competence being, in accordance with Article 37, paragraph 1, ILO Constitution, the International Court of Justice. On this issue, the Committee has reiterated a number of times that: "[I]n order to carry out its function of determining whether the requirements of Conventions are being respected, the Committee has to consider and express its views on the content and meaning of the provisions of Conventions and to determine their legal scope, where appropriate," CEACR, General Report, ILC 77th Session, 1990, p. 8. In response to the objections raised regarding the Committee's interpretative function, the Committee has stated: "[I]nsofar as [the Committee's] views are not contradicted by the International Court of Justice, they are to be considered as valid and generally recognised", noting that "the acceptance of the above considerations is indispensable for the maintenance of the principle of legality and, consequently for the certainty of law required for the proper functioning of the International Labour Organisation." CEACR, General Report, ILC 77th Session, 1990, p.8.

The Committee on Freedom of Association (CFA), a tripartite organ composed from among the ranks of the ILO Governing Body, deals exclusively with complaints concerning violations of trade union rights, irrespective of whether the State, against which a representation or complaint is brought, has ratified the respective ILO Conventions. It thus examines the implementation of ILO standards on an ad hoc basis.

Both the CEACR and the CFA have a long-standing record in monitoring the implementation of international labour standards. This has enabled the supervisory bodies to follow the evolution in the implementation of labour standards in the various member States in a continuous and uninterrupted fashion and has also provided the International Labour Office, the Organisation's secretariat, with a vast amount of information on economic and social rights. In addition, both the CEACR and the CFA conduct their work according to standard procedures.

At the same time, however, one cannot lose sight of certain limitations of the information generated by the ILO supervisory mechanism. Especially with regard to the CEACR, the focus of the present analysis, it must be borne in mind that the Committee's assessment is made on the basis of reports submitted by governments, comments made by workers' and employers' organisations, reports from other bodies as well as other relevant information. This means that, for the most part, the Committee performs its examination on a documentary basis, a fact that sometimes raises reservations as to the accuracy and credibility of the information submitted. This is often amplified by the fact that governments do not respond at all or do not respond promptly to the Committee's questions and requests, as well as the fact that workers' and employers' organisation do not always make use of their right to comment on government reports. When a member-State fails to submit reports, in most cases the Committee refrains from making any substantive finding and reiterates its previous questions and requests. Similarly, in the absence of updated information, the Committee recalls its previous observations and calls upon the government in question to submit all necessary information. In cases where it has available to it information from sources other than the government in question, for example organisations of workers' or employers', non-governmental organisations or other international bodies, the Committee normally communicates that information to the respective government and requests it to either confirm or refute it, or provide additional clarification. The Committee's long-standing record of monitoring the enforcement of international labour standards in the various ILO members States allows it to have a comprehensive account of the particular difficulties facing different State Parties in the application of ILO Conventions and hence to accurately appreciate the available information. This short description shows that, although the system functions more efficiently when governments cooperate, there are yet other mechanisms, which enable the Committee to follow the developments in member States that do not fulfil their reporting obligations. In effect, it is possible that the system does not always provide every piece of information that is relevant to the issues under examination in the timeliest fashion but it is very unlikely that it provides inaccurate information.

It follows from the above, that, notwithstanding the described caveats, the ILO supervisory bodies remain by virtue of their composition, mandate and working methods the most competent institutions in the field of labour rights, and thus are, more than any other international body, in a position to give guarantees of objectivity and accuracy in their assessment of national labour laws and practices in the various ILO member States.

The current research focuses on the CEACR country reports on Conventions 87 and 98 and does not consider the information contained in the reports of the Committee on Freedom of Association. Confining oneself to the reports of the CEACR at this stage serves purely methodological purposes, and is not to be construed as discounting the importance of the reports of the Freedom of Association Committee — or, possibly, other sources — as a valuable information source concerning State compliance with freedom of association.

4.2 Objective of Text Mining within the Project

As already stated in the introduction, the aim of text mining in the present context is to help us identify the nature of issues raised by the CEACR regarding the application of Conventions 87 and 98. Indeed, what we want to know is what types of violations of these two Conventions does the Committee identify in its reports. Using a priori knowledge, we categorised the possible violations in the form of a list of key concepts — types of violations — for each of the two conventions (Table 1). We are thus requested to perform a concept retrieval task for which, since we have a limited list of predefined key concepts, a supervised text categorisation approach is well suited. Drawing on a sample of text labelled with the concerned key concepts, we will learn rules for predicting the key concepts from the text content. The difficulty is that key concepts are not associated to single terms or expressions. The task is thus to find which words or co-occurrences of words best characterise texts reporting a given violation.

The key concepts listed in Table 1 were derived from the more detailed list of 27 key concepts originally defined in Georgiou (2006). The description of these 27 original key concepts is recalled in the appendix, where the interested reader will also find the correspondence table between the retained key concepts and the original list. The first 17 key concepts concern Convention 87 and the remaining 11 Convention 98. Those finally retained for the analysis reported here, were obtained by merging together some of the initial ones.

We are not interested in the outcome — key concept prediction — of text mining per se. Rather, this output will serve as basic material for building some of the missing measurement indicators mentioned in Section 2.

5 The Chosen Text Representation

For the purpose of our analysis, we decided to represent the CEACR comments by means of a limited set of descriptor concepts. This section describes in some details how these concepts were defined in a three-step partially automated process. The first step is closely linked with the extraction of useful terms from the texts. We begin by commenting on this terminology extraction process.

5.1 Extracting the Useful Terminology

The terminology that could be sued for predicting violations reported in the Committees's observations includes not only single words, but also composite expressions such as “trade union” or “right to organise”. It is then essential to find and list the terms useful for the analysis.

Several tools can be used for this. Some of them, such as Xtract (Smadja, 1993), ATR (Frantzi et al., 2000), Lexter (Bourigault and Jacquemin, 1999) proceed automatically either by comparison with a pre-specified lexicon or by seeking frequent sub-sequences of words. Others, such as Exit (Heitz et al., 2005), are semi-automatic and require a domain expert to guide the process.

Table 1. Retained key concepts

	Key concept (violation regarding)	Associated variable name
	Convention 87	
1	Right to life and physical integrity	(not observed)
2	Right to liberty and security of person / Right to a fair trial	(not observed)
3	Right to establish and join workers' organisations	v3_establish_join_tu
4	Trade union pluralism	v4_tu_pluralism
5	Dissolution or suspension of workers' organisations	(not observed)
6	Election of representatives / Eligibility criteria	v6_election_represent
7	Organisation of activities / Protection of property / Financial independence	v7_admin_indep_orga_act
8	Approval and registration of workers' organisations	v8_register_tu
9	Restrictions on the right to industrial action	v9_right_indus_action
	Convention 98	
1	Anti-union discrimination	w1_antiunion_discrim
2	Acts of interference	w2_interference
3	Solidarist association	(not observed)
4	Promotion of free and voluntary collective bargaining	(not observed)
5	Right to collective bargaining	w5_right_coll_bargain
6	Designation of the bargaining partner / Most representative trade union	w6_most_repr_tu_bargain_part
7	Level and scope of collective bargaining	(not observed)
8	Negotiable issues and substantive outcomes of collective bargaining / Permissible restrictions	w8_restr_subst_outcomes
9	Approval and registration of collective agreements / Compulsory arbitration in the context of collective bargaining	(not observed)

The latter are best suited when, as in our case, we do not have access to a lexicon of the considered specialised language. Since we had the possibility to interact with legal experts, the latter approach was retained and we extracted the useful terminology with the aid of the Exit software.

The input data provided to Exit is the grammatically tagged text (the set of all comments merged into a single file). We then select the useful terms in an iterative way. First, we chose successively among pairs of a given type —

noun-noun, noun-adjective, adjective-noun, verb-noun, noun-verb, etc. — that satisfy a minimal frequency criterion those that the expert considers relevant for the analysis. For example, "worker organisation" and "national security" are two retained pairs, the former being of the noun-noun type and the latter of the "adjective-noun" type.

A grammatical tag is assigned to each new retained term according to rules that could be changed by the user. For instance, adjective-noun terms such as "national security" are automatically tagged as noun. Then by iterating the process we single out terms that include themselves previously defined terms. We get thus terms composed of more than two words such as "minimum level of service".

5.2 Descriptor Concepts

As already stated, there is a huge number of different terms — words and composite expressions — used in the CEACR comments and it is not convenient to use all of them as text descriptors. We therefore, decided to represent texts through a small number of descriptor concepts that:

- Characterise the conceptual content of the text;
- Are useful for predicting the issues — violation of key concepts — reported in the observations.

The first step in the learning phase consists in characterising these descriptor concepts. A first entirely statistical possibility (Kumps et al., 2004) would be to seek the words that best discriminate the key concepts we want to predict, and then to group them according to their co-occurrences. Lemmatisation would be necessary in that case.

However, since we had the possibility to interact with legal experts, we preferred to rely on a linguistic approach. Such an approach where terms — words and expressions — are grouped according to both their statistical characteristics and the similarity of their meaning, provide concepts that are semantically better founded.

Thus, the approach followed consists in three steps carried out on the overall corpus: i) a preliminary set of concepts is built during the terminology extraction with Exit; ii) this preliminary set and the concept definitions are refined through an extensional induction process (Kodratoff, 2004) with the legal experts; and iii) the experts' amended list is once again compared with the text content for a final coherence check.

The preliminary concept set is obtained in a semi-automatic way by starting the term extraction process with a high threshold, which provides a relatively short list of terms. Those terms may be considered as initial representatives of the main conceptual axes that can be found inside the texts. We obtain a starting set of concepts after possibly grouping terms with similar semantic meaning. Then, we repeat the process by lowering successively the minimal frequency threshold. At every iteration, we get additional terms and then assign each one of them to the most appropriate preexisting concept. In case there is no reasonable preexisting concept with which the new term could be associated, a new concept is created. At the end of the terminology extraction we get our preliminary list of concepts, where each concept is characterised by its list of associated terms.

This preliminary list of descriptor concepts serves then as a starting list for the experts who may either confirm the relevance of the concepts or change them to fit their overall knowledge of the domain. The preliminary list is thus transformed into an expert's amended list of concepts.

In order to increase even further the coherence of the amended descriptor concepts, we carried out some additional checking. Indeed, we observed that the overall corpus of CEACR comments contains some infrequent terms that clearly belong to one of the retained descriptor concepts. Ignoring them would undoubtedly be a source of errors. The goal of the additional checking is to browse the corpus for such relevant but infrequent terms. More specifically, for terms already associated to a concept, we look for the presence in the corpus of:

– Alternative inflection forms of the term;
– Extended terms obtained by inserting one or more words in the term;
– Synonyms of the term.

