Automated Classification of Legal Cross References Based on Semantic Intent

Nicolas Sannier, Morayo Adedjouma, Mehrdad Sabetzadeh, and Lionel Briand

SnT Centre for Security, Reliability and Trust,
University of Luxembourg, Luxembourg
{nicolas.sannier, morayo.adedjouma, mehrdad.sabetzadeh, lionel.briand}@uni.lu

Abstract. [**Context and motivation**] To elaborate legal compliance requirements, analysts need to read and interpret the relevant legal provisions. An important complexity while performing this task is that the information pertaining to a compliance requirement may be scattered across several provisions that are related via cross references. [**Question/Problem**] Prior research highlights the importance of determining and accounting for the semantics of cross references in legal texts during requirements elaboration, with taxonomies having been already proposed for this purpose. Little work nevertheless exists on automating the classification of cross references based on their semantic intent. Such automation is beneficial both for handling large and complex legal texts, and also for providing guidance to analysts. [**Principal ideas/results**] We develop an approach for automated classification of legal cross references based on their semantic intent. Our approach draws on a qualitative study indicating that, in most cases, the text segments appearing before and after a cross reference contain cues about the cross reference's intent. [**Contributions**] We report on the results of our qualitative study, which include an enhanced semantic taxonomy for cross references and a set of natural language patterns associated with the intent types in this taxonomy. Using the patterns, we build an automated classifier for cross references. We evaluate the accuracy of this classifier through case studies. Our results indicate that our classifier yields an average accuracy (F-measure) of $\approx 84\%$.

Keywords: Compliance Requirements, Legal Cross References, Semantic Taxonomy, Automated Classification.

1 Introduction

In many domains such as public administration, healthcare and finance, software systems need to comply with laws and regulations. To identify and elaborate legal compliance requirements for these systems, requirements analysts typically need to read and interpret the relevant provisions in legal texts. This task is often made difficult by the complexities of legal writing. An important source of complexity is that one cannot consider the legal provisions independently of one another, due to the provisions being inter-dependent. The dependencies between the provisions are captured using *legal cross references* (CR).

The semantic intent of a legal CR directly impacts the way the CR is handled during requirements elaboration [17]: For example, when a provision, say an article, A, cites a provision, B, to state that A does not apply in an exceptional situation described by B, it is best to create a new requirement for the exception. In contrast, when A cites B for a definition, it is more sensible to add the definition to the glossary, rather than creating a new requirement.

A number of useful taxonomies have already been developed to enable the classification of CRs according to their semantic intent [4, 3, 13, 17, 24]. These taxonomies nevertheless consider classification as a manual task, and thus do not provide automation for the task.

In this paper, we develop an automated approach for classifying CRs based on their semantic intent. Such automation has two main benefits: First, the number of CRs that need to be considered by analysts may be large, in the hundreds or thousands [3, 1, 20]. Automated classification helps both to reduce effort, and further to better organize requirements engineering activities, noting that automated classification provides a-priori knowledge about the intent of CRs. Second, research by Massey et al. [15] and Maxwell et al. [16] suggests that software engineers without adequate legal expertise find it difficult to determine the intent of CRs. Automation can provide useful guidance in such situations.

Research Questions (RQs). Our work is motivated by the following RQs:

- **RQ1: What are the possible intents of (legal) CRs?** RQ1 aims at developing a taxonomy of CR intents. This RQ is informed by the existing CR taxonomies, as we explain later.
- **RQ2: Are there natural language (NL) patterns in legal texts that suggest the intent of CRs?** RQ2 aims at investigating whether there are patterns in the text with a direct link to the intent of CRs. Such patterns would enable the automatic classification of CRs.
- **RQ3: How accurately can NL patterns predict CR intent?** Provided that the answer to RQ2 is positive, RQ3 aims at measuring how accurate (in terms of standard accuracy metrics) an automated classification approach based on NL patterns is.

Approach. Fig. 1 outlines our approach. We address RQ1 and RQ2 based on a qualitative study of 1079 CRs from Luxembourg's legislative corpus. Our study is guided by the principles of Grounded Theory (GT) [6] – a systematic methodology for building a theory from data. However, GT normally starts without preconceived knowledge about the theory. In contrast, our study leverages existing CR taxonomies, notably those by Breaux [3], Hamdaqa et al. [13], and Maxwell et al. [17]. The qualitative study yields an enhanced taxonomy (Table 1), along with a collection of NL patterns observed in the text appearing in the vicinity of CRs of each intent type (partially shown in Table 2). We utilize the taxonomy and the identified NL patterns for developing an automated classification solution, and evaluate the accuracy of the solution through case studies.

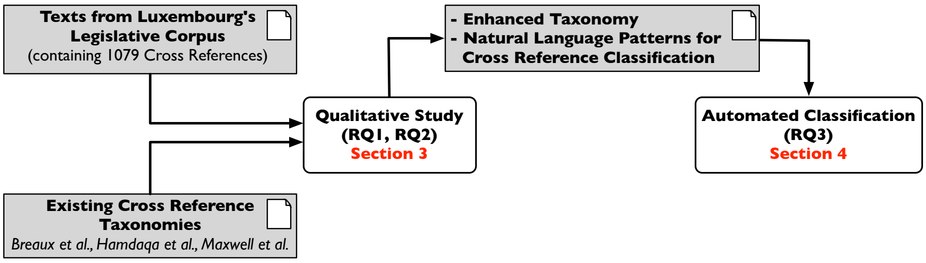

Texts from Luxembourg's Legislative Corpus
(containing 1079 Cross References)
Existing Cross Reference Taxonomies
Breaux et al., Hamdaqa et al., Maxwell et al.
Qualitative Study
(RQ1, RQ2)
Section 3
- Enhanced Taxonomy
- Natural Language Patterns for Cross Reference Classification
Automated Classification
(RQ3)
Section 4

Fig. 1. Overview

Contributions and Key Remarks. Our proposed taxonomy brings together and extends existing taxonomies with the goal of automating CR classification. Our work on NL patterns presents the first systematic attempt we are aware of, where the collocation of CRs and adjacent phrases has been studied for the purpose of determining CR intent. We demonstrate that a rule-based classification approach based on NL patterns is effective. To this end, we report on two case studies. The first case study is over a random sample of pages from various Luxembourgish legislative texts, and the second – over the French editions of the Personal Health Information Protection Act (PHIPA) of Ontario, Canada and the 2014 compilation of the amendment provisions on Canadian consolidated laws [22]. The two case studies collectively include 2585 CRs. Our evaluation of automated classification shows F-measures of 87.57% and 80.59% for the first and second case studies, respectively, yielding an average F-measure of 84.48%.

