Towards a High-Level Controlled Language for Legal Sources on the Semantic Web

Adam Wyner[1(✉)], Adeline Nazarenko[2], and François Lévy[2]

[1] University of Aberdeen, Aberdeen, UK
azwyner@abdn.ac.uk
[2] LIPN, Paris 13 University – Sorbonne Paris Cité & CNRS, Paris, France
{adeline.nazarenko,francois.levy}@lipn.univ-paris13.fr

Abstract. Legislation and regulations are required to be structured and augmented in order to make them serviceable on the Internet. However, it is known that it is complex to accurately parse and semantically represent such texts. Controlled languages have been one approach to adjusting to the complexities, where the source text is rewritten in some systematic form. Such an approach is not only costly, but potentially introduces alternative translations which may be undesirable. To navigate between the requirements and complexities, we propose and exemplify a high-level controlled language that serves as an XML representation for key components of legal content. The language tightly correlates to the source text and also facilitates analysis.

Keywords: Natural language simplification · Semantic annotation · Legal rules · Controlled languages · Semantic web

1 Introduction

The increasing complexity and integration of legal documents and regulations calls for rich legal content management. However, the complexity of legal language and regulations has long been understood to be an obstacle to the development of legal content management tools; for example, as discussed in [15], the complexity and ambiguity of the resulting parses and semantic representations make them difficult to evaluate for correctness as well as to exploit for experts in formal languages, *a fortiori* for legal analysts. The legal semantic web aims at giving a uniform access to legal sources, whatever form they may take or the institution that published them. This is traditionally supported by the definition of a metadata vocabulary and the semantic annotation of the sources. Beyond documents and topic-based annotations, however, legal experts must have direct access to the rules contained in documents and their supported interpretations. This calls for a rich and structured annotation of the rule text fragments.

However, problematics arise from the tensions between the complexities of legal natural language, the requirements of legal professionals, and the specifications of formal or machine-readable languages. In this paper, we attempt to

B. Davis et al. (Eds.): CNL 2016, LNAI 9767, pp. 92–101, 2016.
DOI: 10.1007/978-3-319-41498-0_9

moderate the tensions using a simplified, yet useful controlled language (CL) to mark up the source text. The novel, significant contribution of this paper is the advocation for an analysis and annotation of legal sources using structured annotations, which is our high-level CL (hCL), on the source text. This hCL leaves the source text *in situ*. We claim that the annotations can be semi-automatically associated with NL expressions, and moreover, that the annotations can be associated with representations in XML. In the advocated approach, one rule in the source text can be annotated with different CL statements so as to represent different interpretations, thus leaving it up to the analyst to resolve ambiguities. Furthermore, our hCL focuses on the semantic structure of the rules, providing an abstract representation of the components of a proposition rather than a collection of annotations; as an analysis of sentence components, it is similar to a parser, yet it focuses on semantic annotations that are key to rules. The approach combines the source text for reference, the controlled language annotations for experts, and the semantic representation for Semantic Web querying.

To ground our discussion and provide a running example, we use a corpus that was previously reported in [16], which is a passage from the US Code of Federal Regulations, US Food and Drug Administration, Department of Health and Human Services regulation for blood banks on testing requirements for communicable disease agents in human blood, Title 21 part 610 Section 40. We present a running example from this corpus.

In the remainder of the paper, we outline existing research to contrast with our proposal (Sect. 2). We sketch our annotation approach to our hCL in Sect. 3. We present a specification of the hCL (Sect. 4), then present an incremental methodology by example for annotating legal sources with hCL statements (Sect. 5). The paper closes with a summary and some indications of future work.

2 Related Work

The complexity of legal language and regulations has long been understood to be an obstacle to the development of legal content management tools. Attempts have been made to parse and automatically formalize legal texts. For instance, C&C/Boxer [3] has been applied to fragments of regulations [15]. C&C/Boxer is a wide coverage parser that feeds a tool which generates semantic representations (essentially in First-order Logic). However, as discussed in [15], the complexity and ambiguity of the resulting parses and semantic representations make them difficult to evaluate for correctness as well as to exploit for experts in formal languages, *a fortiori* for legal analysts.