For example, the term "call a strike" is frequent in the corpus and was detected as representing the strike action descriptor concept. The much less frequent expression "calling a strike", which is just a variant inflection of the former, was not detected however. Likewise, the infrequent expression "calling of a strike", which can be derived from the term "calling a strike" by inserting the word "of" in it, clearly denotes also some reference to a strike action. The search of such alternative forms is easily done by browsing the terms found with regular expressions. For example, using the two strong words "call" and "strike", all three aforementioned terms were found with the PCRE regular search expression:[5]

```
"/[^;\.]*call[^;\.\,]{0,45}strike[^;\.]*/i"   .
```

As for synonyms, a lexicon such as the online WordNet (Fellbaum, 1998) may be useful for usual terms. For a specialised corpus such as the one formed by our legal texts, it is most helpful to ask experts in the domain. This is what was done in our analysis. Good sense may also prove useful. For example, we noticed in the reports that experts used independently and equivalently the terms "trade union" and "workers organisation". Hence, each time a concept definition list included a term such as "registration of a trade union", we augmented, when it made sense, the list with "registration of a workers organisation", even when this new expression was infrequent in the corpus.

The final list of descriptor concepts is given in Table 2 and examples of their list of associated terms are given in appendix.

[5] The regular expression searches the text for expressions in which the word "call" is preceded by any sequence of characters other than a semi-column or a dot, the word "strike" is followed by any sequence of characters other than a semi-column or a dot, and the two words are separated by any sequence of at most 45 characters other than a semi-column, a dot or a comma.

Table 2. Retained descriptor concepts

	Descriptor concept
	Convention 87
1	Life and physical integrity
2	Liberty and security of persons
3	Property and financial independence
4	Service
5	Pluralism
6	Election
7	Opinion and expression freedom
8	Restrictions on trade union activities
9	Trade union approval
10	Industrial action
11	Essential service
12	Arbitration
13	Strike action
14	Union establishment limitations
15	Specific workers
16	Number of workers
17	Supervision
	Convention 98
1	Anti-union discrimination
2	Hindering union activities
3	Solidarity and welfare association
4	Promotion of collective bargaining
5	Obstacles to collective bargaining
6	Bargaining partner
7	Level and scope of collective bargaining
8	Negotiable issues and collective agreements
9	Compulsory arbitration in collective bargaining

The designing of the descriptor concepts is clearly a crucial phase of our text mining process. It is also time-consuming and requires clever tuning through individual interventions from both the domain experts and the text mining experts. Furthermore, because of these multiple personal interventions, the resulting descriptor concepts remain somewhat subjective. Improvement and systematisation of the process is possible and would here be necessary. It requires, however, an access to a detailed ontology of the concerned legal domain which does not yet exist. The designing of such an ontology that puts together the characteristic terminology of the domain, organises it in terms of concepts and sub-concepts, and also describes the interrelation between concepts would then be our next development priority.

5.3 The Quantitative Text Representation

Having now defined our descriptor concepts, how can we use them for representing the CEACR comments? Indeed, we have to assign for each document (comment) a load on each concept. A classical way is to use the *tf* × *idf*, which is the term frequency (*tf*) — indeed the term count — in the document weighted by the inverse of the document frequency (*idf*), the document frequency being the number of documents in which the concept has been observed (Salton and Buckley, 1988). The general idea of this *tf* × *idf* is that a term — a concept in our case — is characteristic of a text when it is frequently mentioned in it (high *tf*) and only few other documents mention it (high *idf*). A term frequency is defined for each document i and each concept j, while a document frequency exists for each concept j, but does not depend on i. Hence the notations tf_{ij} and idf_j, for $i = 1, \ldots, d$, and $j = 1, \ldots, c$ with d the number of documents and c the number of concepts. Formally, the inverse document frequency is defined as $\log(d/d_j)$, where d_j is just the number of documents mentioning concept j. The *tf* × *idf* weight of concept j in a document i, is then

$$w_{ij} = tf_{ij}\, idf_j = tf_{ij} \log\left(\frac{d}{d_j}\right) \ . \tag{1}$$

With this formulation, the lengthier a document i the greater chances it has to have large tf_{ij}'s and hence important weights. To avoid this size effect and ensure that there is equal chance that all documents be retrieved, Salton et al. (1992) propose the length normalised form

$$\tilde{w}_{ij} = \frac{tf_{ij} \log(d/d_j)}{\sqrt{\sum_{j=1}^{c} [tf_{ij} \log(d/d_j)]^2}} \ . \tag{2}$$

The document frequency and hence the idf_j is the same for all documents $j = 1, \ldots, c$. It changes proportionally the term frequency of concept j among all documents. As long as the non-normalised formula (1) is used, such scaling of the variables is a concern for distance based methods such as principal component or clustering. It does not, however, affect the fit of regression models nor does it impact on the growing of induced decision trees that we will use later on. For the normalised form (2), the *idf* term is, nonetheless, a concern because of its impact on the normalisation term in the denominator.

Beside these technical and interpretational considerations, the fundamental question is indeed the relevance of the normalisation for our objectives. Our position is that what matters is the absolute place devoted to a given concept in a comment whatever other issues the comment addresses. In that sense, the normalised *tf* × *idf* is not useful in our setting. For the same reason, we prefer using as tf_{ij} the number of occurrences of concept j inside text i rather than the proportion of occurrences. In other words, we consider that the importance of a concept in a text is reflected by its number of occurrences independently of the document's length.

To summarise, we can say that our raw text data are transformed into quantitative data by assigning to each CEACR comment the vector of its *tf* × *idf* of the retained descriptor concepts, the *tf* × *idf* being proportional to the number of occurrences of the concept in the comment. Our text data set can thus be put in the form of a classical quantitative data table as illustrated in Table 3, which exhibits an extract of the data for comments on the application of Convention 87. With this coding, each descriptor concept can be considered as a potential quantitative predictor for the kind of problems reported by the comment.

Table 3. Extract of data representing comments in terms of descriptor concepts

CEACR Comment	Descriptor Concepts								
	c1	c2	c3	c4	c5	c6	c7	c8	· · ·
Algeria 1991	0	0	0	0	2.75	0	0.8	0	· · ·
Antigua and Barbuda 1991	3.0	1.53	0	0	0	0	0	0	· · ·
Argentina 1991	0	0	0	0	20.59	2.39	0.8	0	· · ·
Bangladesh 1991	1.0	0.77	2.35	1.24	0	1.59	5.59	0	· · ·
Belgium 1991	0	0	0	0	0	0	0.8	0	· · ·
Bolivia 1991	0	0.77	0	1.24	1.37	4.77	0	0	· · ·
Bulgaria 1991	0	0	0	0	2.75	0	0	0	· · ·
Burkina Faso 1991	0	0	0	0	0	0	0.8	2.19	· · ·
· · ·									

6 Learning Process

Through the previous steps, i.e. extracting useful terms, organising them into a limited number of relevant descriptor concepts and finally measuring the importance devoted to each descriptor concept by each CEACR comment with the *tf* × *idf* weight, we were able to code the comments numerically. What remains now is to learn the prediction rules.

This learning phase requires a learning sample of texts — comments — previously labelled in accordance with the type of violation they report. The labelling was done by a legal expert for 78 out of 671 CEACR observations concerning Convention 87, and for 101 out of 509 texts concerning Convention 98. The labels are represented by a set of ℓ 0-1 indicator variables v_k, $k = 1, \ldots, \ell$ that take value 1 when the text mentions violation k, and zero otherwise. To avoid confusion between the two conventions considered, we shall use when necessary w_k instead of v_k for denoting label indicator variables for Convention 98. Remember that the violations we are interested in correspond to the key concepts listed in Table 1 page 11.

Using these learning samples the aim is to find rules for predicting each key concept (violation) from the quantified descriptor concepts. We then consider successively each key concept in turn, and build the prediction rule for it. Letting c_j denote the *tf* × *idf* of the jth descriptor concept, we look for each k for a rule such as

$$\hat{v}_k = f_k(c_1, \ldots, c_c) \ . \tag{3}$$

Since our texts are numerically coded, classical supervised statistical or machine learning techniques may be used. These include among others logistic regression, classification trees, neural networks, Bayesian networks, support vector machine (SVM) and k nearest neighbours (k-NN). We use here induced classification trees, which produce usually good classification results and have the advantage of being easily applicable, of detecting automatically interaction effects of the predictors and of providing easily interpretable rules.

Classification trees are grown by seeking, through recursive splits of the learning data set, some optimal partition of the predictor space for predicting the outcome class, i.e. whether the comment does or does not report a violation of type k. Each split is done according to the values of one predictor — descriptor concept —. The process is greedy. At first step, it tries all predictors to find the "best" split using, for quantitative predictors as those we face here (the concept *tf* × *idf*'s), an automatic local optimal discretisation. Then, the process is repeated at each new node until some stopping rule is reached. This requires a local criterion to determine the "best" split at each node. The choice of the criterion is the main difference between the various tree growing methods that have been proposed in the literature, of which CHAID (Kass, 1980), CART (Breiman et al., 1984) and C4.5 (Quinlan, 1993) are perhaps the most popular. We used here the CHAID method which seeks at each step the split that maximises the significance of the independence Chi-Square computed for the table that cross classifies the resulting nodes with the response variable.

Figure 1 shows the tree grown for violation 7 — restrictions on the organisation of trade union activities — using Exhaustive CHAID (the improved CHAID method by Biggs et al., 1991) in Answer Tree (SPSS, 2001) with a significance threshold of 5%, the Bonferroni correction, a minimal leaf size of 10 and a minimal parent node size of 30. The descriptor used is whether the comment explicitly refers to property and financial independence. The optimal threshold for this first binary split is 0, i.e. between at least one mention of this concept or none. Texts with a *tf* × *idf* less or equal to this value, i.e. those without any reference to property and financial independence, are further split using the "election" descriptor concept. This second split is into three groups, the two optimal thresholds being 0 and 3.18.

The tree has 4 terminal nodes, which are called *leaves*. We associate to each of them a rule taking the form *condition* ⇒ *conclusion*. The condition is defined by the path from the root node to the leaf, and the conclusion is, for a classification tree, the most frequent class in the leaf. We may, however, also take the probability distribution as conclusion, i.e. in our case the probability that, given its coding, the text reports difficulties due to restrictions on trade union activities. The four probability rules are given in Table 4.