Our work exclusively considers legal texts in French. The consistency seen between our CR taxonomy and the ones developed previously over English legal corpora provides confidence about our taxonomy being generalizable. Adapting our approach to texts in other languages will nevertheless prompt a re-investigation of RQ2 and RQ3. The observations that we expect to carry over from our work to such adaptations are: (1) There are indeed patterns in legal texts to suggest the intent of CRs; and (2) A reasonably-sized manual investigation of these patterns provides an accurate basis for automated classification.

Structure. The remainder of the paper is organized according to the flow of Fig. 1. Section 2 reviews related work. Sections 3 and 4 present our qualitative study and automated classification solution, respectively. Section 5 discusses practical considerations and threats to validity. Section 6 concludes the paper.

2 Related Work

Several papers address automated detection and resolution of CRs in legal texts. Detection refers to the ability to recognize the complete textual fragment that constitutes a CR, and resolution – to the ability to find a CR's target provision in the right legal text. Work on CR detection and resolution spans several countries and jurisdictions, including the Dutch, Italian, Japanese and Luxembourgish legislation, respectively [8, 24, 23, 20], as well as US regulations [4]. In contrast to the above work, in this paper, we focus on automatically extracting information about the semantics of CRs, once they have been detected. Automated detection (but not resolution) of CRs is a prerequisite to our work; for this, we rely on a tool from our previous research [20].

Work already exists on the semantic classification of CRs. Maxwell et al. [17] propose a CR taxonomy, where they distinguish definitions (the cited provision provides a definition needed by the citing one), constraints (the cited provision imposes additional conditions on the citing one), exceptions (the cited provision restricts the applicability of the citing one), general (generic citations such as to “applicable law”), unrelated (the cited provision is orthogonal to software requirements), incorrect (wrong provision cited), and prioritization (establishing a priority between the citing and the cited provisions).

Breaux & Antón [4, 3] distinguish refinements (the cited provision elaborates upon the citing one), exceptions (same as by Maxwell et al. [17]) and continuations (which, like refinements, elaborate on information in the citing provisions, but through subsequent sub-divisions). Breaux [3] further considers definitions and constraints, but in a more general context than CRs per se.

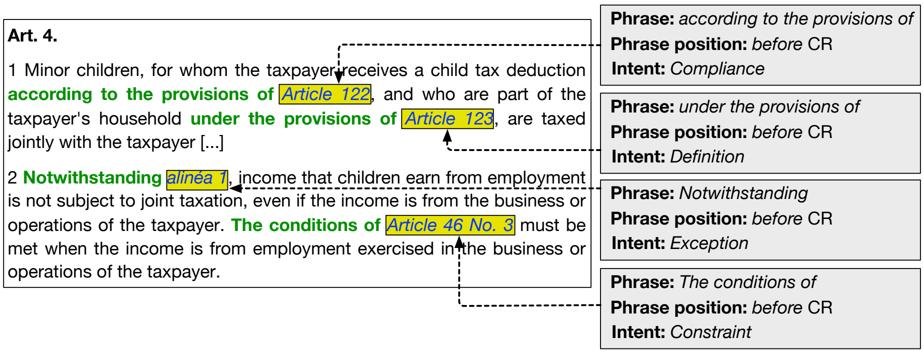
Hamdaqa et al. [13] classify CRs under definitions (same as above), specifications (the cited provision provides more information about the citing one), compliance (the cited provision complies with the citing one in some manner), and amendments. Amendments are further specialized into insertions (amending by adding a new provision), deletions (amending by repealing a provision), striking (amending by replacing the wording within a provision), and redesignation (amending by changing the name of the cited provision).

Our work builds on and is closely guided by the above three taxonomies. A detailed comparison between our taxonomy and these earlier ones is provided in Section 3. Broadly speaking, none of these earlier taxonomies alone provide a complete basis for automated CR classification.

Finally, we note that the general problem of automated classification in legal texts has been studied for a long time. Existing work on this topic mainly address the classification of deontic modalities, e.g., rights, obligations, permissions, and delegations. A number of techniques for this type of classification have been proposed based on machine learning [2], natural language processing [24], and the combination of the two [5]. In contrast to these strands of work, our focus is on automatic classification of CRs.

3 A Qualitative Study of CR Intent Types

We first describe the units of analysis and the analysis procedure in our qualitative study, outlined earlier in Fig. 1. We then address RQ1 and RQ2.



Art. 4.

1 Minor children, for whom the taxpayer receives a child tax deduction **according to the provisions of** *Article 122*, and who are part of the taxpayer's household **under the provisions of** *Article 123*, are taxed jointly with the taxpayer [...]

2 **Notwithstanding** *alinéa 1*, income that children earn from employment is not subject to joint taxation, even if the income is from the business or operations of the taxpayer. **The conditions of** *Article 46 No. 3* must be met when the income is from employment exercised in the business or operations of the taxpayer.

Phrase: *according to the provisions of*
Phrase position: *before* CR
Intent: *Compliance*

Phrase: *under the provisions of*
Phrase position: *before* CR
Intent: *Definition*

Phrase: *Notwithstanding*
Phrase position: *before* CR
Intent: *Exception*

Phrase: *The conditions of*
Phrase position: *before* CR
Intent: *Constraint*

Fig. 2. Examples of Recorded Information for CRs during the Qualitative Study

3.1 Units of Analysis

We manually identified and analyzed CRs from two Luxembourgish legislative texts. These texts are: (1) the 2014 edition of Luxembourg's Income Tax Law [12] and (2) Chamber of Deputies' Draft Law No. 6457 [11]. Both texts are in French.

We chose the Income Tax Law based on advice from legal experts who deemed this law to be among the most complex in terms of CRs. This law, which has been regularly revised since it was first drafted in 1967, further offers a window into several decades of legal writing practices. The second text was chosen to address an a-priori-known limitation posed by the Income Tax Law for our study. In particular, the Income Tax Law is generally not meant to make amendments to other laws, and consequently contains a very small number of amendment CRs. The second text has several such CRs, thus providing more conclusive grounds for studying this class of CRs.

In total, we examined, using the procedure described next, 141 pages from the above legislative texts. These pages collectively contain 1079 CRs: 729 CRs come from the first seven chapters of the Income Tax Law (117 pages) and the remaining 350 CRs – from the first chapter of Draft Law No 6457 (24 pages).