Controlling the legal sources has been proposed as an alternative approach. Efforts are made to clarify and simplify the legal language when drafting (*e.g.* in favor of "Plain English", to ease translation [12], or to avoid ambiguity and clumsiness [8]). More formally, a wide range of controlled languages (CL) has been proposed [9], with the idea that controlled statements would be easier to parse but still be meaningful and manageable. *Attempto Controlled English (ACE)* defines unambiguous readings of quantifier scopes and anaphora as well as prohibits ambiguous syntactic attachments, thus enabling a parse and translation into predicate logical formulae. The *Oracle Policy Modelling* (OPM) system [4] is designed to parse structured sets of controlled sentences and make rulebases available online. *Semantics of Business Vocabulary and Business Rules* (SBVR) has been specifically designed to model business rules [13]: it provides elements of a pattern language and a description of *SBVR-Structured English* to express rules in a form that can be checked by human experts. ACE, OPM, and SBVR try to systematize the NL to CL translation by proposing alternative formulations for unwanted constructions. However, when the source regulations get more complex, the NL to CL translation either fails or gives a formal result, with explicit scopes and qualifiers, which can be difficult to read and even harder to adjudicate for correctness.

A third approach relies on the semantic annotation of legal texts without diving into the detailed syntactic complexity of legal sentences. Annotations are made at the paragraph level, making use both of a high level legal ontology and a specific domain ontology [1]. Provision level annotations are given, which rely on a general model of relationships between normative provisions [6,14]. The provision collection is encoded in RDF-OWL and can be queried using SPARQL. In a similar vein, the LegalRuleML mark-up language is designed to represent legal rules for the semantic web [2].

These approaches share a common disposition – the source legal language must itself be normalized, transformed, and disambiguated in order to be systematically represented. This may not be feasible without unduly constraining the scope of analysis and of interpretation. A pragmatic proposal is to combine the controlled language and the semantic annotation approaches as initially proposed in [10], which provides some content of semantic annotations and fixes the interpretation of underlying fragments of legal sources. This approach builds on SemEx methodology, which was designed to annotate business regulations by business rules through an iterative rewriting process, ideally until a CL form is obtained [7]. However, a full specification of CL seems problematic. In this paper, we focus on key textual components to represent the main legal features rather than the details of domain terminology.

3 Annotating Legal Content with hCL

Formalization of legal documents yields representations that support content management (indexing and search, merge, comparison, update of documents) and legal reasoning (*Is it necessary to test X for Y?* and *If X, then Y*). However, completely formalising the content of legal documents is a distant goal, due to legal and domain-specific terminology, long and complex sentences as well as ambiguities. It would, nonetheless, be useful for legal content management and reasoning to provide a degree of formalization as structured annotations. We develop the formalization *pragmatically* and *partially*; it is pragmatic as we only annotate those components as needed for the analysis of rules, and it is partial in that we allow mixing of annotations and unannotated source text.

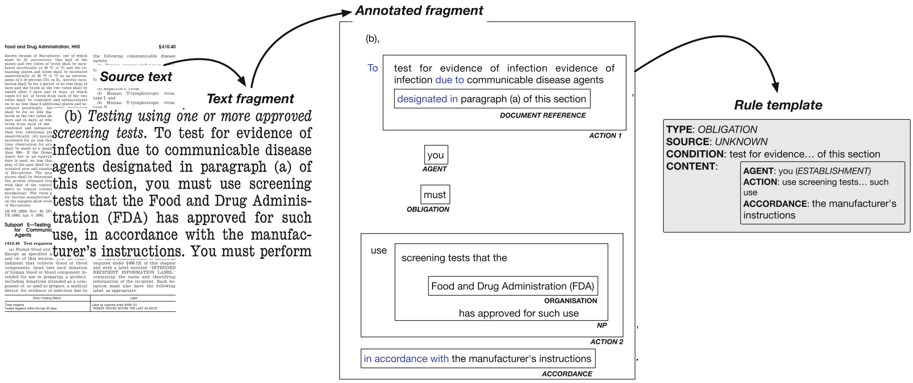
Annotated fragment
Source text
Text fragment
(b) Testing using one or more approved screening tests. To test for evidence of infection due to communicable disease agents designated in paragraph (a) of this section, you must use screening tests that the Food and Drug Administration (FDA) has approved for such use, in accordance with the manufacturer's instructions. You must perform
(b),
To
test for evidence of infection evidence of infection due to communicable disease agents
designated in paragraph (a) of this section
DOCUMENT REFERENCE
ACTION 1
you
AGENT
must
OBLIGATION
use
screening tests that the
Food and Drug Administration (FDA)
ORGANISATION
has approved for such use
NP
ACTION 2
in accordance with the manufacturer's instructions
ACCORDANCE
Rule template
TYPE: OBLIGATION
SOURCE: UNKNOWN
CONDITION: test for evidence… of this section
CONTENT:
AGENT: you (ESTABLISHMENT)
ACTION: use screening tests… such use
ACCORDANCE: the manufacturer's instructions

Fig. 1. Example of an annotated rule statement and of the semantic pattern that can be derived from the hCL annotation.