A similar tree is grown for each type of violation, which results in 6 sets of rules for Convention 87, and 5 sets for Convention 98. Some violations (v_1, v_2

Table 4. Rules for restrictions on the organisation of trade union activities (v_7)

Rule	Condition	$p(v_7 = YES)$
R1	$\textit{property and financial independence} = 0$ **and** $\textit{election} = 0$	15.79%
R2	$\textit{property and financial independence} = 0$ **and** $\textit{election} \in]0, 3.18]$	52.63%
R3	$\textit{property and financial independence} = 0$ **and** $\textit{election} > 3.18$	100%
R4	$\textit{property and financial independence} \geq 0$	100%

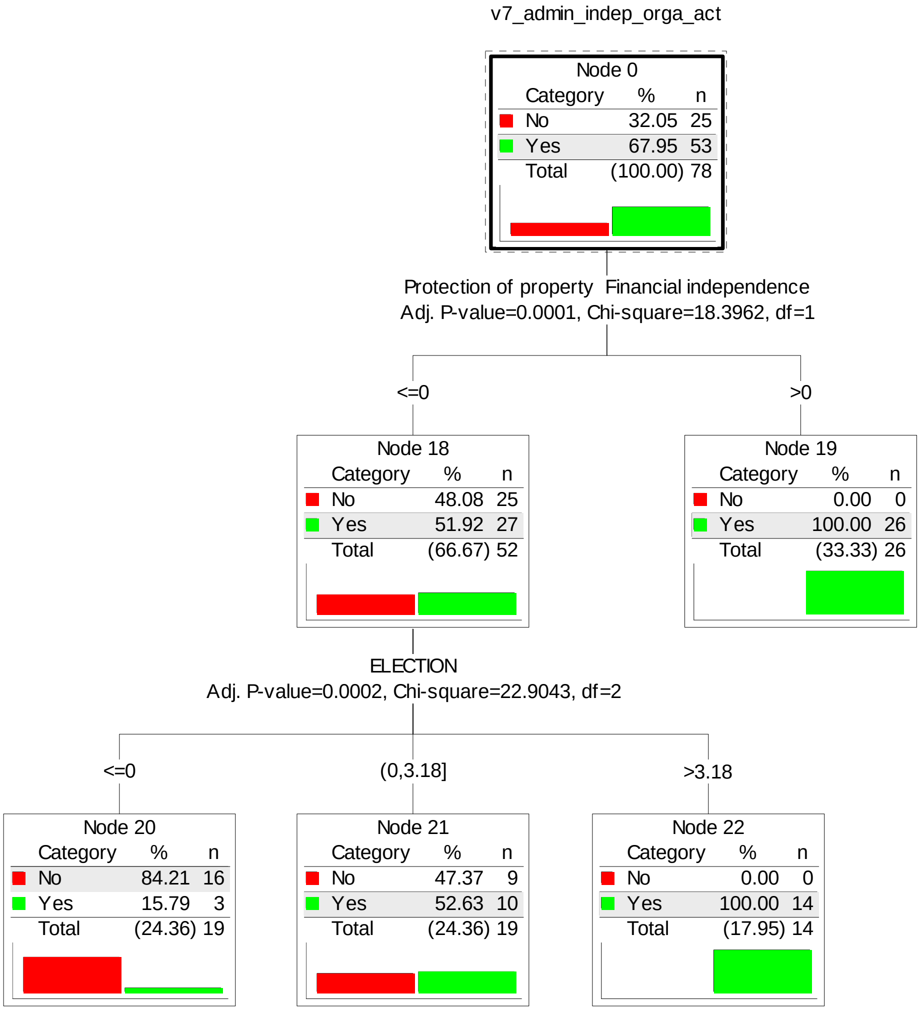

Fig. 1. Induced tree for v_7: Restrictions on organisation of trade union activities

Table 5. Error rates, Convention 87

Key Concept (violation)	Learning error rate	cross-validation error rate	std err	Test sample (size 21) number of errors
v_3	14.10%	n.a.*	n.a.*	3
v_4	5.13%	5.13%	2.50%	0
v_6	12.82%	14.1%	3.94%	4
v_7	15.38%	n.a.*	n.a.*	7
v_8	7.69%	7.69%	3.01%	4
v_9	2.56%	2.56%	1.79%	2

*Cross-validation is not available for v_3 and v_7, because first split is enforced.

and v_5 for instance for Convention 87), are not covered by any comment in the learning sample, and no tree is grown for them. In two cases, we did not rely on the mere statistical criterion and forced the algorithm to split at the first step using the second best variable that seemed theoretically better sounded from our knowledge base.

The classification performance of each tree may be evaluated by means of its classification error, i.e. the percentage of cases which are misclassified by the derived classification rules. Table 5 for Convention 87 and Table 7 for Convention 98 show learning error rates (i.e. rates computed on the learning sample) and 10-fold cross-validation error rates with their standard error. Table 5 gives in addition the number of errors on a small test sample of size 21 of comments about application of Convention 87.

Table 6 exhibits some additional useful indicators for Convention 87. Column ‘True positives’ gives the number of comments classified as reporting a violation of type k that effectively report it, and column ‘Predicted positives’ the total number of comments classified as reporting the violation. For key concept v_7, for example, 50 out of 57 comments classified as reporting the violation actually report it. The number of true and predicted negatives is also shown. Table 6 gives the percentage of the 78 comments that report on the relevant key concept and the percentage of comments that are classified as reporting the key concept.

Table 6. False Positives, False Negatives, Recall and Precision, Convention 87

Key	Positives		Negatives		% with key concept			
Concept	true	predicted	true	predicted	reported	predicted	Recall	Precision
v_3	30	32	37	46	50.0%	41.0%	76.9%	93.8%
v_4	29	31	45	47	39.7%	39.7%	93.5%	93.5%
v_6	35	38	33	40	53.8%	48.7%	83.3%	92.1%
v_7	50	59	16	19	67.9%	75.6%	94.3%	84.7%
v_8	29	30	43	48	43.6%	38.5%	85.3%	96.7%
v_9	57	59	19	19	73.1%	75.6%	100.0%	96.6%

Table 7. Error rates, Convention 98

Key Concept	Learning	cross-validation	
(violation)	error rate	error rate	std err
w_1	4.95%	6.93%	2.53%
w_2	3.96%	3.96%	1.94%
w_5	7.92%	11.88%	3.21%
w_6	12.87%	18.81%	3.89%
w_8	12.87%	16.83%	3.72%

For v_7 again, we may check that $57 = 73.1\% \times 78$, are classified as reporting the violation, while there is actually a total $53 = 67.9\% \times 78$ reporting v_7. The 'Recall' is the percentage of this total that is classified as reporting the violation — true positives —, e.g. $94.7\% = 50/53$ for v_7. The 'Precision' is the ratio of the number of true positives on the number of predicted positives, e.g. $87.7\% = 50/57$ for v_7. Table 8 exhibits similar figures for Convention 98.

These results are quite good when compared with those obtained with other classifiers. For instance, we experimented in Pisetta et al. (2006) with support vector machine (SVM) as well as with neighbouring graphs, which consist in assigning to any unlabelled case the label value (0 or 1) that is most often observed among the say 3 learning cases that are the most similar to it. These methods did not produce significantly better results, while producing much less explicit rules.

Table 8. False Positives, False Negatives, Recall and Precision, Convention 98

Key	Positives		Negatives		% with key concept			
Concept	true	predicted	true	predicted	reported	predicted	Recall	Precision
w_1	40	44	56	57	40.6%	43.6%	97.6%	90.9%
w_2	35	39	62	62	34.7%	38.6%	100.0%	89.7%
w_5	48	53	45	48	50.5%	52.5%	94.1%	90.6%
w_6	12	12	76	89	24.8%	11.9%	48.0%	100.0%
w_8	33	41	55	69	37.6%	40.6%	86.8%	80.5%

7 Text Mining Output

This section presents an analysis of the raw text mining results that were obtained from the application of the rules described above. We first explicate how individual predictions are obtained for each comment. Then, we explain how predictions for country-year comments are transformed into prediction for each country. Afterwards, we present a synthetic analysis of the text mining results using three complementary methods of data analysis. By clustering together countries with similar violation pattern, we build a typology of countries. Typical patterns characterising each cluster provide enlightening knowledge about how violations organise themselves into logical legal systems. Trough a factorial multiple correspondence analysis, we identify then the key dimensions underlying theses systems. Finally, we rely on tools from statistical implicative analysis for gaining additional insight on the relationships between kinds of violations stemmed from the cluster and correspondence analyses.

We focus our presentation on the results for Convention 87 that are the most interesting and only devote short comments to those for Convention 98.

7.1 Violation Prediction for Comments

A small piece of software, developed in Java by Bellal (2006), permits to any non text mining expert to compute the *tf* × *idf* of the descriptor concepts for the whole set of relevant CEACR comments. The program requires two text files as input: A first one with the set of CEACR comments that we want to explore, and a second one with the list of terms that define the descriptor concepts. The generated results are provided in table form with each line corresponding to a comment and each column to a key concept — violation. The comments are labelled with the concerned country name and year, and each cell (i, j) of the table contains the *tf* × *idf* value of descriptor concept j for comment i.

The probability rules derived from the trees grown for each type of violation — key concept — can, at least with the Answer Tree software that we used, easily be exported as SQL queries or in SPSS syntax. It is then rather easy to obtain the probability predictions by copying the *tf* × *idf* values computed by the aforementioned piece of software into SPSS for instance.

7.2 From Comment Predictions to Country Predictions

The labour standard violations in the application of Conventions 87 and 98 in which we are interested display high levels of inertia. When a problem is identified by the CEACR, it usually takes a while before the concerned country takes appropriate measures to correct it if at all. Likewise, the fact that the Committee of Experts reports a violation in a given year, does not necessarily mean that that violation occurred for the first time in that year but rather that the Committee of Experts decided, for various reasons, to comment on it that year. Furthermore, the CEACR does not report every year for all countries. It usually examines and reportson a given country every two years. The absence of comments on a country in a given year t is an issue that requires some attention. Indeed, the absence of comments means either that:

- There is no problem in that country in that particular year t, or that
- A problem exists in year t, but that problem is not reported in that year.

In addition, there is also the case of countries that did not exist as independent states in 1991, or that underwent a major political transformation after this date (as in the case of states that were previously part of the former Soviet Union). In these cases, we do not have CEACR comments over the whole period under consideration.

To overcome these difficulties, we proceeded as follows: We limited our analysis to Committee's comments in the period between 1997-2002. We assigned a zero probability to each type of violation for the years in which there was no comment. Then, we assigned to each country the *maximum* of the predicted probability over the period from 1997 to 2002.

For classification purposes, we decided to predict that there is a violation of type j in a given country when the predicted probability assigned to that country exceeds 80%. Table 9 shows, for each type of violation, the percentage of countries for which we predict the violation. The percents are computed against the total number of concerned countries. For each convention, the sum of the percentages exceeds 100%. This means that several types of violations may be predicted for a same country.

In general, there seem to be fewer violations of Convention 98 than of Convention 87. Indeed, the percentage of countries concerned by each type of violation is generally lower for Convention 98. Moreover, 60% of the countries do not show any violation of Convention 98 against only 34% for Convention 87.

7.3 Typology of Countries

In order to better understand existing legal systems, we carried out a cluster analysis of statistical units — countries in our case — with the intention to build a typology. The aim of this analysis is to cluster together countries for which the text mining indicates similar situations in terms of violations of the

Table 9. Proportion of countries for each predicted type of violation

Key concept		Percentage
	Convention 87	
v_3	Right to establish and join workers' organisations	36.9%
v_4	Trade union pluralism	22.7%
v_6	Election of representatives / Eligibility criteria	31.9%
v_7	Organisation of trade union activities	23.4%
v_8	Approval and registration of workers' organisations	28.4%
v_9	Restrictions on the right to industrial action	61.0%
	Convention 98	
w_1	Anti-union discrimination	25.6%
w_2	Acts of interference	15.4%
w_5	Right to collective bargaining	18.6%
w_6	Designation of bargaining partner / representative trade union	10.3%
w_8	Negotiable issues and substantive outcomes of collective bargaining / Permissible restrictions	6.4%

relevant ILO convention. In doing so, we seek to identify systems. We carried out the analysis with SPSS separately for Conventions 87 and 98, since not all countries have ratified both conventions.

Let us explain in details the analysis for Convention 87. We proceeded in two stages: First, we extracted the principal factors of a multiple correspondence analysis (Greenacre, 1993; Lebart et al., 2000) of the predicted violations. Second, we identified the groups of countries by running an agglomerative hierarchical clustering with the Ward criteria (Anderberg, 1973; Jobson, 1992; Lebart et al., 2000) on the six factorial score variables. The plot of the intra class inertia against the number of clusters shows us that the most significant part of reduction of intra class inertia is achieved with a partition into four groups. The reduction is, with this four cluster solution, of more than 60%.