3.2 Analysis Procedure

Using the judgment of the first two authors, we classified each CR according to the taxonomies by Breaux [3], Hamdaqa et al. [13], and Maxwell et al. [17]. If some CR was not classifiable using any of these taxonomies, we defined a new intent type. After classifying a CR, we considered exclusively the sentence in which the CR appeared and documented any phrase(s) that affected human judgment, along with whether the phrase(s) appeared before or after the CR. No phrase was derived if the judgment happened to rely on information other than the sentence in which the CR appeared (e.g., previous sentences), or if the sentence had no relevant phrase(s) in it. In Fig. 2, we illustrate the information maintained for four CRs from the Income Tax Law (translated from French).

The first two authors, both of whom are native French speakers and have background in legal and regulatory requirements, worked together throughout the procedure explained above. In each case, the intent and the identified phrases (if any) were discussed until an agreement was reached. Once all the CRs had been analyzed, the phrases obtained for each intent type were reviewed. The phrases were then clustered into groups of semantically-equivalent variations. Subsequently, NL patterns were developed to characterize each cluster. A technicality in developing the NL patterns is that some languages, including French, distinguish gender and plurality (and the combinations thereof). To minimize the number of patterns, we defined suitable abstractions over gender and plurality.

We excluded from our analysis an investigation of the content of the provision(s) being cited by a CR. This decision was motivated by two observations: First, the provision(s) cited by a CR seldom refer back to the context in which they are being cited. The provision(s) are therefore unlikely to provide useful information about the intent of the citation. Second, the cited provision(s) may constitute a large amount of text, e.g., several articles and chapters, or even entire laws. Given that potential benefits from considering the content of cited provision(s) is limited, processing this content is not justified in either the qualitative study, or the automated classification solution that builds on the study.

3.3 Results

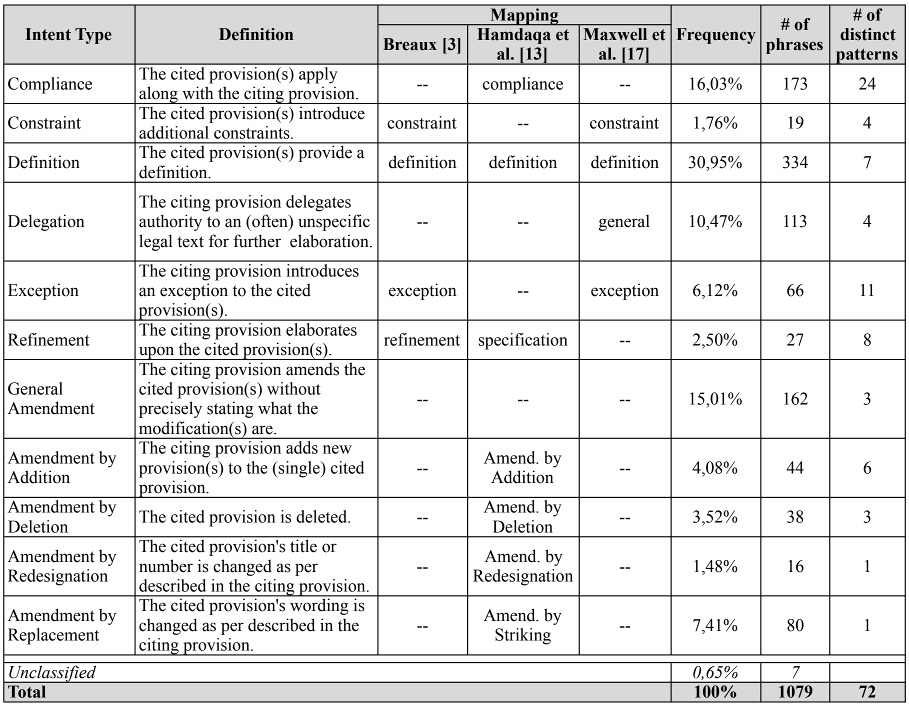
Tables 1 and 2 present the main results from our qualitative study. Specifically, Table 1 lists the intent types of our proposed CR taxonomy and their definitions, along with a mapping of the types to those in the taxonomies of Breaux [3], Hamdaqa et al. [13] and Maxwell et al. [17]. The table further shows, for our qualitative study, the relative frequency of each intent type, the number of phrases retrieved per type, and the number of distinct patterns derived from the phrases.

Table 2 details, for each intent type, the most frequent patterns and the relative frequencies of these patterns. The table further provides illustrating examples for the most frequent patterns in our study. Although the analysis was performed over French texts, we provide (unofficial) English translations to facilitate readability. For each intent type, we provide the frequency of patterns with less than three occurrences, denoted *rare*. We use this notion later in our discussion of RQ2.

Taxonomy of Intent Types (RQ1). Our taxonomy (Table 1) distinguishes eleven intent types for CRs. Except for the *General Amendment* type, all types in our taxonomy have a corresponding type in the taxonomies of Breaux's, Hamdaqa et al.'s, and Maxwell et al.'s. Nevertheless, and as suggested by Table 1, none of the above three taxonomies alone provide, for the purpose of automated classification, adequate coverage of the intent types. At the same time, there are intent types in these three taxonomies that our taxonomy does not cover. Below, we discuss the main differences between our taxonomy and the other three:

Breaux's taxonomy is at a higher level of abstraction than ours. Our taxonomy is primarily an amalgamation of those by Hamdaqa et al. and Maxwell et al.