We focus our analysis on legal rules and their components. To illustrate our approach, we show an analysis of our running example below as in Fig. 1, which is further discussed in Sect. 4:

> *(b)* To test for evidence of infection due to communicable disease agents designated in paragraph (a) of this section, you must use screening tests that the Food and Drug Administration (FDA) has approved for such use, in accordance with the manufacturer's instructions.

The analysis relies on the following intuitions:

- Formalization in hCL adds annotations to the source text, enriching it and leaving it unchanged.
- hCL aims at providing simplified and semantically more explicit versions of the components of rule statements and their integration as rules.
- Formal statements can be expressed through form-based semantic structures for rules. These forms are usually filled with high level annotations, either because there is a straightforward correspondence between the annotations and the semantic roles or because they correspond to a characteristic pattern of annotations. In any case, experts play a key role in the formal labeling of patterns.
- Annotated analyses can be *folded* and *unfolded* so that annotations can accommodate various granularities of formalization. A fully folded analysis is just the annotations (perhaps along with some keywords); a fully unfolded analysis is just the source text.

Interpretation of legal texts is important in legal reasoning, where there always remains room for interpretation. Annotating therefore amounts to specifying an interpretation through the selection of the most important fragments of the source regulation and the clarification of the semantic structure of the rules,

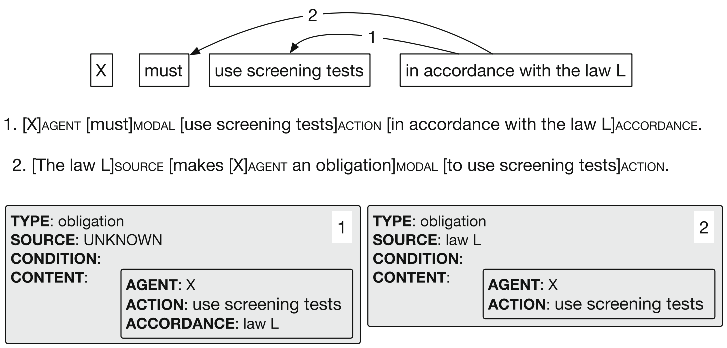
2
1
X
must
use screening tests
in accordance with the law L
1. [X]AGENT [must]MODAL [use screening tests]ACTION [in accordance with the law L]ACCORDANCE.
2. [The law L]SOURCE [makes [X]AGENT an obligation]MODAL [to use screening tests]ACTION.
TYPE: obligation
SOURCE: UNKNOWN
CONDITION:
CONTENT:
AGENT: X
ACTION: use screening tests
ACCORDANCE: law L
1
TYPE: obligation
SOURCE: law L
CONDITION:
CONTENT:
AGENT: X
ACTION: use screening tests
2

Fig. 2. Alternative annotation of an ambiguous source sentence: The ambiguity concerns the attachment of the prepositional phrase *in accordance with the law L* to the modal or main verb (Readings 1 and 2, resp.).

e.g. the relationships amongst the fragments. Figure 2 shows how the annotation in hCL highlights two alternative readings of an ambiguous text fragment.

Note that the original terms (*e.g. screening tests* in Fig. 2) are preserved in the annotations, so that their applicability to actual cases can be discussed in legal terms.

We propose that the content of the annotation be a linguistic expression to preserve readability. In the hCL design, focus is put on the constituent clauses of the rule statements, which are associated with an explicit and unambiguous semantics (see Fig. 2), leaving aside the detailed parsing of the constituents. These may remain unanalyzed (*e.g. use screening tests* in Fig. 2). This approach hides ambiguities and complexities of the lower level of analysis (*e.g.* the anaphoric expression *this section*) to highlight the main structure of the rule statements. Our presumption is that such ambiguities and complexities are either left unanalyzed or are treated by auxiliary processing components. Yet, our approach remains flexible: two different analysts may propose compatible readings even if one is more coarse-grained than the other.