Table 10 gives for each group the percentages of members concerned by each type of violation according to text mining predictions. We observe that the four clusters are relatively balanced, 36% of the cases falling in the larger group and 13% in the smaller.

Table 10. The 4 clusters for Convention 87

		Clusters					Cramer's
Type of violation		1	2	3	4	Total	v
v_3	Right to establish and join workers' organ.	0%	38.9%	51.1%	81.5%	36.9%	0.64
v_4	Trade union pluralism	0%	100.0%	0.0%	51.9%	22.7%	0.85
v_6	Election of representatives /Eligibility criteria	3.9%	33.3%	22.2%	100.0%	31.9%	0.74
v_7	Organisation of trade union activities	2.0%	0.0%	11.1%	100.0%	23.4%	0.89
v_8	Approval and registration of workers' organ.	0%	44.4%	26.7%	74.1%	28.4%	0.60
v_9	Restrictions on the right to industrial action	0%	88.9%	95.6%	100.0%	61.0%	0.94
Percentage of cases in cluster		36.2%	12.8%	31.9%	19.1%	100%	
Mean number of violations		0.06	3.06	2.07	5.07	2.04	

Cramer's v values indicate that the decomposition into four groups is strongly related to each type of violation. We observe that the percentages of countries concerned with each type of violation strongly vary from one group to the other. The violations v_9, v_7, v_6 and v_4 show the strongest association with the selected partition.

The interpretation of the first group is straightforward. It includes countries in there are no violations. We name this cluster "No violation". The next three groups are more interesting, since each characterises a legal system, that is a system of relationships between types of violations of the rights stipulated in Convention 87.

The second group includes countries where trade-union pluralism is non-existent but which workers' organisations do not experience excessive constraints in the organisation of their activities. These countries also seem to impose slightly more frequently excessive restrictions in the procedures of registration of trade unions. This is a logical conclusion since restrictions on registration procedures may be seen as an indirect way of setting up a trade-union monopoly. There is thus a form of control of associations of workers. We label this cluster as "Pluralism restrictions".

The third group is formed by countries imposing restrictions on the right to industrial action (violation v_9) and, to a lesser extent, on the right to establish and join trade unions. Though other violations may also occur, their percentages are generally much lower than that of the main violations. Thus, there is a form of control on the possibilities of resorting to industrial action. We name this cluster "Action restrictions".

Finally, in the last group, we observe high percentages for all violations. The average number of violations accounts for five on a maximum of six. If one compares this group with the general situation, violations v_6 (election of representatives) and v_7 (administrative independence and organisation of trade-union activities) seem to be the most prevalent. Thus, in this group of countries, governments seem to exercise a form of control on the administration and internal activities of trade unions. Generally speaking, this form of control seems to go hand in hand with other types of violations which, with the exception of violation v_4 (trade-union pluralism), yield the highest percentages in this group. We name this cluster "Organisational and action restrictions".

In short, the cluster analysis identifies the four following types of legal systems:

Cluster 1	(Clu_NoViolation):	No violation
Cluster 2	(Clu_Pluralism):	Pluralism restrictions
Cluster 3	(Clu_Action):	Action restrictions
Cluster 4	(Clu_Organisational):	Organisational and action restrictions.

7.4 Factorial Multiple Correspondence Analysis

The previous clusters were derived from the whole set of 6 factors of the multiple correspondence analysis of the 6 (binary) violation indicator variables. The first factors represent the main dimensions of the data space and it is instructive to have a careful look at them. Figures 2 and 3 show how the different types of violations are distributed respectively in the space of the first two dimensions and in that of the first and third dimensions. Solid circles indicate the position of the presence of a given type of violation and empty circles that of their absence. For interpretation purposes some supplementary points were added on the plots. Diamonds indicate the average position of the clusters and squares the average position of regions to which the concerned countries belong.

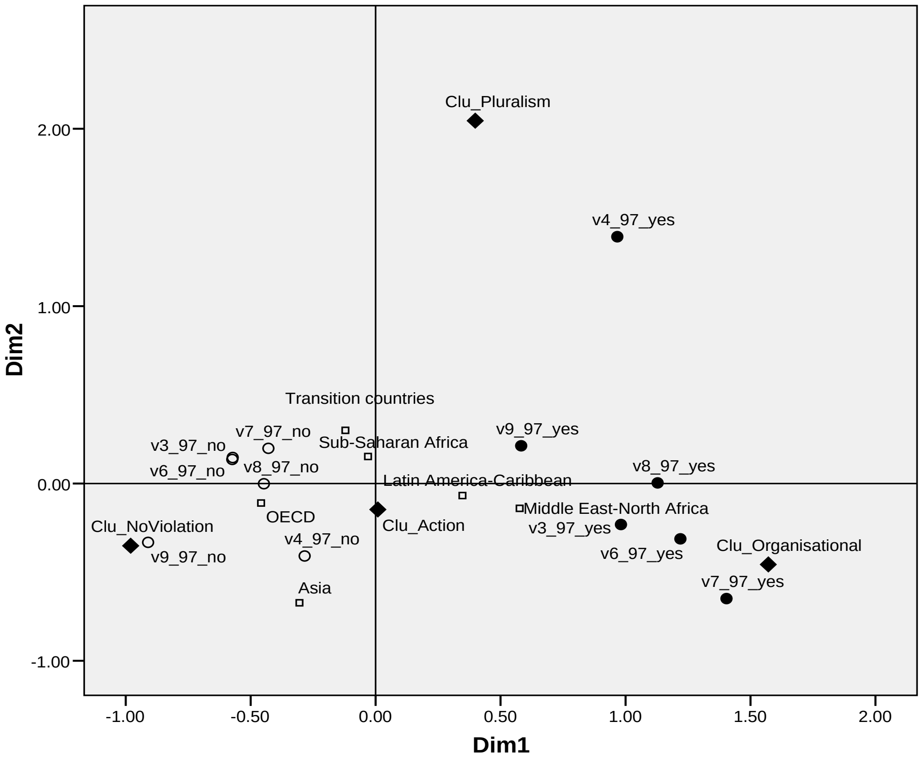

Fig. 2. Violations in the 2 first factor space

The first dimension — horizontal axis in both figures — opposes the mention (on the right) to no mention of violations. Thus, the cluster without violation is the point on the far left and is opposed to the cluster with the most violations, i.e. "Organisational and action restrictions". This dimension is very strongly correlated with the total number of different violations reported for the country ($r = 0.998; p < 0.0001$). It thus indicates some "order" of appearance of violations. Violation v_9, the most frequent, is on the left whereas the less frequent violation v_7 is on the right. This dimension reproduces the greatest proportion of the observed variation (52.9%), while the second dimension explains a less important part of it (14.1%). It highlights an opposition between violations v_3, v_6 and v_7 on the one hand, and violation v_4 on the other. This interpretation is confirmed by the clear opposition on this second axis between cluster "Pluralism" and the other three clusters. In Figure 3, we observe that the third dimension (11.8% of the variation) opposes cluster "Action restrictions" to the others.

It is interesting to examine the average position of the various regions. Two reasons lead us to think that there is some relationship between types of violations and regions. First, certain aspects of regional political cultures can impact upon the types of legal systems in the countries of a given region. Second, it

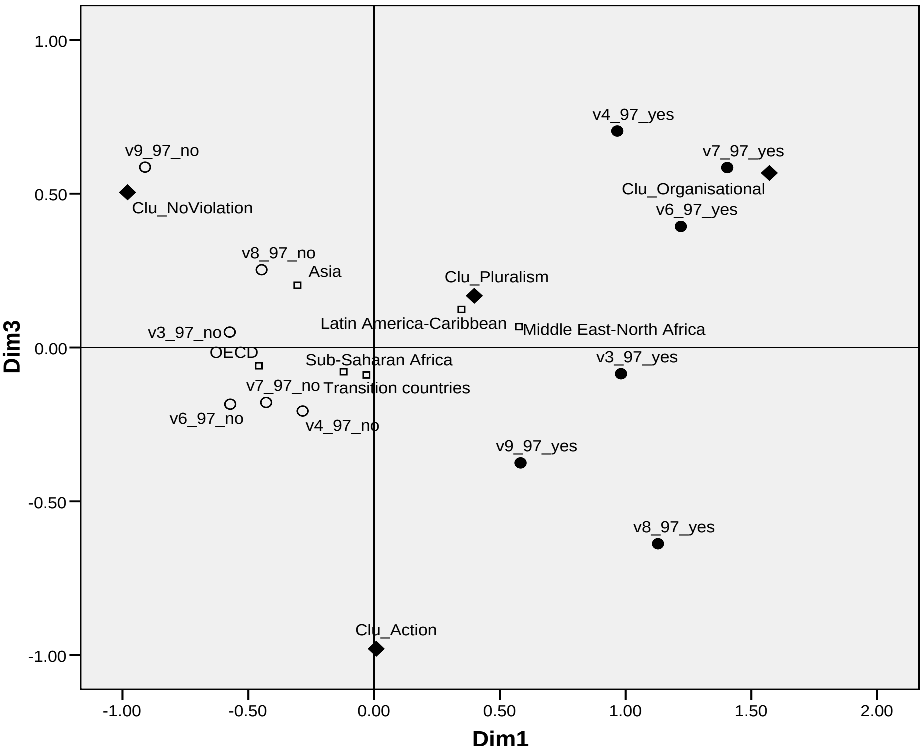

Fig. 3. Violations in the space of factor on and three

is recognised that trade unions in certain regions, in particular in Latin America, resort more commonly to the ILO supervisory system within the framework of their political actions (Kucera, 2004). Thus, we can expect more frequent mention of violations in the CEACR reports. In effect, this is what we observe: Latin American countries (merged here with the Caribbean countries) have a positive average position on the first dimension, which, as shown, is strongly correlated with the number of violations. A similar or even accentuated situation prevails for Middle East and North African countries, which experience on average more violations regarding Convention 87 than other regional groups. These latter regions are closer to the "Organisational" violations than the other groups. Although not surprising, it is interesting to note that the OECD countries are situated on the left of the diagrams (near the "No violation" cluster). It also follows from these figures that Asian countries are less frequently concerned by violations of Convention 87. However, it is worth mentioning here that only countries which have ratified the Convention are included in the analysis and few Asian countries have done so.

7.5 Graph of Implication Between Types of Violation

In order to reach a better understanding of the relationships between the types of violations stemmed from the earlier analyses, we proceed now to a statistical implicative analysis (Gras et al., 1996). Such an analysis focuses on assessing implications of the form $A \Rightarrow B$, according to which when A is true, then B is most likely (though not necessarily) to also be true. We start our analysis with implication graphs, which should help us detect the most relevant two by two relationships among types of violation. Implication strength is measured by means of the implication intensity (Gras et al., 1996; Suzuki and Kodratoff, 1998).[6] An implication graph puts together the strongest statistical implications in the form of a directional acyclic graph.