Table 1. Taxonomy of Semantic Intent Types for CRs



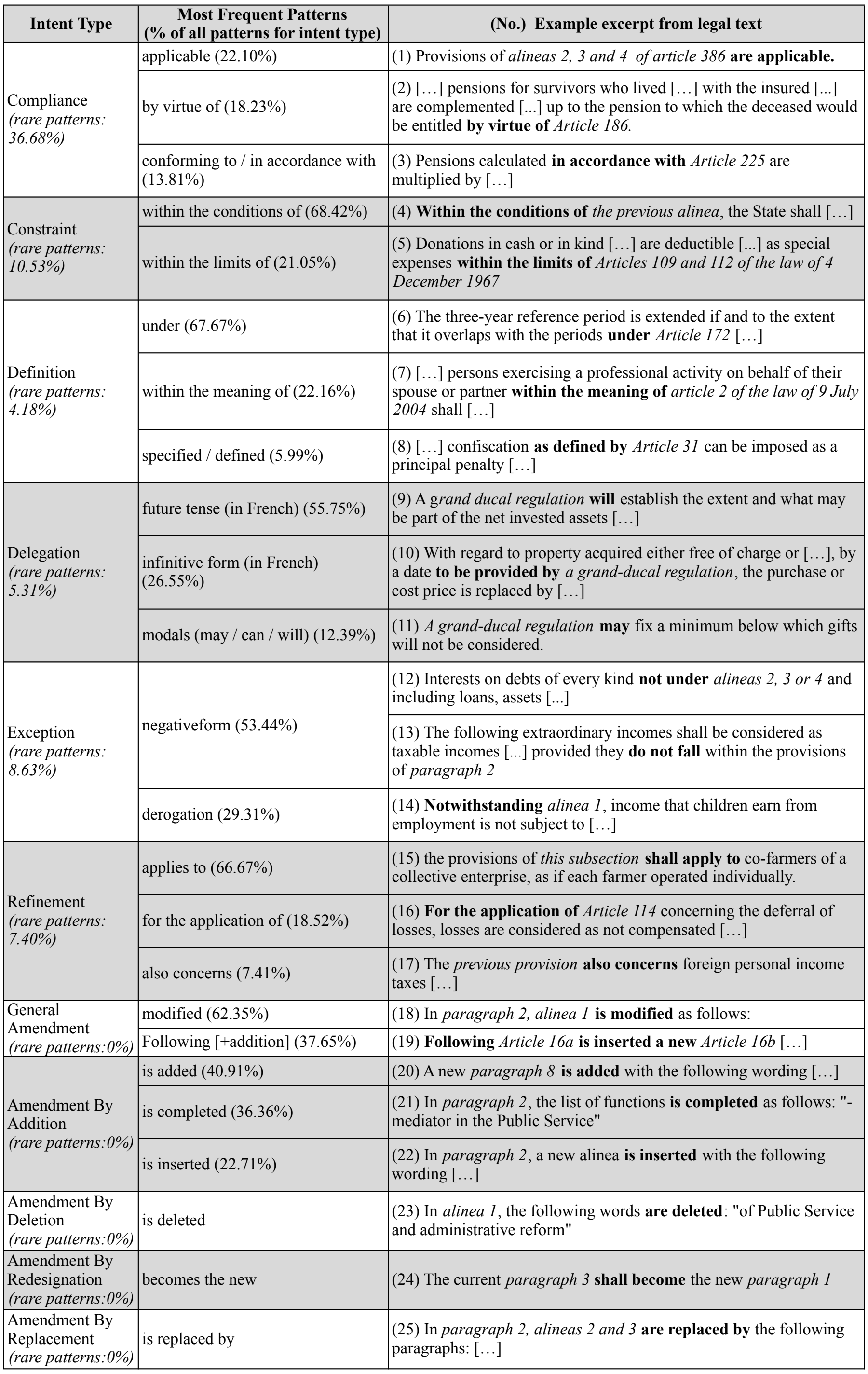
Intent Type	Definition	Mapping			Frequency	# of phrases	# of distinct patterns
		Breaux [3]	Hamdaqa et al. [13]	Maxwell et al. [17]			
Compliance	The cited provision(s) apply along with the citing provision.	--	compliance	--	16,03%	173	24
Constraint	The cited provision(s) introduce additional constraints.	constraint	--	constraint	1,76%	19	4
Definition	The cited provision(s) provide a definition.	definition	definition	definition	30,95%	334	7
Delegation	The citing provision delegates authority to an (often) unspecific legal text for further elaboration.	--	--	general	10,47%	113	4
Exception	The citing provision introduces an exception to the cited provision(s).	exception	--	exception	6,12%	66	11
Refinement	The citing provision elaborates upon the cited provision(s).	refinement	specification	--	2,50%	27	8
General Amendment	The citing provision amends the cited provision(s) without precisely stating what the modification(s) are.	--	--	--	15,01%	162	3
Amendment by Addition	The citing provision adds new provision(s) to the (single) cited provision.	--	Amend. by Addition	--	4,08%	44	6
Amendment by Deletion	The cited provision is deleted.	--	Amend. by Deletion	--	3,52%	38	3
Amendment by Redesignation	The cited provision's title or number is changed as per described in the citing provision.	--	Amend. by Redesignation	--	1,48%	16	1
Amendment by Replacement	The cited provision's wording is changed as per described in the citing provision.	--	Amend. by Striking	--	7,41%	80	1
Unclassified					*0,65%*	*7*	
Total					**100%**	**1079**	**72**

In particular, our intent types for *Compliance*, *Refinement*, and the various notions of amendment are aligned with Hamdaqa et al.'s; whereas, the rest are aligned with Maxwell et al.'s. We note that our choice of names for some intent types differs from those in the above taxonomies. This is mainly to provide better overall contrast between the types in our taxonomy.

Our current taxonomy does not envisage a type for CRs whose intent is *Prioritization*, as proposed by Maxwell et al. We cannot rule out the existence of such CRs in the Luxembourgish legal corpus, but draw attention to an absence of observations in our qualitative study. The main implication of this lack of observations is that our automated classification solution (Section 4) cannot handle CRs whose intent is prioritization. Furthermore and on a different note, our taxonomy does not cover the notions of *Unrelated* and *Incorrect* in the work of Maxwell et al. Determining the relevance and correctness of CRs is outside the scope of our current work.

Finally, as shown in Table 1, we were unable to classify a small fraction (0,65%) of the CRs in our study due to these CRs being too general or vague.

Table 2. NL Patterns Associated with Intent Types along with Examples [*]