Annotations can be exploited for content management and legal content mining with respect to rules. High level annotations homogenize linguistic variation and are used for search, instead of searching by keywords. This allows for answering queries like: *Which rules appear in a document?*, *Do the analysts agree on the interpretation of a specific rule statement or more generally on a section of a document?*, *What are all the rules that concern X?*, and *What are the rules involving a given action?*.

4 Specification of hCL

This section presents hCL and describes how this language can be used to analyze rule statements so as to make explicit the overall semantic structures of the rules. The goal is not to have a complete analysis of the rule statements, but rather to structure and index the rules in a systematic and explicit way so as to enable users to mine the legal sources as semi-structured, semantically annotated documents, yet leaving the source text unchanged.

We assume here (and discuss in Sect. 5) that we can semi-automatically identify semantic annotations, *e.g.* AGENT, with relevant syntactic phrases, *e.g.* noun phrase. Thus, the basic terminology of hCL is: AGENT, THEME, DEONTIC, ACTION, STATE, ACCORDANCE.

Each of these may be realized by a variety of syntactic expressions, so abstracting over linguistic heterogeneity. The AGENT and THEME correlate to noun phrases, DEONTIC correlates to various expressions of deontic modality, *e.g. must, may, prohibited, obligated*, ACTION and STATE are verb phrases, and ACCORDANCE is an adjunct phrase. The ACTION and STATE annotations, as verb phrases, incorporate their verb phrase internal arguments (*i.e.* direct objects and indirect objects). Given the underspecified approach adopted here, modals, actions, and states can be positive or negative.

The simplest rule pattern of hCL is[1]:

$$\text{RULE} \leftarrow \text{AGENT DEONTIC ACTION} \tag{1}$$

In this schema, AGENT, DEONTIC and ACTION elements refer to text fragments that have been annotated as such. Over 30 rule statements, 16 occurrences of this rule schema can be found in our text example. For instance, in *You must use screening tests*, where *you* is the AGENT, *must* is DEONTIC, and *use screening tests* is the ACTION.

This is a simple example. Given complex legal statements, the correspondence is often more complex: in *You must use screening tests that the Food and Drug Administration (FDA) has approved for such use, in accordance with the manufacturer's instructions.*, *you* is the AGENT, *must* is the DEONTIC, *use screening tests that the Food and Drug Administration (FDA) has approved for such use* is the ACTION, and *in accordance with the manufacturer's instructions* is the ACCORDANCE. In the example text, simple rules on average 44 words long (between 22 and 73).

While we can homogenize some linguistic variation, there are other variations we want to explicitly represent such as the active-passive (semantically annotated as ACTION and STATE) and the optionality of adjuncts. We assume that the AGENT of a passivised verb phrase (here indicated as STATE) is not a verb phrase internal argument and is optional. The ACCORDANCE annotation is, as an adjunct phrase, optional. Thus, rule pattern 1 can be revised to:

$$\text{RULE} \leftarrow \text{AGENT DEONTIC ACTION [ACCORDANCE]} \tag{2}$$

$$\text{RULE} \leftarrow \text{THEME DEONTIC STATE [AGENT][ACCORDANCE]} \tag{3}$$

As discussed above, legal rule statements can be complex and of various forms. It is not our intention to incorporate into the hCL all possible constructions, variants, and terminologies. We keep our annotations to a small feasible set and allow key words in the text *in-situ* to co-occur with the annotations (recalling lowercase words, such as *if* and *provided that*, refer to key words in the source documents). Such a mixed approach still allows the annotations to be useful for content management and reasoning without

[1] We have the following typographic conventions: capitalized elements (AGENT) refer to high level annotated textual fragments, brackets indicate optional elements, and lowercase words (*if*) refer to actual key words to be found in the source documents.

fully spelling out all details of the language of the source text. We allow that annotations can be further specified, *e.g.* RULE$_{perm}$ is a rule annotation where the DEONTIC is a *permission*:

$$\text{RULE} \leftarrow \textit{if}\ \text{AGENT ACTION},\ \text{RULE} \tag{4}$$

$$\text{RULE} \leftarrow \textit{if}\ \text{THEME STATE},\ \text{RULE} \tag{5}$$

$$\text{RULE} \leftarrow \text{RULE}_{perm}\ \textit{provided that}\ \text{AGENT ACTION} \tag{6}$$

$$\text{RULE} \leftarrow \text{RULE}_{perm}\ \textit{provided that}\ \text{THEME STATE} \tag{7}$$

All together these patterns cover 4/5 of the rule statements of our source text. We give few examples below:

1. [[you, an establishment that collects blood or blood components]$_{AGENT}$, [must]$_{DEONTIC}$ [test each donation of human blood or blood component intended for use in preparing a product for evidence of infection]$_{ACTION}$]$_{RULE}$
2. [If [you]$_{AGENT}$ [ship autologous donations to another establishment that allows autologous donations to be used for allogeneic transfusion]$_{ACTION}$, [[you]$_{AGENT}$ [must]$_{DEONTIC}$ [assure that all autologous donations shipped to that establishment are tested under this section]$_{ACTION}$]$_{RULE}$]$_{RULE}$
3. [If [a filling]$_{THEME}$ [fails to meet the requirements of the first repeat test]$_{STATE}$, [[a second repeat test [may]$_{DEONTIC\_PERM}$ be conducted on the species which failed the test]$_{ACTION}$]]$_{RULE}$, provided that [50 percent of the total number of animals in that species]$_{THEME}$ [has survived the initial and first repeat tests]$_{STATE}$.
4. [you]$_{AGENT}$ [must]$_{DEONTIC}$ [use screening tests that the Food and Drug Administration (FDA) has approved for such use]$_{ACTION}$, [in accordance with the manufacturer's instructions]$_{ACCORDANCE}$

Of course, some sequences of categories might be ambiguous. For instance, in Example 4 and as discussed in Fig. 2, the accordance phrase could be associated either with the action specifically or with the deontic modal, that is to the rule itself. In the former case, rule 2 applies, but the latter one calls for an additional specification:

$$\text{RULE} \leftarrow \text{RULE ACCORDANCE} \tag{8}$$

hCL is designed to identify the overall semantic structures of the rule statements expressed in legal documents[2]. We make no claim that the few rules above are exhaustive and complete, covering all legal rule statements; but we do claim that, with respect to our corpus, most of the statements can be explained with a limited set of rules. This high level language leaves aside the actual parsing of the texts as well as deep, detailed semantic annotation, since long text fragments can be annotated as single hCL components. This illustrates the pragmatic approach that we have adopted to tackle legal language: even rough annotations are useful in a semantic web perspective. At the high level, a statement is described as a sequence of hCL categories and key words. Attachment ambiguities appear when two different rules apply on the same fragment (as for the ACCORDANCE phrase above).

In this section, we have described basic rule structures using a small set of annotations. The following section explains how the basic elements of the rules can be identified in order to annotate the components of the rule.

[2] These structures can then be transformed into attribute-value structures as the one presented in Fig. 1 but we do not develop that point here.

5 Annotation Support

The initial annotation of the legal texts requires attention from analysts, but a range of resources are available to support that task, which we outline in this section. These cover not only rule identification, but also the identification of relevant components of rules.

Annotation Guidelines. Annotation guidelines must be produced to explain to analysts how legal texts should be annotated. For a given text fragment, all analysts do not have to produce exactly the same annotations but they have to produce compatible annotations which differ only in the granularity of their descriptions. The guidelines must:

- list and define all the allowed semantic categories, with positive and negative examples to illustrate how the categories can or should be used in annotation;
- explain how to handle complex markup issues such as discontinued elements or annotation embeddings;
- specify how the quality of the annotation can be evaluated and assessed.

Terminological Analysis. Key domain terminology can be readily identified from existing terminological list (*e.g. blood transfusion*) or using existing terminological term extraction tools (*e.g.* TermRaider [11] or Termostat [5]). These terms, which are often called "open-textured terms" by legal analysts play an important role in legal interpretation. As legal reasoning consists in applying rules to facts, the classification of the actors, components of a situation, and so on plays a central role. Semantic tagging must therefore identify the key domain terms on which interpretation is based. The goal is not to propose a formal definition of those terms, as the subtleties of the different cases and the open-world assumption usually prevents the complete formalization of legal policies. The goal is simply to highlight them and make explicit their variations.

Interactive Annotation Tool. It is essential to provide analysts with a dedicated, interactive annotation tool. Many different generic tools can be exploited, *e.g.* the GATE, NLTK, or UIMA[3]. Key to using such tools is the application of the finite set of annotations as described in Sect. 4. Once a sequence of categories and keywords matches a right hand part of a rule, the analysts can tag the whole statement as a rule. If the rule is ambiguous, the analyst may be warned that he/she has to choose among alternatives.