The general idea of the implication intensity is as follows: For each pair of binary variables v_r and v_s — violations r and s —, we consider the two implications $(v_r = 1) \Rightarrow (v_s = 1)$ and $(v_s = 1) \Rightarrow (v_r = 1)$. For each implication we count the number of counter-examples, e.g. for the first rule, the number of countries for which we predict violation v_r but not v_s. The implication intensity is the probability that the expected number of counter-examples under independence of the two violations exceeds this observed number. Among the two implication rules, we retain the one with the greater implication intensity. The set of binary variables are the vertices of the implication graph, and arrows between two vertices represent retained implications that exceed a selected threshold.

Using the CHIC software (Couturier and Gras, 2005; Couturier et al., 2006), we produced the two implication graphs shown in Figure 4. The different colours used for the arrows depend on the thresholds for the strength of the represented implications. Red designates implications established with an (entropic) intensity of at least 99%, blue those with at least 95% and grey those with at least 75%. These percentages should not be interpreted as confidences, i.e. as the probabilities with which we would obtain fewer counter-examples than observed under the independence assumption. Since we use the entropic version of the intensity measure, those percentages are, indeed, corrected probabilities preventing them to be compared with the traditional significance levels for statistical tests. The left part of Figure 4 was obtained by including the violation presence indicators and the right part by considering the absence indicators. In both cases we included also the region indicators.

Implications between violations stressed by the left hand graph corroborate the relationships derived from the cluster analysis in Section 7.3. For instance, the implications of violation v_7 on violations v_6 and v_9 confirm that these three violations, which are all three strong characteristics of the "Organisational and action restrictions" cluster, have effectively something in common. The implication graph highlights, however, the structure of their relationship. The direction of the implications is clearly laid down, something that allows a more direct interpretation. The graph shows clearly that restrictions of industrial action (v_9)

[6] Here we used in fact the more discriminating entropic version of the implication intensity (Blanchard et al., 2004).

constitute the most prevalent kind of violation. Other types of violations do not occur alone, each of them implying violation v_9. This result appeared also in the multiple correspondence analysis, though in a much less clear way. Indeed, a similar interpretation could have been drawn from the relatively low score of violation v_9 on the first dimension, which is strongly correlated with the number of different types of violation. More generally, the position of violation v_9 relatively close to the origin in both Figures 2 and 3 indicates that that violation is more common than other types of violations. Beside the system involving violations v_7, v_6 and v_9, no other clear system can be drawn from the implication graph. That confirms the observations made when commenting the "Action restrictions" cluster.

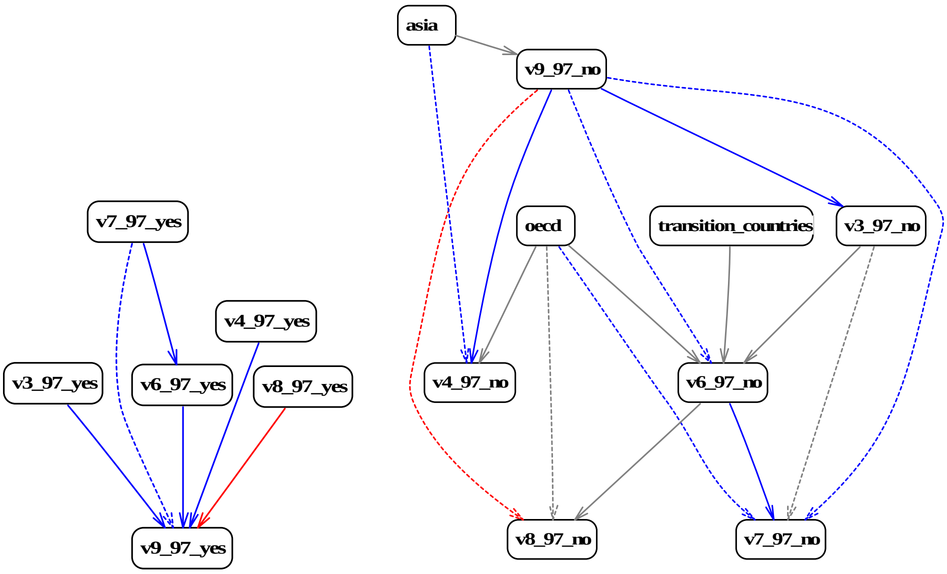

Fig. 4. Graph of implication between types of violation, Convention 87

No relation between regions and actual violations are significant at the various thresholds selected in the first graph. It is more interesting to look at the second graph on the right-hand side for uncovering relationships between regions and types of violation. Remember that this second graph was obtained by considering the absence of each violation instead of its presence. We also retained an additional lower threshold at the 75% level (gray arrows). From the presence-absence of violations standpoint, the two graphs are very similar, the one on the right-hand side being a "reversed" version of the first graph. For instance, while violation v_7 (organization of activities) implies violation v_9 (right to industrial action) in the first graph, the absence of violation v_9 implies the absence of violation v_7 in the second graph. This result is indeed straight as expected since the second graph uses just reversed values of the violation indicator variables.

However, while there is no evidence of relations between regions and types of violations, relationships between regions and non-violations stem from the second graph.[7] For instance, being member of the Organisation for Economic Co-operation and Development (OECD) implies that there are no excessive restrictions on the organisation and the internal administration of trade unions (no violation v_7). There are also less strong (at the 75% level) implication relations of OECD on the absence of trade-union monopolies (no violation v_4), on the procedures of registration (no violation v_8) and on the procedures of election of trade-union representatives (no violation v_6). Transition countries (ex Soviet Union) do not generally set up restrictions on the criteria of election of trade-union representatives (no violation v_6). Likewise, the few Asian countries that were examined do not face violations of the right to industrial action (no violation v_9) nor enforcement of trade-union monopolies (no violation v_4). These results confirm what we already observed on the multiple correspondence factorial maps (Figures 2 and 3). The relationships depicted here are more precise, however, since they are attached to specific violations.

Table 11. Typicality of supplementary variables

Path	Caribbean and Latin America	Middle east and North Africa	Asia	OECD
v7_97_yes ⇒ v6_97_yes ⇒ v9_97_yes	*	**		
v7_97_yes ⇒ v6_97_yes		**		
v7_97_yes ⇒ v9_97_yes		**		
v6_97_yes ⇒ v9_97_yes	**			
v4_97_yes ⇒ v9_97_yes				
v3_97_yes ⇒ v9_97_yes		*		
v8_97_yes ⇒ v9_97_yes		*		
Absence of violations				
v9_97_no ⇒ v4_97_no			**	*
v9_97_no ⇒ v6_97_no ⇒ v7_97_no			**	*
v9_97_no ⇒ v6_97_no			**	*
v6_97_no ⇒ v7_97_no				**
v9_97_no ⇒ v8_97_no			**	**
v9_97_no ⇒ v3_97_no			**	

One star indicate a typicality with a risk of 10% and two stars a risk of 5%.

[7] This dual analysis of positive and negative indicators (of violations in our case) and a same set of contextual co-variables (regions in our case) is an original approach and hence constitutes a significant contribution to the practice of statistical implicative analysis.

As shown, no region effect stems from the graph of implication among positive violation indicators. For finding such effects, it would have been necessary to introduce negative region indicators such as "non OECD", which do not correspond to coherent groups of countries and hence do not make much sense.

However, introducing the regions as supplementary variables rather than as vertexes we found some interesting results. Such supplementary variables do not appear in the implication graph. Instead, CHIC computes typicality indexes (Gras et al., 2006) measuring how typical each of them is for each path in the graph. Table 11 summarises the results obtained for a set of selected paths. The table exhibits that paths stemming from violation v_7 are mostly typical of Middle Eastern and North African countries. The path $v_7 \Rightarrow v_6 \Rightarrow v_9$ is to a less extent also typical of Caribbean and Latin American countries, which are indeed characterised by the sub-path $v_6 \Rightarrow v_9$. Implications $v_3 \Rightarrow v_9$ and $v_8 \Rightarrow v_9$ are also typically observed in the Middle East and North African region. Again, these outcomes can be related with proximities observed in the multiple correspondence maps between the concerned regions and violations. We learn here, however, that these proximities reflect more the typicality of some implications for given regions rather than the typicality of violations themselves.

Similarly, looking at the typicality indexes by regions for paths of the graph between the negative violation indicators, almost all implication paths appear to be typical of both Asian and OECD countries. The main differences between the two regions are non $v_6 \Rightarrow$ non v_7, which is characteristic of OECD only, and non $v_9 \Rightarrow$ non v_3 which characterises Asia only. The former says that in OECD countries there are no issues regarding organisation of trade unions (v_7) when there are no restrictions about the election of representatives (v_6). The second one says that restrictions about establishing or joining worker organisations (v_3) do not exit in Asian countries when there are no restrictions on industrial action v_9. The two most typical paths for OECD countries are the already discussed implication non $v_6 \Rightarrow$ non v_7 and the implication non $v_9 \Rightarrow$ non v_8 according to which no violation v_9 (industrial action) goes in line with no violation v_8 (registration trade unions). Results regarding Asia should be more carefully interpreted because of the relatively small number of non OECD Asian countries that have ratified the Convention. The explanation could be that Asian countries tend to ratify Convention 87 only after all legal issues regarding the convention have been solved.

Typicalities of implication paths among negative violation indicators can clearly not be deduced from those among positive violation indicators. They provide highlighting additional knowledge. Again, this complementarity demonstrates the importance of considering negative indicator variables (variables indicating absence of violations) when they make sense.

7.6 Hierarchical Violation Clustering

With implication graphs, we have obtained nice visual representations of the most prevalent two by two relations among indicator variables. We now look for higher level relationships between groups of violations using different types of hierarchical variable clustering methods.

Figure 5 below shows the cohesive tree Gras and Kuntz (2006) among the violations of Convention 87. This is an oriented hierarchy based on the implication intensity. The first link in the hierarchy is the stronger implication among two violations. Next stages also consider meta rules corresponding to implications among two earlier determined implications or among one earlier implication and a violation. At each step, the most implicative link is retained and the process is halted when there are no more additional significant link. The hierarchy shown was built with the entropic version of the implication intensity. It was obtained with the CHIC software.

The first two levels of the hierarchy in Figure 5 are formed by the two strongest implications depicted in the graph of implication. Thus, the presence of violations v_8 (registration of trade unions) implies the presence also of violation v_9 (right to industrial action), and when violation v_7 (organization of trade union activities) is observed we generally also observe violation v_6 (election). The third level is more interesting since it highlights a meta-rule, namely the relation between the implication of violation v_7 towards v_6, which implies the presence of violation v_3. This result is in agreement with the cluster analysis of countries (Table 10 page 23), where we observe that there are essentially countries belonging to the cluster "Organisational and action restrictions" that experience violation v_3 (join and establish trade unions), and that the simultaneous presence of violations v_6 and v_7 is a prominent characteristic of this cluster. From the relative proximity of violations v_3 and v_9 in the graph of implication (left part of Figure 4), but also in Figures 2 and 3, we could have possibly expected to see v_3 forming a meta-rule with $v_8 \Rightarrow v_9$. Because of the hierarchical structure, this is obviously not possible after the meta-rule $(v_7 \Rightarrow v_6) \Rightarrow v_3$ involving v_3 is formed. Such constraints of hierarchical presentations can sometimes mask interesting lessons. For example, by excluding violation v_3 when growing the cohesive hierarchy, the tree exhibits the meta-rule $(v_7 \Rightarrow v_6) \Rightarrow (v_8 \Rightarrow v_9)$ that is masked in Figure 5 by the presence of v_3. The issue of masked effects is obviously not specific to cohesive tree, but relates to any arborescent presentation.