Intent Type	Most Frequent Patterns (% of all patterns for intent type)	(No.) Example excerpt from legal text
Compliance *(rare patterns: 36.68%)*	applicable (22.10%)	(1) Provisions of *alineas 2, 3 and 4 of article 386* **are applicable.**
	by virtue of (18.23%)	(2) […] pensions for survivors who lived […] with the insured [...] are complemented [...] up to the pension to which the deceased would be entitled **by virtue of** *Article 186.*
	conforming to / in accordance with (13.81%)	(3) Pensions calculated **in accordance with** *Article 225* are multiplied by […]
Constraint *(rare patterns: 10.53%)*	within the conditions of (68.42%)	(4) **Within the conditions of** *the previous alinea*, the State shall […]
	within the limits of (21.05%)	(5) Donations in cash or in kind […] are deductible [...] as special expenses **within the limits of** *Articles 109 and 112 of the law of 4 December 1967*
Definition *(rare patterns: 4.18%)*	under (67.67%)	(6) The three-year reference period is extended if and to the extent that it overlaps with the periods **under** *Article 172* […]
	within the meaning of (22.16%)	(7) […] persons exercising a professional activity on behalf of their spouse or partner **within the meaning of** *article 2 of the law of 9 July 2004* shall […]
	specified / defined (5.99%)	(8) […] confiscation **as defined by** *Article 31* can be imposed as a principal penalty […]
Delegation *(rare patterns: 5.31%)*	future tense (in French) (55.75%)	(9) A g*rand ducal regulation* **will** establish the extent and what may be part of the net invested assets […]
	infinitive form (in French) (26.55%)	(10) With regard to property acquired either free of charge or […], by a date **to be provided by** *a grand-ducal regulation*, the purchase or cost price is replaced by […]
	modals (may / can / will) (12.39%)	(11) *A grand-ducal regulation* **may** fix a minimum below which gifts will not be considered.
Exception *(rare patterns: 8.63%)*	negativeform (53.44%)	(12) Interests on debts of every kind **not under** *alineas 2, 3 or 4* and including loans, assets [...]
		(13) The following extraordinary incomes shall be considered as taxable incomes [...] provided they **do not fall** within the provisions of *paragraph 2*
	derogation (29.31%)	(14) **Notwithstanding** *alinea 1*, income that children earn from employment is not subject to […]
Refinement *(rare patterns: 7.40%)*	applies to (66.67%)	(15) the provisions of *this subsection* **shall apply to** co-farmers of a collective enterprise, as if each farmer operated individually.
	for the application of (18.52%)	(16) **For the application of** *Article 114* concerning the deferral of losses, losses are considered as not compensated […]
	also concerns (7.41%)	(17) The *previous provision* **also concerns** foreign personal income taxes […]
General Amendment *(rare patterns:0%)*	modified (62.35%)	(18) In *paragraph 2, alinea 1* **is modified** as follows:
	Following [+addition] (37.65%)	(19) **Following** *Article 16a* **is inserted a new** *Article 16b* […]
Amendment By Addition *(rare patterns:0%)*	is added (40.91%)	(20) A new *paragraph 8* **is added** with the following wording […]
	is completed (36.36%)	(21) In *paragraph 2*, the list of functions **is completed** as follows: "-mediator in the Public Service"
	is inserted (22.71%)	(22) In *paragraph 2*, a new alinea **is inserted** with the following wording […]
Amendment By Deletion *(rare patterns:0%)*	is deleted	(23) In *alinea 1*, the following words **are deleted**: "of Public Service and administrative reform"
Amendment By Redesignation *(rare patterns:0%)*	becomes the new	(24) The current *paragraph 3* **shall become** the new *paragraph 1*
Amendment By Replacement *(rare patterns:0%)*	is replaced by	(25) In *paragraph 2, alineas 2 and 3* **are replaced by** the following paragraphs: […]

[*] In the examples (column 3), the CRs are *italicized* and the pattern occurrences are **bolded**.

The low incidence of manually-unclassifiable CRs makes it more likely that one can achieve good classification coverage through automation.

NL Patterns for Semantic Classification (RQ2). One of the most interesting observations from our qualitative study is that, for more than 98% of the CRs investigated, we could find a phrase located within the same sentence as a given CR to suggest what the intent of that CR is. As stated in Section 3.2, these phrases are the basis for the NL patterns that we have developed for classification. The patterns are partially listed and exemplified in Table 2. We do not provide in this paper the complete list of the identified phrases and the patterns derived from them. See [19] for details.

To build confidence in the usefulness of our patterns, we need to consider two important factors: (1) whether our qualitative study has covered a reasonably large number of observations for each intent type, and (2) whether the usage frequency of the patterns is reasonably high. A large proportion of patterns with very few occurrences, which we earlier denoted as rare, may indicate a large degree of flexibility in legal writing practices and hence a negative impact on the automatability of CR classification. Below, we discuss these factors for the intent types in our taxonomy based on the information in Tables 1 and 2.

Definition is the most represented intent type constituting nearly 31% of the entire set of CRs in our study. This intent type exhibits a relatively small number of patterns (7 patterns). The three most frequent patterns for this intent type cover more than 95% of the cases, with just over 4% of the patterns being rarely used. Similar observation can be made for the *Delegation* and *General Amendment* types; that is, the types are both well-represented and further have a dominant set of patterns that cover a large majority of cases.

The second most represented intent type is *Compliance*. In contrast to the ones discussed above, this intent type is associated with 24 distinct patterns, with a relatively high rate of rare patterns (≈37%).

The *Refinement* and *Constraint* types have a low representation in our qualitative study. At the same time, the number of rare patterns for these intent types is quite limited (7.53% and 10.53%, respectively),

Finally and with regard to amendment CRs, despite the limited representation of the individual intent types, the CRs are covered by a small number of dominant patterns. This could be either due to the lack of sufficient diversity in our units of analysis (mainly, the portion of Draft Law No. 6457 investigated in our study), or due to legal writing practices being stringent and systematic with regard to amendments.

Our analysis of the NL patterns further led to some technicalities that need to be taken into account for the development of an automated classification tool. First, the occurrences of the NL patterns may not be immediately before or after the CRs. In particular, some *auxiliary phrases*, e.g., "the provisions of", may appear between a pattern occurrence and a CR, e.g., in "[...] as **mentioned in** the provisions of *article 2*". In our qualitative study, we kept track of all the auxiliary phrases encountered, recording a total of 95 of them. Due to the potentially large set of possible auxiliary phrases, providing sufficient coverage of such phrases through patterns seems unlikely to be effective. Nevertheless, we observed that the length of the auxiliary phrases (in terms of tokens) is short. More precisely, the average length of an auxiliary phrase in our study is 2.6 tokens, with the longest phrase observed being five tokens long.

To deal with auxiliary phrases without having to enumerate them all, one can implement a strategy to look back and ahead by a certain number of tokens from where a CR is located when searching for patterns. Based on our study, we recommend that a pattern occurrence as far away as 5 tokens from a given CR should be considered, as long as the occurrence is within the same sentence as the CR and the location of the occurrence matches the before/after property maintained for the underlying pattern (illustrated in Fig. 2). Since this look-back / look-ahead distance is short ($\leq$ 5 tokens), the risk of the CR and the pattern occurrence being in different contexts (and thus, the risk of incorrectly associating the pattern to the CR) is low.