Local Grammars. We have kept our discussion to high level rule components. Clearly, these are not sufficient to cover all of the textual phenomena that may be relevant to a fuller analysis of the text. For this, local grammars, which are grammars designed for subtasks of the overall analysis, may be designed to help analysts identifying the key elements of legal language, such as the markers and phrasal structures that introduce complements and adjuncts, such as *due to* and *except*, or legal terminology (*e.g.* references to textual elements). We may assume that local grammars are applied on

[3] https://gate.ac.uk/
http://www.nltk.org/book/
https://uima.apache.org/index.html.

documents that have already been POS-tagged and chunked so as to identify the borders of the elements that have been identified.

This pipeline of NLP processes support the analyst, but leave the task of semantic tagging under her control.

6 Summary and Future Work

The paper has advocated, motivated and exemplified a pragmatic approach to the analysis and annotation of complex legal texts. The approach combines the benefits of controlled languages – to give manageable although simplified descriptions of legal content – and of semantic annotation – to maintain a tight correlation with the source texts. It was pragmatically designed to help analysts publish legal sources and share interpretations on the semantic web.

For future work, we plan to apply the analysis to the larger regulation from which the sample is drawn, modifying it as required; for instance, the initial fragment could be extended to other constructions, *e.g.* exceptions and conditionals [16]. We would add tool support, *e.g.* contextually relevant pop-up annotation alternatives along with the option to create new, which would be essential to control for annotation variation. Evaluations with respect to inter-annotator agreement and users (e.g. querying) would help to establish the strengths or weaknesses of the approach and tools for the intended texts and user group.

References

1. Asooja, K., Bordea, G., Vulcu, G., O'Brien, L., Espinoza, A., Abi-Lahoud, E., Buitelaar, P., Butler, T.: Semantic annotation of finance regulatory text using multilabel classification. In: LeDA-SWAn (to appear, 2015)
2. Athan, T., Boley, H., Governatori, G., Palmirani, M., Paschke, A., Wyner, A.: OASIS LegalRuleML. In: ICAIL, pp. 3–12. Rome, Italy (2013)
3. Bos, J.: Wide-coverage semantic analysis with boxer. In: Proceedings of Semantics in Text Processing, Research in Computational Semantics, pp. 277–286. College Publications (2008)
4. Dayal, S., Johnson, P.: A web-based revolution in Australian public administration. J. Inf. Law Technol. 1, online (2000)
5. Drouin, P.: Term extraction using non-technical corpora as a point of leverage. Terminology **9**(1), 99–115 (2003)
6. Francesconi, E.: Semantic model for legal resources: annotation and reasoning over normative provisions. Semant. Web **7**(3), 255–265 (2014)
7. Guissé, A., Lévy, F., Nazarenko, A.: From regulatory texts to BRMS: how to guide the acquisition of business rules? In: Bikakis, A., Giurca, A. (eds.) RuleML 2012. LNCS, vol. 7438, pp. 77–91. Springer, Heidelberg (2012)
8. Höfler, S.: Legislative drafting guidelines: how different are they from controlled language rules for technical writing? In: Kuhn, T., Fuchs, N.E. (eds.) CNL 2012. LNCS, vol. 7427, pp. 138–151. Springer, Heidelberg (2012)
9. Kuhn, T.: A survey and classification of controlled natural languages. Comput. Linguist. **40**(1), 121–170 (2014)
10. Lévy, F., Nazarenko, A., Wyner, A.: Towards a high-level controlled language for legal sources on the semantic web. In: LeDA-SWAn (to appear, 2015)

11. Maynard, D., Li, Y., Peters, W.: NLP techniques for term extraction and ontology population. In: The Conference on Ontology Learning and Population, pp. 107–127. IOS Press, Amsterdam (2008)
12. Meunier, M., Charret-Del Bove, M., Damette, E. (eds.): La traduction juridique: points de vue didactiques et linguistiques, 333 pages. Publications du Centre d'Etudes Linguistiques (2013)
13. OMG: Semantics of business vocabulary and business rules. Formal specification, v1.0. Technical report, The Object Management Group (2008)
14. Peters, W., Wyner, A.: Extracting hohfeldian relations from text. In: JURIX, Frontiers in Artificial Intelligence and Applications, vol. 279, pp. 189–190. IOS Press (2015)
15. Wyner, A., Bos, J., Basile, V., Quaresma, P.: An empirical approach to the semantic representation of law. In: JURIX, pp. 177–180. IOS Press, Amsterdam (2012)
16. Wyner, A., Peters, W.: On rule extraction from regulations. In: JURIX, pp. 113–122. IOS Press (2011)