Figure 6 presents the similarity tree of the violations of Convention 87. This hierarchy is built using the "Likelihood of the link" proximity measure (Lerman and Peter, 2003), which unlike the implication intensity is symmetric. The tree was produced also by the CHIC software.

Again, the first two by two merges are similar to those found in the cohesive tree. At the third level, however, violation v_3 joins the conjunction of v_8 and v_9 rather than the group formed by v_6 and v_7 as in the cohesive tree. This difference is indeed attributable to the fact that implication intensity and link likelihood do not measure the same kind of relationship.

For the sake of comparison, we present also a classical cluster analysis of the variables (i.e. violation indicators) realised with SPSS (SPSS Inc., 2003). Figure 7 shows the dendrogram obtained. The clustering is based on proximities among variables evaluated with the Pearson correlation and the method retained

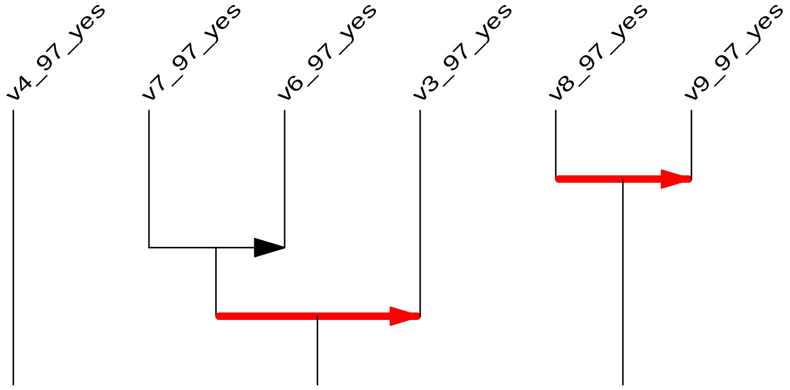

Fig. 5. Cohesive Tree of Violations, Convention 87

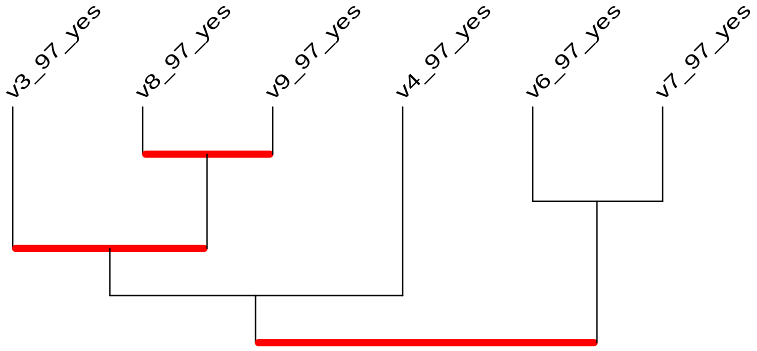

Fig. 6. Similarity Tree of Violations, Convention 87

is an agglomerative hierarchical procedure using the average linkage criterion. The dendrogram highlights again the strength of the two groups $\{v_6, v_7\}$ and $\{v_8, v_9\}$. However, the latter corresponding to the closer similarity in Figure 6 appears here to be less homogeneous than the former. We note also that violation v_3 joins the group formed by v_6 and v_7 before v_8 merges with v_9. Unlike hierarchical trees produced by CHIC, which exhibit only the order in which the successive groupings are done, the dendrogram informs also on the strength of the groupings. Indeed, the length of the branches represents the value of the criterion. However, criteria such as the average linkage used here are aggregations of two by two proximity measures among variables. They are therefore more difficult to interpret than likelihoods or implication intensities, which measure explicitly the link between two groups.

Surprisingly, the structure in Figure 7 resembles more the cohesive tree (Figure 5) built on the basis of directional association measures, than to the similarity tree (Figure 6). We can then ask ourselves whether the highlighted groupings

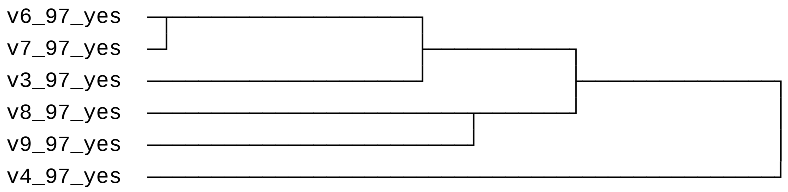

Fig. 7. Dendrogram of Violations, Convention 87 (Average linkage, Pearson correlation proximity)

would resist to small changes in the data. More generally this raises the question of the stability and robustness of the solutions obtained.

7.7 Violations of Convention 98

We carried out a similar analysis for Convention 98, which spells out rules regarding relations between trade unions and employers as well as collective bargaining. No clear structuring in the organisation of the different types of violations stemmed from this analysis. For instance, the partitioning of the countries obtained through the cluster analysis is essentially articulated around the presence of a single type of violations by group. Likewise, multiple correspondence analysis identifies one dimension for each type of violation indicating thus that the violations are almost independent each from the other. This is confirmed by the graph of implication that shows only isolated vertexes, exhibiting thus clearly the absence of any significant relationship among violations.

8 Conclusion

The aim of this paper is to present a study on the mining of legal texts, namely the comments of the Committee of Experts on the Application of Conventions and Recommendations (CEACR). The conventions considered are Conventions 87 and 98. The goal was to produce an automatic system for singling out texts — country-year comments — that report given types of violations. We sought to elaborate a system that could be used by any non text mining expert. We therefore adopted a strategy that does not require any preprocessing — tagging, lemmatisation, etc. — at the application phase.

The text mining process we followed can be divided into two main stages: 1) Find an appropriate quantitative representation of the texts, and 2) Learn prediction rules from the quantitative representation. The first stage is based on the whole corpus. We resorted to a specialised software (EXIT) for extracting the useful terminology. Then, the terms were grouped into a limited number of descriptor concepts. Finally, we obtained the quantitative representation by assigning the *tf* × *idf*'s of the descriptor concepts to each text. The learning phase was done on a sample of texts previously labelled with the violations they reported. It consisted, considering each violation in turn, in finding rules for predicting whether the text does report the violation or not. We used classification trees for that. Growing the trees, we were able to find the relevant descriptor concepts for predicting each kind of violation, as well as the *tf* × *idf* threshold values for designing the conditions of the rules. The outcome of the rule is a probability, i.e. the probability that the comment reports the violation. The classification is done by simply assigning a 'yes, the comment reports the violation' when the probability exceeds 50%, and 'no' otherwise.

The classification rules produced are not error free, error rates ranging approximatively between 3% and 15%. It is worth, however, to recall that the developed prediction system is not intended for making individual predictions for a country-year. We may distinguish between two usages of the rules: 1) As a tool for speeding up the process of searching texts that report violations, and 2) For producing material for a synthetic analysis. In the present paper we have mainly addressed the second one. Concerning the former, it should be clearly stated that the system is intended to help the legal expert, but not to replace him. Detecting the about 10% of errors, especially when errors are just false positives, should be easily done by the expert.

A macro analysis of the outcome of the text mining leaded to some interesting findings. Regarding Convention 87, only about a third of the countries that adhered to the Convention have not experienced difficulties in implementing it during the period 1997-2002. The remaining countries may be categorised into three groups: Those mainly concerned with restrictions on trade-union pluralism, those facing essentially restrictions on industrial action and those confronted with deeper organisational issues regarding trade-union activities. The identified types of violations structure themselves into a hierarchical form. Restrictions on industrial action (v_9) is the most frequent type of violation. It is concomitant upon all types of violations, meaning that when any other type is reported, restrictions on industrial action are also reported. We also observed that difficulties in the organisation of trade-union activities imply problems with the election of representatives and, hence, restrictions on industrial action. We have shown also that the structuring of the different types of violations directly related to the region: OECD and Asian countries are the most respectful of Convention 87. Transition countries are typically concerned by eligibility issues. The combination of difficulties in trade-union organisation and the election process are typical of Middle Eastern and North African countries, while eligibility issues and restrictions on industrial action is a characteristic of Latin American and Caribbean countries.

Appendix

A. The 27 Initial Key Concepts

Group 1: Civil Liberties Pertinent for Freedom of Association

1 *Right to life and physical integrity*

All persons involved directly or indirectly in trade union activities shall be ensured adequate protection through sufficiently deterrent sanctions and effective remedies against all acts of violence, including murder, physical assault, forced disappearance and forced exile.

2 *Right to liberty and security of person / Right to a fair trial*

Persons involved directly or indirectly in trade union activities may not be subject to arbitrary arrest, detention or imprisonment, or other restriction of the right to free movement, nor may they be denied the right to a fair trial by an independent and impartial tribunal.

3 *Protection of property*

Workers' organisations shall enjoy adequate protection of their right to property and independence in the administration of their finances. Trade union property and assets may not be subject to seizure without a judicial warrant.

Group 2: Right to Establish Trade Unions and All Concomitant Rights

4 *Exclusion from the right to establish and join workers' organisations*

All workers without distinction shall enjoy the right to establish and join workers' organisations.

(Note: The extent to which the right to establish trade unions and all concomitant rights shall apply to the armed forces and police personnel shall be determined by national laws).

5 *Trade union pluralism*

Legislation shall not impede trade union pluralism. The direct or indirect imposition by law of a system of trade union monopoly is in breach of Convention No. 87.

6 *Approval and registration of workers' organisations*

The approval and registration of workers' organisations shall not be subject to excessive formalities and restrictive conditions, nor to prior authorisation by public authorities. Any decision refusing the approval and registration of a worker's organisation shall be subject to judicial review.

7 *Establishment of federations and confederations*

First-level organisations shall have the right to establish and join federations and confederations of their own choosing, and to affiliate with international organisations of workers and employers. The establishment and registration of federations and confederations shall not be subject to overly restrictive conditions. Federations and confederations shall enjoy the same rights as first-level organisations.

8 *Dissolution or suspension of workers' organisations*

Workers' organisations shall not be liable to be dissolved or suspended by administrative authority. Any decision ordering the dissolution or suspension of an occupational organisation shall be subject to judicial review.

9 *Approval and registration of Constitutions and by-laws*
Workers' organisations shall have the right to draw up their constitutions and by-laws. The approval of an organisation's constitution and by-laws shall not be subject to overly restrictive conditions and excessive formalities, nor to prior authorisation by public authorities. Any decision refusing the approval of an organisation's constitution or by-laws shall be subject to judicial review.

10 *Election of representatives / Eligibility criteria*
Workers' organisations shall have the right to elect their representatives in full freedom. Eligibility for trade union office shall not be subject to occupational, membership or nationality criteria at least for a reasonable proportion of union representatives. The judiciary shall be the sole authority competent to supervise electoral procedures and pronounce on their legality.

11 *Administrative independence*
Workers' organisations shall enjoy independence in the administration of their internal affairs. The authorities' powers of supervision shall be limited to verifying that the law and the organisations' rules are respected and shall be subject to judicial review.