Second, different grammatical variants of the same phrase may imply different intent types and thus different patterns. For instance, the French phrase "prévu" ("under", in English) suggests a *Definition* (Example 6 in Table 2); whereas the negative form of the phrase, "non prévu" ("not under", in English), suggests an *Exception* (Example 12), and the infinitive form of the phrase, "à prévoir" ("to be provided by", in English), suggests a *Delegation* (Example 10). Similarly, the *Compliance* and *Refinement* intent types have similar associated patterns (Examples 1, 15, 16).

Given what we stated above, one cannot simply use the root forms of terms as the basis for defining patterns. In a similar vein, preprocessing techniques commonly used in Information Retrieval, particularly stemming [18] and similarity measures [14], may yield poor results if applied for CR intent classification.

4 Automated Classification of Cross References (RQ3)

We have developed an automated CR intent classifier based on the results of RQ1 and RQ2 in the previous section. The classifier, which is built as an application within the GATE NLP Workbench [7], works in two steps:

1. It runs our previously-developed CR detector [20] to identify and mark the CRs in a given corpus.
2. Using the NL patterns of RQ2, the classifier attempts to assign an intent type to each detected CR.

To deal with auxiliary phrases, our classifier applies the look-back / look-ahead strategy discussed previously. If multiple overlapping pattern matches are found for a CR, the longest match (in terms of the number of characters in the matching region) determines the CR type.

In the rest of this section, we report on two cases studies aimed at evaluating the accuracy of our classifier. We exclude a re-evaluation of our CR detection technique (the first step), for which we already provided empirical results in our previous work [20].

4.1 Case Study over Luxembourgish Legal Texts

Our first case study is over selected legislative texts from the Luxembourgish legal corpus. The texts cover a long time span –from 1808 to 2014– and several domains, including, among others, the civil code, social security, trade, and data protection. To avoid biasing the results, the two texts in our qualitative study of Section 3 were excluded from the selection. Overall, the selected texts have 1830 pages, excluding non-content pages such as prefaces, tables of contents, and indices. We ran our classifier over these pages. We then randomly picked 10% of the pages (183 pages) for a manual inspection of the classification results.

The random page sample contains a total of 1396 (detected) CRs. The first author reviewed the classification results for all the CRs in the sample and computed, for every intent type X of Table 1, the following four counts:

(c1) *Correctly Classified*: The number of CRs of type X for which automated classification is correct.
(c2) *Incorrectly Classified, Type 1*: The number of CRs that were assigned type X by automated classification, but the correct type is in fact different.
(c3) *Incorrectly Classified, Type 2*: The number of CRs that are of type X, but were assigned a different type by automated classification.
(c4) *Unclassified*: The number of CRs of type X for which automated classification yields no intent type.

Using these counts, we compute the accuracy of automated classification through recall, precision, and F-measure. To do so, we note that c1 denotes *True Positives (TP)*, c2 denotes *False Positives (FP)*, whereas c3 and c4 denote *False Negatives (FN).* Recall is computed as $R = TP/(TP + FN)$, precision as $P = TP/(TP + FP)$, and F-measure as $F = (2 * P * R)/(P + R)$.

The results of automated classification at the level of individual intent types and at an aggregate level are presented in Table 3. Overall, our classifier provided a correct classification for 1113 CRs (c1), an incorrect classification for 33 CRs (c2 and c3), and no classification for 250 CRs (c4). These counts are respectively given in columns 3–6 of the table. We note that c2 and c3 are redistributions of one another; nevertheless, both counts are important, as a false positive for one intent type implies a false negative for another. The classification accuracy metrics are given in columns 7–9. For this case study and at an aggregate level, our classifier has a recall of 79.73% and a precision of 97.12%, thus giving an F-measure of 87.57%.

From the table, we observe that nearly half (16/33) of the incorrect classifications are *Refinement* CRs being erroneously classified as *Compliance* ones. These misclassified CRs are explained by the similarities between the patterns associated with the two intent types in question, as we discussed in Section 3 (under RQ2). A further six classification errors are *Delegation* CRs being classified as *Definition* ones. All these cases were due to an individual variant of an existing pattern for *Delegation* CRs being missing from our pattern catalog.

With regard to unclassified CRs (column 6), 153 cases were due to missing patterns. Our subsequent investigation of these cases resulted in the identification of 75 new patterns. Of these, 60 had less than three occurrences and

Table 3. Classification Results for Luxembourgish Legal Texts

Intent Type	Total CRs	Correctly Classified (TP)	Incorrectly Classified T1 (FP)	Incorrectly Classified T2 (FN)	Unclassified (FN)	Recall	Precision	F-Measure
Compliance	415	334	20	2	79	80,48%	94,35%	86,87%
Constraint	23	4	0	1	18	17,39%	100,00%	29,63%
Definition	548	511	9	3	34	93,25%	98,27%	95,69%
Delegation	93	85	2	6	2	91,40%	97,70%	94,44%
Exception	56	43	2	3	10	76,79%	95,56%	85,15%
Refinement	81	13	0	16	52	16,05%	100,00%	27,66%
General Amendment	61	48	0	1	12	78,69%	100,00%	88,07%
Amend. by Addition	33	28	0	0	5	84,85%	100,00%	91,80%
Amend. by deletion	8	6	0	0	2	75,00%	100,00%	85,71%
Amend. by redesignation	2	1	0	0	1	50,00%	100,00%	66,67%
Amend. by replacement	45	40	0	0	5	88,89%	100,00%	94,12%
Unclassifiable	31	0	0	1	30	NA		
total	1396	1113	33	33	250	79,73%	97,12%	87,57%

would fall under rare patterns, as defined in Section 3. Another 27 unclassified CRs were explained by missing variants of already-known patterns. A further 47 cases where due to the patterns being located more than 5 token away from the CRs, i.e., outside the classifier's look-back / look-ahead range discussed earlier.

During our manual inspection, we encountered 31 CRs whose intent we could not determine due to vagueness. These cases are shown as *Unclassifiable* in Table 3. Our classifier left 30 of these CRs unclassified but matched one to an unrelated pattern (because of our 5-token look-back and look-ahead strategy). When calculating the overall accuracy of our classifier, we take a conservative approach for the unclassifiable cases. In particular, we treat all these cases as false negatives (FN), meaning that we assume a subject matter expert would have been able to determine what the intents of these CRs are.

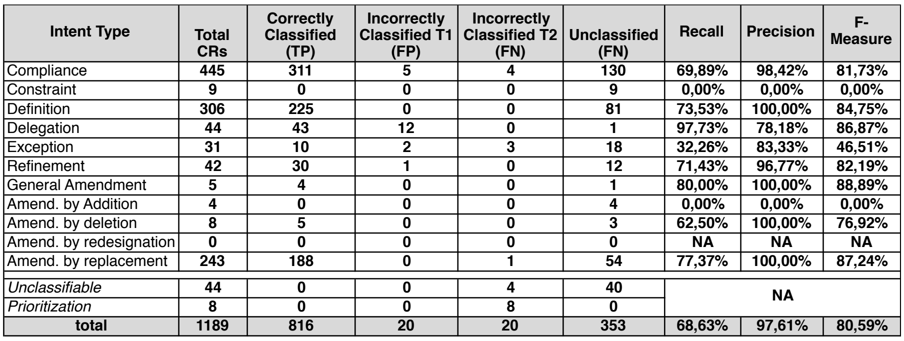
Finally, we observe from Table 3 that recall is low for the *Constraint* and *Refinement* types. This provides evidence for our hypothesis from Section 3 about these two types lacking sufficient representation in our qualitative study.

4.2 Case Study over Canadian Legal Texts

Our second case study is a step towards assessing the generalizability of our approach in other countries where French is an official language of the law. Specifically, we run our classifier *as-is* (i.e., without extending our qualitative study of Section 3) to the French editions of two Canadian legal texts. These texts are: Ontario's Personal Health Information Protection Act (PHIPA) [21] and the 2014 compilation of the amendment provisions on Canadian consolidated laws [22]. PHIPA is a major legal text, which has been already studied in the RE community [9, 10] due to its important implications on software requirements in healthcare systems. The second text is aimed at enabling the evaluation of amendments CRs, which are underrepresented in PHIPA.

We ran our classifier over these two texts, which collectively contain 87 content pages. The first two authors then inspected all the classification results. Our classifier detected a total of 1189 CRs in the texts, of which, it could infer types

Table 4. Classification Results for Canadian Legal Texts



Intent Type	Total CRs	Correctly Classified (TP)	Incorrectly Classified T1 (FP)	Incorrectly Classified T2 (FN)	Unclassified (FN)	Recall	Precision	F-Measure
Compliance	445	311	5	4	130	69,89%	98,42%	81,73%
Constraint	9	0	0	0	9	0,00%	0,00%	0,00%
Definition	306	225	0	0	81	73,53%	100,00%	84,75%
Delegation	44	43	12	0	1	97,73%	78,18%	86,87%
Exception	31	10	2	3	18	32,26%	83,33%	46,51%
Refinement	42	30	1	0	12	71,43%	96,77%	82,19%
General Amendment	5	4	0	0	1	80,00%	100,00%	88,89%
Amend. by Addition	4	0	0	0	4	0,00%	0,00%	0,00%
Amend. by deletion	8	5	0	0	3	62,50%	100,00%	76,92%
Amend. by redesignation	0	0	0	0	0	NA	NA	NA
Amend. by replacement	243	188	0	1	54	77,37%	100,00%	87,24%
Unclassifiable	44	0	0	4	40	NA		
Prioritization	8	0	0	8	0			
total	1189	816	20	20	353	68,63%	97,61%	80,59%

for 816, leaving the remaining 353 unclassified. We calculated the same counts (c1–c4) as in the previous case study (Section 4.1). The results are shown in Table 4. For this case study, the classifier has a recall of 68.63% and a precision of 97.61%, giving an F-measure of 80.59%. We observe that the precision score for this case study is in the same range as that for the previous one; whereas the recall score is lower by ≈11%. Some decrease in recall was to be expected due to the potentially-different legal drafting styles and thus the use of new patterns. In particular, the patterns required for the *Constraint* type were absent from our catalog, resulting in all CRs of this type to go unclassified.

A total of 20 CRs were misclassified. All these cases were caused by unrelated pattern being present in the vicinity of the CRs in question. Our inspection further revealed eight CRs of the *Prioritization* type [17]. As stated earlier, we had not encountered any such CRs in our qualitative study. Consequently, our patterns did not cover this particular type. All the *Prioritization* CRs seen in this case study used the same pattern, which we denote "prevails" (l'emporte), e.g., in "[...] this act and its regulations prevail unless [...]".

With regard to the *Compliance* and *Refinement* types, we observed that, unlike in the first case study, the patterns used for CRs of these types were sufficiently distinct. No misclassification occurred due to our classifier failing to tell apart CRs of these two types.

With regard to the CRs that our tool could not classify, the same observations as those in the previous case study hold, although the proportions differ. A noteworthy difference in the proportions is that we had more CRs not being classified because of long auxiliary phrases. The increase in the length of auxiliary phrases is mainly due to the bilingual context of the Canadian legal corpus, where one has to additionally differentiate between the French and English editions of the laws in the auxiliary phrases. One way to deal with longer auxiliary phrases would be to increase the acceptable distance between the patterns and the CRs (currently 5 tokens, as stated earlier). Doing so however necessitates further investigation because such an increase could lead to reductions in precision caused by the potential presence of unrelated patterns at farther distances.

Lastly, we note that the number of CRs that were deemed *Unclassifiable* by our manual inspection was proportionally larger in this case study than in the previous one (44/1189 versus 20/1396). We believe that this discrepancy is partly due to the more hierarchical nature of Canadian laws, where federal, provincial, and territorial laws co-exist, thus leaving room for more vague citations.

5 Discussion

Usefulness of our approach. The ultimate validation for our approach is whether practitioners who work with legal requirements would benefit from our automatic classification results. Such validation requires a user study which is not tackled in this paper. Nevertheless, the case studies of Section 4 provide some preliminary insights about usefulness. In particular, we observe that, over these case studies, our approach yields an average F-measure of 84.48%, with an average recall and an average precision of 74.62% and 97.33%, respectively. The high precision indicates that users need to spend little effort on finding and correcting errors in the classification recommended by our approach. At the same time, the recall suggests that our approach is capable of classifying nearly three quarters of the CRs. This, in light of the high precision, is expected to lead to significant savings in manual effort.

Considering the limited size of our qualitative study (1079 CRs from two texts), the results are encouraging. We believe that recall can be further improved through additional case studies and iteratively expanding the NL patterns.

Threats to validity. The most important aspects of validity for our work are internal and external validity. Below, we discuss threats to these forms of validity.