12 *Organisation of activities*
Workers' organisations shall be free to organise their activities and formulate their programs without interference by public authorities. Restrictions on the freedom of assembly, demonstration, expression and opinion may not be imposed unless absolutely necessary for the maintenance of public order.

Group 3: Right to Strike

13 *Restrictions on the right to industrial action / Definition of essential services*
The right to industrial action may not be restricted except in cases of an acute national crisis, for workers in the essential services in the strict sense of the term and public servants exercising authority in the name of the State. (Note: Taking into account the special circumstances in the various States Parties to the Convention, national laws shall designate as essential services only those services the interruption of which would endanger the life, personal safety or health of the whole or part of the population).

14 *Conditions for lawful industrial action*
The conditions, which the law requires to be observed in order for industrial action to be lawful, shall not amount to a de facto prohibition of the right to industrial action or to an excessive limitation of its exercise. The judiciary shall be the sole authority competent to pronounce on the legality of a given action.

15 *Minimum service*

The provision for a minimum service as an alternative to a total prohibition of industrial action for workers in the essential services shall be accompanied by the guarantees that it remains minimum, i.e. limited to the operations absolutely necessary to meet the needs of the population, and that workers organisations are able to participate in defining the service, or, failing agreement between the Parties, the task of defining the service is entrusted to an independent body.

16 *Compulsory arbitration in the context of industrial action*
Compulsory arbitration to end a strike may not be imposed save in the case of disputes involving public servants exercising authority in the name of the State or workers in essential services.

17 *Penalties for instigation of, or participation in, industrial action*
Workers and their organisations shall not be subjected to fines or imprisonment, nor shall they be liable for damages in respect of industrial action which conforms to international standards.

Group 4: Anti-Union Discrimination Acts and Acts of Interference by Employers in Trade Union Affairs

18 *Anti-union discrimination*
Legislation shall ensure adequate protection, through sufficiently dissuasive sanctions, against all acts of discrimination at the time of recruitment, during employment and at dismissal for membership or participation in trade union activities.

19 *Acts of interference*
Legislation shall ensure adequate protection, through sufficiently dissuasive sanctions, against all acts of interference by employers and their organisations in the establishment, functioning and administration of trade unions and vice versa.

20 *Solidarist associations*
Solidarist or other organisations set up by both employers and workers for purposes of economic and social welfare may not be treated more favourably than trade unions and shall be prohibited from exercising the rights pertaining to trade unions, in particular the right to collective bargaining by means of direct settlements between employers and non-unionised workers.

Group 5: Collective Bargaining

21 *Promotion of free and voluntary collective bargaining*
Governments should take all necessary measures to encourage and promote free and voluntary collective bargaining.

22 *Exclusion from the right to collective bargaining*
All workers without distinction shall enjoy the right to free and voluntary collective bargaining.
(Note: Convention No. 98 does not determine the position of public servants exercising authority in the name of the State, nor that of the armed forces and police personnel).

23 *Designation of the bargaining partner / Most representative trade union*
The right of workers' organisations to collective bargaining shall not be subject to excessive requirements. The most representative organisation at each level may be granted preferential or exclusive rights, provided that the designation is made according to objective and pre-established criteria.

24 *Level and scope of collective bargaining*
The Parties to collective bargaining shall have the right to freely determine the level at which collective bargaining shall take place and the sectors to be covered.

25 *Negotiable issues and substantive outcomes of collective bargaining / Permissible restrictions*
The Parties to collective bargaining shall have the right to determine the negotiable issues and substantive outcomes of collective bargaining. In cases where imperative economic stabilisation policies require the imposition of restrictions, governments should ensure that these remain exceptional and proportional and that the living standard of those mostly affected is protected.

26 *Approval and registration of collective agreements*
The approval or registration of collective agreements may not be refused save on grounds of form or where their terms do not conform to the minimum standards set out in labour law. Any decision refusing the approval or registration of a collective agreement shall be subject to judicial review.

27 *Compulsory arbitration in the context of collective bargaining*
Compulsory arbitration should not be imposed on the Parties to collective bargaining, except in cases of disputes in public and essential services in the strict sense of the term, to break a deadlock after protracted and fruitless negotiations, or at the initiative of workers' organisations for the conclusion of a first collective agreement.

Table 12. Correspondence between original and retained key concepts (violations)

Convention C87		Convention C98	
retained	original	retained	original
v_1	1	w_1	18
v_2	2	w_2	19
v_3	4	w_3	20
v_4	5	w_4	21
v_5	8	w_5	22
v_6	10	w_6	23
v_7	3, 11, 12	w_7	24
v_8	6, 7, 9	w_8	25
v_9	13, 14, 15, 16, 17	w_9	26, 27

B. Examples of Terms Associated to Descriptor Concepts

Election	**Restrictions on trade union activities**
electoral procedure/	trade union activitie/
representative member/	right to organise/
trade union office/	right of trade unions to organize/
eligibility/	right to publication/
re-election/	right to assembly/
representatives of organization/	right to disseminate information/
representatives of trade union/	freedom of opinion/
vote/	freedom of expression/
elect member/	political opinion/
union leader/	political activity/
elect their representative/	hold meeting/
elect their own representative/	right to organize/
to elect representative/	right of association/
to elect representatives/	right of workers organization/
elect its representatives/	right of workers to organize/
elect their trade union representative/	right to hold trade union/
election of trade union representative/	holding office/
elected representative/	formulate their programmes/
elected workers' representative/	right of organizations to organize/
election/	right of first-level unions to organize/
elect freely their representative/	right of unions to organize/
elected their representative/	right of workers' organizations to organize/
union officer/	right of trade union organizations to organize/
union office/	right of these employees to organize/
representatives of association/	rights of workers' organizations to organize/
representatives of union/	right of workers' trade unions to organize/
representatives of workers' organization/	political activities/
	taking part in political matters/
	holding meetings/
	intervening in political matters/
	political or religious activity/
	political or religious activities/
	freedom of assembly/

C. Examples of CEACR Comments and Related Predictions

For the purpose of illustrating how the text mining works, we present in this appendix two CEACR Comments (Figures 8 and 9) in which we highlight detected terms corresponding to descriptor concepts. Both examples are about the application of Convention 87. The first text is the 2001 report on Tajiskistan. It illustrates a case where our process detected correctly all types of violations reported. The second example is the 1992 report about Gabon, for which the text mining process ends with some errors.

Convention 87: Tajikistan 2001

In Figure 8 terms surrounded by boxes are terms found as belonging to a detector concept list. We use italic for descriptor concept terms used for predicting violation v_6 (election of representative), regular bold for those used for predicting violation v_7 (organisation of activities and administrative independence), and bold italic for those used for predicting violation v_9 (right to industrial action).

When the report is presented to the text mining system, it starts by counting the occurrences of each descriptor concept. These numbers constitute the main component of the $tf \times idf$'s used by the classification trees for predicting the presence or absence of each violation. For instance, there are three occurrences of the "Strike action" descriptors concepts (in bold italic). From it the learned system estimates to 97% the probability that the report raises an issue concerning "restriction on the right to industrial action" (v_9). With such a high probability we predict that there is effectively a violation of that type v_9, which is exactly what the second paragraph points out. The prediction is thus correct.

As another example, we may note that there is only one occurrence of the "Election" descriptor concept (in italic). From this low frequency, the system predicts a probability of only 18% that the text reports problems regarding election of representatives (v_6). With such a low probability, we predict that there is no mention of a violation of type v_6. Again, this is a correct prediction when we look carefully at the text. The advantage is indeed that such predictions are done automatically for all type of violations and all texts processed in the system. Table 13 summarises results for the 2001 Tajikistan report.

Table 13. Actual and predicted violations in CEACR 2001 report for Tajikistan

Violation	3	4	*6*	**7**	8	***9***
Actual	0	0	*0*	**1**	0	***1***
Prediction	9%	4%	*18%*	**53%**	10%	***97%***

Fig. 8. CEACR 2001 report for Tajikistan

Report:
1. Article 3 of the Convention. Right of workers' and employers' organizations to draw up their constitutions and rules, to *elect their representatives* in full freedom and to **organize their administration** and activities. Concerning article 4(1) of the Law on Trade Unions which provides that trade unions shall be independent in their activities and that any **interference** by state authorities shall not be permitted except in cases specified by law, the Committee requests the Government to specify in its next report in which cases the state authorities are allowed to interfere with **trade union activities**.
2. Article 3. Right to ***strike***. Concerning article 211(3) of the Labour Code which provides that restrictions of the right to ***strike*** shall be subject to the provisions of legislation in force in Tajikistan, the Committee requests the Government to provide the text of the provisions relating to such restrictions. Furthermore, the Committee requests the Government to state whether the former provisions of the Penal Code which were at the time applicable in the USSR, and particularly section 190(3), which contained significant restrictions on the exercise of the right to ***strike*** in the transport sector, enforceable by severe sanctions, including sentences of imprisonment for up to three years, have been repealed by a specific text.

The Committee also requests the Government to supply in its next report a copy of the Law of 29 June 1991 regulating the organisation and holding of meetings, gatherings, street processions and demonstrations. In addition, the Committee requests the Government to indicate what are the legal provisions on the **right to organize** of employers.

Convention 87: Gabon 1992

The prediction of violations mentioned in the 1992 CEACR report for Gabon produces some errors as can be shown in Table 14. The report contains two terms related to the "Election" descriptor concept. From this number, the system evaluates a 94% probability for the text to mention problems regarding the election of representatives (v_6). A careful examination of the report shows, however, that the first mention of "election" is within a sentence about new rules adopted by the Government for precisely insuring representative election, and the second one in a sentence for asking the Government to indicate the election results in their next report. This is indeed a positive reference to "election of representatives", which is confused with a violation by the system.

Table 14. Actual and predicted violations in CEACR 1992 report for Gabon

Violation	3	**4**	*6*	7	8	***9***
Actual	0	**1**	*0*	0	0	***1***
Prediction	9%	**94%**	*92%*	53%	10%	***97%***

Fig. 9. CEACR 1992 report for Gabon

Report:
The Committee notes, in particular, the Government representative's statement that the recognition of individual liberties in the new Constitution of Gabon, which came into force on 26 March 1991, has a corollary in the overall social plan, which is the abolition of trade union **monopoly**, that is to say the establishment of genuine and complete freedom of association. It notes that a draft new Labour Code which was discussed during a tripartite meeting from January to April 1991, attended both by the unitary employers' and workers' central organisations and by other organisations of workers and employees, has already been examined by the Government and was to be presented before the end of 1991. According to the Government, the amendment envisaged includes the repeal of section 174 of the present Labour Code which obliges all workers' or employers' organisations to affiliate with the Trade Union Confederation of Gabon (COSYGA) or the Employers' Confederation of Gabon (CPG). The Government also states that Act No. 13/80 of 2 June 1980, establishing a trade union solidarity tax deducted for the COSYGA, is no longer applied and that the tax has not been deducted since March 1990. Legislation is to be adopted for its formal repeal.

With regard to the provisions on compulsory arbitration restricting workers' right to ***strike*** (sections 239, 240, 245 and 249 of the Labour Code), the Government representative stated that a draft law specifically on the right to ***strike***, which takes into account the requirements of the Convention, has been prepared and may be incorporated into the revised Labour Code.

The Committee notes the Government's reply in its last report to the effect that: (1) COSYGA, whose members wish the organisation to continue under the same name, has complied with the laws of the Republic of Gabon and adopted new rules under which it is now protected from any influence on the part of political parties and religions; (2) the new rules of COSYGA settle clearly the problem of the social assets of COSYGA vis-à-vis the new unions; (3) the sole object of occupational organisations is to examine and defend members' economic, industrial, commercial, agricultural and artisanal interests and there are no longer any restrictions on the establishment of these organisations; and (4) future *election*s of staff delegates and members of the Economic and Social Cooperation Committees will demonstrate that the various unions in establishments and enterprises are representative.

In the light of this information, the Committee asks the Government to provide a copy of the new COSYGA rules with its next report and to indicate the results of the above-mentioned *election*s.

References

Anderberg, M. (1973). *Cluster Analysis for Application.* New York: Academic Press.

Baccaro, L., J.-M. Bonvin, J.-P. Laviec, P. O'Donovan, G. Ritschard, and D. A. Zighed (2003). Social dialogue regimes: An investigation in the structural determinants and socioeconomic outcomes of negotiated regulation. Research proposal supported financially by the Geneva International Academic Network (GIAN), IILS and University of Geneva.

Bellal, F. (2006). Catégoristion automatique de textes juridiques. Mémoire du Master recherche en informatique de Lyon, ERIC, Université de Lyon 2.

Biggs, D., B. De Ville, and E. Suen (1991). A method of choosing multiway partitions for classification and decision trees. *Journal of Applied Statistics 18*(1), 49–62.

Blanchard, J., P. Kuntz, F. Guillet, and R. Gras (2004). Mesure de la qualité de règles d'association par l'intensité d'implication entropique. *Revue des nouvelles technologies de l'information RNTI E-1*, 33–43.

Bourigault, D. and C. Jacquemin (1999). Term extraction + term clustering: An integrated platform for computer-aided terminology. In *Conference of the European Chapter of the Association for Computational Linguistics, EACL*, pp. 15–22.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984). *Classification And Regression Trees.* New York: Chapman and Hall.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics 21*(4), 543–565.

Couturier, R., A. Bodin, and R. Gras (2006). CHIC v3.7 Classification Hiérarchique Implicative et Cohésitive. Guide d'utilisation, Ecole Polytechnique, Université, Nantes.

Couturier, R. and R. Gras (2005). CHIC: traitement de données avec l'analyse implicative. In S. Pinson and N. Vincent (Eds.), *Extraction et Gestion des Connaissances (EGC 2005)*, Volume E-3 of *Revue des nouvelles technologies de l'information RNTI*, pp. 679–684. Cépaduès.

Damashek, M. (1995). Gauging similarity with ngrams: Language-independent categorization of text. *Science 267*, 843–848.

Fan, W., L. Wallace, S. Rich, and Z. Zhang (2006). Tapping the power of text mining. *Communications of the ACM 49*(9), 76–82.

Feldman, R. and I. Dagan (1995). Knowledge discovery in textual databases (KDT). In *KDD '95*, pp. 112–117.

Fellbaum, C. (Ed.) (1998). *WordNet: An Electronic Lexical Database.* Cambridge, MA: MIT Press.

Frantzi, K. T., S. Ananiadou, and H. Mima (2000). Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal on Digital Libraries 3*(2), 115–130.

Georgiou, I. (2006). Coding CEACR reports on ILO conventions nos. 87 and 98: A proposed methodology. RUIG social dialogue regimes project, internal report, International Institute of Labour Studies and University of Geneva.

Gras, R., S. Ag Almouloud, M. Bailleul, A. Laher, M. Polo, H. Ratsimba-Rajohn, and A. Totohasina (1996). *L'implication statistique: Nouvelle méthode exploratoire de données*. Recherches en didactique des mathématiques. Grenoble: La pensée sauvage.

Gras, R., J. David, J.-C. Régnier, and F. Guillet (2006). Typicalité et contribution des sujets et des variables supplémentaires en analyse statistique implicative. In G. Ritschard and C. Djeraba (Eds.), *EGC'2006*, Volume RNTI-E-6 (2 volumes) of *Revue des Nouvelles Technologies de l'Information*, pp. 359–370. Cépaduès.

Gras, R. and P. Kuntz (2006). Discovering *R*-rules with a directed hierarchy. *Soft Computing 10*(5), 453–460.

Greenacre, M. (1993). *Correspondence analysis in practice*. London: Academic Press.

Gruber, T. R. (1993). A translation approach to portable ontologies. *Knowledge Acquisition 5*(2), 199–220.

Heitz, T., M. Roche, and Y. Kodratoff (2005). Extraction de termes centrée autour de l'expert. *Revue des nouvelles technologies de l'information RNTI E-5*, 685–690.

Jobson, J. D. (1992). *Applied Multivariate Data Analysis*, Volume II: Categorical and Multivariate Methods. New York: Springer-Verlag.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics 29*(2), 119–127.

Kodratoff, Y. (1999). Knowledge discovery in texts: A definition, and applications. In Z. W. Ras and A. Skowron (Eds.), *Foundations of Intelligent Systems, ISMIS '99*, Volume 1609 of *Lecture Notes in Computer Science*, pp. 16–29. Springer.

Kodratoff, Y. (2004). Induction extensionnelle: définition et application à l'acquisition de concepts à partir de textes. *Revue des nouvelles technologies de l'information RNTI E-2*, 247–252.

Kucera, D. (2004). Measuring trade union rights: A country-level indicator constructed from coding violations recorded in textual sources. Working Paper 50, Policy Integration Department, Statistical Development and Analysis Unit, International Labour Office.

Kumps, N., P. Francq, and A. Delchambre (2004). Création d'un espace conceptuel par analyse de données contextuelles. In G. Purnelle, C. Fairon, and A. Dister (Eds.), *Le Poids des Mots (JADT 2004)*, Volume 2, pp. 683–691. Presse Universitaire de Louvain.

Lebart, L., A. Morineau, and M. Piron (2000). *Statistique exploratoire multivariée* (Troisième ed.). Paris: Dunod.

Lerman, I.-C. and P. Peter (2003). Indice probabiliste de vraisemblance du lien entre objets quelconques ; analyse comparative entre deux approches. *Revue de Statistique Appliquée LI*(1), 3–35.

Liu, B., Y. Ma, and P. S. Yu (2001). Discovering unexpected information from your competitors' web sites. In *KDD '01*, pp. 144–153.

Mayfield, J. and P. McNamee (1998). Indexing using both n-grams and words. In *TREC*, pp. 361–365.

Pisetta, V., H. Hacid, F. Bellal, G. Ritschard, and D. A. Zighed (2006). Traitement automatique de textes juridiques. In R. Lehn, M. Harzallah, N. Aussenac-Gilles, and J. Charlet (Eds.), *Actes de SdC 2006, Semaine de la Connaissance, 26-30 juin 2006*, Nantes, France. (CDrom et sdc2006.org).

Plisson, J., N. Lavrač, and D. Mladenić (2004). A rule based approach to word lemmatization. In *Proceedings of IS04.*

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning.* San Mateo: Morgan Kaufmann.

Riloff, E. and L. A. Hollaar (1996). Text databases and information retrieval. *ACM Compututing Surveys 28*(1), 133–135.

Salton, G., J. Allan, and A. Singhal (1996). Automatic text decomposition and structuring. *Information Processing and Management 32*(2), 127–138.

Salton, G. and C. Buckley (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management 24*(5), 513–523.

Salton, G., C. Buckley, and J. Allan (1992). Automatic structuring of text files. *Electronic Publishing—Origination, Dissemination, and Design 5*(1), 1–17.

Saravanan, M., P. C. Reghu Raj, and S. Raman (2003). Summarization and categorization of text data in high-level data cleaning for information retrieval. *Applied Artificial Intelligence 17*(5-6), 461–474.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing, Manchester.*

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

Smadja, F. A. (1993). Retrieving collocations from text: XTRACT. *Computational Linguistics 19*(1), 143–177.

SPSS (Ed.) (2001). *Answer Tree 3.0 User's Guide.* Chicago: SPSS Inc.

SPSS Inc. (2003). *SPSS 12 Command Syntax Reference.* Chicago, IL: SPSS Inc.

Suzuki, E. and Y. Kodratoff (1998). Discovery of surprising exception rules based on intensity of implication. In J. M. Zytkow and M. Quafafou (Eds.), *Principles of Data Mining and Knowledge Discovery, Second European Symposium, PKDD '98, Nantes, France, September 23-26, Proceedings*, pp. 10–18. Berlin: Springer.

Witten, I. H. and E. Frank (2005). *Data Mining: Practical Machine Learning Tools and Techniques* (2nd ed.). Amsterdam: Morgan Kaufman (Elsevier).

Publications récentes du Département d'économétrie

peuvent être obtenues à l'adresse suivante :

Université de Genève
UNI MAIL
A l'att. de Mme Caroline Schneeberger
Département d'économétrie
40, Bd du Pont-d'Arve
CH - 1211 Genève 4
ou sur
INTERNET : http//www.unige.ch/ses/metri/cahiers

2007.01 KRISHNAKUMAR Jaya and Tobias MÜLLER, Participation and voting behavior in a direct democracy : a structural model of migration policy in Switzerland, Mai 2007, 36 pages.

2006.07 KRISHNAKUMAR Jaya and David NETO, Estimation and Testing in Threshold Cointegrated Systems Using Reduced Rank Regression, November 2006, 23 pages.

2006.06 ZOIA Maria Grazia, A New Algebraic Approach to Reprensentation Theorems for (Co)integrated Processes up to the Second Order, Octobre 2006, 22 pages.

2006.05 MILLS FLEMMING Joanna, Eva CANTONI, Christopher FIELD and Ian MCLAREN, Extracting Long-Term Patterns of Population Changes from Sporadic Counts of Migrant Birds, Juin 2006, 23 pages.

2006.04 LÔ Serigne N. and Elvezio RONCHETTI, Robust Small Sample Accurate Inference in Moment Condition Models, Juin 2006, 33 pages.

2006.03 LÔ Serigne N. and Elvezio RONCHETTI, Robust Second Order Accurate Inference for Generalized Linear Models, Mai 2006, 29 pages.

2006.02 CANTONI Eva, Joanna MILLS FLEMMING and Elvezio RONCHETTI, Variable Selection in Additive Models by Nonnegative Garrote, Avril 2006, 17 pages.

2006.01 COPT Samuel and Stephane HERITIER, Robust MM-Estimation and Inference in Mixed Linear Models, Janvier 2006, 27 pages. Published in *Biometrics* (2007).

2005.04 KRISHNAKUMAR Jaya and David NETO, Partial Cointegration, Novembre 2005 (revised August 2006), 25 pages.

2005.03 VAN BAALEN Brigitte, Tobias MÜLLER, Social Welfare effects of tax-benefit reform under endogenous participation and unemployment, Février 2005, 42 pages.

2005.02 CZELLAR Véronika, G. Andrew KAROLYI, Elvezio RONCHETTI, Indirect Robust Estimation of the Short-term Interest Rate Process, Mars 2005, 29 pages.