Internal validity: The main threat to internal validity is related to the correctness of the taxonomy and the patterns derived from our qualitative study. To mitigate this threat, the first two authors (who are Francophone and further have legal requirements engineering background, as noted earlier), worked closely together throughout the qualitative study. An additional mitigation measure we applied was to build on and align with existing taxonomies as much as possible.

Another potential threat to internal validity is that we may have associated some NL patterns with the wrong intent types. This does not pose a major problem as one can move patterns from one intent type to another, without affecting overall classification accuracy. Finally, we note that the automated classification results in Section 4 were inspected by the authors. To avoid bias, we discussed and developed, based on our experience from the qualitative study, an inspection protocol prior to conducting the inspections.

External validity: We distinguish two dimensions for external validity: (1) generalizability to texts which are written in French, but which come from different countries or jurisdictions than what we considered here, and (2) generalizability to texts written in languages other than French. With regard to (1), external validity mainly has to do with the completeness and relevance of our patterns outside the context in which they were observed. While more case studies are required, the good results from our second case study provide initial evidence for this type of generalizability. With regard to (2), qualitative studies over legal texts written in other languages such as English will be needed. Further investigation of bilingual texts, e.g., from the Canadian legal corpus, will provide an opportunity to study the generalization of our approach to other languages while at the same time establishing a connection to our current results in French.

6 Conclusion

We proposed an approach for the automated classification of cross references in legal texts according to the cross references' semantic intent. Our approach is motivated by providing requirements engineers with tools and support for more efficient and effective elaboration of legal compliance requirements. The basis for our approach is a qualitative study of selected Luxembourgish legislative texts. Through this study, we derived a taxonomy of semantic intent types for cross references along with natural language patterns that enable distinguishing these types in an automated manner. We conducted an empirical evaluation of our automated classification approach over Luxembourgish and Canadian legal texts, demonstrating that the approach yields good accuracy. The promising evaluation results for Canadian legal texts further provides evidence about the generalizability of our approach, noting that the observations in our qualitative study were based exclusively on the Luxembourgish legal corpus.

In the future, we would like to conduct a more thorough evaluation of our approach. In particular, we plan to more closely examine the completeness of our natural language patterns for classification by conducting a series of case studies in succession. This will enable us to have a feedback loop between the case studies and measure whether our catalog of patterns will saturate as it is iteratively extended. Another facet of investigation would be to study legal texts written in other languages, e.g., English, to validate the basic observations behind our approach. Finally, user studies will be necessary to more conclusively determine whether our approach brings about benefits in realistic settings.

Acknowledgments

Financial support for this work was provided by Luxembourg's National Centre for Information Technologies (CTIE) and National Research Fund (FNR) under grant number FNR/P10/03. We thank members of CTIE, particularly Ludwig Balmer and Marc Blau, for their valuable feedback.

References

1. M. Adedjouma, M. Sabetzadeh, and L. Briand. Automated detection and resolution of legal cross references: Approach and a study of Luxembourg's legislation. In *RE'14*, pages 63–72, 2014.

2. C. Biagioli, E. Francesconi, A. Passerini, S. Montemagni, and C. Soria. Automatic semantics extraction in law documents. In *ICAIL'05*, pages 133–140, 2005.
3. T. Breaux. *Legal Requirements Acquisition for the Specification of Legally Compliant Information Systems.* PhD thesis, North Carolina State University, Raleigh, North Carolina, USA, 2009.
4. T. Breaux and A. Antón. Analyzing regulatory rules for privacy and security requirements. *IEEE TSE*, 34(1):5–20, 2008.
5. R. Brighi. An ontology for linkups between norms. In *DEXA Workshops*, pages 122–126, 2004.
6. J. Corbin and A. Strauss. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory.* SAGE Publications, 3rd edition, 2008.
7. Cunningham et al. Developing Language Processing Components with GATE Version 7 (a User Guide).
8. E. de Maat, R. Winkels, and T. van Engers. Automated detection of reference structures in law. In *JURIX'06*, pages 41–50, 2006.
9. S. Ghanavati, D. Amyot, and L. Peyton. Towards a framework for tracking legal compliance in healthcare. In *CAISE'07*, pages 218–232, 2007.
10. S. Ghanavati, A. Rifaut, E. Dubois, and D. Amyot. Goal-oriented compliance with multiple regulations. In *RE'14*, pages 73–82, 2014.
11. Government of Luxembourg. Draft Law No 6457 of the Regular Session 2011-2012 of the Chamber of Deputies, 2012.
12. Government of Luxembourg. Modified Law of Dec. 4, 1967 on Income Taxes, 2014.
13. M. Hamdaqa and A. Hamou-Lhadj. An approach based on citation analysis to support effective handling of regulatory compliance. *Future Generation Computer Systems*, 27(4):395–410, 2011.
14. C. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval.* Cambridge University Press, 2008.
15. A. Massey, B. Smith, P. Otto, and A. Anton. Assessing the accuracy of legal implementation readiness decisions. In *RE'11*, pages 207–216, 2011.
16. J. Maxwell, A. Antón, and J. Earp. An empirical investigation of software engineers' ability to classify legal cross-references. In *RE'13*, pages 24–31, 2013.
17. J. Maxwell, A. Antón, P. Swire, M. Riaz, and C. McCraw. A legal cross-references taxonomy for reasoning about compliance requirements. *REJ*, 17(2):99–115, 2012.
18. M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130137, 1980.
19. N. Sannier, M. Adedjouma, M. Sabetzadeh, and L. Briand. Supplementary Material for Automatic Classification of Legal Cross References Based on Semantic Intent. http://people.svv.lu/sannier/CRSemantics/, 2015.
20. Nicolas Sannier, Morayo Adedjouma, Mehrdad Sabetzadeh, and Lionel Briand. An automated framework for detection and resolution of cross references in legal texts. *Requirements Engineering*, 2015. (in press).
21. The Ontario Ministry of Consumer and Business Services and the Ontario Ministry of Health and Long Term Care. Personal Health Information Protection Act, 2004.
22. The Parliament of Canada. Canada Corrective Act, 2014.
23. O. Tran, N. Bach, M. Nguyen, and A. Shimazu. Automated reference resolution in legal texts. *Artif. Intell. Law*, 22(1):29–60, 2014.
24. N. Zeni, N. Kiyavitskaya, L. Mich, J. Cordy, and J. Mylopoulos. GaiusT: Supporting the extraction of rights and obligations for regulatory compliance. *REJ*, 20(1):1–22, 2015.